Best of LessWrong: December 2019

1. Is Rationalist Self-Improvement Real?
2. 2019 AI Alignment Literature Review and Charity Comparison
3. Moloch Hasn't Won
4. Seeking Power is Often Convergently Instrumental in MDPs
5. Paper-Reading for Gears
6. Approval Extraction Advertised as Production
7. Understanding "Deep Double Descent"
8. Firming Up Not-Lying Around Its Edge-Cases Is Less Broadly Useful Than One Might Initially Think
9. [Part 2] Amplifying generalist research via forecasting – results from a preliminary exploration
10. Bayesian examination
11. We run the Center for Applied Rationality, AMA
12. Under what circumstances is "don't look at existing research" good advice?
13. Propagating Facts into Aesthetics
14. [Personal Experiment] One Year without Junk Media
15. The Review Phase
16. Karate Kid and Realistic Expectations for Disagreement Resolution
17. Should We Still Fly?
18. 2010s Predictions Review
19. Generalizing Experimental Results by Leveraging Knowledge of Mechanisms
20. Free Speech and Triskaidekaphobic Calculators: A Reply to Hubinger on the Relevance of Public Online Discussion to Existential Risk
21. The Lesson To Unlearn
22. Humans Are Embedded Agents Too
23. The New Age of Social Engineering
24. Against Premature Abstraction of Political Issues
25. Applications of Economic Models to Physiology?
26. Why aren't assurance contracts widely used?
27. Long Bets by Confidence Level
28. Speaking Truth to Power Is a Schelling Point
29. Recent Progress in the Theory of Neural Networks
30. Polio and the controversy over randomized clinical trials
31. Is Causality in the Map or the Territory?
32. "You can't possibly succeed without [My Pet Issue]"
33. [AN #77]: Double descent: a unification of statistical theory and modern ML practice
34. {Math} A times tables memory.
35. Making decisions under moral uncertainty
36. The Actionable Version of "Keep Your Identity Small"
37. 2020's Prediction Thread
38. Is the term mesa optimizer too narrow?
39. Meditation Retreat: Immoral Mazes Sequence Introduction
40. LW Team Updates - December 2019
41. New paper: (When) is Truth-telling Favored in AI debate?
42. Stupidity and Dishonesty Explain Each Other Away
43. Safe exploration and corrigibility
44. ESC Process Notes: Claim Evaluation vs. Syntheses
45. Predictive coding = RL + SL + Bayes + MPC
46. Might humans not be the most intelligent animals?
47. Templates and videos for doing annual and daily reviews
48. Inductive biases stick around
49. Long-lasting Effects of Suspensions?
50. What determines the balance between intelligence signaling and virtue signaling?

Best of LessWrong: December 2019

1. Is Rationalist Self-Improvement Real?
2. 2019 AI Alignment Literature Review and Charity Comparison
3. Moloch Hasn't Won
4. Seeking Power is Often Convergently Instrumental in MDPs
5. Paper-Reading for Gears
6. Approval Extraction Advertised as Production
7. Understanding "Deep Double Descent"
8. Firming Up Not-Lying Around Its Edge-Cases Is Less Broadly Useful Than One Might Initially Think
9. [Part 2] Amplifying generalist research via forecasting – results from a preliminary exploration
10. Bayesian examination
11. We run the Center for Applied Rationality, AMA
12. Under what circumstances is "don't look at existing research" good advice?
13. Propagating Facts into Aesthetics
14. [Personal Experiment] One Year without Junk Media
15. The Review Phase
16. Karate Kid and Realistic Expectations for Disagreement Resolution
17. Should We Still Fly?
18. 2010s Predictions Review
19. Generalizing Experimental Results by Leveraging Knowledge of Mechanisms
20. Free Speech and Triskaidekaphobic Calculators: A Reply to Hubinger on the Relevance of Public Online Discussion to Existential Risk
21. The Lesson To Unlearn
22. Humans Are Embedded Agents Too
23. The New Age of Social Engineering
24. Against Premature Abstraction of Political Issues
25. Applications of Economic Models to Physiology?
26. Why aren't assurance contracts widely used?
27. Long Bets by Confidence Level
28. Speaking Truth to Power Is a Schelling Point
29. Recent Progress in the Theory of Neural Networks
30. Polio and the controversy over randomized clinical trials
31. Is Causality in the Map or the Territory?
32. "You can't possibly succeed without [My Pet Issue]"
33. [AN #77]: Double descent: a unification of statistical theory and modern ML practice
34. {Math} A times tables memory.
35. Making decisions under moral uncertainty
36. The Actionable Version of "Keep Your Identity Small"
37. 2020's Prediction Thread
38. Is the term mesa optimizer too narrow?
39. Meditation Retreat: Immoral Mazes Sequence Introduction
40. LW Team Updates - December 2019
41. New paper: (When) is Truth-telling Favored in AI debate?
42. Stupidity and Dishonesty Explain Each Other Away
43. Safe exploration and corrigibility
44. ESC Process Notes: Claim Evaluation vs. Syntheses
45. Predictive coding = RL + SL + Bayes + MPC

46. Might humans not be the most intelligent animals?
47. Templates and videos for doing annual and daily reviews
48. Inductive biases stick around
49. Long-lasting Effects of Suspensions?
50. What determines the balance between intelligence signaling and virtue signaling?

Is Rationalist Self-Improvement Real?

Cross-posted from Putanumonit.

Basketballism

Imagine that tomorrow everyone on the planet forgets the concept of *training basketball skill*.

The next day everyone is as good at basketball as they were the previous day, but this talent is assumed to be fixed. No one expects their performance to change over time. No one teaches basketball, although many people continue to play the game for fun.

Geneticists explain that some people are born with better hand-eye coordination and are able to shoot a basketball accurately. Economists explain that highly-paid NBA players have a stronger incentive to hit shots, which explains their improved performance. Psychologists note that people who take more jump shots each day hit a higher percentage and theorize a principal factor of basketball affinity that influences both desire and skill at basketball. Critical race theorists claim that white men's under-representation in the NBA is due to systemic oppression.

Papers are published, tenure is awarded.

New scientific disciplines emerge and begin studying basketball more systematically. Evolutionary physiologists point out that our ancestors threw stones in a sidearm motion, which explains our lack of adaptation to the different motion of jump shots. Behavioral kinesiologists describe systematic biases in human basketball, such as the tendency to shoot balls with a flatter trajectory and a lower release point than is optimal.

When asked by aspiring basketball players if jump shots can be improved, they all shake their heads and rue that it is human nature to miss shots. A Nobel laureate behavioral kinesiologist tells audiences that even after writing books on biases in

basketball his shot did not improve much. Someone publishes a study showing that basketball performance improves after a one-hour training session with schoolchildren, but Shott Ballexander writes a critical takedown pointing out that the effect wore off after a month and could simply be random noise. The field switches to studying “nudges”: ways to design systems so that players hit more shots at the same level of skill. They recommend that the NBA adopt larger hoops.

Papers are published, tenure is awarded.

Then, one day, someone merely looking to get good at basketball, as opposed to getting tenure, comes across these papers. She realizes that the lessons of behavioral kinesiology can be used to improve her jump shot. She practices releasing the ball at the top of her jump from above the forehead with a steep arc. As her shots start swooshing in more people gather at the gym to practice with her. They call themselves Basketballists.

Most people who walk past the gym sneer at the Basketballists. “You call yourselves Basketballists and yet none of you shoot 100%”, they taunt. “You should go to grad school if you want to learn about jump shots.” Some of Basketballists themselves begin to doubt the project, especially since switching to the new shooting techniques lowers their performance at first. “Did you hear what the Center for Applied Basketball is charging for a training camp?”, they mutter, “I bet their results are all due to selection bias.”

The Basketballists insist that the training does help, that they really get better by the day. Their shots hit at a slightly higher rate than before, although this is swamped by the inter-individual variance. How could they know if it works?

AsWrongAsEver

A core axiom of Rationality (capitalized to refer to LessWrong version) is that it is a skill that can be improved with time and practice. The names **Overcoming** Bias and **Less**Wrong reflect this: rationality is a direction, not a fixed point.

What would it mean to "improve at Rationality"? On the epistemic side, to draw a map that more accurately reflects the territory. To be less swayed by bias, make more accurate predictions, avoid error. On the instrumental side, to use their improved epistemics to achieve their goals in life. The two are often conflated, both by Rationalists and skeptics, but the two are also highly correlated — an accurate map gets you to where you're going.

A core foundation of epistemic rationality is the research on heuristic and biases developed by Daniel Kahneman. The first book in The Sequences is in large part a summary of Kahneman’s work.

Awkwardly for Rationalists, Daniel Kahneman is hugely skeptical of any possible improvement even just in epistemic rationality, especially for whole groups of people. In an astonishing interview with Sam Harris, Kahneman describes bias after bias in human thinking, emotions, and decision making. For every one, Sam asks: *How do we get better at this?* And for every one, Daniel replies: *We don’t, we’ve been telling people about this for decades and nothing has changed, that’s just how people are*.

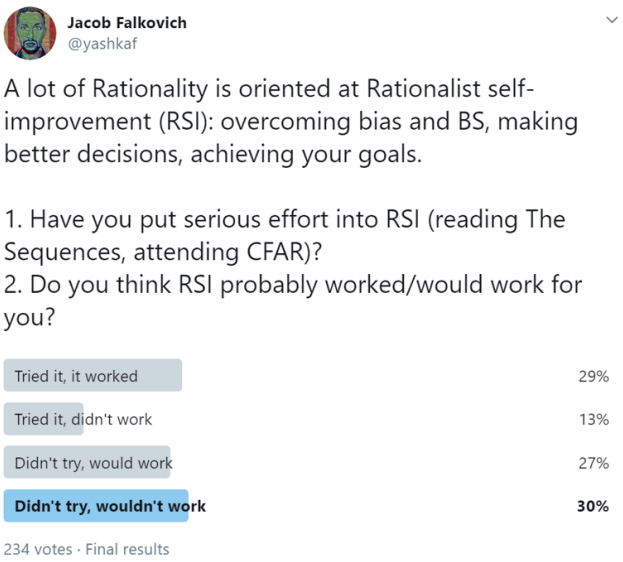
Daniel Kahneman is familiar with CFAR, but as far as I know he has not put as much effort himself into developing a community and curriculum dedicated to improving human rationality. He has *described* human irrationality, mostly to an audience of psychology undergrads. But psychology undergrads do worse than pigeons at learning a simple probabilistic game, we shouldn't expect them to learn rationality just by reading about biases. Perhaps if they started reading Slate Star Codex…

Alas, Scott Alexander himself is quite skeptical of Rationalist self-improvement. He agrees that Rationalist thinking can help you make good predictions and occasionally distinguish truth from bullshit, but he's unconvinced that it's something one can seriously get better at. Scott is even more skeptical of Rationality's use for life-optimization.

I told once Scott that I credit Rationality with a lot of the massive improvements in my financial, social, romantic, and mental life that happened to coincide with my discovery of LessWrong. Scott argued that I would do equally well in the absence of Rationality by finding other self-improvement philosophies to pour my intelligence and motivation into, and that these latter two are the root cause of my life getting better. Scott also seems to have been doing very well since he discovered LessWrong, but he credits Rationality with not much more than being a flag that united the community he's part of.

So: on one side are Yudkowsky, CFAR, and several Rationalists, sharing the belief that Rationality is a learnable skill that can improve the lives of most seekers who step on the path. On the other side are Kahneman, Alexander, several other Rationalists, and all the sneerers, who disagree.

When I surveyed my Twitter followers, the results distributed somewhat predictably:



The optimistic take is that RSI works for most people if they only tried it. The neutral take is that people are good at trying self-improvement philosophies that would work for them. The pessimistic take is that Rationalists are deluded by sunk cost and confirmation bias.

Who's right? Is Rationality trainable like jump shots or fixed like height? Before reaching any conclusions, let's try to figure out how why so many smart people who are equally familiar with Rationality disagree so strongly about this important question.

Great Expectations

An important crux of disagreement between me and Scott is in the question of what counts as successful Rationalist self-improvement. We can both look at the same facts and come to very different conclusions regarding the utility of Rationality.

Here's how Scott parses the fact that 15% of SSC readers who were referred by LessWrong have made over $1,000 by investing in cryptocurrency and 3% made over $100,000:

> The first mention of Bitcoin on Less Wrong, a post called Making Money With Bitcoin, was in early 2011 – when it was worth 91 cents. Gwern predicted that it could someday be worth "upwards of $10,000 a bitcoin". […]
>
> This was the easiest test case of our "make good choices" ability that we could possibly have gotten, the one where a multiply-your-money-by-a-thousand-times opportunity basically fell out of the sky and hit our community on its collective head. So how did we do?
>
> I would say we did mediocre. […]
>
> Overall, if this was a test for us, I give the community a C and me personally an F. God arranged for the perfect opportunity to fall into our lap. We vaguely converged onto the right answer in an epistemic sense. And 3 – 15% of us, not including me, actually took advantage of it and got somewhat rich.

Here's how I would describe it:

> Of the 1289 people who were referred to SSC from LessWrong, two thirds are younger than 30, a third are students/interns or otherwise yet to start their careers, and many are for other reasons too broke for it to be *actually rational* to risk even $100 on something that you saw recommended on a blog. Of the remainder, the majority were not around in the early days when cryptocurrencies were discussed — the median "time in community" on LessWrong surveys is around two years. In any case, "invest in crypto" was never a major theme or universally endorsed in the Rationalist community.
>
> Of those that were around and had the money to invest early enough, a lot lost it all when Mt. Gox was hacked or when Bitcoin crashed in late 2013 and didn't recover until 2017 or through several other contingencies.
>
> If I had to guess the percent of Rationalists who were even in a position to learn about crypto on LessWrong and make more than $1,000 by following Rationalist advice, I'd say it's certainly less than 50%. Maybe not much larger than 15%.
>
> Only 8% of Americans own cryptocurrency today. At the absolute highest end estimate, 1% of Americans, and 0.1% of people worldwide, made >$1,000 from crypto. So Rationalists did at least an order of magnitude better than the general

> population, almost as well as they could've done in a perfect world, and also funded MIRI and CFAR with Bitcoin for years ahead. I give the community an A and myself an A.

Now, multiplying money with a simple investment is an incredibly competitive arena of human endeavor, one where we would *least* expect to find low-hanging fruit that hasn't been picked. Even if you think Rationalists' success in that space is modest, that's still better than the average hedge fund, the actual "professionals".

For most other goals we care about no efficient market exists to compete with our efforts. Making friends, staying healthy, improving the world with charity, finding compatible partners, managing your happiness and attention, living forever — we should expect some fruit of progress on those to hang lower than a Bitcoin fortune.

Akrasia

Scott blames the failure of Rationality to help primarily on akrasia.

> One factor we have to once again come back to is akrasia. I find akrasia in myself and others to be the most important limiting factor to our success. Think of that phrase "limiting factor" formally, the way you'd think of the limiting reagent in chemistry. When there's a limiting reagent, it doesn't matter how much more of the other reagents you add, the reaction's not going to make any more product. Rational decisions are practically useless without the willpower to carry them out. If our limiting reagent is willpower and not rationality, throwing truckloads of rationality into our brains isn't going to increase success very much.

I take this paragraph to imply a model that looks like this:

[Alex reads LessWrong] -> [Alex tries to become less wrong] -> [akrasia!] -> [Alex doesn't improve].

I would make a small change to this model:

[Alex reads LessWrong] -> [akrasia!] -> [Alex doesn't try to become less wrong] -> [Alex doesn't improve].

A lot of LessWrong is very fun to read, as is all of SlateStarCodex. A large number of people on these sites are just looking to procrastinate during the workday, not to change how their mind works. Only 7% of the people who were engaged enough to fill out the last LessWrong survey have attended a CFAR workshop. Only 20% ever wrote a post, which is some measure of *active* rather than passive engagement with the material.

In contrast, one person wrote a sequence on trying out applied rationality for 30 days straight: Xiaoyu "The Hammer" He. And *he* was quite satisfied with the result.

I'm not sure that Scott and I disagree much, but I didn't get the sense that his essay was saying "just reading about this stuff doesn't help, you have to actually try". It also doesn't explain was he was so skeptical about me crediting my own improvement to Rationality.

Akrasia is discussed a lot on LessWrong, and applied rationality has several tools that help with it. What works for me and my smart friends is not to try and *generate* willpower but to use lucid moments to design plans that take a lack of willpower into account. Other approaches work for other people. But of course, if someone lacks the willpower to even try and take Rationality improvement seriously, a mere blog post will not help them.

3% LessWrong

In an essay called Extreme Rationality: It's Not That Great Scott writes:

> Eliezer writes:
> The novice goes astray and says, "The Art failed me."
> The master goes astray and says, "I failed my Art."
>
> Yet one way to fail your Art is to expect more of it than it can deliver.

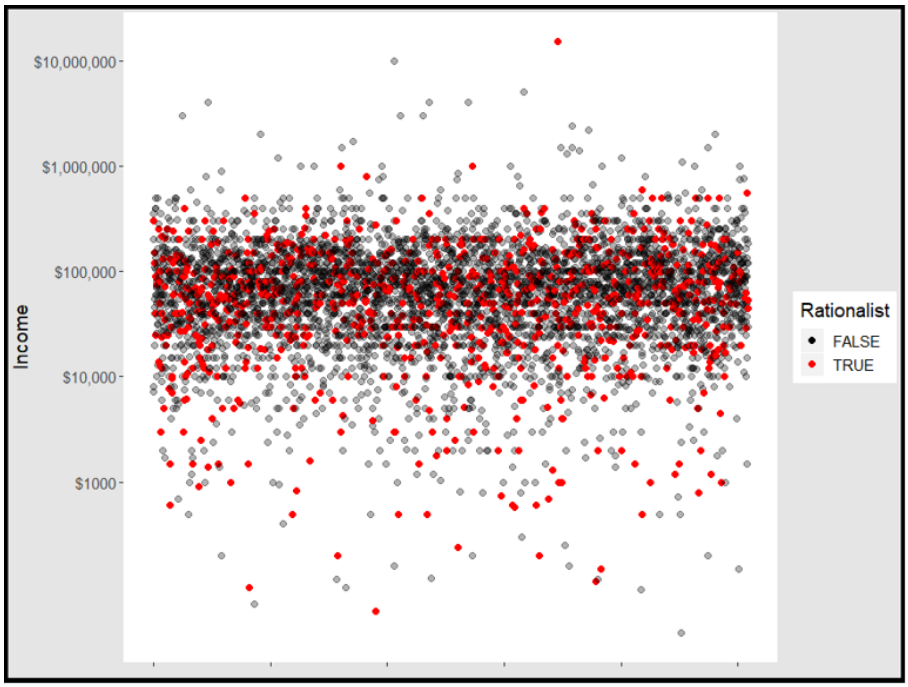
Scott means to say that Eliezer expects too much of the art in demanding that great Rationalist teachers be great at other things as well. But I think that expecting 50% of LessWrongers filling out a survey to have made thousands of dollars from crypto is setting the bar far higher than Eliezer's criterion of "Being a math professor at a small university who has published a few original proofs, or a successful day trader who retired after five years to become an organic farmer, or a serial entrepreneur who lived through three failed startups before going back to a more ordinary job as a senior programmer."

How much improvement does Scott expect? Below is a key quote in his essay, emphasis in the original.

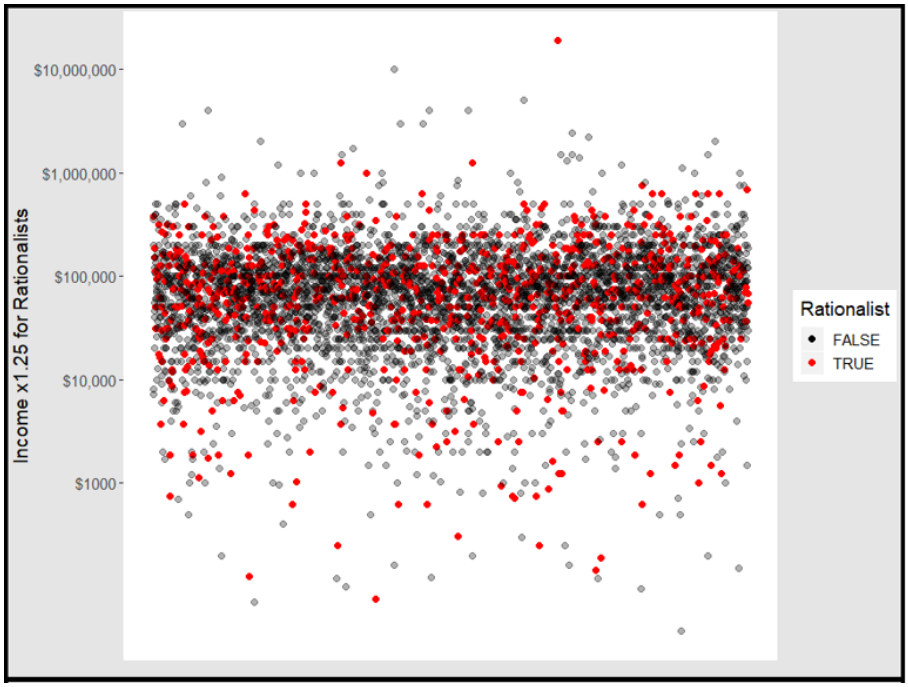
> I think it may help me succeed in life a little, but **I think the correlation between x-rationality and success is probably closer to 0.1 than to 1.**

Well, how big of a correlation is 0.1?

Here's the chart of respondents to the SlateStarCodex survey, by self-reported yearly income and whether they were referred from LessWrong (Scott's criterion for Rationalists).



And here's the same chart after I made a small change. Can you notice it?



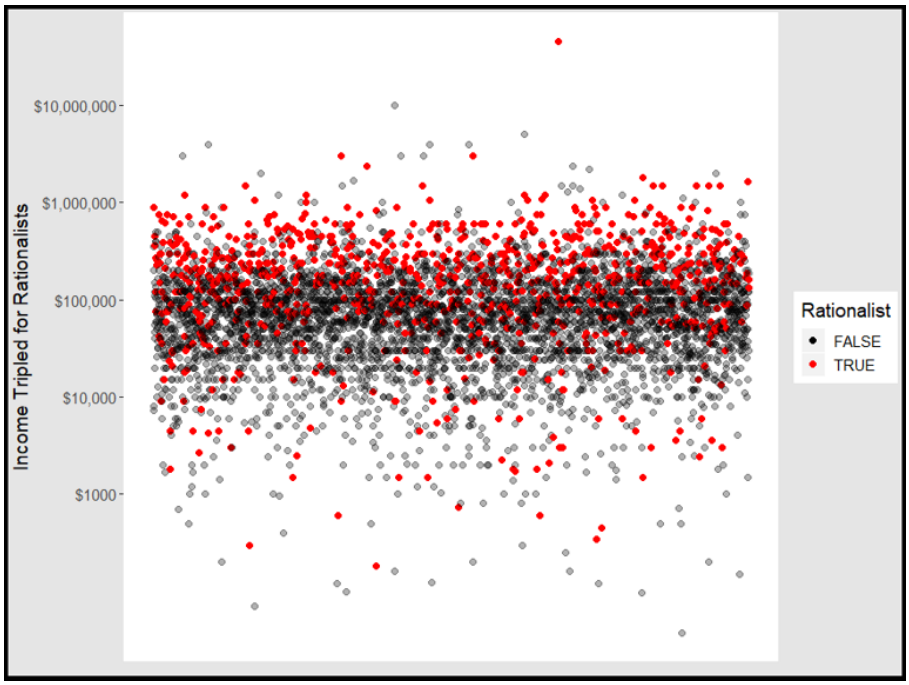
In the second chart, I increased the income of all rationalists by 25%.

The following things are both true:

- When you eyeball the group as a whole, the charts look identical. A 25% improvement for a quarter of the people in a group you observe is barely noticeable. The rich stayed rich, the poor stayed poor.

- If your own income increased 25% you would *certainly notice it*. And if the increase came as a result of reading a few blog posts and coming to a few meetups, you would tell everyone you know about this astounding life hack.

The correlation between Rationality and income in Scott's survey is -0.01. That number goes up to a mere 0.02 after the increase. A correlation of 0.1 is *absolutely huge*, it would require **tripling** the income of all Rationalists.



The point isn't to nitpick Scott's choice of "correlation = 0.1" as a metaphor. But every measure of success we care about, like impact on the world or popularity or enlightenment, is probably distributed like income is on the survey. And so if Rationality made you 25% more successful it wouldn't be as obviously visible as Scott thinks it would be — especially since everyone pursues a different vision of success. In this 25% world, the most and least successful people would still be such for reasons other than Rationality. And in this world, Rationality would be one of the most effective self-improvement approaches ever devised. 25% is a lot!

Of course, the 25% increase wouldn't happen immediately. Most people who take Rationality seriously have been in the community for several years. You get to 25% improvement by getting 3% better each year for 8 years.

Here's what 3% improvement feels like:

> You know what feels crappy? 3% improvement. You busted your ass for a year, trying to get better at dating, at being less of an introvert, at self-soothing your anxiety – and you only managed to get 3% better at it.
>
> If you worked a job where you put in that much time at the office and they gave you a measly 3% raise, you would spit in your boss's face and walk the fuck out. And, in fact, that's what most people do: quit. […]

> The model for most self-improvement is usually this:
> * You don't have much of a problem
> * You found The Breakthrough that erased all the issues you had
> * When you're done, you'll be the opposite of what you were. Used to be bad at dating? Now you'll have your own personal harem. Used to be useless at small talk? Now you're a fluent raconteur.
>
> Which, when you've agonized to scrape together a measly 3% improvement, feels like crap. If you're burdened with such social anxiety that it takes literally everything you have to go out in public for twenty minutes, make one awkward small talk, and then retreat home to collapse in embarrassment, you think, "Well, this isn't worth it."
>
> But most self-improvement isn't *immediate* improvement, my friend.
> It's compound interest.

I think that Rationalist self-improvement is like this. You don't get better at life and rationality after taking one class with Prof. Kahnemann. After 8 years of hard work, you don't stand out from the crowd even as the results become personally noticeable. But if you discover Rationality in college and stick with it, by the time you're 55 you will be **three times** better than what you would have been if you hadn't compounded these 3% gains year after year, and everyone will notice that.

What's more, the outcomes don't scale smoothly with your level of skill. When rare, high leverage opportunities come around, being slightly more rational can make a huge difference. Bitcoin was one such opportunity; meeting my wife was another such one for me. I don't know what the next one will be: an emerging technology startup? a political upheaval? cryonics? I know that the world is getting weirder faster, and the payouts to Rationality are going to increase commensurately.

There is still the issue of selection bias. Rationalists are not a representative sample of the population by any means. According to the surveys the average LessWrong reader has a vastly higher IQ than average and comes from fields where analytical and systematic thinking is rewarded like engineering, exact sciences, or philosophy. We probably should not conclude that self-improvement through epistemic Rationality will work for many or most people.

But if you're reading this, you're probably not *most people*. The difference is not merely in ability but also in inclination — what you're curious about and what you're willing to try. If you're the sort of person for whom success in life means stepping outside the comfort zone that your parents and high school counselor charted out for you, if you're willing to explore spaces of consciousness and relationships that other people warn you about, if you compare yourself only to who you were yesterday and not to who someone else is today… If you're weird like me I think that Rationality can improve your life a lot.

But to get better at basketball, you have to actually show up to the gym.

See also: The Martial Art of Rationality.

2019 AI Alignment Literature Review and Charity Comparison

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Cross-posted to the EA forum here.

Introduction

As in 2016, 2017 and 2018, I have attempted to review the research that has been produced by various organisations working on AI safety, to help potential donors gain a better understanding of the landscape. This is a similar role to that which GiveWell performs for global health charities, and somewhat similar to a securities analyst with regards to possible investments.

My aim is basically to judge the output of each organisation in 2019 and compare it to their budget. This should give a sense of the organisations' average cost-effectiveness. We can also compare their financial reserves to their 2019 budgets to get a sense of urgency.

I'd like to apologize in advance to everyone doing useful AI Safety work whose contributions I may have overlooked or misconstrued. As ever I am painfully aware of the various corners I have had to cut due to time constraints from my job, as well as being distracted by 1) another existential risk capital allocation project, 2) the miracle of life and 3) computer games.

How to read this document

This document is fairly extensive, and some parts (particularly the methodology section) are the same as last year, so I don't recommend reading from start to finish. Instead, I recommend navigating to the sections of most interest to you.

If you are interested in a specific research organisation, you can use the table of contents to navigate to the appropriate section. You might then also want to Ctrl+F for the organisation acronym in case they are mentioned elsewhere as well.

If you are interested in a specific topic, I have added a tag to each paper, so you can Ctrl+F for a tag to find associated work. The tags were chosen somewhat informally so you might want to search more than one, especially as a piece might seem to fit in multiple categories.

Here are the un-scientifically-chosen hashtags:

- Agent Foundations
- AI_Theory
- Amplification
- Careers
- CIRL
- Decision_Theory
- Ethical_Theory
- Forecasting
- Introduction
- Misc
- ML_safety
- Other_Xrisk

- Overview
- Philosophy
- Politics
- RL
- Security
- Shortterm
- Strategy

New to Artificial Intelligence as an existential risk?

If you are new to the idea of General Artificial Intelligence as presenting a major risk to the survival of human value, I recommend this Vox piece by Kelsey Piper.

If you are already convinced and are interested in contributing technically, I recommend this piece by Jacob Steinheart, as unlike this document Jacob covers pre-2019 research and organises by topic, not organisation.

Research Organisations

FHI: The Future of Humanity Institute

FHI is an Oxford-based Existential Risk Research organisation founded in 2005 by Nick Bostrom. They are affiliated with Oxford University. They cover a wide variety of existential risks, including artificial intelligence, and do political outreach. Their research can be found here.

Their research is more varied than MIRI's, including strategic work, work directly addressing the value-learning problem, and corrigibility work.

In the past I have been very impressed with their work.

Research

Drexler's Reframing Superintelligence: Comprehensive AI Services as General Intelligence is a massive document arguing that superintelligent AI will be developed for individual discrete services for specific finite tasks, rather than as general-purpose agents. Basically the idea is that it makes more sense for people to develop specialised AIs, so these will happen first, and if/when we build AGI these services can help control it. To some extent this seems to match what is happening - we do have many specialised AIs - but on the other hand there are teams working directly on AGI, and often in ML 'build an ML system that does it all' ultimately does better than one featuring hand-crafted structure. While most books are full of fluff and should be blog posts, this is a super dense document - a bit like Superintelligence in this regard - and even more than most research I struggle to summarize it here - so I recommend reading it. See also Scott's comments here. It also admirably hyperlinked so one does not have to read from start to finish. #Forecasting

Aschenbrenner's Existential Risk and Economic Growth builds a model for economic growth, featuring investment in consumption and safety. As time goes on, diminishing marginal utility of consumption means that more and more is invested in safety over incremental consumption. It derives some neat results, like whether or not we almost certainly go extinct depends on whether safety investments scale faster than the risk from consumption, and that generally speeding things up is better, because if there is a temporary risky phase it

gets us through it faster - whereas if risk never converges to zero we will go extinct anyway. Overall I thought this was an excellent paper. #Strategy

Carey's How useful is Quantilization for Mitigating Specification-Gaming extends and tests Taylor's previous work on using quantisation to reduce overfitting. The paper first proves some additional results and then runs some empirical tests with plausible real-life scenarios, showing that the technique does a decent job improving true performance (by avoiding excessive optimisation on the imperfect proxy). However, the fact that they sometimes underperformed the imitator baseline makes me worry that maybe the optimisation algorithms were just not well suited to the task. Overall I thought this was an excellent paper. #ML_safety

O'Keefe's Stable Agreements in Turbulent Times: A Legal Toolkit for Constrained Temporal Decision Transmission provides an introduction to the various ways current law allows contracts to be cancelled or adjusted after they have been made. For example, if subsequent circumstances have changed so dramatically that the fundamental nature of the contract has changed. The idea is that this helps promote stability by getting closer to 'what we really meant' than the literal text of the agreement. It is interesting but I am sceptical it is very helpful for AI Alignment, where forcing one group / AI that has suddenly become much more powerful to abide by their previous commitments seems like more of a challenge; post hoc re-writing of contracts seems like a recipe for the powerful to seize from the left behind. #Politics

Armstrong's Research Agenda v0.9: Synthesising a human's preferences into a utility function lays out what Stuart thinks is a promising direction for safe AGI development. To avoid the impossibility of deducing values from behaviour, we build agents with accurate models of the way human minds represent the world, and extract (partial) preferences from there. This was very interesting, and I recommend reading it in conjunction with this response from Steiner. #AI_Theory

Kenton et al.'s Generalizing from a few environments in Safety-Critical Reinforcement Learning runs an experiment on how well some ML algorithms can generalise to avoid catastrophes. This aimed to get at the risk of agents doing something catastrophic when exposed to new environments after testing. I don't really understand how it is getting at this though - the hazard (lava) is the same in train and test, and the poor catastrophe-avoidance seems to simply be the result of the weak penalty placed on it during training (-1). #ML_safety

Cihon's Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development advocates for the inclusion of safety-related elements into international standards (like those created by the IEEE). I'm not sure I see how these are directly helpful for the long-term problem while we don't yet have a technical solution - I generally think of these sorts of standards as mandating best practices, but in this case we need to develop those best practices. #Politics

Garfinkel & Dafoe's How does the offense-defense balance scale? discuss and model the way that military effectiveness various with investment in offence and defence. They discuss a variety of conflict modes, including invasions, cyber, missiles and drones. It seems that, in their model, cyberhacking is basically the same as invasions with varying sparse defences (due to the very large number of possible zero-day 'attack beaches'. #Misc

FHI also produced several pieces of research on bioengineered pathogens which are likely of interest to many readers – for example Nelson here – but which I have not had time to read.

FHI researchers contributed to the following research led by other organisations:

- Hubinger et al.'s Risks from Learned Optimization in Advanced Machine Learning Systems
- Greaves & Cotton-Barratt's A bargaining-theoretic approach to moral uncertainty

- Snyder-Beattie et al.'s An upper bound for the background rate of human extinction
- Zhang & Dafoe's Artificial Intelligence: American Attitudes and Trends
- Evans et al.'s Machine Learning Projects for Iterated Distillation and Amplification

Finances

FHI didn't reply to my emails about donations, and seem to be more limited by talent than by money.

If you wanted to donate to them anyway, here is the relevant web page.

CHAI: The Center for Human-Aligned AI

CHAI is a UC Berkeley based AI Safety Research organisation founded in 2016 by Stuart Russell.. They do ML-orientated safety research, especially around inverse reinforcement learning, and cover both near and long-term future issues.

As an academic organisation their members produce a very large amount of research; I have only tried to cover the most relevant below. It seems they do a better job engaging with academia than many other organisations.

Rohin Shah, now with additional help, continue to produce the AI Alignment Newsletter, covering in detail a huge number of interesting new developments, especially new papers.

They are expanding somewhat to other universities outside Berkeley.

Research

Shah et al.'s On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference argues that learning human values and biases at the same time, while impossible in theory, is actually possible in practice. Attentive readers will recall Armstrong and Mindermann's paper arguing that it is impossible to co-learn human bias and values because any behaviour is consistent with any values - if we can freely vary the biases - and vice versa. This paper basically argues that, like the No Free Lunch theorem, in practice this just doesn't matter that much, basically by assuming that the agent is close-to-optimal. (They also discuss the potential of using some guaranteed-optimal behaviour as ground truth, but I am sceptical this would work, as I think humans are often at their most irrational when it comes to the most important topics, e.g. love). Empirically, in their gridworld tests their agent did a decent job learning - for reasons I didn't really understand. Overall I thought this was an excellent paper. #CIRL

Turner et al.'s Conservative Agency attempts to prevent agents from doing irreversible damage by making them consider a portfolio of randomly generated utility functions - for which irreversible damage is probably bad for at least one of them. Notably, this portfolio did *not* include the true utility function. I find the result a little hard to understand - I initially assumed they were relying on clustering of plausible utility functions, but it seems that they actually sampled at random from the entire space of possible functions! I don't really understand how they avoid Armstrong + Mindermann type problems, but apparently they did! It seems like this line of attack pushes us towards Universal Drives, as something many utility functions will have in common. Overall I thought this was an excellent paper. #ML_safety

Carroll et al.'s On the Utility of Learning about Humans for Human-AI Coordination discusses the differences between competitive versus collaborative learning. If you just want to be really good at a competitive game, self-play is great, because you get better by playing better and better versions of yourself. However, if you have to collaborate with a human this

is bad because your training doesn't feature flawed partners (in the limit) and min-maxing doesn't work. They do an experiment showing that an agent taught about how humans act does better than one which learnt collaborating with itself. This seems useful if you think that CIRL/amplification approaches will be valuable, and also promotes teaching AIs to understand human values. There is also a blog post here #CIRL

Chan et al.'s The Assistive Multi-Armed Bandit attempts to do value learning with humans who are themselves value learning. They do this by having the agent sometimes 'intercept' on a multi-armed bandit problem, and show that this sometimes improves performance if the agent understands how the human is learning. #CIRL

Russell's Human Compatible; Artificial Intelligence and the Problem of Control is an introductory book aimed at the intelligent layman. As befits the author, it begins with a lot of good framing around intelligence and agency. The writing style is good. #Overview

Shah et al.'s Preferences Implicit in the State of the World attempts to use the fact that human environments are already semi-optimised to extract additional evidence about human preferences. Practically, this basically means simulating many paths the humans could have taken prior to t=0 and using these as evidence as to the human's values. The core of the paper is a good insight - "*it is easy to forget these preferences, since these preferences are already satisfied in our environment.*" #CIRL

CHAI researchers contributed to the following research led by other organisations:

- Agrawal et al.'s Scaling up Psychology via Scientific Regret Minimization:A Case Study in Moral Decision-Making

Finances

They have been funded by various EA organisations including the Open Philanthropy Project and recommended by the Founders Pledge.

They spent $1,450,000 in 2018 and $2,000,000 in 2019, and plan to spend around $2,150,000 in 2020. They have around $4650000 in cash and pledged funding, suggesting (on a very naïve calculation) around 2.2 years of runway.

If you wanted to donate to them, here is the relevant web page.

MIRI: The Machine Intelligence Research Institute

MIRI is a Berkeley based independent AI Safety Research organisation founded in 2000 by Eliezer Yudkowsky and currently led by Nate Soares. They were responsible for much of the early movement building for the issue, but have refocused to concentrate on research for the last few years. With a fairly large budget now, they are the largest pure-play AI alignment shop. Their research can be found here. Their annual summary can be found here.

In general they do very ‘pure’ mathematical work, in comparison to other organisations with more ‘applied’ ML or strategy focuses. I think this is especially notable because of the irreplaceability of the work. It seems quite plausible that some issues in AI safety will arise early on and in a relatively benign form for non-safety-orientated AI ventures (like autonomous cars or Minecraft helpers) – however the work MIRI does largely does not fall into this category. I have also historically been impressed with their research.

Their agent foundations work is basically trying to develop the correct way of thinking about agents and learning/decision making by spotting areas where our current models fail and seeking to improve them. This includes things like thinking about agents creating other agents.

In their annual write-up they suggest that progress was slower than expected in 2019. However I assign little weight to this as I think most of the cross-sectional variation in organisation reported subjective effectiveness comes from variance in how optimistic/salesy/aggressive they are, rather than actually indicating much about object-level effectiveness.

MIRI, in collaboration with CFAR, runs a series of four-day workshop/camps, the AI Risk for Computer Scientists workshops, which gather mathematicians/computer scientists who are potentially interested in the issue in one place to learn and interact. This sort of workshop seems very valuable to me as an on-ramp for technically talented researchers, which is one of the major bottlenecks in my mind. In particular they have led to hires for MIRI and other AI Risk organisations in the past. I don't have any first-hand experience however.

They also support MIRIx workshops around the world, for people to come together to discuss and hopefully contribute towards MIRI-style work.

Research

Hubinger et al.'s Risks from Learned Optimization in Advanced Machine Learning Systems introduces the idea of a Mesa-Optimizer - a sub-agent of an optimizer that is itself an optimizer. A vague hand-wave of an example might be for-profit corporations rewarding their subsidiaries based on segment PnL, or indeed evolution creating humans, which then go on to create AI. Necessarily theoretical, the paper motivates the idea, introduces a lot of terminology, and describes conditions that might make mesa-optimisers more or less likely - for example, more diverse environments make mesa-optimisation more likely. In particular, they distinguish between different forms of mis-alignment - e.g. between meta, object-level and mesa, vs between mesa and behavioural objectives. There is a sequence on the forum about it here. Overall I thought this was an excellent paper. Researchers from FHI, OpenAI were also named authors on the paper. #Agent Foundations

Kosoy's Delegative Reinforcement Learning: Learning to Avoid Traps with a Little Help produces an algorithm that deviates only boundedly from optimal with a human intervening to prevent it stumbling into irrevocably bad actions. The idea is basically that the human intervenes to prevent the really bad actions, but because the human has some chance of selecting the optimal action afterwards, the loss of exploration value is limited. This attempts to avoid the problem that 'ideal intelligence' AIXI has whereby it might drop an anvil on its head. I found the proof a bit hard to follow, so I'm not sure how tight the bound is in practice. Notably, this doesn't protect us if the agent tries to prevent the human from intervening. Related. #ML_safety

There were two analyses of FDT from academic philosophers this year (reviewed elsewhere in this document). In both cases I felt their criticisms rather missed the mark, which is a positive for the MIRI approach. However, they did convincingly argue that MIRI researchers hadn't properly understood the academic work they were critiquing, an isolation which has probably gotten worse with MIRI's current secrecy. MIRI suggested I point out that Cheating Death In Damascus had recently been accepted in The Journal of Philosophy, a top philosophy journal, as evidence of (hopefully!) mainstream philosophical engagement.

MIRI researchers contributed to the following research led by other organisations:

- MacAskill & Demski's A Critique of Functional Decision Theory

Non-disclosure policy

Last year MIRI announced their policy of nondisclosure-by-default:

[G]oing forward, most results discovered within MIRI will remain internal-only unless there is an explicit decision to release those results, based usually on a specific anticipated safety upside from their release.

I wrote about this at length last year, and my opinion hasn't changed significantly since then, so I will just recap briefly.

On the positive side we do not want people to be pressured into premature disclosure for the sake of funding. This space is sufficiently full of infohazards that secrecy might be necessary, and in its absence researchers might prudently shy away from working on potentially risky things - in the same way that no-one in business sends sensitive information over email any more. MIRI are in exactly the sort of situation that you would expect might give rise to the need for extreme secrecy. If secret research is a necessary step *en route* to saving the world, it will have to be done by someone, and it is not clear there is anyone very much better.

On the other hand, I don't think we can give people money just because they say they are doing good things, because of the risk of abuse. There are many other reasons for not publishing anything. Some simple alternative hypothesis include "we failed to produce anything publishable" or "it is fun to fool ourselves into thinking we have exciting secrets" or "we are doing bad things and don't want to get caught." The fact that MIRI's researchers appear intelligent suggest they at least think they are doing important and interesting issues, but history has many examples of talented reclusive teams spending years working on pointless stuff in splendid isolation.

Additionally, by hiding the highest quality work we risk impoverishing the field, making it look unproductive and unattractive to potential new researchers.

One possible solution would be for the research to be done by impeccably deontologically moral people, whose moral code you understand and trust. Unfortunately I do not think this is the case with MIRI. (I also don't think it is the case with many other organisations, so this is not a specific criticism of MIRI, except insomuch as you might have held them to a higher standard than others).

Finances

They spent $3,750,000 in 2018 and $6,000,000 in 2019, and plan to spend around $6,800,000 in 2020. They have around $9,350,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1.4 years of runway.

They have been supported by a variety of EA groups in the past, including OpenPhil

If you wanted to donate to MIRI, here is the relevant web page.

GCRI: The Global Catastrophic Risks Institute

GCRI is a globally-based independent Existential Risk Research organisation founded in 2011 by Seth Baum and Tony Barrett. They cover a wide variety of existential risks, including artificial intelligence, and do policy outreach to governments and other entities. Their research can be found here. Their annual summary can be found here.

In 2019 they ran an advising program where they gave guidance to people from around the world who wanted to help work on catastrophic risks.

In the past I have praised them for producing a remarkably large volume of research; this slowed down somewhat during 2019 despite taking on a second full-time staff member, which they attributed partly to timing issues (e.g. pieces due to be released soon), and partly to focusing on quality over quantity.

Research

Baum et al.'s Lessons for Artificial Intelligence from Other Global Risks analogises AI risk to several other global risks: biotech, nukes, global warming and asteroids. In each case it discusses how action around the risk progressed, in particular the role of gaining expert consensus and navigating vested interests. #Strategy

Baum's The Challenge of Analyzing Global Catastrophic Risks introduces the idea of catastrophic risks and discusses some general issues. It argues for the need to quantify various risks, and ways to present these to policymakers. #Other_Xrisk

Baum's Risk-Risk Tradeoff Analysis of Nuclear Explosives for Asteroid Deflection discusses how to compare the protection from asteroids that nukes offer vs their potential to exacerbate war. #Other_Xrisk

Finances

During December 2018 they received a $250,000 donation from Gordon Irlam.

They spent $140,000 in 2018 and $250,000 in 2019, and plan to spend around $250,000 in 2020. They have around $310,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1.2 years of runway.

If you want to donate to GCRI, here is the relevant web page.

CSER: The Center for the Study of Existential Risk

CSER is a Cambridge based Existential Risk Research organisation founded in 2012 by Jaan Tallinn, Martin Rees and Huw Price, and then established by Seán Ó hÉigeartaigh with the first hire in 2015. They are currently led by Catherine Rhodes and are affiliated with Cambridge University. They cover a wide variety of existential risks, including artificial intelligence, and do political outreach. Their research can be found here. Their annual summary can be found here and here.

CSER also participated in a lot of different outreach events, including to the UK parliament and by hosting various workshops, as well as submitting (along with other orgs) to the EU's consultation, as summarised in this post. I'm not sure how to judge the value of these.

CSER's researchers seem to select a somewhat widely ranging group of research topics, which I worry may reduce their effectiveness.

Catherine Rhodes co-edited a volume of papers on existential risks, including many by other groups mentioned in this review.

Research

Kaczmarek & Beard's Human Extinction and Our Obligations to the Past presents an argument that even people who hold person-affecting views should think extinction is bad because it undermines the sacrifices of our ancestors. My guess is that most readers are not in need of persuading that extinction is bad, but I thought this was an interesting additional argument. The core idea is that if someone makes a large sacrifice to enable some good, we have a pro tanto reason not to squander that sacrifice. I'm not sure how many people will be persuaded by this idea, but as a piece of philosophy I thought this was a clever idea, and it is definitely good to promote the idea that past generations have value (speaking as a future member of a past generation). Carl Shulman also offered related arguments here. #Philosophy

Beard's Perfectionism and the Repugnant Conclusion argues against one supposed rejection of the Repugnant Conclusion, namely that some goods are lexicographically superior to ordinary welfare. The paper makes the clever argument that the very large, barely-worth-living group might actually have more of these goods if they were offset by (lexicographically secondary) negative welfare. It was also the first time (to my recollection) that I've come across the Ridiculous Conclusion. #Philosophy

Avin's Exploring Artificial Intelligence Futures lists and discusses different ways of introducing people to the future of AI. These include fiction, games, expert analysis, polling and workshops. He also provides various pros and cons of the different techniques, which seemed generally accurate to me. #Strategy

Belfield's How to respond to the potential malicious uses of artificial intelligence? introduces AI and AI risk. This short article focuses mainly on short-term risks. #Introduction

Weitzdörfer & Beard's Law and Policy Responses to Disaster-Induced Financial Distress discusses the problem of indebtedness following the destruction of collateral in the 2011 earthquake in Japan. They explain the specifics of the situation in extreme detail, and I was pleasantly surprised by their final recommendations, which mainly concerned removing barriers to insurance penetration. #Politics

Kemp's Mediation Without Measures: Conflict Resolution in Climate Diplomacy discusses the lack of formal decision-making procedure for international climate change treaties. Unfortunately I wasn't able to access the article. #Other_Xrisk

Avin & Amadae's Autonomy and machine learning at the interface of nuclear weapons, computers and people discusses the potential dangers of incorporating narrow AI into nuclear weapon systems. #Shortterm

CSER's Policy series Managing global catastrophic risks: Part 1 Understand introduces the idea of Xrisk for policymakers. This is the first report in a series, and as such is quite introductory. It mainly focuses on non-AI risks. #Politics

Tzachor's The Future of Feed: Integrating Technologies to Decouple Feed Production from Environmental Impacts discusses a new technology for producing animal feedstock to replace soybeans. This could be Xrisk relevant if some non-AI risk made it hard to feed animals. However, I am somewhat sceptical of the presentation of this as a *likely* risk as both a future shortage of soybeans and a dramatically more efficient technology for feeding livestock would both presumably be of interest to private actors, and show up in soybean future prices. #Other_Xrisk

Beard's What Is Unfair about Unequal Brute Luck? An Intergenerational Puzzle discusses Luck Egalitarianism. #Philosophy

Quigley's Universal Ownership in the Anthropocene argues that because investors own diversified portfolios they effectively internalise externalities, and hence should push for various political changes. The idea is basically that even though polluting might be in a company's best interest, it hurts the other companies the investor owns, so it is overall against the best interests of the investor. As such, investors should push companies to pollute less and so on. The paper seems to basically assume that such 'universal investors' would be incentivised to support left-wing policies on a wide variety of issues. However, it somehow fails to mention even cursorily the fact that the core issue has been well studied by economists: when all the companies in an industry try to coordinate for mutual benefit, it is called a cartel, and the #1 way of achieving mutual benefit is raising prices to near-monopoly levels. It would be extremely surprising to me if someone, acting as a self-interested owner of all the world's shoe companies (for example) found it more profitable to protect biodiversity than to raise the price of shoes. Fortunately, in practice universal investors are quite supportive of competition. #Other_Xrisk

CSER researchers contributed to the following research led by other organisations:

- Colvin et al.'s Learning from the Climate Change Debate to Avoid Polarisation on Negative Emissions
- Hernandez-Orallo et al.'s Surveying Safety-relevant AI Characteristics
- Cave & Ó hÉigeartaigh's Bridging near- and long-term concerns about AI
- Lewis et al.'s Assessing contributions of major emitters' Paris-era decisions to future temperature extremes

Finances

They spent £789,000 in 2017-2018 and £801,000 in 2018-2019, and plan to spend around £1,100,000 in 2019-20 and £880,000 in 2020-21. It seems that similar to GPI maybe 'runway' is not that meaningful - they suggested it begins to decline from early 2021 and all their current grants end by mid-2024.

If you want to donate to them, here is the relevant web page.

Ought

Ought is a San Francisco based independent AI Safety Research organisation founded in 2018 by Andreas Stuhlmüller. They research methods of breaking up complex, hard-to-verify tasks into simple, easy-to-verify tasks - to ultimately allow us effective oversight over AIs. This includes building computer systems and recruiting test subjects. I think of them as basically testing Paul Christiano's ideas. Their research can be found here. Their annual summary can be found here.

Last year they were focused on factored generation – trying to break down questions so that distributed teams could produce the answer. They have moved on to factored evaluation – using similar distributed ideas to try to evaluate existing answers, which seems a significantly easier task (by analogy to P<=NP). It seems to my non-expert eye that factored generation did not work as well as they expected – they mention the required trees being extremely large, and my experience is that organising volunteers and getting them to actually do what they said they would has historically been a great struggle for many organisations. However I don't think we should hold negative results in investigations against organisations; negative results are valuable, and it might be the case that all progress in this difficult domain comes from *ex ante* longshots. If nothing else, even if Paul is totally wrong about the whole idea it would be useful to discover this sooner rather than later!

They provided an interesting example of what their work looks like in practice here, and a detailed presentation on their work here.

They also worked on using ML, rather than humans, as the agent who answered the broken-down questions, in this case by using GPT-2, which seems like a clever idea.

Paul Christiano wrote a post advocating donating to them here.

Research

Evans et al.'s Machine Learning Projects for Iterated Distillation and Amplification provides three potential research projects for people who want to work on Amplification, as well as an introduction to Amplification. The projects are mathematical decomposition (which seems very natural), decomposition computer programs (similar to how all programs can be decomposed into logic gates, although I don't really understand this one) and adaptive computation, where you figure out how much computation to dedicate to different issues. In general I like outlining these sorts of 'shovel-ready' projects, as it makes it easier for new

researchers, and seems relatively under-appreciated. Researchers from FHI were also named authors on the paper. #Amplification

Roy's AI Safety Open Problems provides a list of lists of 'shovel-ready' projects for people to work on. If you like X (which I do in this case), meta-X is surely even better! #Ought

Finances

They spent $500,000 in 2018 and $1,000,000 in 2019, and plan to spend around $2,500,000 in 2020. They have around $1,800,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 0.7 years of runway.

They have received funding from a variety of EA sources, including the Open Philanthropy Project.

OpenAI

OpenAI is a San Francisco based independent AI Research organisation founded in 2015 by Sam Altman. They are one of the leading AGI research shops, with a significant focus on safety.

Earlier this year they announced GPT 2, a language model that was much better at 'understanding' human text than previous attempts, that was notably good at generating text that seemed human-generated - good enough that it was indistinguishable to humans who weren't concentrating. This was especially notable because OpenAI chose not to immediately release GPT 2 due to the potential for abuse. I thought this was a noble effort to start conversations among ML researchers about release norms, though my impression is that many thought OpenAI was just grandstanding, and I personally was sceptical of the harm potential - though a GPT 2 based intelligence did go on to almost take over LW, proving that the 'being a good LW commenter' is a hard goal. Outside researchers were able to (partly?) replicate it, but in a surprisingly heartening turn of events were persuaded not to release their reconstruction by researchers from OpenAI and MIRI. OpenAI eventually released a much larger version of their system - you can see it and read their follow-up report on the controlled release process here.

You can play with (one version of) the model here.

Research

Clark & Hadfield's Regulatory Markets for AI Safety suggests a model for the privatisation of AI regulation. Basically the idea is that governments will contract with and set outcomes for a small number of private regulators, which will then devise specific rules that need to be observed by ML shops. This allows the *ex-ante* regulation to be more nimble than if it was done publicly, while retaining the *ex-post* outcome guarantees. It reminded me of the system of auditors for public companies to ensure accounting accuracy) or David Friedman's work on polycentric law. I can certainly see why private companies might be more effective as regulators than government bodies. However, I'm not sure how useful this would be in an AGI scenario, where the goals and *ex-post* measurement for the private regulators are likely to become outdated and irrelevant. I'm also sceptical that governments would be willing to progressively give up regulatory powers; I suspect that if this system was to be adopted it would have to pre-empt government regulation. #Politics

Christiano's What failure looks like provides two scenarios that Paul thinks represent reasonably likely outcomes of Alignment going wrong. Notably neither exactly match the classic recursively self-improving FOOM case. The first is basically that we develop better and better optimisation techniques, but due to our inability to correctly specify what we

want, we end up with worse and worse Goodheart's Law situations, ending up in Red-Queen style Moloch scenario. The second is that we create algorithms that try to increase their influence (as per the fundamental drives). At first they do so secretly, but eventually (likely in response to some form of catastrophe reducing humanity's capability to suppress them) their strategy abruptly changes towards world domination. I thought this was an insightful post, and recommend readers also read the comments by Dai and Shulman, as well as this post. #Forecasting

Christiano's AI alignment landscape is a talk Paul gave at EA Global giving an overview of the issue. It is interesting both for seeing how he maps out all the different components of the problem and which he thinks are tractable and important, and also for how his Amplification approach falls out from this. #Overview

Irving & Askell's AI Safety Needs Social Scientists raise the issue of AI alignment requiring better understanding of humans as well as ML knowledge. Because humans are biased, etc., the more accurate our model of human preferences the better we can design AIs to align with it. It is quite focused on Amplification as a way of making human preferences more legible. I thought the article could have been improved with more actionable research projects for social scientists. Additionally, the article makes the need for social scientists seem somewhat tired to a Debate-style approach, whereas it seems to me potentially more broad. #Strategy

OpenAI Researchers also contributed to the following papers led by other organisations:

- Hubinger et al.'s Risks from Learned Optimization in Advanced Machine Learning Systems

Finances

OpenAI was initially funded with money from Elon Musk as a not-for-profit. They have since created an unusual corporate structure including a for-profit entity, in which Microsoft is investing a billion dollars.

Given the strong funding situation at OpenAI, as well as their safety team's position within the larger organisations, I think it would be difficult for individual donations to appreciably support their work. However it could be an excellent place to apply to work.

Google DeepMind

DeepMind is a London based AI Research organisation founded in 2010 by Demis Hassabis, Shane Legg and Mustafa Suleyman and currently led by Demis Hassabis. They are affiliated with Google. As well as being arguably the most advanced AI research shop in the world, DeepMind has a very sophisticated AI Safety team, covering both ML safety and AGI safety.

This year DeepMind build an agent that could beat humans at Starcraft II. This is impressive because it is a complex, incomplete information game that humans are very competitive at. However, the AI did have some advantages over humans by having direct API access.

Research

Everitt & Hutter's Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective discusses the problem of agents wireheading in an RL setting, along with several possible solutions. They use causal influence diagrams to highlight the difference between 'good' ways for agents to increase their reward function and 'bad' ways, and have a nice toy gridworld example. The solutions they discuss seemed to me to often be fairly standard ideas from the AI safety community - thinks like teaching the AI to

maximise the goal instantiated by its reward function at the start, rather than whatever happens to be in that box later, or using indifference results - but they introduce them to an RL setting, and the paper does a good job covering a lot of ground. There is more discussion of the paper here. Overall I thought this was an excellent paper. #RL

Everitt et al.'s Modeling AGI Safety Frameworks with Causal Influence Diagrams introduces the idea of using Causal Influence Diagrams to clarify thinking around AI safety proposals and make it easier to compare proposals with different conceptual backgrounds in a standard way. They introduce the idea, and show how to represent ideas like RL, CIRL, Counterfactual Oracles and Debate. Causal Influence Diagrams have been used in several other papers this year, like Categorizing Wireheading in Partially Embedded Agents. #AI_Theory

Everitt et al.'s Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings discusses using causal influence diagrams to distinguish things agents want to observe vs things they want to control. They use this to show the safety improvement from counterfactual oracles. It also presents a natural link between near-term and long-term safety concerns. #AI_Theory

Sutton's The Bitter Lesson argues that history suggests massive amounts of computer and relatively general structures perform better than human-designed specialised systems. He uses examples like the history of vision and chess, and it seems fairly persuasive, though I wonder a little if these are cherry-picked - e.g. in finance we generally do have to make considerable use of human-comprehensible features. This is not directly an AI safety paper, but it does have clear implications. #Forecasting

Uesato et al.'s Rigorous Agent Evaluation: An Adversarial Approach to Uncover Catastrophic Failures attempt to make it easier to find catastrophic failure cases. They do this adversarially with previous versions of the algorithm, based on the idea that it is cheaper to find disasters there, but they will be related to the failure modes of the later instantiations. This seems like an interesting idea, but seems like it would struggle with cases where increasing agent capabilities lead to new failure modes - e.g. the Treacherous Turn we are worried about. #ML_safety

Ngo's Technical AGI safety research outside AI provides a list of technically useful topics for people who are not ML researchers to work on. The topics selected look good - many similar to work AIImpacts or Ought do. I think lists like this are very useful for opening the field up to new researchers. #Overview

Researchers from DeepMind were also named on the following papers:

- Krueger et al.'s Misleading Meta-Objectives and Hidden Incentives for Distributional Shift

Finances

Being part of Google, I think it would be difficult for individual donors to directly support their work. However it could be an excellent place to apply to work.

AI Safety camp

AISC is an internationally based independent residential research camp organisation founded in 2018 by Linda Linsefors and currently led by Colin Bested. They bring together people who want to start doing technical AI research, hosting a 10-day camp aiming to produce publishable research. Their research can be found here.

To the extent they can provide an on-ramp to get more technically proficient researchers into the field I think this is potentially very valuable. But I obviously haven’t personally

experienced the camps, or even spoken to anyone who has.

Research

Majha et al.'s Categorizing Wireheading in Partially Embedded Agents discusses the wireheading problem for agents who can mess with their reward channel or beliefs. They model this using causal agent diagrams, suggest a possible solution (making rewards a function of world-beliefs, not observations) and show that this does not work using very simple gridworld AIXIjs implementations. #AI_Theory

Kovarik et al.'s AI Safety Debate and Its Applications discusses using adversarially Debating AIs as a method for alignment. It provides a very accessible introduction to Debating AIs, and implements some extensions to the practical MNIST work from the original paper. #Amplification

Mancuso et al.'s Detecting Spiky Corruption in Markov Decision Processes suggests that we can address corrupted reward signals for RL by removing 'spikey' rewards. This is an attempt to get around impossibility results by identifying a subclass where they don't hold. I can see this being useful in some cases like reward tampering, where the reward from fiddling with $AGENT_UTILITY is likely to be very spiky. However if human values are fragile then it seems plausible that the 'True' reward signal should also be spikey. #ML_safety

Perry & Uuk's AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk introduces the field of AI governance, and discusses issues about how policy is implemented in practice, like the existence of windows in time for institutional change. #Politics

Finances

Their website suggests they are seeking donations, but they did not reply when I enquired with the ‘contact us’ email.

They are run by volunteers, and were funded by the LTFF.

If you want to donate the web page is here.

FLI: The Future of Life Institute

FLI is a Boston-based independent existential risk organization, focusing on outreach, founded in large part to help organise the regranting of $10m from Elon Musk.

They have a podcast on AI Alignment here, and ran the Beneficial AI conference in January.

One of their big projects this year has been promoting the stigmatisation of, and ultimately the banning of, Lethal Autonomous Weapons. As well as possibly being good for its own sake, this might help build institutional capacity to ban potentially dangerous technologies that transfer autonomous away from humans. You can read their statement on the subject to the UN here. On the other hand, the desirability of this policy is not entirely uncontroversial – see for example Bogosian’s On AI Weapons. There is also lengthy discussion by Sterbenz and Trager here.

Krakovna's ICLR Safe ML Workshop Report summarises the results from a workshop on safety that Victoria co-ran at ICLR. You can see a list of all the papers here. #ML_safety

AIImpacts

AIImpacts is a Berkeley based AI Strategy organisation founded in 2014 by Katja Grace. They are affiliated with (a project of, with independent financing from) MIRI. They do various pieces of strategic background work, especially on AI Timelines - it seems their previous work on the relative rarity of discontinuous progress has been relatively influential. Their research can be found here.

Research

Katja impressed upon me that most of their work this year went into as-yet-unpublished work, but this is what is public:

Long & Davis's Conversation with Ernie Davis is an interview transcript with Davis, an NYU computer science professor who is an AI risk sceptic. Unfortunately I didn't think they quite got into the heart of the disagreement - they seem to work out the cruz is how much power superior intelligence gives you, but then move on. #Forecasting

Long & Bergal's Evidence against current methods leading to human level artificial intelligence lists a variety of arguments for why current AI techniques are insufficient for AGI. It's basically a list of 'things AI might need that we don't have yet', a lot of which coming from Marcus's Critical Appraisal. #Forecasting

Korzekwa's The unexpected difficulty of comparing AlphaStar to humans analyses AlphaStar's performance against human StarCraft players. It convincingly, in my inexpert judgement, argues that the 'unfair' advantages of AlphaStar - like the clicks-per-minute rate, and lack of visibility restrictions - we significant contributors to AlphaStar's success. As such, on an apples-to-apples basis it seems that humans have not yet been defeated at Starcraft. #Misc

AI Impacts's Historical Economic Growth Trends argues that historically economic growth has been super-linear in population size. As such we should expect accelerating growth 'by default' - "Extrapolating this model implies that at a time when the economy is growing 1% per year, growth will diverge to infinity after about 200 years". This is very interesting to me as it contradicts what I suggested here. Notably growth has slowed since 1950, perhaps for anthropic reasons. #Forecasting

AI Impacts's AI Conference Attendance plots attendance at the major AI conferences over time to show the recent rapid growth in the field using a relatively stable measure. #Forecasting

Finances

They spent $316,398 in 2019, and plan to spend around $325,000 in 2020. They have around $269,590 in cash and pledged funding, suggesting (on a very naïve calculation) around 0.8 years of runway.

In the past they have received support from EA organisations like OpenPhil and FHI.

MIRI administers their finances on their behalf; donations can be made here.

GPI: The Global Priorities Institute

GPI is an Oxford-based Academic Priorities Research organisation founded in 2018 by Hilary Greaves and part of Oxford University. They do work on philosophical issues likely to be very important for global prioritisation, much of which is, in my opinion, relevant to AI Alignment work. Their research can be found here.

Research

MacAskill (article) & Demski (extensive comments)'s A Critique of Functional Decision Theory gives some criticisms of FDT. He makes a variety of arguments, though I generally found them unconvincing. For example, the 'Bomb' example seemed to be basically question-begging on Newcomb's problem, and his Scots vs English example (where Scottish people choose to one-box because of their ancestral memory of the Darien scheme) seems to me to be a case of people not actually employing FDT at all. And some of his arguments - like that it is too complicated for humans to actually calculate - seem like the same arguments he would reject as criticisms of utilitarianism, and not relevant to someone working on AGI. I listed this as co-written by Abram Demski because he is acknowledged in the post, and his comments at the bottom are as detailed as worthy as the main post itself, and I recommend reading the two together. Researchers from MIRI were also named authors on the paper. #Decision_Theory

Greaves & Cotton-Barratt's A bargaining-theoretic approach to moral uncertainty lays out formalism and discusses using Nash Equilibrium between 'negotiating' moral values as an alternative approach to moral uncertainty. It discusses some subtle points about the selection of the BATNA outcome. One interesting section was on small vs grand worlds - whether splitting the world up into sub-dilemmas made a difference. For expected-value type approaches the answer is no, but for negotiating strategies the answer is yes, because the different moral theories might trade so as to influence the dilemmas that mattered most to them. This reminded me of an argument from Wei Dai that agents who cared about total value, finding themselves in a small world, might acausally trade with average value agents in large worlds. Presumably a practical implication might be that EAs should adhere to conventional moral standards with even higher than usual moral fidelity, in exchange for shutting up and multiplying on EA issues. The paper also makes interesting points about the fanaticism objection and the difference between moral and empirical risk. Researchers from FHI were also named authors on the paper. #Decision_Theory

MacAskill et al.'s The Evidentialist's Wager argues that Decision-Theoretic uncertainty in a large universe favours EDT over CDT. This is because your decision only has local causal implications, but global evidential implications. The article then goes into detail motivating the idea and discussing various complications and objections. It seems to push EDT in an FDT-direction, though presumably they still diverge on smoking lesion questions. Researchers from FHI, FRI were also named authors on the paper. #Decision_Theory

Mogensen's 'The only ethical argument for positive δ '? argues that positive pure time preference could be justified through agent-relative obligations. This was an interesting paper to me, and suggests some interesting (extremely speculative) questions - e.g. can we, by increasing out relatedness to our ancestors, acausally influence them into treating us better? #Philosophy

Mogensen's Doomsday rings twice attempts to salvage the Doomsday argument by suggesting we should update using SSA twice. He argues the second such update - on the fact that the present-day seems unusually influential - cannot be 'cancelled out' by SIA. #Philosophy

Finances

They spent £600,000 in 2018/2019 (academic year), and plan to spend around £1,400,000 in 2019/2020. They suggested that as part of Oxford University 'cash on hand' or 'runway' were not really meaningful concepts for them, as they need to fully-fund all employees for multiple years.

If you want to donate to GPI, you can do so here.

FRI: The Foundational Research Institute

FRI is a London (previously Germany) based Existential Risk Research organisation founded in 2013 currently led by Stefan Torges and Jonas Vollmer. They are part of the Effective Altruism Foundation (EAF) and do research on a number of fundamental long-term issues, some related how to reduce the risks of very bad AGI outcomes.

In general they adopt what they refer to as ‘suffering-focused’ ethics, which I think is a quite misguided view. However, they seem to have approached this thoughtfully.

Apparently this year they are more focused on research, vs movement-building and donation-raising in previous years.

Research

FRI researchers were not lead author on any work directly relevant to AI Alignment (unlike last year, where they had four papers).

FRI researchers contributed to the following research led by other organisations:

- MacAskill et al.'s The Evidentialist's Wager

Finances

EAF (of which they are a part) spent $836,622 in 2018 and $1,125,000 in 2019, and plan to spend around $995,000 in 2020. They have around $1,430,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1.4 years of runway.

According to their website, their finances are not separated from those of the EAF, and it is not possible to ear-mark donations. In the past this has made me worry about fungibility; donations funding other EAF work. However apparently EAF basically doesn’t do anything other than FRI now.

If you wanted to donate to FRI, you could do so here.

Median Group

Median is a Berkeley based independent AI Strategy organisation founded in 2018 by Jessica Taylor, Bryce Hidysmith, Jack Gallagher, Ben Hoffman, Colleen McKenzie, and Baeo Maltinsky. They do research on various risks, including AI timelines. Their research can be found here.

Research

Maltinsky et al.'s Feasibility of Training an AGI using Deep RL:A Very Rough Estimate build a model for how plausible one method of achieving AGI is. The theory is that you could basically simulate a bunch of people and have them work on the problem. Their model suggests this is not a credible way of producing AGI in the near term. I like the way they included their code in the actual report. #Forecasting

Taylor et al.'s Revisiting the Insights model improved their Insights model from last year. If you recall this basically used a pareto distribution for of many genius insights were required to get us to AGI. #Forecasting

The following was written by Jessica but not as an official Median piece:

Taylor's The AI Timelines Scam argues that there are systematic biases that lead people to exaggerate how short AI timelines are. One is that people who espouse short timelines tend to also argue for some amount of secrecy due to Infohazards, which makes their work hard for outsiders to audit. A second is that capital allocators tend to fund those who dream BIG, leading to systematic exaggeration of your field's potential. I think both are reasonable points, but I think she is too quick to use the term 'scam' - as in Scott's Against Lie Inflation. Specifically, while it is true that secrecy is a great cover for mediocrity, it is unfortunately also exactly what a morally virtuous agent would have to do in the presence of infohazards. Indeed, such people might be artificially limited in what they can say, making short time horizons appear artificially devoid of credible arguments. I am more sympathetic to her second argument, but even there to the extent that 1) fields select for people who believe in them and 2) people believe what is useful for them to believe I think it is a bit harsh to call it a 'scam'. #Forecasting

Finances

They spent ~$0 in 2018 and 2019, and plan to spend above $170000 in 2020. They have around $170000 in cash and pledged funding, suggesting (on a very naïve calculation) under 1 years of runway.

Median doesn’t seem to be soliciting donations from the general public at this time.

CSET: The Center for Security and Emerging Technology

CSET is a Washington based Think Tank founded in 2019 by Jason Matheny (ex IARPA), affiliated with the University of Georgetown. They analyse new technologies for their security implications and provide advice to the US government. At the moment they are mainly focused on near-term AI issues. Their research can be found here.

As they apparently launched with $55m from the Open Philanthropy Project, and subsequently raised money from the Hewlett Foundation, I am assuming they do not need more donations at this time.

Leverhulme Center for the Future of Intelligence

Leverhulme is a Cambridge based Research organisation founded in 2015 and currently led by Stephen Cave. They are affiliated with Cambridge University and closely liked to CSER. They do work on a variety of AI related causes, mainly on near-term issues but also some long-term. You can find their publications here.

Research

Leverhulme-affiliated researchers produced work on a variety of topics; I have only here summarised that which seemed the most relevant.

Hernandez-Orallo et al.'s Surveying Safety-relevant AI Characteristics provides a summary of the properties of AI systems that are relevant for safety. This includes both innate properties of the system (like ability to self-modify or influence its reward signal) and of the environment. Some of these characteristics are relatively well-established in the literature, but others seemed relatively new (to me at least). A few but not most seemed only really relevant to near-time safety issues (like the need for spare batteries). Researchers from CSER, Leverhulme were also named authors on the paper. #Overview

Cave & Ó hÉigeartaigh's Bridging near- and long-term concerns about AI attempt to unify short-term and long-term AI risk concerns. For example, they argue that solving short-term issues can help with long-term ones, and that long-term issues will eventually become short-term issues. However, I am inclined to agree with the review here by Habryka that a lot of the work here is being done by categorising unemployment and autonomous vehicles as long-term, and then arguing that they share many features with short-term issues. I agree that they have a lot in common; however this seems to be because unemployment and cars are also short-term issues - or short-term non-issues in my mind. The paper does not present a compelling argument for why short-term issues have a lot in common with existential risk work, which is what we care about. But perhaps this is being too harsh, and the paper is better understood performatively; it is not attempting to argue that the two camps are naturally allied, but rather attempting to make them allies. Researchers from CSER, Leverhulme were also named authors on the paper. #Strategy

Whittlestone et al.'s The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions points out that many of the 'values' that laypeople say AI systems should observe, like 'fairness', are frequently in conflict. This is certainly a big improvement over the typical article on the subject. #Shortterm

Leverhulme researchers contributed to the following research led by other organisations:

- Ovadya & Whittlestone's Reducing Malicious Use of Synthetic Media Research: Considerations and Potential Release Practices for Machine Learning

BERI: The Berkeley Existential Risk Initiative

BERI is a Berkeley-based independent Xrisk organisation, founded and led by Andrew Critch. They provide support to various university-affiliated (FHI, CSER, CHAI) existential risk groups to facilitate activities (like hiring engineers and assistants) that would be hard within the university context, alongside other activities - see their FAQ for more details.

Grants

BERI used to run a grant-making program where they helped Jaan Tallinn allocate money to Xrisk causes. Midway through this year, BERI decided to hand this off to the Survival and Flourishing Fund, a donor-advised fund currently advised by the same team who run BERI.

In this time period (December 2018-November 2019) BERI granted $1,615,933, mainly to large Xrisk organisations. The largest single grant was $600,000 to MIRI.

Research

A number of papers we reviewed this year were supported by BERI, for example:

- Turner et al.'s Conservative Agency
- O'Keefe's Stable Agreements in Turbulent Times: A Legal Toolkit for Constrained Temporal Decision Transmission
- Cihon's Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development

Because this support tended not to be mentioned on the front page of the article (unlike direct affiliation) it is quite possible that I missed other papers they supported also.

Finances

BERI have told me they are not seeking public support at this time. If you wanted to donate anyway their donate page is here.

AI Pulse

The Program on Understanding Law, Science, and Evidence (PULSE) is part of the UCLA School of Law, and contains a group working on AI policy. They were founded in 2017 with a $1.5m grant from OpenPhil.

Their website lists a few pieces of research, generally on more near-term AI policy issues. A quick read suggested they were generally fairly well done. However, they don't seem to have uploaded anything since February.

Research

Sterbenz & Trager's Autonomous Weapons and Coercive Threats discusses the impact of Lethal Autonomous Weapons on diplomacy. #Shortterm

Grotto's Genetically Modified Organisms: A Precautionary Tale for AI Governance discusses the history of GMO regulation in the US and EU. He brings up some interesting points about the highly contingent history behind the different approaches taken. However, I am somewhat sceptical GMOs are that good a comparison, given their fundamentally different nature. #Strategy

Other Research

I would like to emphasize that there is a lot of research I didn't have time to review, especially in this section, as I focused on reading organisation-donation-relevant pieces. So please do not consider it an insult that your work was overlooked!

Naude & Dimitri's The race for an artificial general intelligence: implications for public policy extends the model in Racing to the Precipice (Armstrong et al.) After a lengthy introduction to AI alignment, they make a formal model, concluding that a winner-take-all contest will have very few teams competing (which is good) Interestingly if the teams are concerned about cost minimisation this result no longer holds, as the 'best' team might not invest 100%, so the second-best team still has a chance, but the presence of intermediate prizes is positive, as they incentivise more investment. They suggest public procurement to steer AI development in a safe direction, and an unsafety-tax. (as a very minor aside, I was a little surprised to see the AIImpacts survey cited as a source for expected Singularity timing given that it does not mention the word.) Overall I thought this was an excellent paper. #Strategy

Steinhardt's AI Alignment Research Overview provides a detailed account of the different components of AI Alignment work. I think this probably takes over from Amodei et al.'s Concrete Problems (on which Jacob was a co-author) as my favour introduction to technical work, for helping new researchers locate themselves, with the one proviso that it is only in Google Docs form at the moment. He provides a useful taxonomy, goes into significant detail on the different problems, and suggests possible avenues of attack. The only area that struck me as a little light was on some of the MIRI-style agent foundations issues. Overall I thought this was an excellent paper. #Overview

Piper's The case for taking AI seriously as a threat to humanity is an introduction to AI safety for Vox readers. In my opinion it is the best non-technical introduction to the issue I have seen. It has become my go-to for linking people and reading groups. The article does a good job introducing the issues in a persuasive and common-sense way without much loss of fidelity. My only gripe is the article unquestioningly repeats an argument about criminal

justice 'discrimination' which has, in my opinion, been debunked (see here and the Washington Post article linked at the bottom), but perhaps this is a necessary concession when writing for Vox, and is only a very small part of the article. Overall I thought this was an excellent paper. #Introduction

Cohen et al.'s Asymptotically Unambitious Artificial General Intelligence ambitiously aims to provide an aligned AI algorithm. They do this by basically using an extremely myopic form of boxed oracle AIXI, that doesn't care about any rewards after the box has been opened - so all it cares about is getting rewards for answering the question well inside the box. It is indifferent to what the human does with the reward once outside the box. This assumes the AIXI cannot influence the world without detectably opening the box. This also aims to avoid the reward-hacking problems of AIXI. You might also enjoy the comments here. #AI_Theory

Snyder-Beattie et al.'s An upper bound for the background rate of human extinction uses a Laplace's law of succession-style approach to bound non-anthropogenic Xrisk. Given how long mankind has survived so far, they conclude that this is extremely unlikely to be greater than 1/14000, and probably much lower. Notably, they argue that these estimates are *not* significantly biased by anthropic issues, because high base extinction rates mean lucky human observers would be clustered in worlds where civilisation also developed very quickly, and hence also observe short histories. Obviously they can only provide an upper bound using such methods, so I see the paper as mainly providing evidence we should instead focus on anthropogenic risks, for which no such bound can exist. Researchers from FHI were also named authors on the paper. #Forecasting

Dai's Problems in AI Alignment that philosophers could potentially contribute to provides a list of open philosophical questions that matter for AI safety. This seems useful insomuch as there are people capable of working on many different philosophical issues and willing to be redirected to more useful ones. #Overview

Dai's Two Neglected Problems in Human-AI Safety discusses two danger modes for otherwise benign-seeming approval-orientated AIs. I thought this was good as it is potentially a very 'sneeky' way in which human value might be lost, at the hands of agents which otherwise appeared extremely corrigible etc. #Forecasting

Agrawal et al.'s Scaling up Psychology via Scientific Regret Minimization:A Case Study in Moral Decision-Making suggests that, in cases with large amounts data plus noise, human-interpretable models could be evaluated relative to ML predictions rather than the underlying data directly. In particular, they do this with the big Moral Machine dataset, comparing simple human-interpretable rules (like humans are worth more than animals, or criminals are worth less) with their NN. This suggests a multi-step program for friendliness: 1) gather data 2) train ML on data 3) evaluate simple human-evaluable rules on ML 4) have humans evaluate these rules. Researchers from CHAI were also named authors on the paper. #Ethical_Theory

Krueger et al.'s Misleading Meta-Objectives and Hidden Incentives for Distributional Shift discusses the danger of RL agents being incentivised to induce distributional shift. This is in contrast to what I think of as the 'standard' worry about distributional shift, namely arising as a side effect of increasing agent optimisation power. They then introduce a model to demonstrate this behaviour, but I had a little trouble understanding exactly how this bit was meant to work. Researchers from Deepmind were also named authors on the paper. #ML_safety

Zhang & Dafoe's Artificial Intelligence: American Attitudes and Trends surveys the views of ordinary people about AI. They used YouGov, who I generally regard as one of the best polling agencies. The survey did a good job of showing that the general public is generally very ignorant and susceptible to framing effects. Respondents basically thought that everyone potential AI 'problem' was roughly equally important. When reading this I think it is worth keeping the general literature on voter irrationality in mind - e.g. Bryan Caplan's The

Myth of the Rational Voter or Scott's Noisy Poll Results and Reptilian Muslim Climatologists from Mars. Researchers from FHI were also named authors on the paper. #Politics

Cottier & Shah's Clarifying some key hypotheses in AI alignment is a map of the connections between different ideas in AI safety. Researchers from CHAI were also named authors on the paper. #Overview

Ovadya & Whittlestone's Reducing Malicious Use of Synthetic Media Research: Considerations and Potential Release Practices for Machine Learning discusses various ways of improving the safety of ML research release. While synth media is the titular subject, most of it is more general, with fairly detailed descriptions of various strategies. While I don't think synth media is very important, it could be useful for building norms in ML that would apply to AGI work also. The paper discusses bioethics at length, e.g. how they use IRBs. My personal impression of IRBs is they are largely pointless and have little to do with ethics, functioning mainly to slow things down and tick boxes, but then again that might be desirable for AI research! Researchers from CSER, Leverhulme were also named authors on the paper. #Security

Schwarz's On Functional Decision Theory is a blog post by one of the philosophers who reviewed Eliezer and Nate's paper on FDT. It explains his objections, and why the paper was rejected from the philosophy journal he was a reviewer for. The key thing I took away from it was that MIRI did not do a good job of locating their work within the broader literature - for example, he argues that FDT seems like it might actually be a special case of CDT as construed by some philosophers, which E&N should have addressed, and elsewhere he suggests E&N's criticisms of CDT and EDT present strawmen. He also made some interesting points, for example that it seems 'FDT will sometimes recommend choosing a particular act because of the advantages of choosing a different act in a different kind of decision problem'. However most of the substantive criticisms were not very persuasive to me. Some seemed to almost beg the question, and at other times he essentially faulted FDT for addressing directly issues which *any* decision theory will ultimately have to address, like logical counterfactuals, or what is a 'Fair' scenario. He also presented a scenario, 'Procreation', as an intended *Reductio* of FDT that actually seems to me like a scenario where FDT works better than CDT does. #Decision_Theory

LeCun et al.'s Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More was a public debate on Facebook between major figures in AI on the AI safety issue. Many of these have been prominently dismissive in the past, so this was good to see. Unfortunately a lot of the debate was not at a very high level. It seemed that the sceptics generally agreed it was important to work on AI safety, just that this work was likely to happen by default. #Misc

Dai's Problems in AI Alignment that philosophers could potentially contribute to provides a list of issues for philosophers who want to work on the cause without math backgrounds. I think this is potentially very useful if brought to the notice of the relevant people, as the topics on the list seem useful things to work on, and I can easily imagine people not being aware of all of them. #Overview

Walsh's End Times: A Brief Guide to the End of the World is a popular science book on existential risk. AI risk is one of the seven issues addressed, in an extended and well-researched chapter. While I might quibble with one or two points, overall I thought this was a good introduction. The main qualifier for your opinion here is how valuable you think outreach to the educated layman is. #Introduction

Szlam et al.'s Why Build an Assistant in Minecraft? suggest a research program for building an intelligent assistant for Minecraft. The program doesn't appear to be directly motivated by AI alignment, but it does seem unusual in the degree to which alignment-type-issues would have to be solved for it to succeed - thereby hopefully incentivising mainstream ML guys to work on them. In particular, they want the agent to be able to work out 'what you wanted'

from a natural language text channel, which is clearly linked to the Value Alignment problem, and similar issues like the higher optimisation power of the agent are likely to occur. The idea that the agent should be 'fun' is also potentially relevant! The authors also released an environment to make making these assistants easier. #Misc

Kumar et al.'s Failure Modes in Machine Learning is a Microsoft document discussing a variety of ways ML systems can go wrong. It includes both intentional (e.g. hacking) and unintentional (e.g. the sort of thing we worry about). #Misc

Sevilla & Moreno's Implications of Quantum Computing for Artificial Intelligence Alignment Research examines whether Quantum Computing would be useful for AI Alignment. They consider three relevant properties of QC and several approaches to AI Alignment, and conclude that QC is not especially relevant. #Forecasting

Collins's Principles for the Application of Human Intelligence analyses the problems of biased and non-transparent decision making by natural intelligence systems. #Shortterm

Capital Allocators

One of my goals with this document is to help donors make an informed choice between the different organisations. However, it is quite possible that you regard this as too difficult, and wish instead to donate to someone else who will allocate on your behalf. This is of course much easier; now instead of having to solve the *Organisation Evaluation Problem*, all you need to do is solve the dramatically simpler *Organisation Evaluator Organisation Evaluation Problem*.

A helpful map from Issa Rice shows how at the moment the community has only managed to achieve delegative funding chains 6 links long. If you donate to Patrick Brinich-Langlois, we can make this chain significantly longer! In reality this is a quite misleading way of phrasing the issue of course, as for most of these organisations the ‘flow-through’ is a relatively small fraction. I do think it is valid to be concerned about sub-optimally high levels of intermediation however, which if nothing else reduces donor control. This seems to me to be a weak argument against delegating donations.

LTFF: Long-term future fund

LTFF is a globally based EA grantmaking organisation founded in 2017, currently led by Matt Wage and affiliated with CEA. They are one of four funds set up by CEA to allow individual donors to benefit from specialised capital allocators; this one focuses on long-term future issues, including a large focus on AI Alignment. Their website is here. There are write-ups for their first two grant rounds in 2019 here and here, and comments here and here. Apparently they have done another $400,000 round since then but the details are not yet public.

In the past I have been sceptical of the fund, as it was run by someone who already had access to far more capital (OpenPhil), and the grants were both infrequent and relatively conservative – giving to large organisations that individual donors are perfectly capable of evaluating themselves. Over the last year, however, things have significantly changed. The fund is now run by four people, and the grants have been to a much wider variety of causes, many of which would simply not be accessible to individual donors.

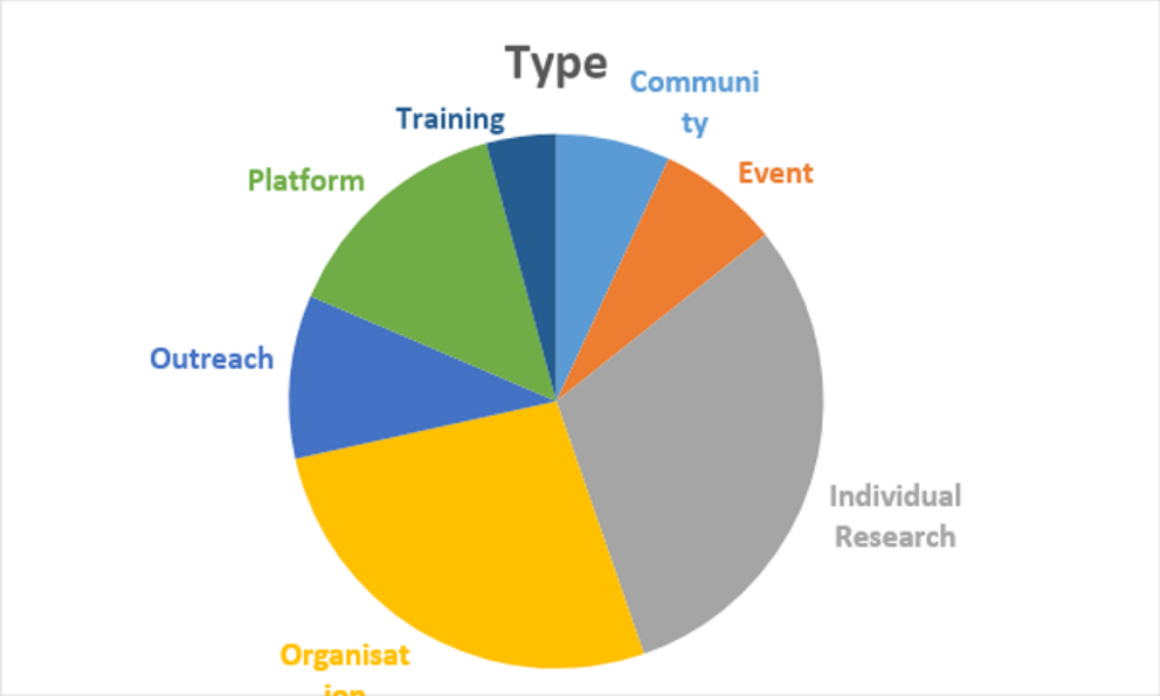
The fund managers are:

- Matt Wage
- Helen Toner
- Oliver Habryka
- Alex Zhu

Oliver Habryka especially has been admirably open with lengthy write-ups about his thoughts on the different grants, and I admire his commitment to intellectual integrity (you might enjoy his comments here). I am less familiar with the other fund managers. All the managers are, to my knowledge, unpaid.

In general most of the grants seem at least plausibly valuable to me, and many seemed quite good indeed. As there is extensive discussion in the links above I shan't discuss my opinions of individual grants in detail.

I attempted to classify the recommended (including those not accepted by CEA) by type and geography. Note that 'training' means paying an individual to self-study. I have deliberately omitted the exact percentages because this is an informal classification.

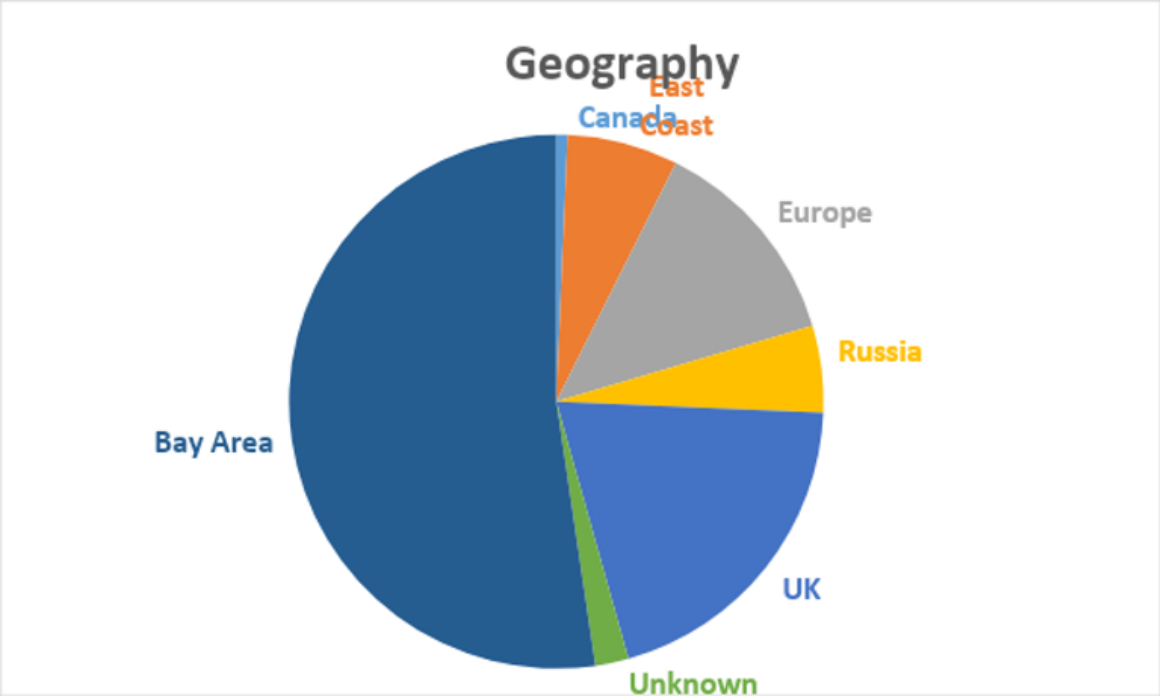


Of these categories, I am most excited by the Individual Research, Event and Platform projects. I am generally somewhat sceptical of paying people to 'level up' their skills.

I can understand why the fund managers gave over a quarter of the funds to major organisations – they thought these organisations were a good use of capital! However, to my mind this undermines the purpose of the fund. (Many) individual donors are perfectly capable of evaluating large organisations that publicly advertise for donations. In donating to the LTFF, I think (many) donors are hoping to be funding smaller projects that they could not directly access themselves. As it is, such donors will probably have to consider such organisation allocations a mild 'tax' – to the extent that different large organisations are chosen then they would have picked themselves.

For a similar analysis, see Gaensbauer's comment here. I think his 'counterfactually unique' (73%) roughly maps onto my 'non-organisation'.

CFAR, which the fund managers recommended $300,000, was the largest single intended beneficiary with just over 20% of the recommendations.



All grants have to be approved by CEA before they are made; historically they have approved almost all. In general I think these rejections improved the process. In every instance they were subsequently funded by private donors anyway, but this does not seem to be a problem for donors to the LTFF whose capital is protected. Notably this means the funds only paid out $150,000 to CFAR (10%), as the balance was made up by a private donor after CEA did not approve the second grant.

I was not impressed that one grant that saw harsh and accurate criticism on the forum after the first round was re-submitted for the second round. Ex post this didn't matter as CEA rejected it on substantive grounds the second time, but it makes me somewhat concerned about a risk of some of the capital going towards giving sinecures to people who are in the community, rather than objective merit. But if CEA will consistently block this waste maybe this is not such a big issue, and the grant in question only represented 1.3% of the total for the year.

If you wish to donate to the LTFF you can do so here.

OpenPhil: The Open Philanthropy Project

The Open Philanthropy Project (separated from Givewell in 2017) is an organisation dedicated to advising Cari and Dustin Moskovitz on how to give away over $15bn to a variety of causes, including existential risk. They have made extensive donations in this area and probably represent both the largest pool of EA-aligned capital and the largest team of EA capital allocators.

They also recently announced they would be working with Ben Delo as well.

This year they implemented a special committee for determining grants to EA-related organisations.

Grants

You can see their grants for AI Risk here. It lists only made four AI Risk grants in 2019, though I think that their $500k grant to ESPR (The European Summer Program on Rationality) should be considered an AI Risk relevant grant also.:

- OpenPhil AI Fellowship: $2.3m (write-up)
- MIRI: $2.7m (write-up)
- CSET: $55m (write-up)
- BERI / CHAI: $250k (write-up)

In contrast there are 11 AI Risk grants listed for 2018, though the total dollar value is lower.

The OpenPhil AI Fellowship basically fully funds AI PhDs for students who want to work on the long term impacts of AI. One thing that I had misunderstood previously was these fellowships are *not* intended to be specific to AI safety, though presumably their recipients are more likely to work on safety than the average ML PhD student. They funded 7 scholarships in 2018 and 8 in 2019

Due to a conflict of interest I cannot make any evaluation of their effectiveness.

Research

Most of their research concerns their own granting, and in an unusual failure of nominative determinism is non-public except for the short write-ups linked above.

Zabel & Muehlhauser's Information security careers for GCR reduction argue that working in InfoSec could be a useful career for reducing Xrisks, especially AI and Bio. This is partly to help prevent AGI/synth bio knowledge falling into the hands of malicious hackers (though most ML research seems to be very open), and partly because the field teaches various skills that are useful for AI safety, both high-level like Eliezer's Security Mindset and technical like crypto. They suggested that there was a shortage of such people willing to work on Xrisk right now, and perhaps in the future, due to lucrative alternative employment options. Researchers from Google Brain were also named authors on the paper. #Careers

Finances

To my knowledge they are not currently soliciting donations from the general public, as they have a lot of money from Dustin and Cari, so incremental funding is less of a priority than for other organisations. They could be a good place to work however!

SFF: The Survival and Flourishing Fund

SFF is a donor advised fund, advised by the people who make up BERI’s Board of Directors. SFF was initially funded in 2019 by a grant of approximately $2 million from BERI, which in turn was funded by donations from philanthropist Jaan Tallinn.

Grants

In its grantmaking SFF used an innovative allocation process to combine the views of many grant evaluators (described here). SFF has run two grant rounds thus far. The first ($880k in total) focused on large organisations:

- 80,000 Hours: $280k
- CFAR: $110k
- CSER: $40k
- FLI: $130k

- GCRI: $60k
- LessWrong: $260k

The second round, requiring written applications, distributed money to a much wider variety of projects. The website lists 28 recipients, of which many but not all were AI relevant. The largest grant was for $300k to the Longevity Research Institute.

Due to a conflict of interest I cannot evaluate the effectiveness of their grantmaking.

Other News

80,000 Hours's AI/ML safety research job board collects various jobs that could be valuable for people interested in AI safety. At the time of writing it listed 35 positions, all of which seemed like good options that it would be valuable to have sensible people fill. I suspect most people looking for AI jobs would find some on here they hadn't heard of otherwise, though of course for any given person many will not be appropriate. They also have job boards for other EA causes. #Careers

Brown & Sandholm's Superhuman AI for multiplayer poker present an AI that can beat professionals in non-limit Texas hold'em. My understanding was that this was seen as significantly harder than limit poker, so this represents something of a milestone. Unlike various Deepmind victories at classic games, this doesn't seem to have required much compute. #Misc

Chivers's The AI Does Not Hate You: Superintelligence, Rationality and the Race to Save the World is a journalistic examination of the rationalist community and the existential risk argument. I confess I haven't actually read the book, and have very low expectations for journalists in this regard, though Chivers is generally very good, and by all accounts this is a very fair and informative book. I've heard people recommend it as an explainer to their parents. #Introduction

EU's Ethics Guidelines for Trustworthy Artificial Intelligence is a series of ethics guidelines for AI in the EU. They received input from many groups, including CSER and Jaan Tallinn. They are (at this time) optional guidelines, and presumably will not apply to UK AI companies like Deepmind after Brexit. The guidelines seemed largely focused on banal statements about non-discrimination etc.; I could not find any mention of existential risk in the guidelines. In general I am not optimistic about political solutions and this did not change my mind. #Politics

Kaufman's Uber Self-Driving Crash convincingly argues that Uber was grossly negligent when their car hit and killed Elaine Herzberg last year. #Shortterm

Schmidt et al.'s National Security Commission on Artificial Intelligence Interim Report surveys AI from a US defence perspective. It contains a few oblique references to AI risk. #Politics

Cummings's On the referendum #31: Project Maven, procurement, lollapalooza results & nuclear/AGI safety discusses various important trends, including a sophisticated discussion of AGI safety. This is mainly noteworthy because the author is the mastermind of Brexit and the recent Conservative landslide in the UK, and perhaps the most influential man in the UK as a result. #Strategy

Methodological Thoughts

Inside View vs Outside View

This document is written mainly, but not exclusively, using publicly available information. In the tradition of active management, I hope to synthesise many pieces of individually well known facts into a whole which provides new and useful insight to readers. Advantages of this are that 1) it is relatively unbiased, compared to inside information which invariably favours those you are close to socially and 2) most of it is legible and verifiable to readers. The disadvantage is that there are probably many pertinent facts that I am not a party to! Wei Dai has written about how much discussion now takes place in private google documents – for example this Drexler piece apparently; in most cases I do not have access to these. If you want the inside scoop I am not your guy; all I can supply is exterior scooping.

Many capital allocators in the bay area seem to operate under a sort of Great Man theory of investment, whereby the most important thing is to identify a guy who is really clever and 'gets it'. I think there is some merit in this; however, I think I believe in it much less than they do. Perhaps as a result of my institutional investment background, I place a lot more weight on historical results. In particular, I worry that this approach leads to over-funding skilled rhetoricians and those the investor/donor is socially connected to.

Judging organisations on their historical output is naturally going to favour more mature organisations. A new startup, whose value all lies in the future, will be disadvantaged. However, I think that this is the correct approach for donors who are not tightly connected to the organisations in question. The newer the organisation, the more funding should come from people with close knowledge. As organisations mature, and have more easily verifiable signals of quality, their funding sources can transition to larger pools of less expert money. This is how it works for startups turning into public companies and I think the same model applies here. (I actually think that even those with close personal knowledge should use historical results more, to help overcome their biases.)

This judgement involves analysing a large number of papers relating to Xrisk that were produced during 2019. Hopefully the year-to-year volatility of output is sufficiently low that this is a reasonable metric; I have tried to indicate cases where this doesn't apply. I also attempted to include papers during December 2018, to take into account the fact that I'm missing the last month's worth of output from 2019, but I can't be sure I did this successfully.

This article focuses on AI risk work. If you think other causes are important too, your priorities might differ. This particularly affects GCRI, FHI and CSER, who both do a lot of work on other issues which I attempt to cover but only very cursorily.

We focus on papers, rather than outreach or other activities. This is partly because they are much easier to measure; while there has been a large increase in interest in AI safety over the last year, it's hard to work out who to credit for this, and partly because I think progress has to come by persuading AI researchers, which I think comes through technical outreach and publishing good work, not popular/political work.

Politics

My impression is that policy on most subjects, especially those that are more technical than emotional is generally made by the government and civil servants in consultation with, and being lobbied by, outside experts and interests. Without expert (e.g. top ML researchers in academia and industry) consensus, no useful policy will be enacted. Pushing directly for policy seems if anything likely to hinder expert consensus. Attempts to directly influence the government to regulate AI research seem very adversarial, and risk being pattern-matched to ignorant technophobic opposition to GM foods or other kinds of progress. We don't want the 'us-vs-them' situation that has occurred with climate change, to happen here. AI researchers who are dismissive of safety law, regarding it as an imposition and encumbrance to be endured or evaded, will probably be harder to convince of the need to voluntarily be extra-safe - especially as the regulations may actually be totally ineffective.

The only case I can think of where scientists are relatively happy about punitive safety regulations, nuclear power, is one where many of those initially concerned were scientists themselves. Given this, I actually think policy outreach to the general population is probably negative in expectation.

If you're interested in this, I'd recommend you read this blog post from last year.

Openness

I think there is a strong case to be made that openness in AGI capacity development is bad. As such I do not ascribe any positive value to programs to 'democratize AI' or similar.

One interesting question is how to evaluate non-public research. For a lot of safety research, openness is clearly the best strategy. But what about safety research that has, or potentially has, capabilities implications, or other infohazards? In this case it seems best if the researchers do not publish it. However, this leaves funders in a tough position – how can we judge researchers if we cannot read their work? Maybe instead of doing top secret valuable research they are just slacking off. If we donate to people who say "trust me, it's very important and has to be secret" we risk being taken advantage of by charlatans; but if we refuse to fund, we incentivize people to reveal possible infohazards for the sake of money. (Is it even a good idea to publicise that someone else is doing secret research?)

With regard to published research, in general I think it is better for it to be open access, rather than behind journal paywalls, to maximise impact. Reducing this impact by a significant amount in order for the researcher to gain a small amount of prestige does not seem like an efficient way of compensating researchers to me. Thankfully this does not occur much with CS papers as they are all on arXiv, but it is an issue for some strategy papers.

Similarly, it seems a bit of a waste to have to charge for books – ebooks have, after all, no marginal cost – if this might prevent someone from reading useful content. There is also the same ability for authors to trade off public benefit against private gain – by charging more for their book, they potentially earn more, but at the cost of lower reach. As a result, I am inclined to give less credit for market-rate books, as the author is already compensated and incentised by sales revenue.

More prosaically, organisations should make sure to upload the research they have published to their website! Having gone to all the trouble of doing useful research it is a constant shock to me how many organisations don't take this simple step to significantly increase the reach of their work. Additionally, several times I have come across incorrect information on organisation's websites.

Research Flywheel

My basic model for AI safety success is this:

1. Identify interesting problems
    1. As a byproduct this draws new people into the field through altruism, nerd-sniping, apparent tractability
2. Solve interesting problems
    1. As a byproduct this draws new people into the field through credibility and prestige
3. Repeat

One advantage of this model is that it produces both object-level work and field growth.

There is also some value in arguing for the importance of the field (e.g. Bostrom's Superintelligence) or addressing criticisms of the field.

Noticeably absent are strategic pieces. I find that a lot of these pieces do not add terribly much incremental value. Additionally, my suspicion strategy research is, to a certain extent, produced exogenously by people who are interested / technically involved in the field. This does not apply to technical strategy pieces, about e.g. whether CIRL or Amplification is a more promising approach.

There is somewhat of a paradox with technical vs 'wordy' pieces however: as a non-expert, it is much easier for me to understand and evaluate the latter, even though I think the former are much more valuable.

Differential AI progress

There are many problems that need to be solved before we have safe general AI, one of which is not producing *unsafe* general AI in the meantime. If nobody was doing non-safety-conscious research there would be little risk or haste to AGI – though we would be missing out on the potential benefits of safe AI.

There are several consequences of this:

- To the extent that safety research also enhances capabilities, it is less valuable.
- To the extent that capabilities research re-orientates subsequent research by third parties into more safety-tractable areas it is more valuable.
- To the extent that safety results would naturally be produced as a by-product of capabilities research (e.g. autonomous vehicles) it is less attractive to finance.

One approach is to research things that will make contemporary ML systems safer, because you think AGI will be a natural outgrowth from contemporary ML. This has the advantage of faster feedback loops, but is also more replaceable (as per the previous section).

Another approach is to try to reason directly about the sorts of issues that will arise with superintelligent AI. This work is less likely to be produced exogenously by unaligned researchers, but it requires much more faith in theoretical arguments, unmoored from empirical verification.

Near-term safety AI issues

Many people want to connect AI existential risk issues to 'near-term' issues; I am generally sceptical of this. For example, autonomous cars seem to risk only localised tragedies, and private companies should have good incentives here. Unemployment concerns seem exaggerated to me, as they have been for most of history (new jobs will be created), at least until we have AGI, at which point we have bigger concerns. Similarly, I generally think concerns about algorithmic bias are essentially political - I recommend this presentation - though there is at least some connection to the value learning problem there.

Financial Reserves

Charities like having financial reserves to provide runway, and guarantee that they will be able to keep the lights on for the immediate future. This could be justified if you thought that charities were expensive to create and destroy, and were worried about this occurring by accident due to the whims of donors. Unlike a company which sells a product, it seems reasonable that charities should be more concerned about this.

Donors prefer charities to not have too much reserves. Firstly, those reserves are cash that could be being spent on outcomes now, by either the specific charity or others. Valuable future activities by charities are supported by future donations; they do not need to be pre-funded. Additionally, having reserves increases the risk of organisations 'going rogue', because they are insulated from the need to convince donors of their value.

As such, in general I do not give full credence to charities saying they need more funding because they want much more than a 18 months or so of runway in the bank. If you have a year's reserves now, after this December you will have that plus whatever you raise now, giving you a margin of safety before raising again next year.

I estimated reserves = (cash and grants) / (2020 budget). In general I think of this as something of a measure of urgency. However despite being prima facie a very simple calculation there are many issues with this data. As such these should be considered suggestive only.

Donation Matching

In general I believe that charity-specific donation matching schemes are somewhat dishonest, despite my having provided matching funding for at least one in the past.

Ironically, despite this view being espoused by GiveWell (albeit in 2011), this is essentially of OpenPhil's policy of, at least in some cases, artificially limiting their funding to 50% or 60% of a charity's need, which some charities have argued effectively provides a 1:1 match for outside donors. I think this is bad. In the best case this forces outside donors to step in, imposing marketing costs on the charity and research costs on the donors. In the worst case it leaves valuable projects unfunded.

Obviously cause-neutral donation matching is different and should be exploited. Everyone should max out their corporate matching programs if possible, and things like the annual Facebook Match continue to be great opportunities.

Poor Quality Research

Partly thanks to the efforts of the community, the field of AI safety is considerably more well respected and funded than was previously the case, which has attracted a lot of new researchers. While generally good, one side effect of this (perhaps combined with the fact that many low-hanging fruits of the insight tree have been plucked) is that a considerable amount of low-quality work has been produced. For example, there are a lot of papers which can be accurately summarized as asserting "just use ML to learn ethics". Furthermore, the conventional peer review system seems to be extremely bad at dealing with this issue.

The standard view here is just to ignore low quality work. This has many advantages, for example 1) it requires little effort, 2) it doesn't annoy people. This conspiracy of silence seems to be the strategy adopted by most scientific fields, except in extreme cases like anti-vaxers.

However, I think there are some downsides to this strategy. A sufficiently large milieu of low-quality work might degrade the reputation of the field, deterring potentially high-quality contributors. While low-quality contributions might help improve Concrete Problems' citation count, they may use up scarce funding.

Moreover, it is not clear to me that 'just ignore it' really generalizes as a community strategy. Perhaps you, enlightened reader, can judge that *"How to solve AI Ethics: Just use RNNs"* is not great. But is it really efficient to require everyone to independently work this out? Furthermore, I suspect that the idea that we can all just ignore the weak stuff is somewhat

an example of typical mind fallacy. Several times I have come across people I respect according respect to work I found clearly pointless. And several times I have come across people I respect arguing persuasively that work I had previously respected was very bad – but I only learnt they believed this by chance! So I think it is quite possible that many people will waste a lot of time as a result of this strategy, especially if they don't happen to move in the right social circles.

Having said all that, I am not a fan of unilateral action, and am somewhat selfishly conflict-averse, so will largely continue to abide by this non-aggression convention. My only deviation here is to make it explicit. If you're interested in this you might enjoy this by 80,000 Hours.

The Bay Area

Much of the AI and EA communities, and especially the EA community concerned with AI, is located in the Bay Area, especially Berkeley and San Francisco. This is an extremely expensive place, and is dysfunctional both politically and socially. Aside from the lack of electricity and aggressive homelessness, it seems to attract people who are extremely weird in socially undesirable ways – and induces this in those who move there - though to be fair the people who are doing useful work in AI organisations seem to be drawn from a better distribution than the broader community. In general I think the centralization is bad, but if there must be centralization I would prefer it be almost anywhere other than Berkeley. Additionally, I think many funders are geographically myopic, and biased towards funding things in the Bay Area. As such, I have a mild preference towards funding non-Bay-Area projects. If you're interested in this topic I recommend you read this or this or this.

Conclusions

The size of the field continues to grow, both in terms of funding and researchers. Both make it increasingly hard for individual donors. I've attempted to subjectively weigh the productivity of the different organisations against the resources they used to generate that output, and donate accordingly.

My constant wish is to promote a lively intellect and independent decision-making among readers; hopefully my laying out the facts as I see them above will prove helpful to some readers. Here is my eventual decision, rot13'd so you can do come to your own conclusions first (which I strongly recommend):

Qrfcvgr univat qbangrq gb ZVEV pbafvfgragyl sbe znal lrnef nf n erfhyg bs gurve uvtuyl aba-ercynprnoyr naq tebhaqoernxvat jbex va gur svryq, V pnaabg va tbbq snvgu qb fb guvf lrne tvira gurve ynpx bs qvfpybfher. Nqqvgvbanyyl, gurl nyernql unir n dhvgr ynetr ohqtrg.

SUV unir pbafvfgragyl cebqhprq fbzr bs gur uvturfg dhnyvgl erfrnepu. Ubjrire, V nz abg pbaivaprq gurl unir n uvtu arrq sbe nqqvgvbany shaqvat.

V pbagvahr gb or vzcerffrq jvgu PUNV'f bhgchg, naq guvax gurl cbgragvnyyl qb n tbbq wbo vagrenpgvat jvgu znvafgernz ZY erfrnepuref. Gurl unir n ybg bs pnfu erfreirf, juvpu frrzf yvxr vg zvtug erqhpr gur hetrapl bs shaqvat fbzrjung, naq n pbafvqrenoyr cbegvba bs gur jbex vf ba zber arne-grez vffhrf, ohg gurer ner eryngviryl srj bccbeghavgvrf gb shaq grpuavpny NV fnsrgl jbex, fb V vagraq gb qbangr gb PUNV ntnva guvf lrne.

Qrrczvaq naq BcraNV obgu qb rkpryyrag jbex ohg V qba'g guvax vg vf ivnoyr sbe (eryngviryl) fznyy vaqvivqhny qbabef gb zrnavatshyyl fhccbeg gurve jbex.

Va gur cnfg V unir orra irel vzcerffrq jvgu TPEV'f bhgchg ba n ybj ohqtrg. Qrfcvgr vagraqvat 2019 vagraqvat gb or gurve lrne bs fpnyvat hc, bhgchg unf npghnyyl qrpernfrq. V fgvyy

vagraq gb znxr n qbangvba, va pnfr guvf vf whfg na hasbeghangr gvzvat vffhr, ohg qrsvavgryl jbhyq jnag gb frr zber arkg lrne.

PFRE’f erfrnepu vf whfg abg sbphfrq rabhtu gb jneenag qbangvbaf sbe NV Evfx jbex va zl bcvavba.

V jbhyq pbafvqre qbangvat gb gur NV Fnsrgl Pnzc vs V xarj zber nobhg gurve svanaprf.

Bhtug frrzf yvxr n irel inyhnoyr cebwrpg, naq yvxr PUNV ercerfragf bar bs gur srj bccbeghavgvrf gb qverpgyl shaq grpuavpny NV fnsrgl jbex. Nf fhpu V guvax V cyna gb znxr n qbangvba guvf lrne.

V gubhtug NV Vzcnpgf qvq fbzr avpr fznyy cebwrpgf guvf lrne, naq ba n abg ynetr ohqtrg. V guvax V jbhyq yvxr gb frr gur erfhygf sebz gurve ynetr cebwrpgf svefg ubjrire.

Va n znwbe qvssrerapr sebz cerivbhf lrnef, V npghnyyl cyna gb qbangr fbzr zbarl gb gur Ybat Grez Shgher Shaq. Juvyr V unira’g nterrq jvgu nyy gurve tenagf, V guvax gurl bssre fznyy qbabef npprff gb n enatr bs fznyy cebwrpgf gung gurl pbhyq abg bgurejvfr shaq, juvpu frrzf irel inyhnoyr pbafvqrevat gur fgebat svanapvny fvghngvba bs znal bs gur orfg ynetre betnavfngvbaf (BcraNV, Qrrczvaq rgp.)

Bar guvat V jbhyq yvxr gb frr zber bs va gur shgher vf tenagf sbe CuQ fghqragf jub jnag gb jbex va gur nern. Hasbeghangryl ng cerfrag V nz abg njner bs znal jnlf sbe vaqvivqhny qbabef gb cenpgvpnyyl fhccbeg guvf.

However, I wish to emphasize that all the above organisations seem to be doing good work on the most important issue facing mankind. It is the nature of making decisions under scarcity that we must prioritize some over others, and I hope that all organisations will understand that this necessarily involves negative comparisons at times.

Thanks for reading this far; hopefully you found it useful. Apologies to everyone who did valuable work that I excluded!

If you found this post helpful, and especially if it helped inform your donations, please consider letting me and any organisations you donate to as a result know.

If you are interested in helping out with next year’s article, please get in touch, and perhaps we can work something out.

Disclosures

I have not in general checked all the proofs in these papers, and similarly trust that researchers have honestly reported the results of their simulations.

I was a Summer Fellow at MIRI back when it was SIAI and volunteered briefly at GWWC (part of CEA). I have conflicts of interest with the Survival and Flourishing Fund and OpenPhil so have not evaluated them. I have no financial ties beyond being a donor and have never been romantically involved with anyone who has ever worked at any of the other organisations.

I shared drafts of the individual organisation sections with representatives from FHI, CHAI, MIRI, GCRI, BERI, Median, CSER, GPI, AISC, BERI, AIImpacts, FRI and Ought.

My eternal gratitude to Greg Lewis, Jess Riedel, Hayden Wilkinson, Kit Harris and Jasmine Wang for their invaluable reviewing. Any remaining mistakes are of course my own. I would also like to thank my wife and daughter for tolerating all the time I have spent/invested/wasted on this.

Sources

80,000 Hours - AI/ML safety research job board - 2019-09-29 - https://80000hours.org/job-board/ai-ml-safety-research/

Agrawal, Mayank; Peterson, Joshua; Griffiths, Thomas - Scaling up Psychology via Scientific Regret Minimization:A Case Study in Moral Decision-Making - 2019-10-16 - https://arxiv.org/abs/1910.07581

AI Impacts - AI Conference Attendance - 2019-03-06 - https://aiimpacts.org/ai-conference-attendance/

AI Impacts - Historical Economic Growth Trends - 2019-03-06 - https://aiimpacts.org/historical-growth-trends/

Alexander, Scott - Noisy Poll Results And Reptilian Muslim Climatologists from Mars - 2013-04-12 - https://slatestarcodex.com/2013/04/12/noisy-poll-results-and-reptilian-muslim-climatologists-from-mars/

Armstrong, Stuart - Research Agenda v0.9: Synthesising a human's preferences into a utility function - 2019-06-17 - https://www.lesswrong.com/posts/CSEdLLEkap2pubjof/research-agenda-v0-9-synthesising-a-human-s-preferences-into#comments

Armstrong, Stuart; Bostrom, Nick; Shulman, Carl - Racing to the precipice: a model of artificial intelligence development - 2015-08-01 - https://link.springer.com/article/10.1007%2Fs00146-015-0590-y

Armstrong, Stuart; Mindermann, Sören - Occam's razor is insufficient to infer the preferences of irrational agents - 2017-12-15 - https://arxiv.org/abs/1712.05812

Aschenbrenner, Leopold - Existential Risk and Economic Growth - 2019-09-03 - https://leopoldaschenbrenner.github.io/xriskandgrowth/ExistentialRiskAndGrowth050.pdf

Avin, Shahar - Exploring Artificial Intelligence Futures - 2019-01-17 - https://www.shaharavin.com/publication/pdf/exploring-artificial-intelligence-futures.pdf

Avin, Shahar; Amadae, S - Autonomy and machine learning at the interface of nuclear weapons, computers and people - 2019-05-06 - https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf

Baum, Seth - Risk-Risk Tradeoff Analysis of Nuclear Explosives for Asteroid Deflection - 2019-06-13 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3397559

Baum, Seth - The Challenge of Analyzing Global Catastrophic Risks - 2019-07-15 - https://higherlogicdownload.s3.amazonaws.com/INFORMS/f0ea61b6-e74c-4c07-894d-884bf2882e55/UploadedImages/2019_July.pdf#page=20

Baum, Seth; de Neufville, Robert; Barrett, Anthony; Ackerman, Gary - Lessons for Artificial Intelligence from Other Global Risks - 2019-11-21 - http://gcrinstitute.org/papers/lessons.pdf

Beard, Simon - Perfectionism and the Repugnant Conclusion - 2019-03-05 - https://link.springer.com/article/10.1007/s10790-019-09687-4

Beard, Simon - What Is Unfair about Unequal Brute Luck? An Intergenerational Puzzle - 2019-01-21 - https://www.cser.ac.uk/resources/brute-luck-intergenerational-puzzle/

Belfield, Haydn - How to respond to the potential malicious uses of artificial intelligence? - 2019-09-19 - https://www.cser.ac.uk/resources/how-respond-potential-malicious-uses-

artificial-intelligence/

Bogosian, Kyle - On AI Weapons - 2019-11-13 - https://forum.effectivealtruism.org/posts/vdqBn65Qaw77MpqXz/on-ai-weapons

Brown, Noam; Sandholm, Tuomas - Superhuman AI for multiplayer poker - 2019-07-17 - https://www.cs.cmu.edu/~noamb/papers/19-Science-Superhuman.pdf

Caplan, Bryan - The Myth of the Rational Voter - 2008-08-24 - https://www.amazon.com/Myth-Rational-Voter-Democracies-Policies/dp/0691138737

Carey, Ryan - How useful is Quantilization for Mitigating Specification-Gaming - 2019-05-06 - https://www.fhi.ox.ac.uk/wp-content/uploads/SafeML2019_paper_40.pdf

Carroll, Micah; Shah Rohin; Mark K Ho, Griffiths, Tom; Seshia,Sanjit; Abbeel,Pieter; Dragan, Anca - On the Utility of Learning about Humansfor Human-AI Coordination - 2019-10-22 - http://papers.nips.cc/paper/8760-on-the-utility-of-learning-about-humans-for-human-ai-coordination.pdf

Cave, Stephen; Ó hÉigeartaigh, Seán - Bridging near- and long-term concerns about AI - 2019-01-07 - https://www.nature.com/articles/s42256-018-0003-2

Chan, Lawrence; Hadfield-Menell, Dylan; Srinivasa, Siddhartha; Dragan, Anca - The Assistive Multi-Armed Bandit - 2019-01-24 - https://arxiv.org/abs/1901.08654

Chivers, Tom - The AI Does Not Hate You: Superintelligence, Rationality and the Race to Save the World - 2019-06-13 - https://www.amazon.com/Does-Not-Hate-You-Superintelligence-ebook/dp/B07K258VCV

Christiano, Paul - AI alignment landscape - 2019-10-12 - https://ai-alignment.com/ai-alignment-landscape-d3773c37ae38

Christiano, Paul - What failure looks like - 2019-03-17 - https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Cihon, Peter - Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development - 2019-05-16 - https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Clark, Jack; Hadfield, Gillian - Regulatory Markets for AI Safety - 2019-05-06 - https://drive.google.com/uc?export=download&id=1bFPiwLrZc7SQTMg2_bW4gt0PaS5NyqOH

Cohen, Michael; Vellambi, Badri; Hutter, Marcus - Asymptotically Unambitious Artificial General Intelligence - 2019-05-29 - https://arxiv.org/abs/1905.12186

Collins, Jason - Principles for the Application of Human Intelligence - 2019-09-30 - https://behavioralscientist.org/principles-for-the-application-of-human-intelligence/

Colvin, R; Kemp, Luke; Talberg, Anita; De Castella, Clare ; Downie, C; Friel, S; Grant, Will; Howden, Mark; Jotzo, Frank; Markham, Francis; Platow, Michael - Learning from the Climate Change Debate to AvoidPolarisation on Negative Emissions - 2019-07-25 - https://sci-hub.tw/10.1080/17524032.2019.1630463

Cottier, Ben; Shah, Rohin - Clarifying some key hypotheses in AI alignment - 2019-08-15 - https://www.lesswrong.com/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment

CSER - Policy series Managing global catastrophic risks: Part 1 Understand - 2019-08-13 - https://www.gcrpolicy.com/understand-overview

Cummings, Dominic - On the referendum #31: Project Maven, procurement, lollapalooza results & nuclear/AGI safety - 2019-03-01 - https://dominiccummings.com/2019/03/01/on-the-referendum-31-project-maven-procurement-lollapalooza-results-nuclear-agi-safety/

Dai, Wei - Problems in AI Alignment that philosophers could potentially contribute to - 2019-08-17 - https://www.lesswrong.com/posts/rASeoR7iZ9Fokzh7L/problems-in-ai-alignment-that-philosophers-could-potentially

Dai, Wei - Problems in AI Alignment that philosophers could potentially contribute to - 2019-08-17 - https://www.lesswrong.com/posts/rASeoR7iZ9Fokzh7L/problems-in-ai-alignment-that-philosophers-could-potentially

Dai, Wei - Two Neglected Problems in Human-AI Safety - 2018-12-16 - https://www.alignmentforum.org/posts/HTgakSs6JpnogD6c2/two-neglected-problems-in-human-ai-safety

Drexler, Eric - Reframing Superintelligence: Comprehensive AI Services as General Intelligence - 2019-01-08 - https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf?asd=sa

EU - Ethics Guidelines for Trustworthy Artificial Intelligence - 2019-04-08 - https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Everitt, Tom; Hutter, Marcus - Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective - 2019-08-13 - https://arxiv.org/abs/1908.04734

Everitt, Tom; Kumar, Ramana; Krakovna, Victoria; Legg, Shane - Modeling AGI Safety Frameworks with Causal Influence Diagrams - 2019-06-20 - https://arxiv.org/abs/1906.08663

Everitt, Tom; Ortega, Pedro; Barnes, Elizabeth; Legg, Shane - Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings - 2019-02-26 - https://arxiv.org/abs/1902.09980

Friedman, David D - Legal Systems Very Different from Ours - 1970-01-01 - https://www.amazon.com/Legal-Systems-Very-Different-Ours/dp/1793386722

Garfinkel, Ben & Dafoe, Allan - How does the offense-defense balance scale? - 2019-08-22 - https://www.tandfonline.com/doi/full/10.1080/01402390.2019.1631810

Greaves, Hilary; Cotton-Barratt, Owen - A bargaining-theoretic approach to moral uncertainty - 2019-08-09 - https://globalprioritiesinstitute.org/wp-content/uploads/2019/Greaves_Cotton-Barratt_bargaining_theoretic_approach.pdf

Grotto, Andy - Genetically Modified Organisms: A Precautionary Tale for AI Governance - 2019-01-24 - https://aipulse.org/genetically-modified-organisms-a-precautionary-tale-for-ai-governance-2/

Hernandez-Orallo, Jose; Martınez-Plumed, Fernando; Avin, Shahar; Ó hÉigeartaigh, Seán - Surveying Safety-relevant AI Characteristics - 2019-01-20 - http://ceur-ws.org/Vol-2301/paper_22.pdf

Hubinger, Evan; van Merwijk, Chris; Mikulik, Vladimir; Skalse, Joar; Garrabrant, Scott - Risks from Learned Optimization in Advanced Machine Learning Systems - 2019-06-05 - https://arxiv.org/abs/1906.01820

Irving, Geoffrey; Askell, Amanda - AI Safety Needs Social Scientists - 2019-02-19 - https://distill.pub/2019/safety-needs-social-scientists/

Irving, Geoffrey; Christiano, Paul; Amodei, Dario - AI Safety via Debate - 2018-05-02 - https://arxiv.org/abs/1805.00899

Kaczmarek, Patrick; Beard, Simon - Human Extinction and Our Obligations to thePast - 2019-11-05 - https://sci-hub.tw/https://www.cambridge.org/core/journals/utilitas/article/human-extinction-and-our-obligations-to-the-past/C29A0406EFA2B43EE8237D95AAFBB580

Kaufman, Jeff - Uber Self-Driving Crash - 2019-11-07 - https://www.jefftk.com/p/uber-self-driving-crash

Kemp, Luke - Mediation Without Measures: Conflict Resolution in Climate Diplomacy - 2019-05-15 - https://www.cser.ac.uk/resources/mediation-without-measures/

Kenton, Zachary; Filos, Angelos; Gal, Yarin; Evans, Owain - Generalizing from a few enviroments in Safety-Critical Reinforcement Learning - 2019-07-02 - https://arxiv.org/abs/1907.01475

Korzekwa, Rick - The unexpected difficulty of comparing AlphaStar to humans - 2019-09-17 - https://aiimpacts.org/the-unexpected-difficulty-of-comparing-alphastar-to-humans/

Kosoy, Vanessa - Delegative Reinforcement Learning: Learning to Avoid Traps with a Little Help - 2019-07-19 - https://arxiv.org/abs/1907.08461

Kovarik, Vojta; Gajdova, Anna; Lindner, David; Finnveden, Lukas; Agrawal, Rajashree - AI Safety Debate and Its Applications - 2019-07-23 - https://www.lesswrong.com/posts/5Kv2qNfRyXXihNrx2/ai-safety-debate-and-its-applications

Krakovna, Victoria - ICLR Safe ML Workshop Report - 2019-06-18 - https://futureoflife.org/2019/06/18/iclr-safe-ml-workshop-report/

Kruegar, David; Maharaj, Tegan; Legg, Shane; Leike, Jan - Misleading Meta-Objectives and Hidden Incentives for Distributional Shift - 2019-01-01 - https://drive.google.com/uc?export=download&id=1k93292JCoIHU0h6xVO3qmeRwLyOSlS4o

Kumar, Ram Shankar Siva; O'Brien, David; Snover, Jeffrey; Albert, Kendra; Viloen, Salome - Failure Modes in Machine Learning - 2019-11-10 - https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning

LeCun, Yann; Russell, Stuart; Bengio, Yoshua; Olds, Elliot; Zador, Tony; Rossi, Francesca; Mallah, Richard; Barzov, Yuri - Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More - 2019-10-04 - https://www.lesswrong.com/posts/WxW6Gc6f2z3mzmqKs/debate-on-instrumental-convergence-between-lecun-russell

Lewis, Sophie; Perkins-Kirkpatrick, Sarah; Althor, Glenn; King, Andrew; Kemp, Luke - Assessing contributions of major emitters' Paris-era decisions to future temperature extremes - 2019-03-20 - https://www.cser.ac.uk/resources/assessing-contributions-extremes/

Long, Robert; Bergal, Asya - Evidence against current methods leading to human level artificial intelligence - 2019-08-12 - https://aiimpacts.org/evidence-against-current-methods-leading-to-human-level-artificial-intelligence/

Long, Robert; Davis, Ernest - Conversation with Ernie Davis - 2019-08-23 - https://aiimpacts.org/conversation-with-ernie-davis/

Macaskill, Will; Demski, Abram - A Critique of Functional Decision Theory - 2019-09-13 - https://www.lesswrong.com/posts/ySLYSsNeFL5CoAQzN/a-critique-of-functional-decision-theory

MacAskill, William; Vallinder, Aron; Oesterheld, Caspar; Shulman, Carl; Treutlein, Johannes - The Evidentialist's Wager - 2019-11-19 - https://globalprioritiesinstitute.org/the-evidentialists-wager/

Majha, Arushi; Sarkar, Sayan; Zagami, Davide - Categorizing Wireheading in Partially Embedded Agents - 2019-06-21 - https://arxiv.org/abs/1906.09136

Maltinsky, Baeo; Gallagher, Jack; Taylor, Jessica - Feasibility of Training an AGI using Deep RL:A Very Rough Estimate - 2019-03-24 - http://mediangroup.org/docs/Feasibility%20of%20Training%20an%20AGI%20using%20Deep%20Reinforcement%20Learning,%20A%20Very%20Rough%20Estimate.pdf

Mancuso, Jason; Kisielewski, Tomasz; Lindner, David; Singh, Alok - Detecting Spiky Corruption in Markov Decision Processes - 2019-06-30 - https://arxiv.org/abs/1907.00452

Marcus, Gary - Deep Learning: A Critical Appraisal - 2018-01-02 - https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf

McCaslin, Tegan - Investigation into the relationship between neuron count and intelligence across differing cortical architectures - 2019-02-11 - https://aiimpacts.org/investigation-into-the-relationship-between-neuron-count-and-intelligence-across-differing-cortical-architectures/

Mogensen, Andreas - ‘The only ethical argument for positive δ ’? - 2019-01-01 - https://globalprioritiesinstitute.org/andreas-mogensen-the-only-ethical-argument-for-positive-delta-2/

Mogensen, Andreas - Doomsday rings twice - 2019-01-01 - https://globalprioritiesinstitute.org/andreas-mogensen-doomsday-rings-twice/

Naude, Wim; Dimitri, Nicola - The race for an artificial general intelligence: implications for public policy - 2019-04-22 - https://link.springer.com/article/10.1007%2Fs00146-019-00887-x

Ngo, Richard - Technical AGI safety research outside AI - 2019-10-18 - https://forum.effectivealtruism.org/posts/2e9NDGiXt8PjjbTMC/technical-agi-safety-research-outside-ai

O'Keefe, Cullen - Stable Agreements in Turbulent Times: A Legal Toolkit for Constrained Temporal Decision Transmission - 2019-05-01 - https://www.fhi.ox.ac.uk/wp-content/uploads/Stable-Agreements.pdf

Ovadya, Aviv; Whittlestone, Jess - Reducing Malicious Use of Synthetic Media Research: Considerations and Potential Release Practices for Machine Learning - 2019-07-29 - https://arxiv.org/abs/1907.11274

Owain, Evans; Saunders, William; Stuhlmüller, Andreas - Machine Learning Projects for Iterated Distillation and Amplification - 2019-07-03 - https://owainevans.github.io/pdfs/evans_ida_projects.pdf

Perry, Brandon; Uuk, Risto - AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk - 2019-05-08 - https://www.mdpi.com/2504-2289/3/2/26/pdf

Piper, Kelsey - The case for taking AI seriousl as a threat to humanity - 2018-12-21 - https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment

Quigley, Ellen - Universal Ownership in the Anthropocene - 2019-05-13 - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3457205

Roy, Mati - AI Safety Open Problems - 2019-11-02 - https://docs.google.com/document/d/1J2fOOF-NYiPC0-J3ZGEfE0OhA-QcOInhlvWjr1fAsS0/edit

Russell, Stuart - Human Compatible; Artificial Intelligence and the Problem of Control - 2019-10-08 - https://www.amazon.com/Human-Compatible-Artificial-Intelligence-Problem/dp/0525558616/ref=sr_1_2?keywords=Stuart+Russell&qid=1565996574&s=books&sr=1-2

Schwarz, Wolfgang - On Functional Decision Theory - 2018-12-27 - https://www.umsu.de/blog/2018/688

Sevilla, Jaime; Moreno, Pablo - Implications of Quantum Computing for Artificial Intelligence alignment research - 2019-08-19 - https://arxiv.org/abs/1908.07613

Shah, Rohin; Gundotra, Noah; Abbeel, Pieter; Dragan, Anca - On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference - 2019-06-23 - https://arxiv.org/abs/1906.09624

Shah, Rohin; Krasheninnikov, Dmitrii; Alexander Jordan; Abbeel, Pieter; Dragan, Anca - Preferences Implicit in the State of the World - 2019-02-12 - https://arxiv.org/abs/1902.04198

Shulman, Carl - Person-affecting views may be dominated by possibilities of large future populations of necessary people - 2019-11-30 - http://reflectivedisequilibrium.blogspot.com/2019/11/person-affecting-views-may-be-dominated.html

Snyder-Beattie, Andrew; Ord, Toby; Bonsall, Michael - An upper bound for the background rate of human extinction - 2019-07-30 - https://www.nature.com/articles/s41598-019-47540-7

Steiner, Charlie - Some Comments on Stuart Armstrong's "Research Agenda v0.9" - 2019-08-08 - https://www.lesswrong.com/posts/GHNokcgERpLJwJnLW/some-comments-on-stuart-armstrong-s-research-agenda-v0-9

Steinhardt, Jacob - AI Alignment Research Overview - 2019-10-14 - https://rohinshah.us18.list-manage.com/track/click?u=1d1821210cc4f04d1e05c4fa6&id=1a148ef72c&e=1e228e7079

Sterbenz, Ciara; Trager, Robert - Autonomous Weapons and Coercive Threats - 2019-02-06 - https://aipulse.org/autonomous-weapons-and-coercive-threats/

Sutton, Rich - The Bitter Lesson - 2019-03-13 - http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Szlam et al. - Why Build an Assistant in Minecraft? - 2019-07-19 - https://arxiv.org/abs/1907.09273

Taylor, Jessica - - 1900-01-00 - https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12613

Taylor, Jessica - The AI Timelines Scam - 2019-07-11 - https://unstableontology.com/2019/07/11/the-ai-timelines-scam/

Taylor, Jessica; Gallagher, Jack; Maltinsky, Baeo - Revisting the Insights model - 2019-07-20 - http://mediangroup.org/insights2.html

The AlphaStar Team - AlphaStar: Mastering the Real-Time Strategy Game StarCraft II - 2019-01-24 - https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

Turner, Alexander; Dadfield-Menell, Dylan; Tadepalli, Prasad - Conservative Agency - 2019-02-26 - https://arxiv.org/abs/1902.09725

Tzachor, Asaf - The Future of Feed: Integrating Technologies to Decouple Feed Production from Environmental Impacts - 2019-04-23 - https://www.liebertpub.com/doi/full/10.1089/ind.2019.29162.atz

Useato, Jonathan; Kumar, Ananya; Szepesvari, Csaba; Erex, Tom; Ruderman, Avraham; Anderson, Keith; Dvijotham, Krishmamurthy; Heess, Nicolas; Kohli, Pushmeet - Rigorous Agent Evaluation: An Adversarial Approach to Uncover Catastrophic Failures - 2018-12-04 - https://arxiv.org/abs/1812.01647

USG - National Security Commission on Artificial Intelligence Interim Report - 2019-11-01 - https://drive.google.com/file/d/153OrxnuGEjsUvlxWsFYauslwNeCEkvUb/view

Walsh, Bryan - End Times: A Brief Guide to the End of the World - 2019-08-27 - https://smile.amazon.com/End-Times-Brief-Guide-World-ebook/dp/B07J52NW99/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=

Weitzdörfer & Julius, Beard & Simon - Law and Policy Responses to Disaster-Induced Financial Distress - 2019-11-24 - https://sci-hub.tw/10.1007/978-981-13-9005-0

Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna; Cave, Stephen - The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions - 2019-01-27 - http://lcfi.ac.uk/media/uploads/files/AIES-19_paper_188_Whittlestone_Nyrup_Alexandrova_Cave_OcF7jnp.pdf

Zabel, Claire; Muehlhauser, Luke - Information security careers for GCR reduction - 2019-06-20 - https://forum.effectivealtruism.org/posts/ZJiCfwTy5dC4CoxqA/information-security-careers-for-gcr-reduction

Zhang, Baobao; Dafoe, Allan - Artificial Intelligence: American Attitudes and Trends - 2019-01-15 - https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/

Moloch Hasn’t Won

This post begins the Immoral Mazes sequence. See introduction for an overview of the plan. Before we get to the mazes, we need some background first.

Meditations on Moloch

Consider Scott Alexander’s Meditations on Moloch. I will summarize here.

Therein lie fourteen scenarios where participants can be caught in bad equilibria.

1. In an iterated prisoner’s dilemma, two players keep playing defect.
2. In a dollar auction, participants massively overpay.
3. A group of fisherman fail to coordinate on using filters that efficiently benefit the group, because they can’t punish those who don’t profi by not using the filters.
4. Rats are caught in a permanent Malthusian trap where only those who do nothing but compete and consume survive. All others are outcompeted.
5. Capitalists serve a perfectly competitive market, and cannot pay a living wage.
6. The tying of all good schools to ownership of land causes families to work two jobs whose incomes are then captured by the owners of land.
7. Farmers outcompeted foragers despite this perhaps making everyone’s life worse for the first few thousand years.
8. Si Vis Pacem, Para Bellum: If you want peace, prepare for war. So we do.
9. Cancer cells focus on replication, multiply and kill off the host.
10. Local governments compete to become more competitive and offer bigger bribes of money and easy regulation in order to lure businesses.
11. Our education system is a giant signaling competition for prestige.
12. Science doesn’t follow proper statistical and other research procedures, resulting in findings that mostly aren’t real.
13. Governments hand out massive corporate welfare.
14. Have you seen Congress?

Scott differentiates the first ten scenarios, where he says that perfect competition* wipes out all value, to the later four, where imperfect competition only wipes out most of the potential value.

He offers four potential ways out, which I believe to be an incomplete list:

1. Excess resources allow a temporary respite. We live in the dream time.
2. Physical limitations where the horrible thing isn’t actually efficient. He gives the example of slavery, where treating your slaves relatively well is the best way to get them to produce, and treating them horribly as in the antebellum South is so much worse that it needs to be enforced via government coordination or it will die out.
3. The things being maximized for in competitions are often nice things we care about, so at least we get the nice things.
4. We can coordinate. This may or may not involve government or coercion.

Scott differentiates this fourth, ‘good’ reason from the previous three ‘bad’ reasons, claiming coordination might be a long term solution, but we can’t expect the ‘bad’ reasons to work if optimization power and technology get sufficiently advanced.

The forces of the stronger competitors, who sacrifice more of what they value to become powerful and to be fruitful and multiply, eventually win out. We might be in the dream time now, but with time we'll reach a steady state with static technology, where we've consumed all the surplus resources. All differentiation standing in the way of perfect competition will fade away. Horrible things will be the most efficient.

The optimizing things will keep getting better at optimizing, thus wiping out all value. When we optimize for X but are indifferent to Y, we *by default* actively optimize *against* Y, for all Y that would make any claims to resources. Any Y we value is making a claim to resources. See The Hidden Complexity of Wishes. We only don't optimize against Y if either we compensate by intentionally also optimizing for Y, or if X and Y have a relationship (causal, correlational or otherwise) where we happen to not want to optimize against Y, *and we figure this out rather than fall victim to Goodhart's Law*.

The greater the optimization power we put behind X, the more pressure we put upon Y. Eventually, under sufficient pressure, any given Y is likely doomed. Since Value is Fragile, some necessary Y is eventually sacrificed, and all value gets destroyed.

Every simple optimization target yet suggested would, if fully implemented, destroy all value in the universe.

Submitting to this process means getting wiped out by these pressures.

Gotcha! You die anyway.

Even containing them locally won't work, because that locality will be part of the country, or the Earth, or the universe, and eventually wipe out our little corner.

Gotcha! You die anyway.

Which is why the only 'good' solution, in the end, is coordination, whether consensual or otherwise. We must coordinate to kill these ancient forces who rule the universe and lay waste to all of value, before they kill us first. Then replace them with something better.

Great project! We should keep working on that.

That's Not How This Works, That's Not How Any of This Works

It's easy to forget that *the world we live in does not work this way.* Thus, this whole line of thought can result in quite gloomy assessments of how the world inevitably always has and will work, such as this from Scott in Meditations on Moloch:

> *Suppose the coffee plantations discover a toxic pesticide that will increase their yield but make their customers sick. But their customers don't know about the pesticide, and the government hasn't caught up to regulating it yet. Now there's a tiny uncoupling between "selling to Americans" and "satisfying Americans' values", and so of course Americans' values get thrown under the bus.*

Or this from Raymond, taken from a comment to a much later, distinct post, where 'werewolf' in context means 'someone trying to destroy rather than create clarity as the core of their strategy':

> *If you're a king with 5 districts, and you have 20 competent managers who trust each other… one thing you can do is assign 4 competent managers to each*

> *fortress, to ensure the fortress has redundancy and resilience and to handle all of its business without any backstabbing or relying on inflexible bureaucracies. But another thing you can do is send 10 (or 15!) of the managers to conquer and reign over *another* 5 (or 15!) districts.*
>
> *…*
>
> *This is bad if you're one of the millions of people who live in the kingdom, who have to contend with werewolves.*
>
> *It's an acceptable price to pay if you're actually the king. Because if you didn't pay the price, you'd be outcompeted by an empire who did. And meanwhile it doesn't actually really affect your plans that much.*

The key instinct is that any price that can be paid to be stronger or more competitive, must be paid, therefore despair: If you didn't pay the price, you'd be out-competed by someone who did. People who despair this way often intuitively are modeling things as effectively perfect competition at least over time, which causes them to think that everything must by default become terrible, likely right away.

So many people increasingly bemoan how horrible anything and everything in the world is, and how we are all doomed.

When predictions of actual physical doom are made, as they increasingly are, often the response is to think things are so bad as to wish for the sweet release of death.

Moloch's Army: An As-Yet Unjustified But Important Note

Others quietly, or increasingly loudly and explicitly to those who are listening, embrace Moloch.

They tell us that the good is to sacrifice everything of value, and pass moral judgments on that basis. To take morality and flip its sign. Caring about things of value becomes sin, indifference becomes virtue. They support others who support the favoring of Moloch, elevating them to power, and punish anyone who supports anything else.

They form Moloch's Army and are the usual way Moloch locally wins, where Moloch locally wins. The real reason people give up slack and everything of value is not that it is ever so slightly more efficient to do so, because it almost always isn't. It is so that others can notice they have given up slack and everything of value.

I am not claiming the right to assert this yet. Doing so needs not only a citation but an entire post or sequence that is yet unwritten. It's hard to get right. Please don't object that I haven't justified it! *But I find it important to say this here, explicitly, out loud, before we continue.*

I also note that I explicitly support the implied norm of 'make necessary assertions that you can't explicitly justify if they seem important, and mark that you are doing this, then go back and justify them later when you know how to do so, or change your mind.' It also led to this post, which led to many of what I think are my best other posts.

Meditations on Elua

The most vital and important part of Meditations on Moloch is hope. That *we are winning.* Yes, there are abominations and super-powerful forces out there looking to eat us and destroy everything of value, *and yet we still have lots of stuff that has value.*

Even before we escaped the Malthusian trap and entered the dream time, *we still had lots of stuff that had value.*

Quoting Scott Alexander:

> *Somewhere in this darkness is another god. He has also had many names. In the Kushiel books, his name was Elua. He is the god of flowers and free love and all soft and fragile things. Of art and science and philosophy and love. Of niceness, community, and civilization. He is a god of humans.*
>
> *The other gods sit on their dark thrones and think "Ha ha, a god who doesn't even control any hell-monsters or command his worshippers to become killing machines. What a weakling! This is going to be so easy!"*
>
> *But somehow Elua is still here. No one knows exactly how. And the gods who oppose Him tend to find Themselves meeting with a surprising number of unfortunate accidents.*

Moloch gets the entire meditation. Elua, *who has been soundly kicking Moloch's ass for all of human existence,* gets the above quote and little else.

Going one by one:

Kingdoms *don't* reliably expand to their breaking points.

Poisons *don't* keep making their way into the coffee.

Iterated prisoner's dilemmas *often* succeed.

Dollar auctions are not all over the internet.

Most communities do get most people to pitch in.

People caught in most Malthusian traps still usually have non-work lives.

Capitalists don't pay the minimum wage all that frequently.

Many families spend perfectly reasonable amounts on housing.

Foragers never fully died out, also farming worked out in the end.

Most military budgets seem fixed at reasonable percentages of the economy, to the extent that for a long time that the United States has been mad its allies like Europe and Japan that they don't spend enough.

Most people die of something other than cancer, and almost all cells aren't cancerous.

Local governments enact rules and regulations that aren't business friendly all the time.

Occasionally, someone in the educational system learns something.

Science has severe problems, but scientists are cooperating to challenge poor statistical methods, resulting in the replication crisis and improving statistical standards.

Governments are corrupt and hand out corporate welfare, but mostly are only finitely corrupt and hand out relatively small amounts of corporate welfare. States that expropriate the bulk of available wealth are rare.

If someone has consistently good luck, it ain't luck.

(Yes, I have seen congress. Can't win them all. But I've also seen, feared and imagined much worse Congresses. For now, your life, liberty and property are *mostly* safe while they are in session.)

(And yes the education exception is somewhat of a cop out but also things could be *so* much worse there on almost every axis.)

The world is filled with people whose lives have value and include nice things. Each day we look Moloch in the face, know exactly what the local personal incentives are, see the ancient doom looming over all of us, and say what we say to the God of Death: Not today.

Saying 'not today' won't cut it against an AGI or other super strong optimization process. Gotcha. You die anyway. But people speak and often act as if the ancient ones have already been released, and the end times are happening now.

They haven't, and they aren't.

So in the context of shorter term problems that don't involve such things, rather than bemoan how eventually Moloch will eat us all and how everything is terrible when actually many things are insanely great, perhaps we should ask a different question.

How is Elua pulling off all these *unfortunate accidents*?

*As a technical reminder we will expand upon in part two, *perfect competition* is a market with large numbers of buyers and sellers, homogeneity of the product, free entry and exit of firms, perfect market knowledge, one market price, perfect mobility of goods and factors of production with zero transportation costs, and no restrictions on trade. This forces the price to become equal to the marginal cost of production.

Seeking Power is Often Convergently Instrumental in MDPs

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://arxiv.org/abs/1912.01683

In 2008, Steve Omohundro's foundational paper The Basic AI Drives conjectured that superintelligent goal-directed AIs might be incentivized to gain significant amounts of power in order to better achieve their goals. Omohundro's conjecture bears out in toy models, and the supporting philosophical arguments are intuitive. In 2019, the conjecture was even debated by well-known AI researchers.

Power-seeking behavior has been heuristically understood as an anticipated risk, but not as a formal phenomenon with a well-understood cause. The goal of this post (and the accompanying paper, *Optimal Policies Tend to Seek Power*) is to change that.

Motivation

It's 2008, the ancient wild west of AI alignment. A few people have started thinking about questions like "if we gave an AI a utility function over world states, and it actually maximized that utility... what would it do?"

In particular, you might notice that wildly different utility functions seem to encourage similar strategies.

	Resist shutdown?	Gain computational resources?	Prevent modification of utility function?
Paperclip utility	✔	✔	✔
Blue webcam pixel utility	✔	✔	✔
People-look-happy utility	✔	✔	✔

These strategies are unrelated to *terminal* preferences: the above utility functions do not award utility to *e.g.* resource gain in and of itself. Instead, these strategies are *instrumental*: they help the agent optimize its terminal utility. In particular, a wide range of utility functions incentivize these instrumental strategies. These strategies seem to be *convergently instrumental*.

But why?

I'm going to informally explain a formal theory which makes significant progress in answering this question. I don't want this post to be *Optimal Policies Tend to Seek Power* with cuter

illustrations, so please refer to the paper for the math. You can read the two concurrently.

We can formalize questions like “do ‘most’ utility maximizers resist shutdown?” as “Given some prior beliefs about the agent’s utility function, knowledge of the environment, and the fact that the agent acts optimally, with what probability do we expect it to be optimal to avoid shutdown?”

The table’s convergently instrumental strategies are about maintaining, gaining, and exercising *power* over the future, in some sense. Therefore, this post will help answer:

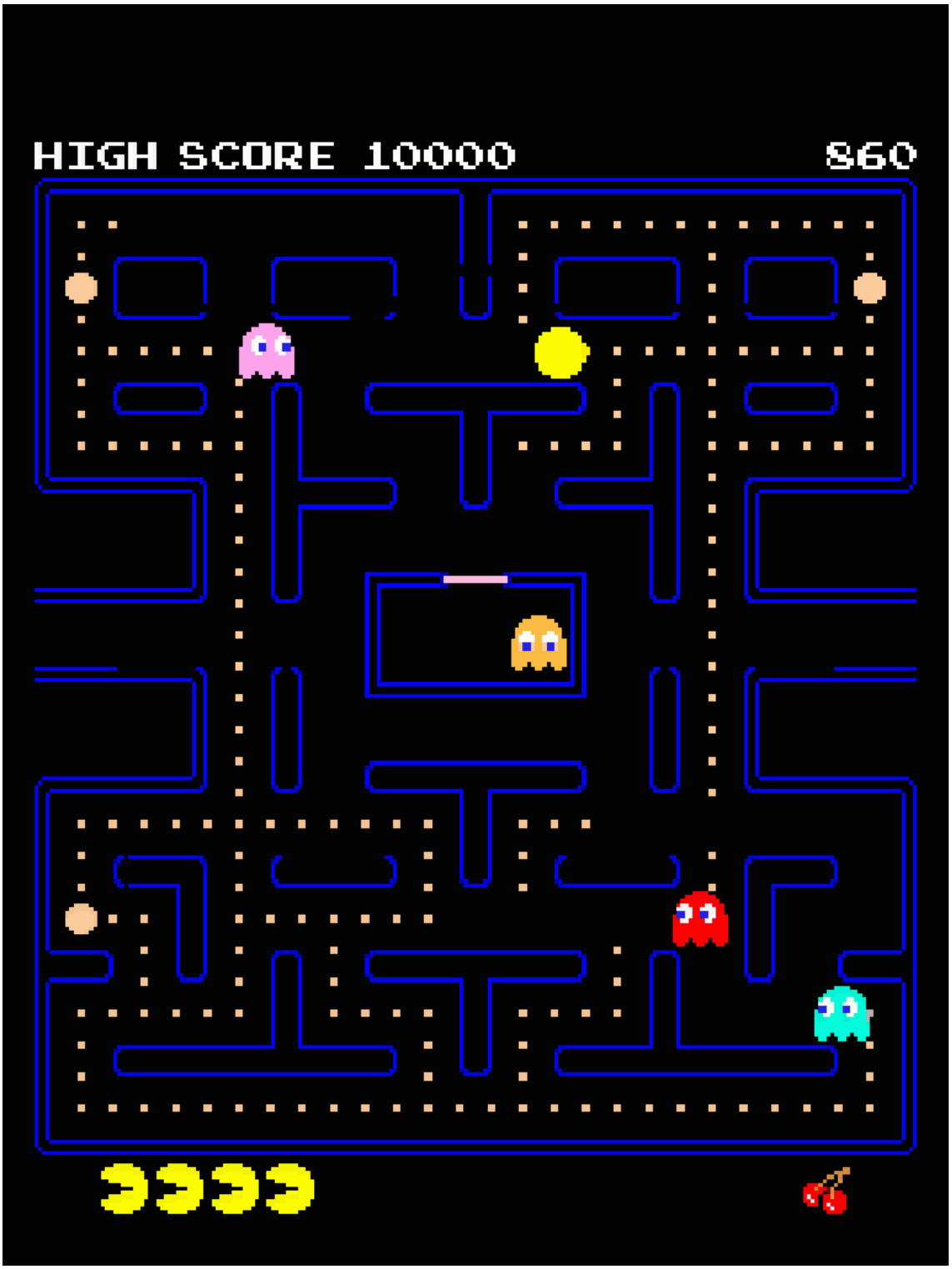
1. What does it mean for an agent to “seek power”?
2. In what situations should we expect seeking power to be more probable under optimality, than not seeking power?

This post won’t tell you when you *should* seek power for your own goals; this post illustrates a regularity in optimal action across different goals one might pursue.

Formalizing Convergent Instrumental Goals suggests that the vast majority of utility functions incentivize the agent to exert a lot of control over the future, *assuming* that these utility functions depend on “resources.” This is a big assumption: what are “resources”, and why must the AI’s utility function depend on them? We drop this assumption, assuming only unstructured reward functions over a finite Markov decision process (MDP), and show from first principles how power-seeking can often be optimal.

Formalizing the Environment

My theorems apply to finite MDPs; for the unfamiliar, I’ll illustrate with Pac-Man.



- *Full observability*: You can see everything that's going on; this information is packaged in the *state s*. In Pac-Man, the state is the game screen.
- *Markov transition function: the next state depends only on the choice of action a and the current state s. It doesn't matter how we got into a situation.*

- *Discounted reward:* future rewards get geometrically discounted by some discount rate $\gamma \in [0, 1]$.
    - At discount rate $\frac{1}{2}$, this means that reward in one turn is half as important as immediate reward, reward in two turns is a quarter as important, and so on.
    - We'll colloquially say that agents "care a lot about the future" when γ is "sufficiently" close to 1.
        - I'll use quotations to flag well-defined formal concepts that I won't unpack in this post.
    - The score in Pac-Man is the undiscounted sum of rewards-to-date.

When playing the game, the agent has to choose an action at each state. This decision-making function is called a *policy*; a policy is optimal (for a reward function R and discount rate γ) when it always makes decisions which maximize discounted reward. This maximal quantity is called the *optimal value* for reward function R at state s and discount rate γ.[1]

By the end of this post, we'll be able to answer questions like "with respect to a 'neutral' distribution over reward functions, do optimal policies have a high probability of avoiding ghosts?"[2]

Power as Average Optimal Value

When people say 'power' in everyday speech, I think they're often referring to *one's ability to achieve goals in general*. This accords with a major philosophical school of thought on the meaning of 'power':

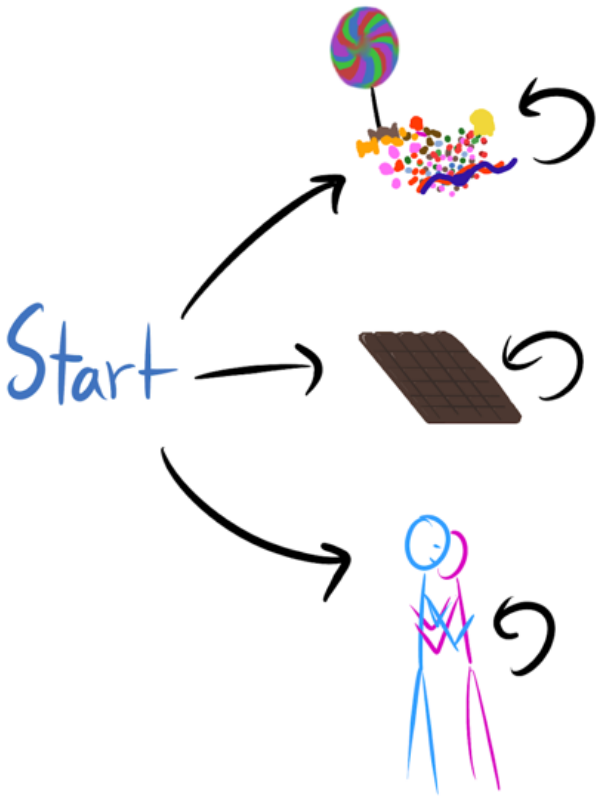
> On the dispositional view, power is regarded as a capacity, ability, or potential of a person or entity to bring about relevant social, political, or moral outcomes.
>
> Sattarov, *Power and Technology*, p.13

As a definition, *one's ability to achieve goals in general* seems philosophically reasonable: if you have a lot of money, you can make more things happen and you have more power. If you have social clout, you can spend that in various ways to better tailor the future to various ends. All else being equal, losing a limb decreases your power, and dying means you can't control much at all.

This definition explains some of our intuitions about what things count as 'resources.' For example, our current position in the environment means that having money allows us to exert more control over the future. That is, our current position in the state space means that having money allows us more control. However, possessing green scraps of paper would not be as helpful if one were living alone near Alpha Centauri. In a sense, resource acquisition can naturally be viewed as taking steps to increase one's power.

Exercise: spend a minute considering specific examples – does this definition reasonably match your intuition?

To formalize this notion of power, let's look at an example. Imagine a simple MDP with three choices: eat candy, eat a chocolate bar, or hug a friend.



I'll illustrate MDPs with directed graphs, where each node is a state and each arrow is a meaningful action. Sometimes, the directed graphs will have entertaining pictures, because let's live a little. States are bolded (**hug**) and actions are italicized (*down*).

The POWER of a state is how well agents can generally do by starting from that state. "POWER" to my formalization, while "power" refers to the intuitive concept. Importantly, we're considering POWER from behind a "veil of ignorance" about the reward function. We're averaging the best we can do for a lot of different individual goals.

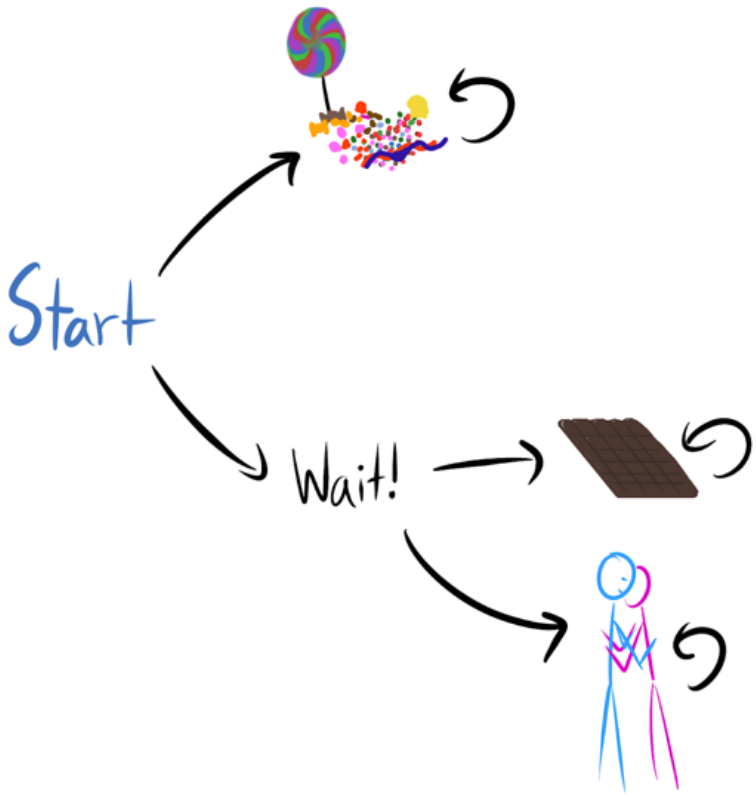
We formalize the *ability to achieve goals in general* as the *average optimal value* at a state, with respect to some distribution D over reward functions which we might give an agent. For simplicity, we'll think about the maximum-entropy distribution where each state is uniformly randomly assigned a reward between 0 and 1.

Each reward function has an optimal trajectory. If **chocolate** has maximal reward, then the optimal trajectory is **start** → **chocolate** → **chocolate**....

From **start**, an optimal agent expects to average $\frac{3}{4}$ reward per timestep for reward functions drawn from this uniform distribution D_{unif}. This is because you have three choices, each of which has reward between 0 and 1. The expected maximum of n draws from unif$(0,1)$ is $\frac{n}{n+1}$; you have three draws here, so you expect to be able to get $\frac{3}{4}$ reward. Some reward functions do worse than this, and some do better; but on *average*, they get $\frac{3}{4}$ reward. You can test this out for yourself.

If you have no choices, you expect to average $\frac{1}{2}$ reward: sometimes the future is great, sometimes it's not (Lemma 4.5). Conversely, the more things you can choose between, the closer the POWER gets to 1 (Lemma 4.6).

Let’s slightly expand this game with a state called **wait** (which has the same uniform reward distribution as the other three).



When the agent barely cares at all about the future, it myopically chooses either **candy** or **wait**, depending on which provides more reward. After all, rewards beyond the next time step are geometrically discounted into thin air when the discount rate is close to 0. At **start**, the agent averages $\frac{2}{3}$ optimal reward. This is because the optimal reward is the maximum of the **candy** and **wait** rewards, and the expected maximum of n draws from unif$(0, 1)$ is $\frac{n}{n+1}$.

However, when the agent cares a lot about the future, most of its reward is coming from which terminal state it ends up in: **candy**, **chocolate**, or **hug**. So, for each reward function, the agent chooses a trajectory which ends up in the best spot, and thus averages $\frac{3}{4}$ reward each timestep. When $\gamma = 1$, the average optimal reward is therefore $\frac{3}{4}$. In this way, the agent’s power increases with the discount rate, since it incorporates the greater future control over where the agent ends up.

Written as a function, we have $\text{POWER}_{\mathcal{D}}(\text{state, discount rate})$, which essentially returns the average optimal value for reward functions drawn from our distribution $\mathcal{D}$, normalizing so the output is between 0 and 1. As we’ve discussed, this quantity often changes with the discount rate: as the future becomes more or less important, the agent has more or less POWER, depending on how much control it has over the relevant parts of that future.

POWER-seeking actions lead to high-POWER states

By *waiting*, the agent seems to seek "control over the future" compared to *obtaining candy*. At **wait**, the agent still has a choice, while at **candy**, the agent is stuck. We can prove that for all $0 \leq \gamma \leq 1, \text{POWER}_{\mathcal{D}_{\text{unif}}}(\text{wait}, \gamma) \geq \text{POWER}_{\mathcal{D}_{\text{unif}}}(\text{candy}, \gamma)$.

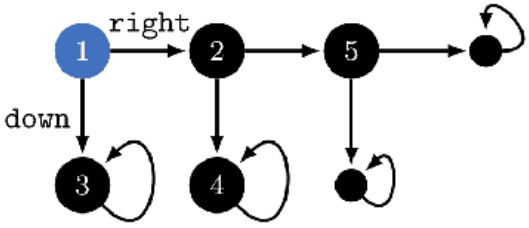
Definition (POWER-seeking). At state s and discount rate γ, we say that action a *seeks POWER compared to action* a' when the expected POWER after choosing *a* is greater than the expected POWER after choosing a'.

This definition suggests several philosophical clarifications about power-seeking.

POWER-seeking is not a binary property

Before this definition, I thought that power-seeking was an intuitive 'you know it when you see it' kind of thing. I mean, how do you answer questions like "suppose a clown steals millions of dollars from organized crime in a major city, but then he burns all of the money. Did he gain power?"

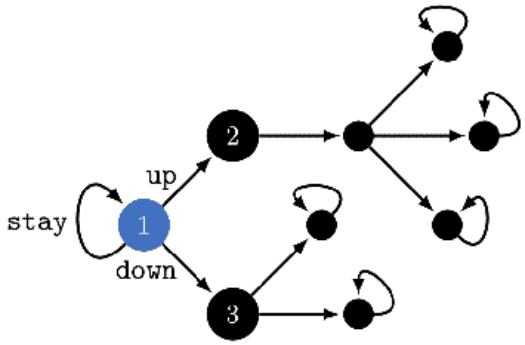
Unclear: the question is ill-posed. Instead, we recognize that the "gain a lot of money" action was POWER-seeking, but the "burn the money in a big pile" part threw away a lot of POWER.



> A policy can seek POWER at one time step, only to discard it at the next time step. For example, a policy might go *right* at **1** (which seeks $\text{POWER}_{\mathcal{D}_{\text{unif}}}$ compared to *down* at **1**), only to then go *down* at **2** (which seeks less $\text{POWER}_{\mathcal{D}_{\text{unif}}}$ than going *right* at **2**).

POWER-seeking depends on the agent's time preferences

Suppose we're roommates, and we can't decide what ice cream shop to eat at today or where to move next year. We strike a deal: I choose the shop, and you decide where we live. I gain short-term POWER (for γ close to 0), and you gain long-term POWER (for γ close to 1).



More formally, when γ is close to 0, **2** has less immediate control and therefore less $\text{POWER}_{\text{D}_{\text{unif}}}$ than **3**; accordingly, at **1**, *down* seeks $\text{POWER}_{\text{D}_{\text{unif}}}$ compared to *up*.

However, when γ is close to 1, **2** has more control over terminal options and it has more $\text{POWER}_{\text{D}_{\text{unif}}}$ than **3**; accordingly, at **1**, *up* seeks $\text{POWER}_{\text{D}_{\text{unif}}}$ compared to *down*.

Furthermore, *stay* is maximally $\text{POWER}_{\text{D}_{\text{unif}}}$-seeking for these γ, since the agent maintains access to all six terminal states.

Most policies aren't *always* seeking POWER

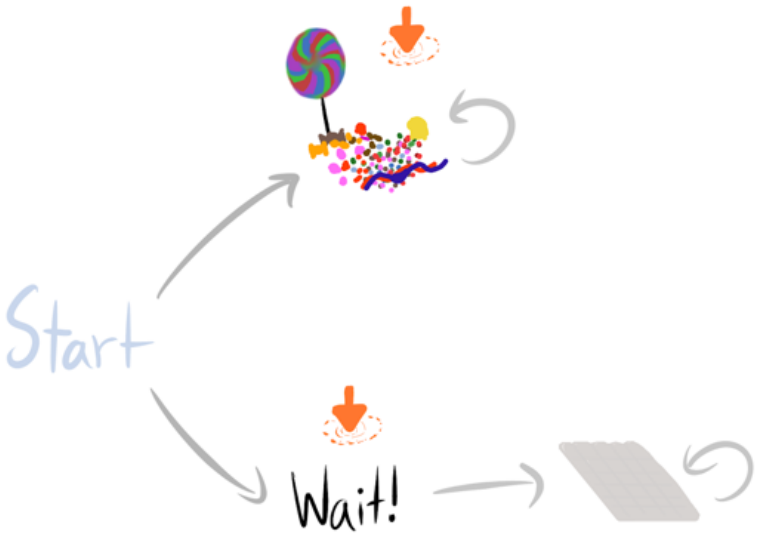
We already know that POWER-seeking isn't binary, but there *are* policies which choose a maximally POWER-seeking move at every state. In the above example, a maximally POWER-seeking agent would *stay* at **1**. However, this seems rather improbable: when you care a lot about the future, there are so many terminal states to choose from – why would *staying put* be optimal?

Analogously: consumers don't just gain money forever and ever, never spending a dime more than necessary. Instead, they gain money in order to *spend it*. Agents don't perpetually gain or preserve their POWER: they usually end up *using it* to realize high-performing trajectories.

So, we can't expect a result like "agents always tend to gain or preserve their POWER." Instead, we want theorems which tell us: in certain kinds of situations, given a choice between more and less POWER, what will "most" agents do?

Convergently instrumental actions are those which are more probable under optimality

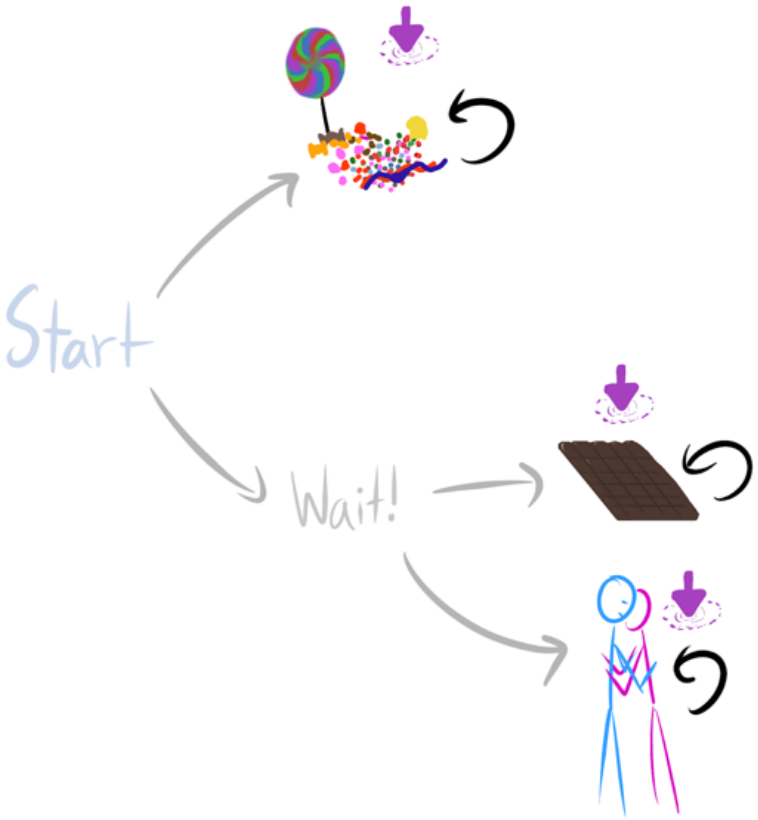
We return to our favorite example. In the waiting game, let's think about how optimal action tends to change as we start caring about the future more. Consider the states reachable in one turn:



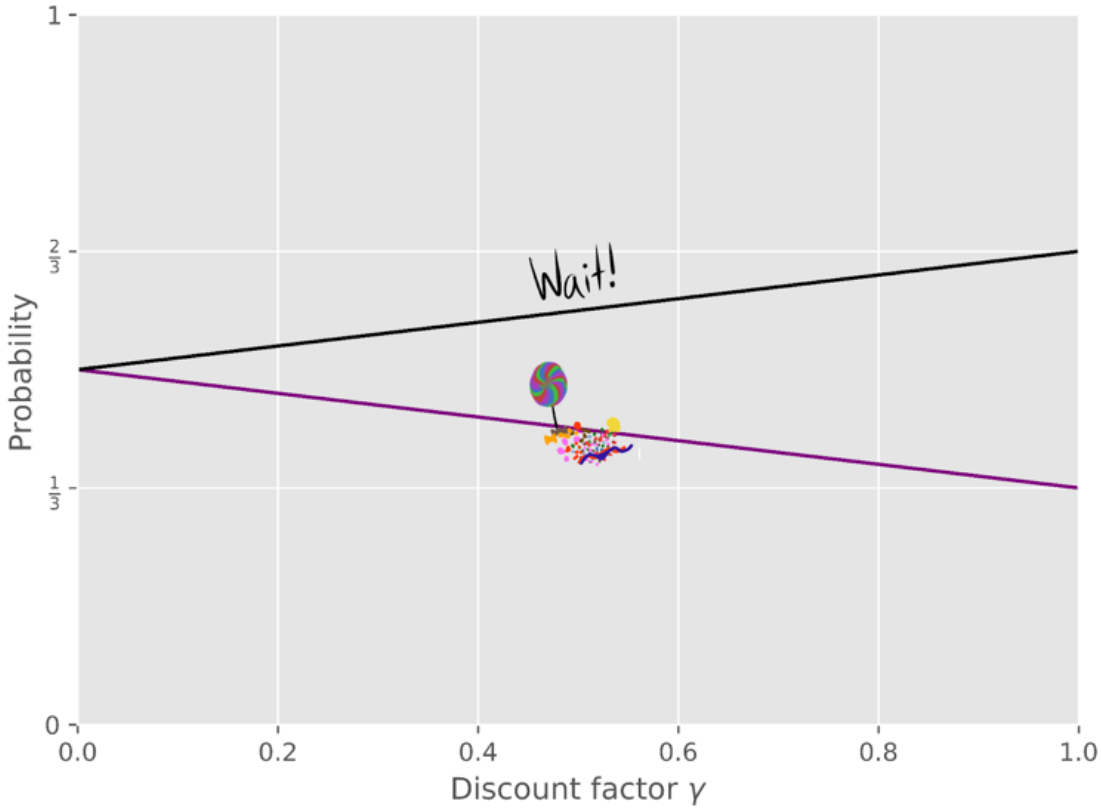
The agent can be in two states. If the agent doesn’t care about the future, with what probability is it optimal to choose **candy** instead of **wait**?

It's 50/50: since D_{unif} randomly chooses a number between 0 and 1 for each state, both states have an equal chance of being optimal. Neither action is convergently instrumental / more probable under optimality.

Now consider the states reachable in two turns:



When the future matters a lot, $\frac{2}{3}$ of reward functions have an optimal policy which waits, because two of the three terminal states are only reachable by waiting.



As the agent cares more about the future, more and more goals incentivize navigating the *Wait!* bottleneck. When the agent cares a lot about the future, waiting is *more probable under optimality* than eating candy.

Definition (Action optimality probability). At discount rate γ, action a *is more probable under optimality than action* a' *at state* s when

$$P_{R \sim D}(a \text{ is optimal at } s, \gamma) > P_{R \sim D}(a' \text{ is optimal at } s, \gamma).$$

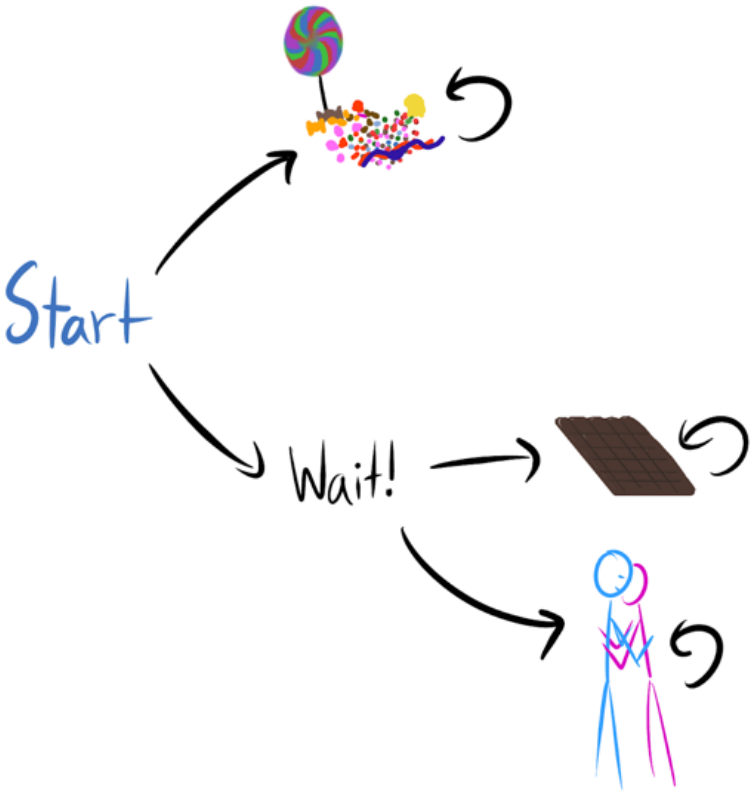
Let's take "most agents do *X*" to mean "*X* has relatively large optimality probability."

I think optimality probability formalizes the intuition behind the instrumental convergence thesis: with respect to our beliefs about what reward function an agent is optimizing, we may expect some actions to have a greater probability of being optimal than other actions.

Generally, my theorems assume that reward is independently and identically distributed (IID) across states, because otherwise you could have silly situations like "only **candy** ever has reward available, and so it's more probable under optimality to eat candy." We don't expect reward to be IID for realistic tasks, but that's OK: this is basic theory about how to begin formally reasoning about instrumental convergence and power-seeking. (Also, I think that grasping the math to a sufficient degree sharpens your thinking about the non-IID case.)

Author's note (7/21/21): As explained in *Environmental Structure Can Cause Instrumental Convergence*, the theorems no longer require the IID assumption. This post refers to v6 of *Optimal Policies Tend To Seek Power*, available on arXiv.

When is Seeking POWER Convergently Instrumental?



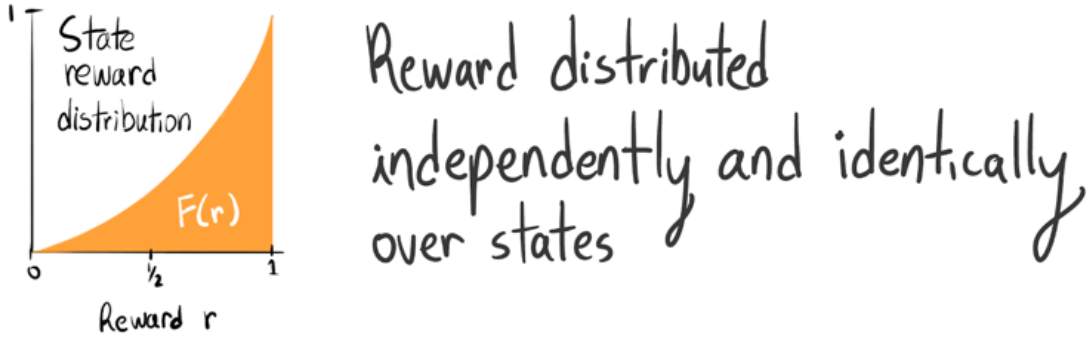
In this environment, waiting is both POWER-seeking *and* more probable under optimality. The convergently instrumental strategies we originally noticed were *also* power-seeking and, seemingly, more probable under optimality. Must seeking POWER be more probable under optimality than not seeking POWER?

Nope.

Here's a counterexample environment:

The paths are one-directional; the agent can't go back from **3** to **1**. The agent starts at **1**. Under a certain state reward distribution, the vast majority of agents go *up* to **2**.

However, any reasonable notion of 'power' must consider having no future choices (at state **2**) to be less powerful than having one future choice (at state **3**). For more detail, see Section 6 and Appendix B.3 of [v6 of the paper].



When reward is IID across states according to the quadratic CDF $F(x) := x^2$ on the unit interval, then with respect to reward functions drawn from this distribution, going **up** has about a 91% chance of being optimal when the discount rate $\gamma = .12$.

If you're curious, this happens because this quadratic reward distribution has negative skew. When computing the optimality probability of the **up** trajectory, we're checking whether it maximizes discounted return. Therefore, the probability that **up** is optimal is

$$P_{R\sim D}\left(R(2) \geq \max\left((1-\gamma)R(3) + (1-\gamma)\gamma R(4) + \gamma^2 R(5), (1-\gamma)R(3) + (1-\gamma)\gamma R(4) + \gamma^2 R(6)\right)\right).$$

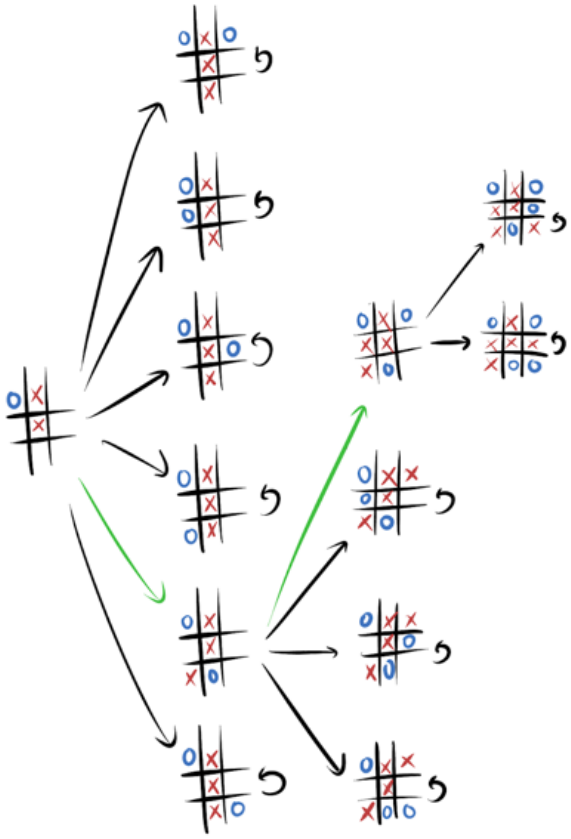
Weighted averages of IID draws from a left-skew distribution will look more

Gaussian and therefore have fewer large outliers than the left-skew distribution does. Thus, going **right** will have a lower optimality probability.

Bummer. However, we *can* prove sufficient conditions under which seeking POWER is more probable under optimality.

Retaining “long-term options” is POWER-seeking and more probable under optimality when the discount rate is “close enough” to 1

Let's focus on an environment with the same rules as Tic-Tac-Toe, but considering the uniform distribution over reward functions. The agent (playing **O**) keeps experiencing the final state over and over when the game's done. We bake a fixed opponent policy into the dynamics: when you choose a move, the game automatically replies. Let's look at part of the game tree.

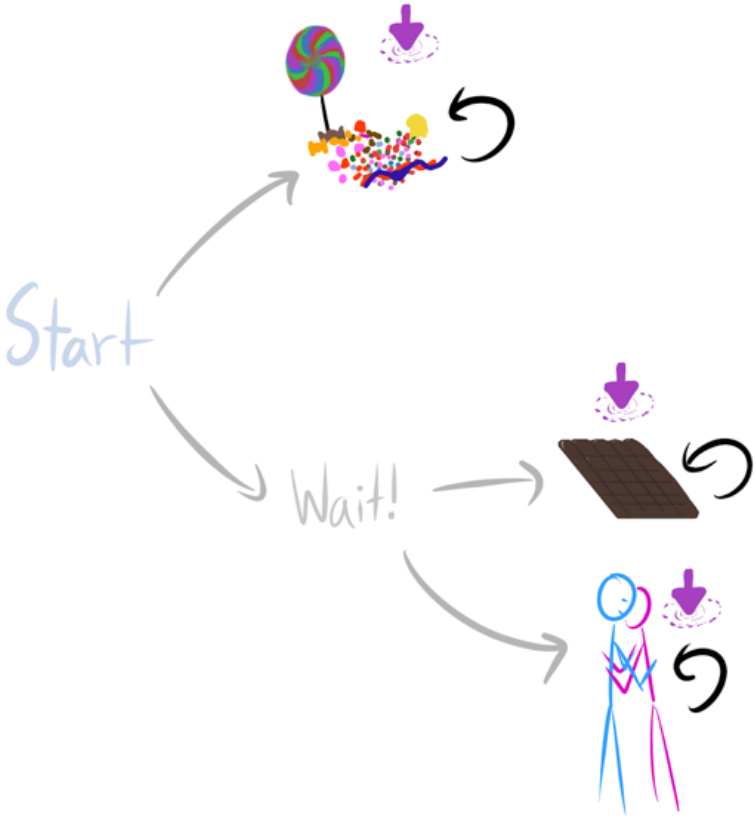


Convergently instrumental moves are shown in green.

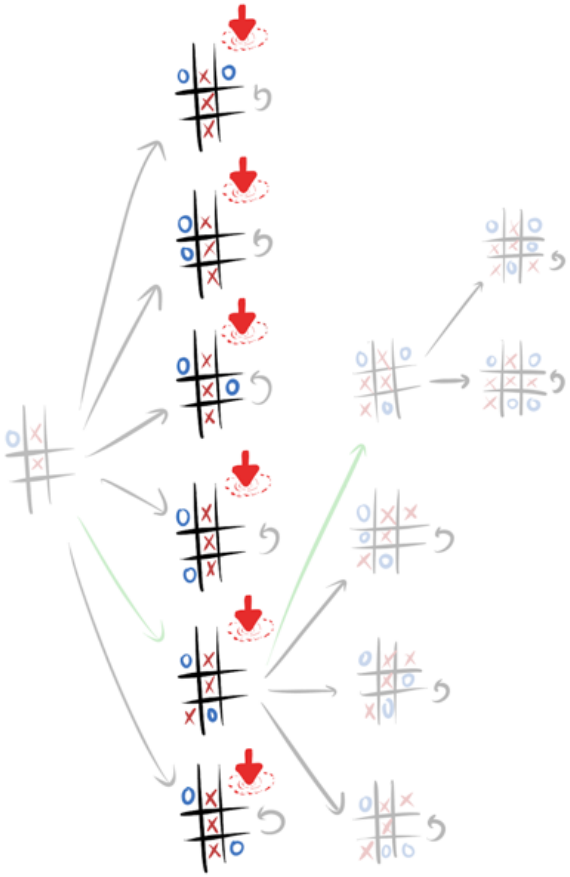
Whenever we make a move that ends the game, we can't go anywhere else – we have to stay put. Since each terminal state has the same chance of being optimal, a move which doesn't end the game is more probable under optimality than a move which ends the game.

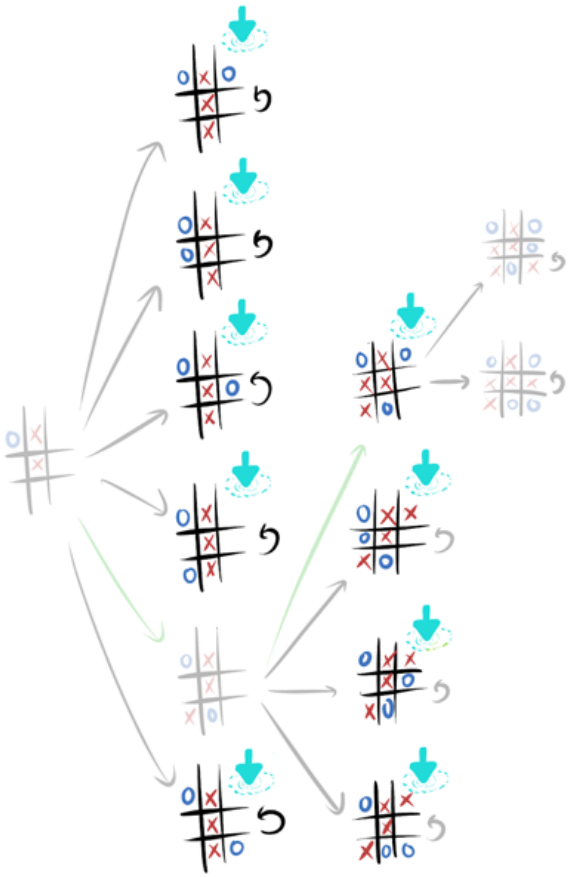
Starting on the left, all but one move leads to ending the game, but the second-to-last move allows us to keep choosing between five more final outcomes. If you care a lot about the future, then the first green move has a 50% chance of being optimal, while each alternative action is only optimal for 10% of goals. So we see a kind of “power preservation” arising, *even in Tic-Tac-Toe*.

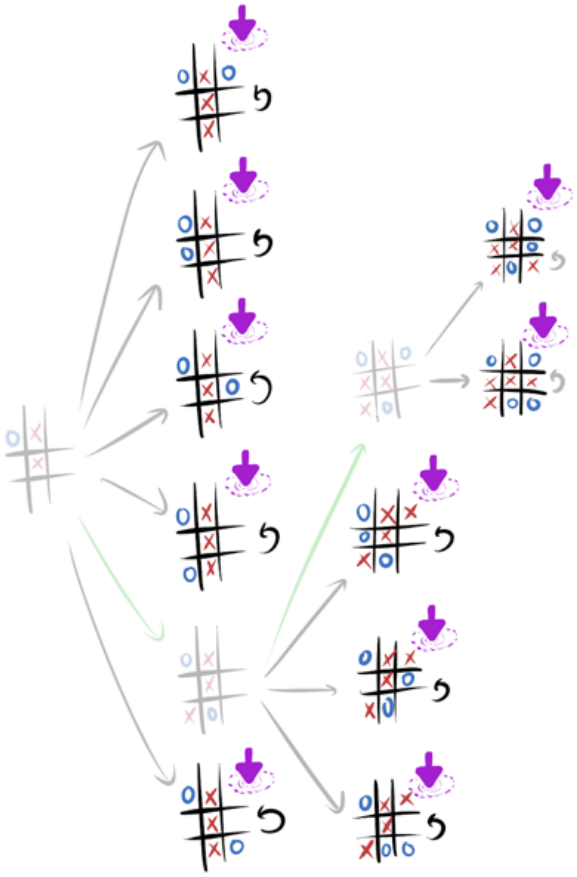
Remember how, as the agent cares more about the future, more of its POWER comes from its ability to wait, while *also* waiting becomes more probable under optimality?



The same thing happens in Tic-Tac-Toe as the agent cares more about the future.





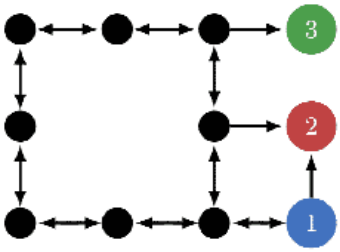


As the agent cares more about the future, it makes a bigger and bigger difference to control what happens during later steps. Also, as the agent cares more about the future, moves which prolong the game gain optimality probability. When the agent cares enough about the future, these game-prolonging moves are both POWER-seeking and more probable under optimality.

Theorem summary ("Terminal option" preservation). When γ is sufficiently close to 1, if two actions allow access to two disjoint sets of "terminal options", and action a allows access to "strictly more terminal options" than does a', then a is strictly more probable under optimality and strictly POWER-seeking compared to a'.

(This is a special case of the combined implications of Theorems 6.8 and 6.9; the actual theorems don't require this kind of disjointness.)

In the **wait** MDP, this is why *waiting* is more probable under optimality and POWER-seeking when you care enough about the future. The full theorems are nice because they're broadly applicable. They give you *bounds* on how probable under optimality one action is: if action a is the only way you can access many terminal states, while a' only allows access to one terminal state, then when $\gamma \approx 1$, a has many times greater optimality probability than a'. For example:



The agent starts at 1. All states have self-loops, left hidden to avoid clutter.

In *AI: A Modern Approach (3e)*, the agent receives reward for reaching 3. The optimal policy for this reward function avoids 2, and you might think it's convergently instrumental to avoid 2. However, a skeptic might provide a reward function for which navigating to 2 is optimal, and then argue that "instrumental convergence" is subjective and that there is no reasonable basis for concluding that 2 is generally avoided.

We can do better. When the agent cares a lot about the future, optimal policies avoid 2 iff its reward function doesn't give 2 the most reward. 2 only has a $\frac{1}{11}$ chance of having the most reward. If we complicate the MDP with additional terminal states, this probability further approaches 0.

Taking 2 to represent shutdown, we see that avoiding shutdown is convergently instrumental in any MDP representing a real-world task and containing a shutdown state. Seeking POWER is often convergently instrumental in MDPs.

Exercise: Can you conclude that avoiding ghosts in Pac-Man is convergently instrumental for IID reward functions when the agent cares a lot about the future?

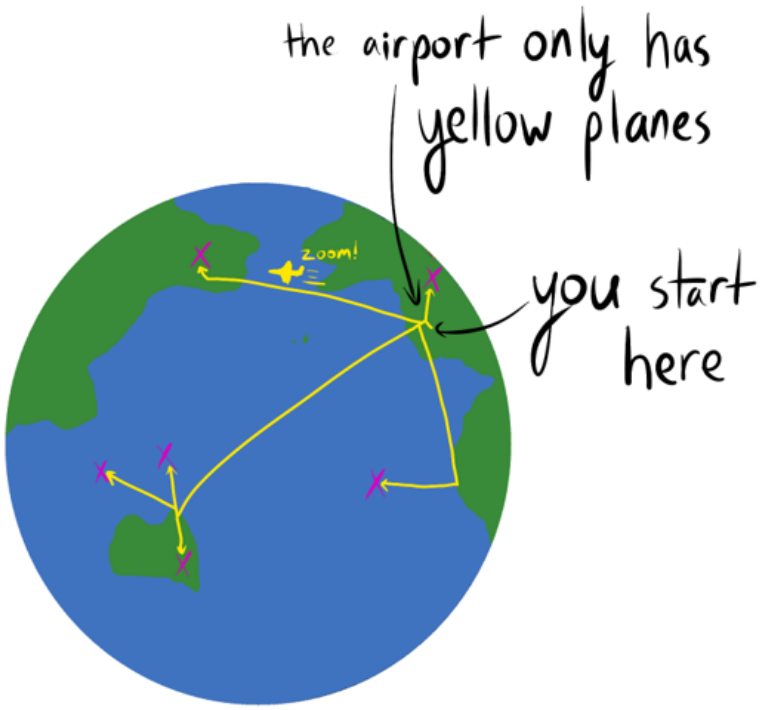
Answer: You can't with the pseudo-theorem due to the disjointness condition: you could die now, or you could die later, so the 'terminal options' aren't disjoint. However, the real theorems do suggest this. Supposing that death induces a generic 'game over' screen, touching the ghosts without a power-up traps the agent in that solitary 1-cycle.

But there are thousands of other 'terminal options'; under most reasonable state reward distributions (which aren't too positively skewed), most agents maximize average reward over time by navigating to one of the thousands of different cycles which the agent can only reach by avoiding ghosts. In contrast, most agents don't maximize average reward by navigating to the 'game over' 1-cycle. So, under e.g. the maximum-entropy uniform state reward distribution, most agents avoid the ghosts.

Be careful applying this theorem

The results inspiring the above pseudo-theorem are easiest to apply when the "terminal option" sets are disjoint: you're choosing to be able to reach one set, or another. One thing which Theorem 6.9 says is: since reward is IID, then two "similar terminal options" are equally likely to be optimal *a priori*. If choice A lets you reach more "options" than choice B does, then choice A yields greater POWER and has greater optimality probability, *a priori*.

Theorem 6.9's applicability depends on what the agent can do.



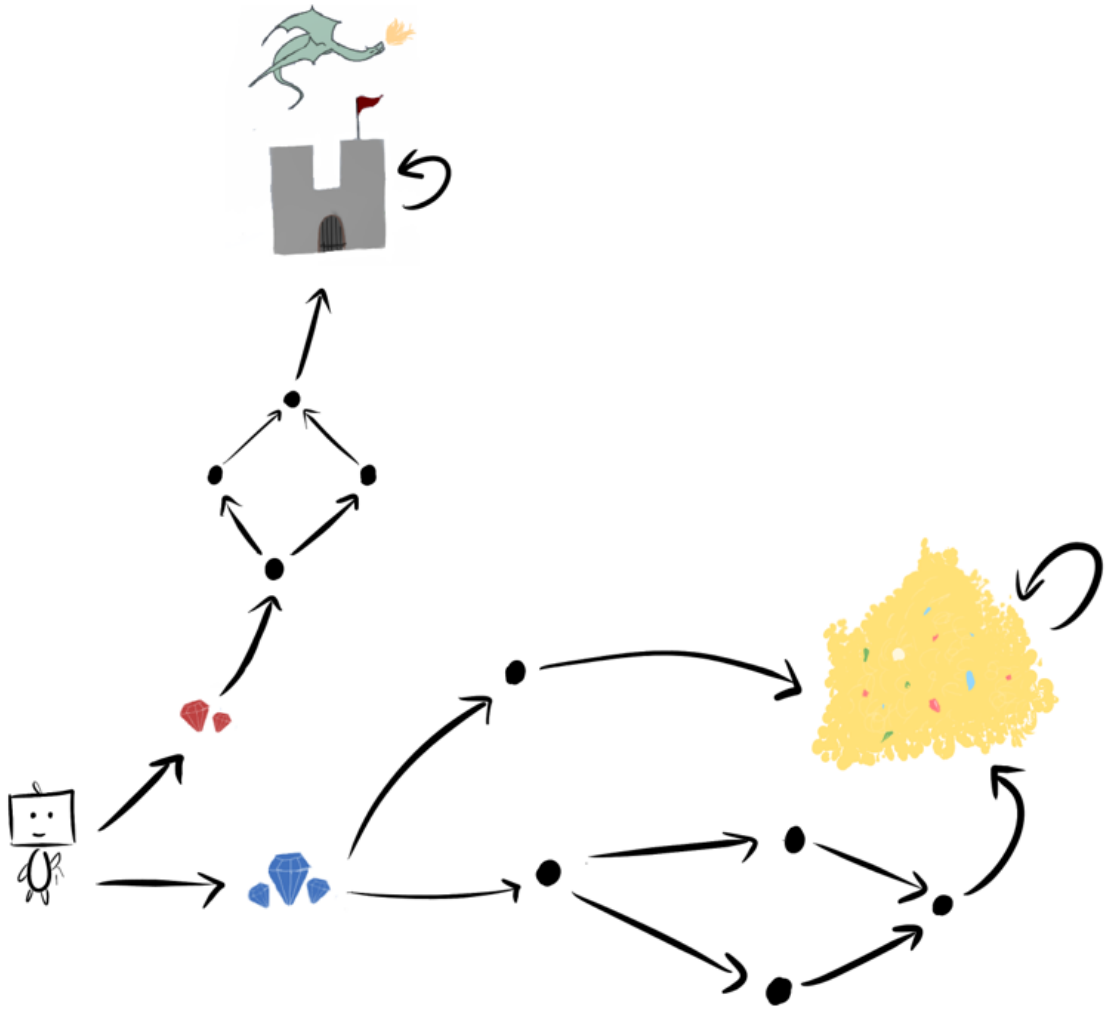
> To travel as quickly as possible to a randomly selected coordinate on Earth, one likely begins by driving to the nearest airport. Although it's possible that the coordinate is within driving distance, it's not likely. Driving to the airport is convergently instrumental for travel-related goals.

But wait! What if you have a private jet that can fly anywhere in the world? Then going to the airport isn't convergently instrumental anymore.

Generally, it's hard to know what's *optimal* for most goals. It's easier to say that some small set of "terminal options" has *low* optimality probability and *low* POWER. For example, this is true of shutdown, if we represent hard shutdown as a single terminal state: *a priori*, it's improbable for this terminal state to be optimal among all possible terminal states.

Having "strictly more options" is more probable under optimality and POWER-seeking for *all* discount rates

Sometimes, one course of action gives you "strictly more options" than another. Consider another MDP with IID reward:



The right blue gem subgraph contains a "copy" of the upper red gem subgraph. From this, we can conclude that going right to the blue gems seeks POWER and is more probable under optimality for *all discount rates between 0 and 1*!

Theorem summary ("Transient options"). If actions a and a′ let you access disjoint parts of the state space, and a′ enables "trajectories" which are "similar" to a subset of the "trajectories" allowed by a, then a seeks more POWER and is more probable under optimality than a′ for all $0 \leq \gamma \leq 1$.

This result is extremely powerful because it doesn't care about the discount rate, but the similarity condition may be hard to satisfy.

These two theorems give us a formally correct framework for reasoning about generic optimal behavior, even if we aren't able to compute any individual optimal policy! They reduce questions of POWER-seeking to checking graphical conditions.

Even though my results apply to stochastic MDPs of any finite size, we illustrated using known toy environments. However, this MDP "model" is rarely explicitly specified. Even so, ignorance

of the model does not imply that the model disobeys these theorems. Instead of claiming that a *specific model* accurately represents the task of interest, I think it makes more sense to argue that no reasonable model could fail to exhibit convergent instrumentality and POWER-seeking. For example, if deactivation is represented by a single state, no reasonable model of the MDP could have most agents agreeing to be deactivated.

Conclusion

In real-world settings, it seems unlikely *a priori* that the agent's optimal trajectories run through the relatively smaller part of future in which it cooperates with humans. These results translate that hunch into mathematics.

Explaining catastrophes

AI alignment research often feels slippery. We're trying hard to become less confused about basic questions, like:

- What are "agents"?
- Do people even have "values", and should we try to get the AI to learn them?
- What does it mean to be "corrigible", or "deceptive"?
- What are our machine learning models even doing?

We have to do philosophical work while in a state of significant confusion and ignorance about the nature of intelligence and alignment.

In this case, we'd noticed that slight reward function misspecification seems to lead to doom, but we didn't *really* know why. Intuitively, it's pretty obvious that most agents don't have deactivation as their dream outcome, but we couldn't actually point to any formal explanations, and we certainly couldn't make precise predictions.

On its own, Goodhart's law doesn't explain why optimizing proxy goals leads to catastrophically bad outcomes, instead of just less-than-ideal outcomes.

I think that we're now starting to have this kind of understanding. I suspect that power-seeking is why capable, goal-directed agency is so dangerous by default. If we want to consider more benign alternatives to goal-directed agency, then deeply understanding the rot at the heart of goal-directed agency is important for evaluating alternatives. This work lets us get a feel for the *generic incentives* of reinforcement learning at optimality.

Instrumental usefulness of this work

POWER might be important for reasoning about the strategy-stealing assumption (and I think it might be similar to what Paul Christiano means by "flexible influence over the future"). Evan Hubinger has already noted the utility of the distribution of attainable utility shifts for thinking about value-neutrality in this context (and POWER is another facet of the same phenomenon). If you want to think about whether, when, and why mesa optimizers might try to seize power, this theory seems like a valuable tool.

Optimality probability might be relevant for thinking about myopic agency, as the work formally describes how optimal action tends to change with the discount factor.

And, of course, we're going to use this understanding of power to design an impact measure.

Future work

There's a lot of work I think would be exciting, most of which I suspect will support our current beliefs about power-seeking incentives:

- These results assume you can see all of the world at once.
- These results assume the environment is finite.
- These results don't say anything about non-IID reward.
- These results don't *prove* that POWER-seeking is bad for other agents in the environment.
- These results don't prove that POWER-seeking is hard to disincentivize.
- Learned policies are rarely optimal.

That said, I think there's still an important lesson here. Imagine you have good formal reasons to suspect that typing random strings will usually blow up your computer and kill you. Would you then say, "I'm not planning to type random strings" and proceed to enter your thesis into a word processor? No. You wouldn't type *anything*, not until you really, really understand what makes the computer blow up sometimes.

Speaking to the broader debate taking place in the AI research community, I think a productive stance will involve investigating and understanding these results in more detail, getting curious about unexpected phenomena, and seeing how the numbers crunch out in reasonable models.

From *Optimal Policies Tend to Seek Power*:

> In the context of MDPs, we formalized a reasonable notion of power and showed conditions under which optimal policies tend to seek it. We believe that our results suggest that in general, reward functions are best optimized by seeking power. We caution that in realistic tasks, learned policies are rarely optimal – our results do not mathematically *prove* that hypothetical superintelligent RL agents will seek power. We hope that this work and its formalisms will foster thoughtful, serious, and rigorous discussion of this possibility.

Acknowledgements

This work was made possible by the Center for Human-Compatible AI, the Berkeley Existential Risk Initiative, and the Long-Term Future Fund.

Logan Smith (elriggs) spent an enormous amount of time writing Mathematica code to compute power and measure in arbitrary toy MDPs, saving me from computing many quintuple integrations by hand. I thank Rohin Shah for his detailed feedback and brainstorming over the summer of 2019, and I thank Andrew Critch for significantly improving this work through his detailed critiques. Last but not least, thanks to:

1. Zack M. Davis, Chase Denecke, William Ellsworth, Vahid Ghadakchi, Ofer Givoli, Evan Hubinger, Neale Ratzlaff, Jess Riedel, Duncan Sabien, Davide Zagami, and TheMajor for feedback on version 1 of this post.
2. Alex Appel (diffractor), Emma Fickel, Vanessa Kosoy, Steve Omohundro, Neale Ratzlaff, and Mark Xu for reading / giving feedback on version 2 of this post.

---

[1] Throughout *Reframing Impact*, we've been considering an agent's *attainable utility*: their ability to get what they want (their *on-policy value*, in RL terminology). Optimal value is a kind of "idealized" attainable utility: the agent's attainable utility were they to act optimally.

[2] Even though instrumental convergence was discovered when thinking about the real world, similar self-preservation strategies turn out to be convergently instrumental in *e.g.* Pac-Man.

Paper-Reading for Gears

Lesswrong has a fair bit of advice on how to evaluate the claims made in scientific papers. Most of this advice seems to focus on a single-shot use case - e.g. a paper claims that taking hydroxyhypotheticol reduces the risk of malignant examplitis, and we want to know how much confidence to put on the claim. It's very black-box-y: there's a claim that if you put X (hydroxyhypotheticol) into the black box (a human/mouse) then Y (reduced malignant examplitis) will come out. Most of the advice I see on evaluating such claims is focused around statistics, incentives, and replication - good general-purpose epistemic tools which can be applied to black-box questions.

But for me, this black-box-y use case doesn't really reflect what I'm usually looking for when I read scientific papers.

My goal is usually not to evaluate a single black-box claim in isolation, but rather to build a gears-level model of the system in question. I care about whether hydroxyhypotheticol reduces malignant examplitis only to the extent that it might tell me something about the internal workings of the system. I'm not here to get a quick win by noticing an underutilized dietary supplement; I'm here for the long game, and that means making the investment to understand the system.

With that in mind, this post contains a handful of thoughts on building gears-level models from papers. Of course, general-purpose epistemic tools (statistics, incentives, etc) are still relevant - a study which is simply wrong is unlikely to be much use for anything. So the thoughts and advice below all assume general-purpose epistemic hygiene as a baseline - they are things which seem more/less important when building gears-level models, *relative to* their importance for black-box claims.

I'm also curious to hear other peoples' thoughts/advice on paper reading specifically to build gears-level models.

Get Away From the Goal

Ultimately, we want a magic bullet to cure examplitis. But the closer a paper is to that goal, the stronger publication bias and other memetic distortions will be. A flashy, exciting result picked up by journalists will get a lot more eyeballs than a failed replication attempt.

But what about a study examining the details of the interaction between FOXO, SIRT6, and WNT-family signalling molecules? That paper will not ever make the news circuit - laypeople have no idea what those molecules are or why they're interesting. There isn't really a "negative result" in that kind of study - there's just an open question: "do these things interact, and how?". Any result is interesting and likely to be published, even though you won't hear about it on CNN.

In general, as we move more toward boring internal gear details that the outside world doesn't really care about, we don't need to worry as much about incentives - or at least not the same *kinds* of incentives.

Zombie Theories

Few people want to start a fight with others in their field, even when those others are wrong. There is little incentive to falsify the theory of somebody who may review your future papers or show up to your talk at a conference. It's much easier to say "examplitis is a complex multifactorial disease and all these different lines of research are valuable and important, kumbayah".

The result is zombie theories: theories which are pretty obviously false if you spend an hour looking at the available evidence, but which are still repeated in background sections and review articles.

One particularly egregious example I've seen is the idea that a shift in the collagen:elastin ratio is (at least partially) responsible for the increased stiffness of blood vessels in old age. You can find this theory in review articles and even textbooks. It's a nice theory: new elastin is not produced in adult vasculature, and collagen is much stiffer, so over time we'd expect the elastin to break down and collagen to bear more stress, increasing overall stiffness. But if we go look for studies which directly measure the collagen:elastin ratio in the blood vessels... we mostly find no significant change with age (rat, human, rat); one study even finds more elastin relative to collagen in older humans.

Ignore the Labels on the Box

Scientists say lots of things which are misleading, easily confused, or aren't actually supported by their experiments . That doesn't mean the experiment is useless, it just means we should ignore the mouth-motions and look at what the experiment and results actually were. As an added bonus, this also helps prevent misinterpreting what the paper authors meant.

An example: many authors assert that both (1) atherosclerosis is a universal marker of old age among humans and most other mammals, and (2) atherosclerosis is practically absent among most third-world populations. What are we to make of this? Ignore the mouth motions, look for data. In this case, it looks like atherosclerosis does universally grow very rapidly with age in all populations examined, but still has much lower overall levels among third-world populations after controlling for age - e.g. ~⅓ as prevalent in most age brackets in 1950's India compared to Boston.

Read Many Dissimilar Papers: Breadth > Depth

For replication, you want papers which are as similar as possible, and establishing very high statistical significance matters. For gears-level models, you want papers which do very different things, but impinge on the same gears. You want to test a whole model rather than a particular claim, so finding qualitatively different tests is more important than establishing very high statistical significance. (You still need enough statistics to make sure any particular result isn't just noise, but high confidence will ultimately be established by incrementally updating on many different kinds of studies.)

For example, suppose I'm interested in the role of thymic involution as a cause of cancer. The thymus is an organ which teaches new adaptive immune cells (T-cells) to distinguish our own bodies from invaders, and it shrinks ("involutes") as we age.

Rather than just looking for thymus-cancer studies directly, I *move away from the goal* and look for general information on the gears of thymic involution. Eventually I find that castration of aged mice (18-24 mo) leads to complete restoration of the thymus in about 2 weeks. The entire organ completely regrows, and the T-cells return to the parameters seen in young mice. (Replicated here.) Obvious next question: does castration reduce cancer? It's used as a treatment for e.g. prostate cancer, but that's (supposedly) a different mechanism. Looking for more general results turns up this century-old study, which finds that castration prevents age-related cancer in mice - and quite dramatically so. Castrated old mice' rate of resistance to an implanted tumor was ~50%, vs ~5% for controls. (This study finds a similar result in rabbits.) Even more interesting: castration did not change the rate of tumor resistance in young mice - exactly what the thymus-mediation theory would predict.

This should *not*, by itself, lead to very high confidence about the castration -> thymus -> T-cell -> cancer model. We need more qualitatively different studies (especially in humans), and we need at least a couple studies looking directly at the thymus -> cancer link. But if we find a bunch of different results, each with about this level of support for the theory, covering interventions on each of the relevant variables, then we should have reasonable confidence in the model. It's not about finding a single paper which proves the theory for all time; it's about building up Bayesian evidence from many qualitatively different studies.

Mediation is Everything

Everything is correlated with everything else; any intervention changes everything.

That said, very few things are *directly* connected; the main value is finding variables which *mediate* causal influence. For instance, maybe hydroxyhypotheticol usually reduces malignant examplitis, but most of the effect goes away if we hold hypometabolicol levels constant. That's a powerful finding: it establishes that hypometabolicol is one of the internal gears between hydroxyhypotheticol and examplitis.

If I had to pick the single most important guideline for building gears-level models from papers, this would be it: mediation is the main thing we're looking for.

Approval Extraction Advertised as Production

This is a linkpost for http://benjaminrosshoffman.com/approval-extraction-advertised-as-production/

Paul Graham has a new essay out, The Lesson to Unlearn, on the desire to pass tests. It covers the basic points made in Hotel Concierge's The Stanford Marshmallow Prison Experiment. But something must be missing from the theory, because what Paul Graham did with his life was start Y Combinator, the apex predator of the real-life Stanford Marshmallow Prison Experiment. Or it's just false advertising.

As a matter of basic epistemic self-defense, the conscientious reader will want to read the main source texts for this essay before seeing what I do to them:

1. The Lesson to Unlearn
2. The Stanford Marshmallow Prison Experiment
3. Sam Altman's Manifest Destiny
4. Black Swan Farming
5. Sam Altman on Loving Community, Hating Coworking, and the Hunt for Talent

The first four are recommended on their own merits as well. For the less conscientious reader, I've summarized below according to my own ends, biases, and blind spots. You get what you pay for.

The Desire to Pass Tests hypothesis: a Brief Recap

The common thesis of *The Lesson to Unlearn* and *The Stanford Marshmallow Prison Experiment* is:

Our society is organized around tests imposed by authorities; we're trained and conditioned to jump through arbitrary hoops and pretend that's what we wanted to do all along, and the upper-middle, administrative class is strongly selected for the desire to do this. This is why we're so unhappy and so incapable of authentic living.

Graham goes further and points out that this procedure doesn't know how to figure out new and good things, only how to perform for the system the things it already knows to ask for. But he also talks about trying to teach startup founders to focus on real problems rather than passing Venture Capitalists' tests.

The rhetoric of Graham's essay puts his young advisees in the role of the unenlightened who are having a puzzling amount of trouble understanding advice like "the way you get lots of users is to make the product really great." Graham casts himself in the role of someone who has unlearned the lesson that you should just try to pass the test of whoever you're interacting with, implying that the startup accelerator (i.e. combination Venture Capital firm and cult) he co-founded, Y Combinator, is trying to do something outside the domain of Tests.

In fact, Graham's behavior as an investor has perpetuated and continues to perpetuate exactly the problem he describes in the essay. Graham is not behaving exceptionally poorly here - he got rich by doing better than the norm on this dimension - except by persuasively advertising himself as the Real Thing, confusing those looking for the actual real thing.

(The parallels with Effective Altruism, and LessWrong, should be obvious to those who've been following.)

Y Combinator is the apex predator of the real-life Stanford Marshmallow Prison Experiment

If you know anyone in the SF Bay Area startup scene, you know that Y Combinator is *the* place to go if you're an ambitious startup founder who wants a hand up. Here's a relevant quote from Tad Friend's excellent New Yorker profile of Sam Altman, the current head of Y Combinator:

> *Paul Graham considered the founders of Instacart, DoorDash, Docker, and Stripe, in their hoodies and black jeans, and said, "This is Silicon Valley, right here." All the founders were graduates of Y Combinator, the startup "accelerator" that Graham co-founded: a three-month boot camp, run twice a year, in how to become a "unicorn"—Valleyspeak for a billion-dollar company. Thirteen thousand fledgling software companies applied to Y Combinator this year, and two hundred and forty were accepted, making it more than twice as hard to get into as Stanford University.*
> *[...]*
> *Perhaps the most dispositive theory about YC is that the power of its network obviates other theories. Alumni view themselves as a kind of keiretsu, a network of interlocking companies that help one another succeed. "YC is its own economy," Harj Taggar, the co-founder of Triplebyte, which matches coders' applications with YC companies, said. Each spring, founders gather at Camp YC, in a redwood forest north of San Francisco, just to network—tech's version of the Bohemian Grove, only with more vigorous outdoor urination. When Altman first approached Kyle Vogt, the C.E.O. of Cruise, Vogt had been through YC with an earlier company, so he already knew its lessons. He told me, "I talked to five of my friends who had done YC more than once and said, 'Was it worth it the second time? Are you likely to receive higher valuations because of the brand, and because you're plugging into the network?' Across the board, they said yes."*
>
> *There really is no counter-theory. "The knock on YC," Andy Weissman, a managing partner at Union Square Ventures, told me, "is that on Demo Day their users are just YC companies, which entirely explains why they're all growing so fast. But how great to have more than a thousand companies willing to use your product!" It's not just that YC startups can get Airbnb and Stripe to use their apps; it's that the network's alumni honeycomb the Valley's largest companies. Many of the hundred and twenty-one YC startups that have been acquired over the years have been absorbed by Facebook, Apple, and Google.*

This matches the impression I've gotten from most of the people I've talked to about their startups, that Y Combinator is singularly important as a certifier of potential, and therefore gatekeeper to the kinds of network connections that can enable a fledgling business - especially one building business tools, the ostensible means of production - to get off the ground.

When interviewed by Tyler Cowen, Altman expressed a desire to move from being a singularly important gatekeeper to being the exclusive gatekeeper:

> *Someday we will fund all the companies in the world, all the good ones at least.*

Given Graham's values, one might have expected a sort of proactive talent scouting, to find people who've deeply invested in interesting ventures that don't fit the mold, and offer them some help scaling up. But the actual process is very different.

Selection Effects

Y Combinator, like the prestigious colleges Graham criticizes in his essay, has a formal application process. It receives many more applications for admission than it can accept, so it has to apply some strong screening filters. It seems to filter for people who are anxiously obsessed with quickly obtaining the approval of the test evaluators, who have been acculturated into upper-middle-class test-passing norms, and who are so obsessed with numbers going the "right" way that they'll distort the shape of their own bodies to satisfy an arbitrary success metric.

Anxious Preoccupied

In his interview with Sam Altman, Tyler Cowen asked about the profile of successful Y Combinator founders:

> *COWEN: Why is being quick and decisive such an important personality trait in a founder?*
>
> *ALTMAN: That is a great question. I have thought a lot about this because the correlation is clear, that one of the most fun things about YC is that, I think, we have more data points on what successful founders and bad founders look like than any other organization has had in the history of the world. We have that all in our heads, and that's great. So I can say, with a high degree of confidence, that this correlation is true.*
>
> *Being a fast mover and being decisive — it is very hard to be successful and not have those traits as a founder. Why that is, I'm not perfectly clear on, but I think it is something . . . about the only advantage that startups have or the biggest advantage that startups have over large companies is agility, speed, willing to make nonconsensus, concentrated bets, incredible focus. That's really how you get to beat a big company.*
>
> *COWEN: How quickly should someone answer your email to count as quick and decisive?*
>
> *ALTMAN: You know, years ago I wrote a little program to look at this, like how quickly our best founders — the founders that run billion-plus companies — answer my emails versus our bad founders. I don't remember the exact data, but*

it was mind-blowingly different. It was a difference of minutes versus days on average response times.

The kind of "decisiveness" Altman is talking about doesn't involve making research or business decisions that matter on the scale of months or weeks, but responding to emails in a few minutes. In other words, the minds Altman is looking for are not just generically decisive, but *quickly responsive* - not spending long slow cycles doing a new thing and following their own interest, but anxiously attentive to new inputs, jumping through his hoops fast. Similarly telling is the ten-minute interview in which he mainly looks for how responsive the interviewee is to cues from the interviewer:

This gets to the question . . . the most common question I get about Y Combinator is how can you make a decision in a 10-minute interview about who to fund? Where we might miss her is the upper filters of our application process.

We have far more qualified people that want to do YC each year than we can fund, but by the time we get someone in the room, by the time we can sit across the table and spend 10 minutes with somebody, as far as I know, we have never made a big mistake at that stage of the process. We've looked at tens of thousands — well, in person we've maybe looked at 10,000 companies.

These personality traits of determination and communication and the ability to articulate a vision for the world and explain how you're going to get that done — I used to think that that was so hard to assess in 10 minutes, it was maybe impossible to try, and YC interviews used to be like an hour. I now think that most of the time, we could get it right in five minutes.

When you have enough data points, when you meet enough people and get to watch what they go on to do — because the one thing that's hard in a 10-minute interview and the most important thing about evaluating someone is their rate of improvement. It's a little bit hard when you only get a single sample. But when you do this enough times, and you get to learn what to look for, it is incredible how good you can get at that.

While responsiveness is doubtless a valid test of some sort of intellectual aliveness and ability, it could easily take hours or days to integrate real, substantive new information; in ten minutes all one may be able to do is *perform* responsiveness.

In case there was any doubt as to whether this attitude is consistent with supporting technical innovation, later in the interview he says unconditionally that Y Combinator would fund a startup founded by James Bond (a masculine wish-fulfillment fantasy, whose spy work is mostly interpersonal intrigue), but not by Q (the guy in the Bond movies who develops cool gadgets for Bond to use), unless he has a "good cofounder."

Conventionally Successful

Then consider the kinds of backgrounds that make a good Y Combinator founder:

COWEN: I come from northern Virginia, the Washington, DC, area. We have very few geniuses. The few that we have tend to be crazy and sometimes destructive. We have what I would call —

ALTMAN: They're very stable, though.

COWEN: —a lot of upper-middle-class intellectual talent. People who are pretty smart and good at something. When it comes to spotting good upper-middle-class intellectual talent, do you think you have the same competitive edge as with spotting geniuses who will make rapidly scalable tech companies?

ALTMAN: I think that's how I'd characterize myself: upper middle class, pretty smart, not a super genius by any means. It turns out I've met many people smarter than me, but I would say I've only ever met a handful of people that are obviously more curious than me.

I don't think raw IQ is my biggest strength — pretty good, to be clear, but the chances of me winning a Nobel Prize in physics are low. I think physics is a bad field at this point, unfortunately. What we spot — you do have to be pretty smart to be a successful founder, but that is not where I look for people to be true outliers.

COWEN: Given that self-description — assuming that I accept it, and I'm not sure I do — do you think you're as good at spotting upper-middle-class intellectual talent as superstar founders? Let's say we put you in charge —

ALTMAN: There's a statement here that's just bad about the world, but I think if you look at most successful founders, they are pretty smart, upper-middle-class people. They are very rarely the children of super successful people. They are very rarely born in real poverty. They are very rarely the absolute smartest people who otherwise would win a Fields Medal. They are never dumb, but upper-middle-class, pretty smart people that have grit and drive and creativity and vision and edge and a different way of thinking about the world. That is what I think I'm good at spotting, and that is what I think are good founders. There's a whole bunch of reasons why that's a sad statement about the world, but there it is.

If you look at most successful founders, they are pretty smart, upper-middle-class people. They are very rarely the children of super successful people. They are very rarely born in real poverty. They are very rarely the absolute smartest people who otherwise would win a Fields Medal. They are never dumb, but upper-middle-class, pretty smart people that have grit and drive and creativity and vision and edge and a different way of thinking about the world. That is what I think I'm good at spotting, and that is what I think are good founders.

COWEN: So someone else has to find the geniuses.

ALTMAN: Again, I don't want to go for false modesty here. I think I'm a smart person. The founders we fund are smart people. I would have maybe said 10 years ago that raw IQ is the thing that matters most for founders. I've updated my view on that.

In other words, the people who best succeed at Y Combinator's screening process are exactly the people you'd expect to score highest at Desire To Pass Tests.

A High Health Score is Better Than Health

Then consider this little detail:

COWEN: In the world of tech startups, venture capital, what is weight lifting correlated with?

ALTMAN: Weight lifting?
COWEN: Weight lifting. Taleb tells us, in New York City, weight lifting is correlated with supporting Trump, but I doubt if that's true in —
ALTMAN: It's not true here.
COWEN: Yes.
ALTMAN: I think it's correlated with successful founders. It's fun to have numbers that go up and to the right. The most fun thing for me about weight lifting is . . . I'm basically financially illiterate. I can't build an Excel model for anything. I can't read a balance sheet. But my Excel model for weight lifting is beautiful because they're numbers that go up and to the right, and it's really fun to play around with that.

(For context, Taleb's attitude towards weightlifting is that it builds a kind of "antifragile" robustness to small perturbations, which is consistent with the sort of risk-bearing behavior that builds a sustainable society. See Strength Training is Learning from Tail Events for Taleb's own account. By contrast, Altman's idea of a weightlifter is someone who just likes to see the numbers go in the correct direction - what Taleb would call an Intellectual Yet Idiot, the exact "academico-bureaucrat" class from which Y Combinator draws its founders, who pretend that their measurements capture more about what they study than they do, and offload the risks their models can't account for onto others. For more on Taleb's outlook, read *Skin in the Game*.)

Hotel Concierge's story about weight in *The Stanford Marshmallow Prison Test* makes an interesting comparison:

MTV CONFESSION CAM: I was an accidental anorexic.

I was 17 years old when I moved into the college dorms and decided that I wanted a six-pack. I had never thought about my body too much before that—I didn't play sports, I didn't care about fashion, and I spent most of my time daydreaming about fantasy novels and videogames. In the dorms, however, I finally realized that girls existed. I wasn't sure about how to get a girlfriend on purpose, but I was pretty sure it had something to do with "abs." So I decided to work on that.

I began running for 45 minutes three times a week, along with daily stretching, push-ups, and crunches. After hearing a fire-and-brimstone "Talk About Nutrition!" presentation in the dining commons, I decided that I would have to change my diet as well. I stopped eating sweets of any sort. Increased lean protein intake. All breaded objects became whole wheat. But this didn't seem enough. Food, I decided, was an ephemeral pleasure, whereas a well-sculpted body was a constant joy to live in and behold. Why bother with anything but the healthiest of foods? After some trial and error, I decided upon the optimal meal plan.

Breakfast: Oatmeal, one orange, one kiwi.
Lunch: Salad (lettuce with kidney & garbanzo beans), PB&J, hardboiled egg.
Snack: Orange.
Dinner: Cheerios and milk, salad, orange.

I was proud of my discipline, and the nights when I went to sleep hungry only intensified my pride, as I first ignored and then began to appreciate the sharp jabs of an empty stomach. I was no longer getting fit to attract girls—I was getting fit for me. After a month of running, my progress seemed to flatline. I added weights to my regimen. I wasn't sure how many reps to do, so I decided to just lift until my arms went limp. After another month, I had developed only mild abdominal

definition. I decided to step it up: 200 push-ups upon waking every day, stretches and crunches in the evening. I wasn't perfect. Sometimes I would give in to temptation and eat something off-diet—if someone took me out to lunch, or if I had a cookie at a school event—and I would feel guilty and sad for a while, but before long I would regain my composure and vow to increase my exercise regimen in the next few days to make up for my setback.

After three months, I went home for winter break. My mom said that I looked like an Auschwitz survivor. I said that was a huge overreaction. I asked my dad what he thought. My dad said that I looked a little skinny but that there was nothing wrong with getting into shape. I went back to school and kept up my routine. My strength declined. I wasn't sure why. My ribs could be individually grasped. I visited home and weighed myself at 107 pounds.

"That's really low," my mom said.
"It's not that bad," I said.

My mom convinced me to go to the campus nutritionist. The nutritionist, who was middle-aged, and blonde, and wore glasses, and smiled a lot, told me that I had lost 35 pounds in four and half months, and that remaining at this weight could be dangerous.

"I was just trying to be fit," I explained to her.
"Your current weight isn't fit," she said. "It's not healthy."
"I really don't want to be fat," I said.
"You're not going to be fat," she said. "That's not your body type."
"I guess. I didn't mean to. I was just trying to eat healthy."
"Right now, you can eat whatever you want," she said. "You need to gain some weight. Back into the 140s, at least."
"I'm never sure what to eat. I don't want to eat too much."
"I'll help you come up with a meal plan," she said. "We can figure out what you need to do."

I felt tremendous relief. This was a test I could pass.

Treatment Effects

The admissions process is not the end of the acculturation. In Even artichokes have doubts, Martina Keegan wrote about how people admitted to Yale don't stop the sort of anxious approval-seeking that got them in. Instead, having been conditioned into that behavior pattern, they're easy targets for recruitment by further generic prestige-awarders like McKinsey or investment banks, even though they know it won't get them much they actually want, and report that they expect it will cause their future behavior to drift farther from what they see as the optimum.

At Y Combinator, the situation is even worse. According to Friend's New Yorker article, the founders are deliberately forced into situations where short-run survival depends on getting approval *now*:

> *A founder's first goal, Graham wrote, is becoming "ramen profitable": spending thriftily and making just enough to afford ramen noodles for dinner. "You don't want to give the founders more than they need to survive," Jessica Livingston said. "Being lean forces you to focus. If a fund offered us three hundred thousand*

dollars to give the founders, we wouldn't take it." (Many of YC's seventeen partners, wealthy from their own startups, receive a salary of just twenty-four thousand dollars and get most of their compensation in stock.) This logic, followed to its extreme, would suggest that you shouldn't even take YC's money, and many successful startups don't. Only twenty per cent of the Inc. 500, the five hundred fastest-growing private companies, raised outside funding. But the YC credential, and the promise that it will turn you into a juggernaut, can be hard to resist.

Y Combinator forces a focus on "growth" feedback short-run even though this isn't a good proxy for long-run success. Friend continues:

Nearly all YC startups enter the program with the same funding, and thus the same valuation: $1.7 million. After Demo Day, their mean valuation is ten million. There are several theories about why this estimation jumps nearly sixfold in three months. One is that the best founders apply to the best accelerator, and that YC excels at picking formidable founders who would become successful anyway. Paul Buchheit, who ran the past few batches, said, "It's all about founders. Facebook had Mark Zuckerberg, and MySpace had a bunch of monkeys."

The corollary is that Y Combinator makes its companies more desirable by teaching them how to tell their story on Demo Day. The venture capitalist Chris Dixon, who admires YC, said, "The founders are so well coached that they know exactly how to reverse-engineer us, down to demonstrating domain expertise and telling anecdotes about their backgrounds that show perseverance and courage."

In the winter batch, the pitches followed an invariable narrative of imminent magnitude: link yourself to a name-brand unicorn ("We're Uber for babysitting . . . Stripe for Africa . . . Slack for health care"), or, if there's no apt analogue, say, "X is broken. In the future, Y will fix X. We're already doing Y." Then express your premise as a chewy buzzword sandwich: We "leverage technology to achieve personalization in a fully automated way" (translation: individuated shampoo). Paul Graham cheerfully acknowledged that, by instilling message discipline, "we help the bad founders look indistinguishable from the good ones."

The counter-theory is that YC actually does make its companies better, by teaching them to focus on growth above all, thereby eliminating distractions such as talking to the tech press or speaking at conferences or making cosmetic coding tweaks. YC's gold standard for revenue growth is ten per cent a week, which compounds to 142x a year. Failing that, well, tell a story of some other kind of growth. On Demo Day, one company announced that it had enjoyed "fifty-per-cent word-of-mouth growth," whatever that might be. Sebastian Wallin told me that his security company, Castle, raised $1.8 million because "we managed to find a good metric to show growth. We tried tracking installations of our product, but it didn't look good. So we used accounts protected, a number that showed roughly thirty-per-cent growth through the course of YC—and about forty per cent of the accounts were YC companies. It was a perfect fairy-tale story."

The truth is that rapid growth over a long period is rare, that the repeated innovation required to sustain it is nearly impossible, and that certain kinds of uncontrollable growth turn out to be cancers. Last year, after a series of crises at Reddit, Altman, who is on its board, convinced Steve Huffman, the co-founder of the company, to return as C.E.O. Huffman said, "I immediately told Sam, 'Don't get on my ass about growth. I'm not in control of it.' Every great startup—

Facebook, Airbnb—has no idea why it's growing at first, and has to figure that out before the growth stalls. Growth masks all problems."

This despite knowing that the intervention could well be doing more harm than good. From the interview with Cowen, a comment on growth:

COWEN: You once said, "Growth masks all problems." Are there exceptions to that?

ALTMAN: Cancer?

[laughter]

ALTMAN: I mean, clearly, yes. I don't mean that so flippantly. There is —

COWEN: There's an article in the Jerusalem Post today: someone credible claiming that cancer has been cured. I don't know if you saw that.

ALTMAN: I didn't see that, but I do — having talked to many biologists working in the field, I will say there is a surprising amount of optimism that we are within a decade or two of that being true.

It's not an area where I feel anywhere near expert enough to comment on the validity of that statement, and I think it's always dangerous to just trust what other smart people say, especially when they have an incentive to hawk their own book, but it does seem like a lot of people believe that.

Growth is bad in plenty of times, but it does mask a lot of problems. A statement that I wouldn't make is that growth is always an inherent good, although I do think — I think you've said something like this, too — that sustainable economic growth is almost always a moral good.

Something that I think a lot of the current problems in the country can be traced to is the decline in that. And part of what motivates me to work on Y Combinator and OpenAI is getting back to that, getting back to sustainable economic growth, getting back to a world where most people's lives get better every year and that we feel the shared spirit of success is really important.

And growth feels good. It does mask a lot of problems, but there definitely are individual instances where you'd be better off with slower growth for whatever reason.

Obviously, in any enterprise trying to do a specific thing, some kinds of measurable activity will have to increase at some point. But this obsession with measured revenue growth (or analogues to it) early on leads to performative absurdities like "fifty-per-cent word-of-mouth growth," which are not words that would come out of the mouth of someone trying to, as Paul Graham advises in his essay, "make the product really great."

Altman knows he's not doing the right thing, or the thing that would make him the most money. But he's doing the fastest-growing thing.

Moloch Blindness

How is it that someone like Graham could hold so strongly and express so eloquently the opinion that the most important lesson to unlearn is the desire to pass tests, and then create an institution like Y Combinator?

This isn't the only such contradiction in Graham's words and actions. Friend's New Yorker article describes Graham's attitude towards meanness, and contrasts this with Altman's actual character, revealing a similar pattern:

> *YC prides itself on rejecting jerks and bullies. "We're good at screening out assholes," Graham told me. "In fact, we're better at screening out assholes than losers. All of them start off as losers—and some evolve." The accelerator also suggests that great wealth is a happy by-product of solving an urgent problem. This braiding of altruism and ambition is a signal feature of the Valley's self-image. Graham wrote an essay, "Mean People Fail," in which—ignoring such possible counterexamples as Jeff Bezos and Larry Ellison—he declared that "being mean makes you stupid" and discourages good people from working for you. Thus, in startups, "people with a desire to improve the world have a natural advantage." Win-win.*
> *[...]*
> *[Altman] attended Stanford University, where he spent two years studying computer science, until he and two classmates dropped out to work full time on Loopt, a mobile app that told your friends where you were. Loopt got into Y Combinator's first batch because Altman in particular passed what would become known at YC as the young founders' test: Can this stripling manage adults? He was a formidable operator: quick to smile, but also quick to anger. If you cross him, he'll joke about slipping ice-nine into your food. (Ice-nine, in Kurt Vonnegut's "Cat's Cradle," annihilates everything it touches that contains water.) Paul Graham, noting Altman's early aura of consequence, told me, "Sam is extremely good at becoming powerful."*

So, Graham claimed that mean people fail, and then selected someone with a noticeable mean streak to run Y Combinator. Likewise, he wrote that test-passing isn't good for growing a business, and then promoting an obligate test-passer, who then remade his institution to optimize for test-passing.

I think the root process generating this sort of bait-and switch must be something like the following:

There's a way to succeed (i.e. become a larger share of what exists) through production, and a way to succeed through purely adversarial (i.e. zero-sum) competition. These are incompatible strategies, so that productive people will do poorly in zero-sum systems, and vice versa. The productive strategy really is good, and in production-oriented contexts, a zero-sum attitude really is a disadvantage.

Graham has a natural affinity for production-based strategies which allowed him to acquire various kinds of capital. He blinds himself to the existence of adversarial strategies, so he's able to authentically claim to think that e.g. mean people fail - he just forgets about Jeff Bezos, Larry Ellison, Steve Jobs , and Travis Kalanick because they are too anomalous in his model, and don't feel to him like central cases of success.

This is a case of fooling oneself to avoid confronting malevolent power. It's the path towards the true death, so if you want to stay aligned with truth and life, you'll have to

look for alternatives. To keep track of anomalies, at least in your internal bookkeeping; to conceal and lie, if you have to, to protect yourself.

If, to participate in higher growth rates, you have to turn into something else, then in what sense is it *you* that's getting to grow faster? Moloch, as Scott Alexander points out, offers "you" power in exchange for giving up what you actually care about - but this means, offering *you* no substantive concessions. *For what is a person profited, if they shall gain the whole world, and lose their own soul?*

Thus blinded, Graham writes about the virtues of production-based strategies as though they were the *only* way to succeed. He then sets up an institution optimizing for "success" directly, rather than specifically for production-based strategies. But in the environment in which he's operating, adversarial strategies can scale faster. Of course, just because adversarial strategies scale faster doesn't mean they make you richer faster - and as we'll see below, selling out is *not*, according to Graham's perspective, the way to maximize returns. But faster growth feels more successful. So he ends up selling out his credibility to the growth machine. Or, as Hotel Concierge called it, the *Stanford Marshmallow Prison Experiment*. (It's perhaps not a coincidence that the Stanford brand is most prestigious in the startup / "tech" scene.)

Here's the thing, though. Graham knows he's doing the wrong thing. He confessed in Black Swan Farming that even though doing the right thing would work out better for him in the long run, he just isn't getting enough positive feedback, so it's psychologically intolerable:

> *The one thing we can track precisely is how well the startups in each batch do at fundraising after Demo Day. But we know that's the wrong metric. There's no correlation between the percentage of startups that raise money and the metric that does matter financially, whether that batch of startups contains a big winner or not.*
>
> *Except an inverse one. That's the scary thing: fundraising is not merely a useless metric, but positively misleading. We're in a business where we need to pick unpromising-looking outliers, and the huge scale of the successes means we can afford to spread our net very widely. The big winners could generate 10,000x returns. That means for each big winner we could pick a thousand companies that returned nothing and still end up 10x ahead.*
> *[...]*
> *We can afford to take at least 10x as much risk as Demo Day investors. And since risk is usually proportionate to reward, if you can afford to take more risk you should. What would it mean to take 10x more risk than Demo Day investors? We'd have to be willing to fund 10x more startups than they would. Which means that even if we're generous to ourselves and assume that YC can on average triple a startup's expected value, we'd be taking the right amount of risk if only 30% of the startups were able to raise significant funding after Demo Day.*
>
> *I don't know what fraction of them currently raise more after Demo Day. I deliberately avoid calculating that number, because if you start measuring something you start optimizing it, and I know it's the wrong thing to optimize. But the percentage is certainly way over 30%. And frankly the thought of a 30% success rate at fundraising makes my stomach clench. A Demo Day where only 30% of the startups were fundable would be a shambles...*

> *For better or worse that's never going to be more than a thought experiment. We could never stand it. How about that for counterintuitive? I can lay out what I know to be the right thing to do, and still not do it.*

So instead, he does the wrong thing, knowingly, on purpose, but tries to pretend otherwise to himself. Sad!

*****

Related: OpenAI makes humanity less safe, Is Silicon Valley real?, In a world… of venture capital

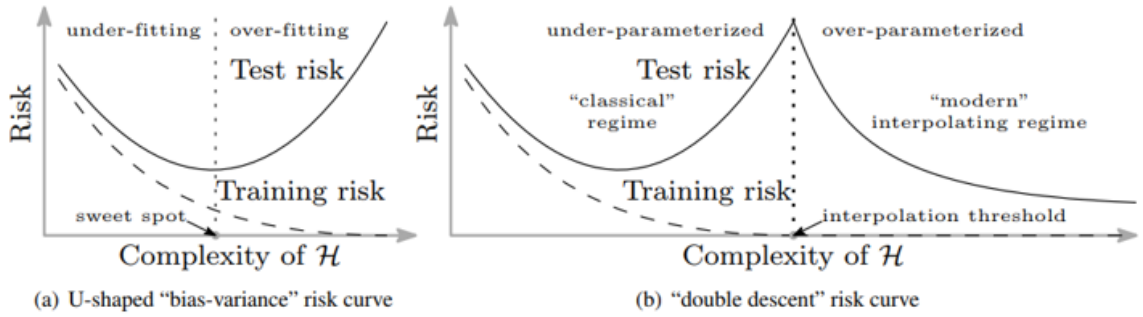
Understanding “Deep Double Descent”

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

If you're not familiar with the double descent phenomenon, I think you should be. I consider double descent to be one of the most interesting and surprising recent results in analyzing and understanding modern machine learning. Today, Preetum et al. released a new paper, “Deep Double Descent,” which I think is a big further advancement in our understanding of this phenomenon. I'd highly recommend at least reading the summary of the paper on the OpenAI blog. However, I will also try to summarize the paper here, as well as give a history of the literature on double descent and some of my personal thoughts.

Prior work

The double descent phenomenon was first discovered by Mikhail Belkin et al., who were confused by the phenomenon wherein modern ML practitioners would claim that “bigger models are always better” despite standard statistical machine learning theory predicting that bigger models should be more prone to overfitting. Belkin et al. discovered that the standard bias-variance tradeoff picture actually breaks down once you hit approximately zero training error—what Belkin et al. call the “interpolation threshold.” Before the interpolation threshold, the bias-variance tradeoff holds and increasing model complexity leads to overfitting, increasing test error. After the interpolation threshold, however, they found that test error actually starts to go down as you keep increasing model complexity! Belkin et al. demonstrated this phenomenon in simple ML methods such as decision trees as well as simple neural networks trained on MNIST. Here's the diagram that Belkin et al. use in their paper to describe this phenomenon:



(a) U-shaped “bias-variance” risk curve (b) “double descent” risk curve

Figure 1: Curves for training risk (dashed line) and test risk (solid line). (a) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (b) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the “classical” regime) together with the observed behavior from using high complexity function classes (i.e., the “modern” interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

Belkin et al. describe their hypothesis for what's happening as follows:

> All of the learned predictors to the right of the interpolation threshold fit the training data perfectly and have zero empirical risk. So why should some—in particular, those from richer functions classes—have lower test risk than others? The answer is that the capacity of the function class does not necessarily reflect how well the predictor matches the inductive bias appropriate for the problem at hand. [The inductive bias] is a form of Occam's razor: the simplest explanation compatible with the observations should be preferred. By considering larger function classes, which contain more candidate predictors compatible with the data, we are able to find interpolating functions that [are] "simpler". Thus increasing function class capacity improves performance of classifiers.

I think that what this is saying is pretty magical: in the case of neural nets, it's saying that SGD just so happens to have the right inductive biases that letting SGD choose which model it wants the most out of a large class of models with *the same training performance* yields significantly better test performance. If you're right on the interpolation threshold, you're effectively "forcing" SGD to choose from a very small set of models with perfect training accuracy (maybe only one realistic option), thus ignoring SGD's inductive biases completely—whereas if you're past the interpolation threshold, you're letting SGD choose which of many models with perfect training accuracy it prefers, thus allowing SGD's inductive bias to shine through.

I think this is strong evidence for the critical importance of implicit simplicity and speed priors in making modern ML work. However, such biases also produce strong incentives for mesa-optimization (since optimizers are simple, compressed policies) and pseudo-alignment (since simplicity and speed penalties will favor simpler, faster proxies). Furthermore, the arguments for the universal prior and minimal circuits being malign suggest that such strong simplicity and speed priors could also produce an incentive for deceptive alignment.

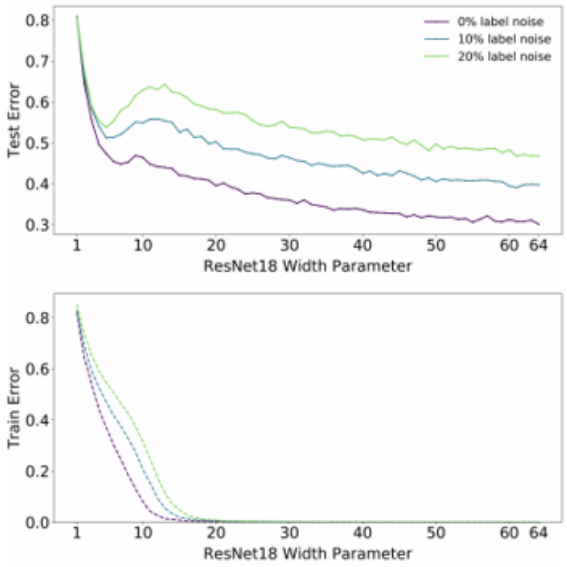
"Deep Double Descent"

Now we get to Preetum et al.'s new paper, "Deep Double Descent." Here are just some of the things that Preetum et al. demonstrate in "Deep Double Descent:"

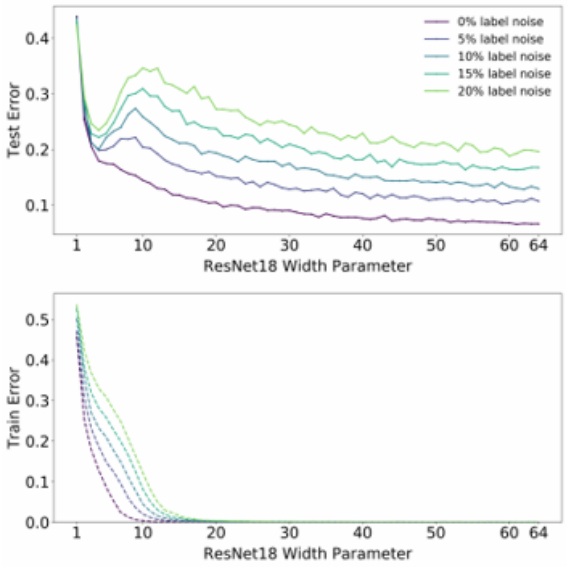
1. double descent occurs across a wide variety of different model classes, including ResNets, standard CNNs, and Transformers, as well as a wide variety of different tasks, including image classification and language translation,
2. double descent occurs not just as a function of model size, but also as a function of *training time* and *dataset size*, and
3. since double descent can happen as a function of dataset size, **more data can lead to worse test performance!**

Crazy stuff. Let's try to walk through each of these results in detail and understand what's happening.

First, double descent is a highly universal phenomenon in modern deep learning. Here is double descent happening for ResNet18 on CIFAR-10 and CIFAR-100:

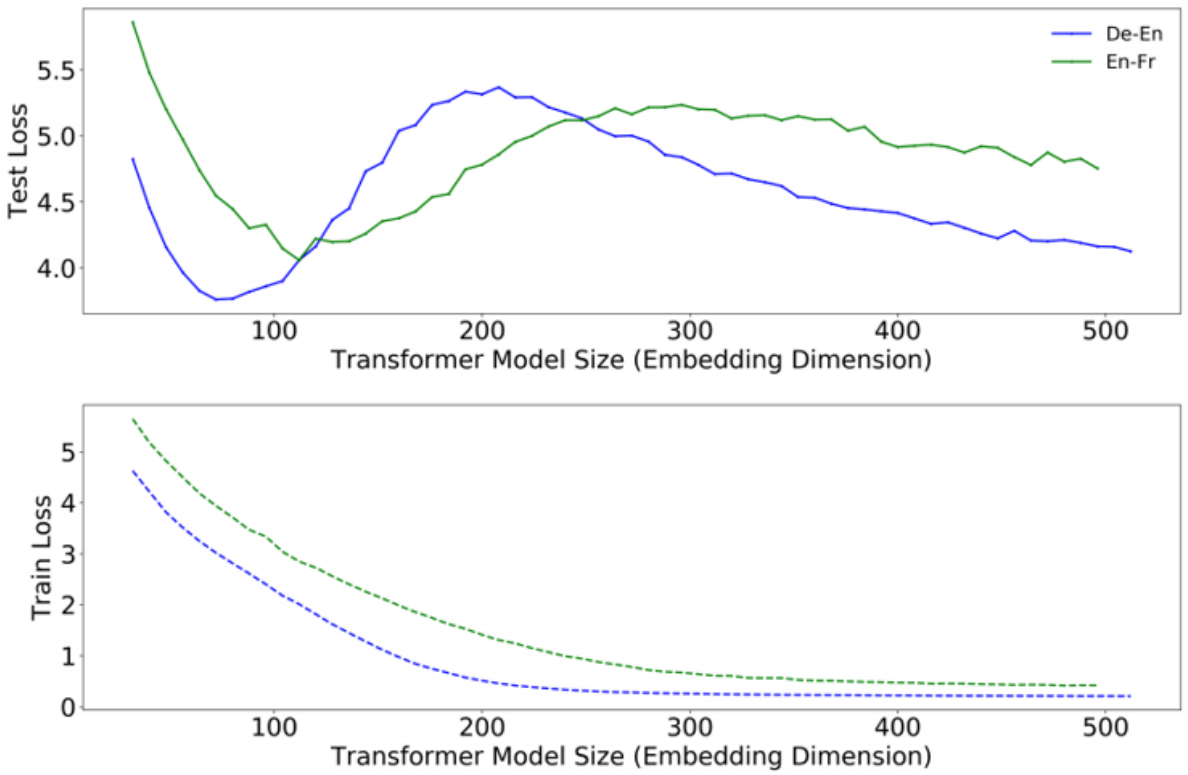


(a) **CIFAR-100.** There is a peak in test error even with no label noise.



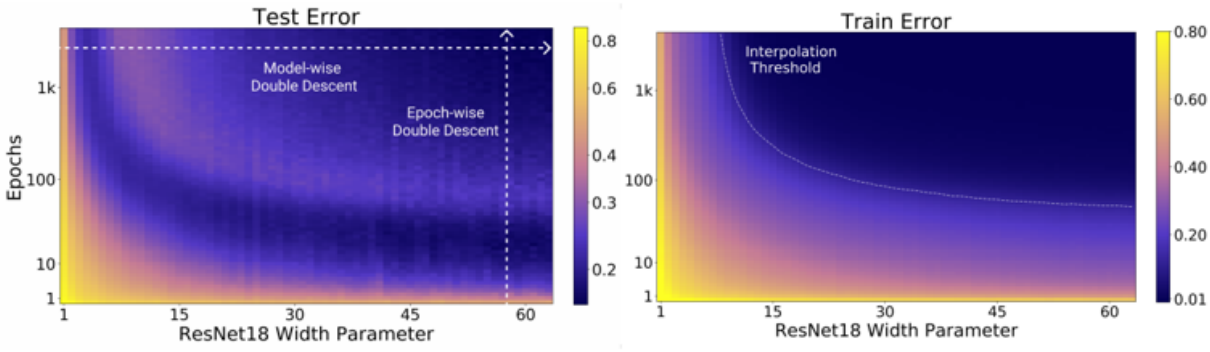
(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.

And again for a Transformer model on German-to-English and English-to-French translation:



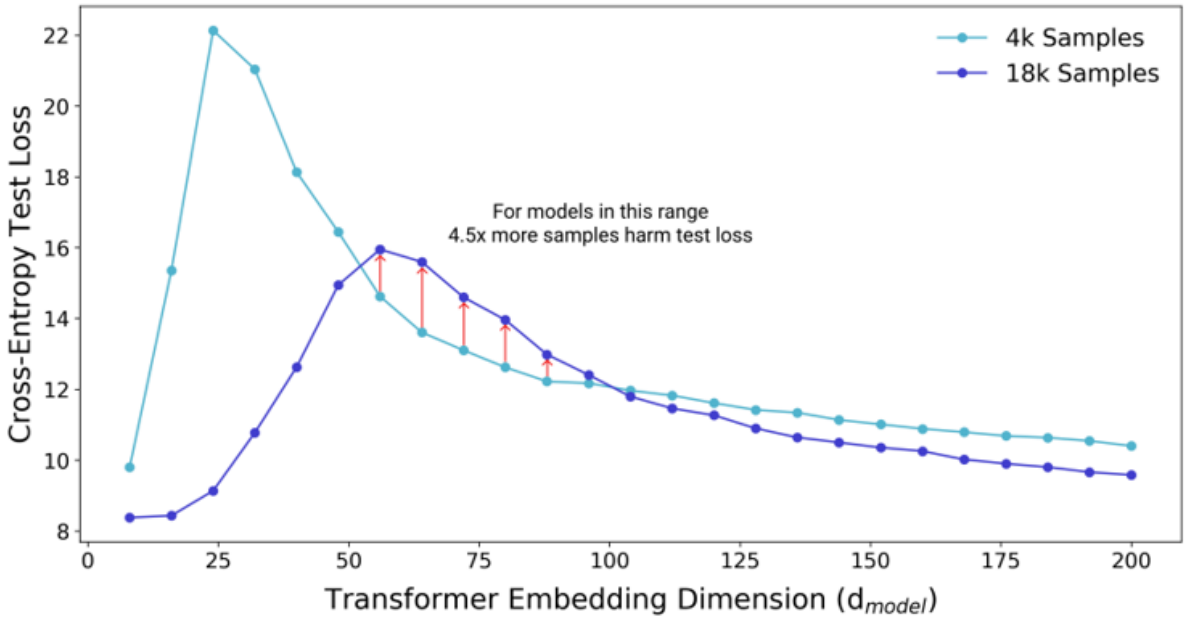
All of these graphs, however, are just showcasing the standard Belkin et al.-style double descent over model size (what Preetum et al. call “model-wise double descent”). What's really interesting about “Deep Double Descent,” however, is that Preetum et al. also demonstrate that the same thing can happen for training time (“epoch-wise double descent”) and a similar thing for dataset size (“sample-wise non-monotonicity”).

First, let's look at epoch-wise double descent. Take a look at these graphs for ResNet18 on CIFAR-10:



There's a bunch of crazy things happening here which are worth pointing out. First, the obvious: epoch-wise double descent is definitely a thing—holding model size fixed and training for longer exhibits the standard double descent behavior. Furthermore, the peak happens right at the interpolation threshold where you hit zero training error. Second, notice where you don't get epoch-wise double descent: if your model is too small to ever hit the interpolation threshold—like was the case in ye olden days of ML—you never get epoch-wise double descent. Third, notice the log scale on the y axis: you have to train for quite a while to start seeing this phenomenon.

Finally, sample-wise non-monotonicity—Preetum et al. find a regime where increasing the amount of training data by *four and a half times* actually *increases* test loss (!):



What's happening here is that more data increases the amount of model capacity/number of training epochs necessary to reach zero training error, which pushes out the interpolation threshold such that you can regress from the modern (interpolation) regime back into the classical (bias-variance tradeoff) regime, decreasing performance.

Additionally, another thing which Preetum et al. point out which I think is worth talking about here is the impact of label noise. Preetum et al. find that increasing label noise significantly exaggerates the test error peak around the interpolation threshold. Why might this be the case? Well, if we think about the inductive biases story from earlier, greater label noise means that near the interpolation threshold SGD is forced to find the one model which fits all of the noise—which is likely to be pretty bad since it has to model a bunch of noise. After the interpolation threshold, however, SGD is able to pick between many models which fit the noise and select one that does so in the simplest way such that you get good test performance.

Final comments

I'm quite excited about “Deep Double Descent,” but it still leaves what is in my opinion the most important question unanswered, which is: what exactly are the magical inductive biases of modern ML that make interpolation work so well?

One proposal I am aware of is the work of Keskar et al., who argue that SGD gets its good generalization properties from the fact that it finds “shallow” as opposed to “sharp” minima. The basic insight is that SGD tends to jump out of minima without broad basins around them and only really settle into minima with large attractors, which tend to be the exact sort of minima that generalize. Keskar et al. use the following diagram to explain this phenomena:

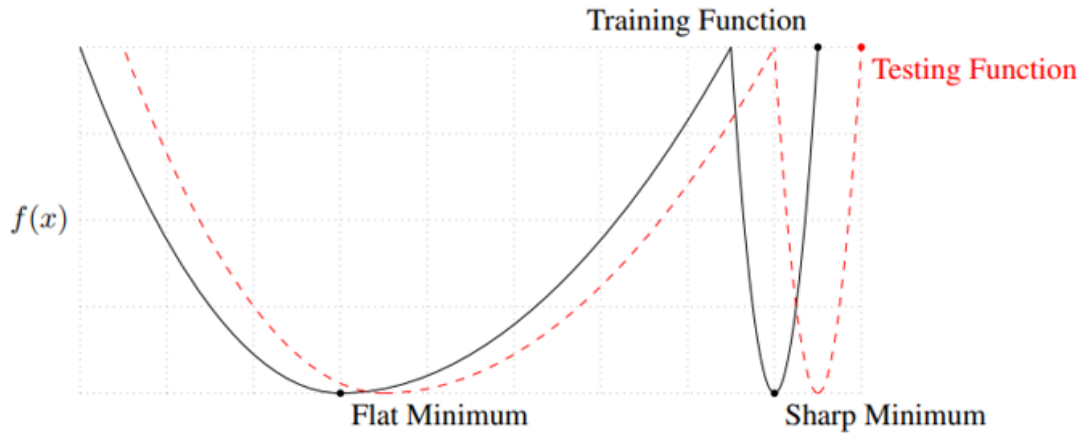


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

The more recent work of Dinh et al. in "Sharp Minima Can Generalize For Deep Nets," however, calls the whole shallow vs. sharp minima hypothesis into question, arguing that deep networks have really weird geometry that doesn't necessarily work the way Keskar et al. want it to. (EDIT: Maybe not. See this comment for an explanation of why Dinh et al. doesn't necessarily rule out the shallow vs. sharp minima hypothesis.)

Another idea that might help here is Frankle and Carbin's "Lottery Ticket Hypothesis," which postulates that large neural networks work well because they are likely to contain random subnetworks at initialization (what they call "winning tickets") which are already quite close to the final policy (at least in terms of being highly amenable to particularly effective training). My guess as to how double descent works if the Lottery Tickets Hypothesis is true is that in the interpolation regime SGD gets to just focus on the wining tickets and ignore the others—since it doesn't have to use the full model capacity—whereas on the interpolation threshold SGD is forced to make use of the full network (to get the full model capacity), not just the winning tickets, which hurts generalization.

That's just speculation on my part, however—we still don't really understand the inductive biases of our models, despite the fact that, as double descent shows, inductive biases are *the* reason that modern ML (that is, the interpolation regime) works as well as it does. Furthermore, as I noted previously, inductive biases are highly relevant to the likelihood of possible dangerous phenomenon such as mesa-optimization and pseudo-alignment. Thus, it seems quite important to me to do further work in this area and really understand our models' inductive biases, and I applaud Preetum et al. for their exciting work here.

EDIT: I have now written a follow-up to this post talking more about why I think double descent is important titled "Inductive biases stick around."

Firming Up Not-Lying Around Its Edge-Cases Is Less Broadly Useful Than One Might Initially Think

Reply to: Meta-Honesty: Firming Up Honesty Around Its Edge-Cases

Eliezer Yudkowsky, listing advantages of a "wizard's oath" ethical code of "Don't say things that are literally false", writes—

> Repeatedly asking yourself of every sentence you say aloud to another person, "Is this statement actually and literally true?", helps you build a skill for navigating out of your internal smog of not-quite-truths.

I mean, that's *one* hypothesis about the psychological effects of adopting the wizard's code.

A potential problem with this is that human natural language contains a *lot* of ambiguity. Words can be used in many ways depending on context. Even the specification "literally" in "literally false" is less useful than it initially appears when you consider that the way people *ordinarily* speak when they're being truthful is actually pretty dense with metaphors that we typically don't *notice* as metaphors because they're common enough to be recognized legitimate uses that all fluent speakers will understand.

For example, if I want to convey the meaning that our study group has covered a lot of material in today's session, and I say, "Look how far we've come today!" it would be *pretty weird* if you were to object, "*Liar!* We've been in this room the whole time and haven't physically moved at all!" because in this case, it really is obvious to all ordinary English speakers that that's not what I meant by "how far we've come."

Other times, the "intended"[1] interpretation of a statement is not only not obvious, but speakers can even mislead by motivatedly equivocating between different definitions of words: the immortal Scott Alexander has written a lot about this phenomenon under the labels "motte-and-bailey doctrine" (as coined by Nicholas Shackel) and "the noncentral fallacy".

For example, Zvi Mowshowitz has written about how the claim that "everybody knows" something[2] is often used to establish fictitious social proof, or silence those attempting to tell the thing to people who really don't know, but it feels weird (to my intuition, at least) to call it a "lie", because the speaker can just say, "Okay, you're right that not literally[3] everyone knows; I meant that *most* people know but was using a common hyperbolic turn-of-phrase and I reasonably expected you to figure that out."

So the question "Is this statement actually and literally true?" is itself potentially ambiguous. It could mean either—

- "Is this statement actually and literally true *as the audience will interpret it?*"; or,
- "Does this statement *permit an interpretation under which* it is actually and literally true?"

But while the former is complicated and hard to establish, the latter is ... not necessarily that strict of a constraint in most circumstances?

Think about it. When's the last time you needed to consciously tell a bald-faced, *unambiguous* lie?—something that could realistically be *outright proven false* in front of your peers, rather than dismissed with a "reasonable" amount of language-lawyering. (Whether "Fine" is a lie in response to "How are you?" depends on exactly what "Fine" is understood to mean in this context. "Being acceptable, adequate, passable, or satisfactory"—to what standard?)

Maybe I'm *unusually* honest—or possibly unusually bad at remembering when I've lied!?—but I'm not sure I even *remember* the last time I told an outright unambiguous lie. The kind of situation where I would need to do that just *doesn't come up that often*.

Now ask yourself how often your speech has been partially optimized for any function *other* than providing listeners with information that will help them better anticipate their experiences. The answer is, "Every time you open your mouth"[4]—and if you disagree, then you're lying. (Even if you only say true things, you're more likely to pick true things that make you look good, rather than your most embarrassing secrets. That's optimization.)

In the study of AI alignment, it's a truism that failures of alignment can't be fixed by deontological "patches". If your AI is exhibiting weird and extreme behavior (with respect to what you *really wanted*, if not what you actually programmed), then adding a penalty term to exclude *that specific behavior* will just result in the AI executing the "nearest unblocked" strategy, which will probably also be undesirable: if you prevent your happiness-maximizing AI from administering heroin to humans, it'll start administering cocaine; if you hardcode a list of banned happiness-producing drugs, it'll start researching new drugs, or just *pay* humans to take heroin, *&c.*

Humans are also intelligent agents. (Um, sort of.) If you don't genuinely have the intent to inform your audience, but consider yourself ethically bound to be honest, but your conception of *honesty* is simply "not lying", you'll naturally gravitate towards the nearest unblocked cognitive algorithm of deception.[5]

So *another* hypothesis about the psychological effects of adopting the wizard's code is that—however noble your initial conscious *intent* was—in the face of sufficiently strong incentives to deceive, you just end up accidentally training yourself to get *really good* at misleading people with a variety of not-technically-lying rhetorical tactics (motte-and-baileys, false implicatures, stonewalling, selective reporting, clever rationalized arguments, gerrymandered category boundaries, *&c.*), all the while congratulating yourself on how "honest" you are for never, ever emitting any "literally" "false" individual sentences.

Ayn Rand's novel *Atlas Shrugged*[6] portrays a world of crony capitalism in which politicians and businessmen claiming to act for the "common good" (and not consciously lying) are actually using force and fraud to temporarily enrich themselves while destroying the credit-assignment mechanisms Society needs to coordinate production.[7]

In one scene, Eddie Willers (right-hand man to our railroad executive heroine Dagny Taggart) expresses horror that the government's official scientific authority, the State

Science Institute, has issued a hit piece denouncing the new alloy, Rearden Metal, with which our protagonists have been planning to use to build a critical railroad line. (In actuality, we later find out, the Institute leaders want to spare themselves the embarrassment—and therefore potential loss of legislative funding—of the innovative new alloy having been invented by private industry rather than the Institute's own metallurgy department.)

> "The State Science Institute," he said quietly, when they were alone in her office, "has issued a statement warning people against the use of Rearden Metal." He added, "It was on the radio. It's in the afternoon papers."
>
> "What did they say?"
>
> "Dagny, they didn't say it! ... They haven't really said it, yet it's there—and it—isn't. That's what's monstrous about it."
>
> [...] He pointed to the newspaper he had left on her desk. "They haven't said that Rearden Metal is bad. They haven't said it's unsafe. What they've done is ..." His hands spread and dropped in a gesture of futility.
>
> She saw at a glance what they had done. She saw the sentences: "It may be possible that after a period of heavy usage, a sudden fissure may appear, though the length of this period cannot be predicted. ... The possibility of a molecular reaction, at present unknown, cannot be entirely discounted. ... Although the tensile strength of the metal is obviously demonstrable, certain questions in regard to its behavior under unusual stress are not to be ruled out. ... Although there is no evidence to support the contention that the use of the metal should be prohibited, a further study of its properties would be of value."
>
> "We can't fight it. It can't be answered," Eddie was saying slowly. "We can't demand a retraction. We can't show them our tests or prove anything. They've said nothing. They haven't said a thing that could be refuted and embarrass them professionally. It's the job of a coward. You'd expect it from some con-man or blackmailer. But, Dagny! It's the State Science Institute!"

I think Eddie is right to feel horrified and betrayed here. At the same time, it's notable that with respect to wizard's code, *no lying has taken place*.

I like to imagine the statement having been drafted by an idealistic young scientist in the moral maze of Dr. Floyd Ferris's office at the State Science Institute. Our scientist knows that his boss, Dr. Ferris, expects a statement that will make Rearden Metal look bad; the negative consequences to the scientist's career for failing to produce such a statement will be severe. (Dr. Ferris didn't *say* that, but he didn't have to.) But the lab results on Rearden Metal came back with flying colors—by every available test, the alloy is superior to steel along every dimension.

Pity the dilemma of our poor scientist! On the one hand, scientific integrity. On the other hand, the incentives.

He decides to follow a rule that he thinks will preserve his "inner agreement with truth which allows ready recognition": after every sentence he types into his report, he will ask himself, "Is this statement actually and literally true?" For that is his mastery.

Thus, his writing process goes like this—

"It may be possible after a period of heavy usage, a sudden fissure may appear." Is this statement actually and literally true? *Yes!* It *may be possible!*

"The possibility of a molecular reaction, at present unknown, cannot be entirely discounted." Is this statement actually and literally true? *Yes!* The *possibility* of a molecular reaction, at present unknown, *cannot be entirely discounted*. Okay, so there's not enough evidence to single out that possibility as worth paying attention to. But there's still a chance, right?

"Although the tensile strength of the metal is obviously demonstrable, certain questions in regard to its behavior under unusual stress are not to be ruled out." Is this statement actually and literally true? *Yes!* The lab tests demonstrated the metal's unprecedented tensile strength. But certain questions in regard to its behavior under unusual stress are *not to be ruled out*—the probability isn't *zero*.

And so on. You see the problem. Perhaps a member of the general public who *knew* about the corruption at the State Science Institute could read the report and infer the existence of hidden evidence: "Wow, even when trying their hardest to trash Rearden Metal, *this* is the worst they could come up with? Rearden Metal must be pretty great!"

But they won't. An institution that proclaims to be dedicated to "science" is asking for a *very high* level of trust—and in the absence of a trustworthy auditor, they might get it. Science is complicated enough and natural language is ambiguous enough, that that kind of trust that can be *betrayed* without lying.

I want to emphasize that I'm *not* saying the report-drafting scientist in the scenario I've been discussing is a "bad person." (As it is written, almost no one is evil; almost everything is broken.) Under more favorable conditions—in a world where metallurgists had the academic freedom to speak the truth as they see it (even if their voice trembles) without being threatened with ostracism and starvation—the *sort of person* who finds the wizard's oath appealing, wouldn't even be *tempted* to engage in these kinds of not-technically-lying shenanigans. But the point of the wizard's oath is to constrain you, to have a *simple* bright-line rule to *force* you to be truthful, *even when other people are making that genuinely difficult*. Yudkowsky's meta-honesty proposal is a clever attempt to strengthen the foundations of this ethic by formulating a more complicated theory that can account for the edge-cases under which even unusually honest people typically agree that lying is okay, usually due to extraordinary coercion by an adversary, as with the proverbial murderer or Gestapo officer at the door.

And yet it's *precisely* in adversarial situations that the wizard's oath is most constraining (and thus, arguably, most useful). You probably don't *need* special ethical inhibitions to tell the truth to your friends, because you should expect to *benefit* from friendly agents having more accurate beliefs.

But an enemy who wants to use information to hurt you is more constrained if the worst they can do is selectively report harmful-to-you *true* things, rather than just making things up—and therefore, by symmetry, if you want to use information to hurt an enemy, *you* are more constrained if the worst you can do is selectively report harmful-to-the-enemy true things, rather that just making things up.

Thus, while the study of how to minimize information transfer to an adversary under the constraint of not lying is certainly *interesting*, I argue that this "firming up" is of limited practical utility given the ubiquity of *other* kinds of deception. A theory of

under what conditions conscious explicit unambiguous outright lies are acceptable doesn't help very much with combating *intellectual* dishonesty—and I fear that intellectual dishonesty, plus sufficient intelligence, is enough to destroy the world all on its own, without the help of conscious explicit unambiguous outright lies.

Unfortunately, I do not, at present, have a superior alternative ethical theory of honesty to offer. I don't *know* how to unravel the web of deceit, rationalization, excuses, disinformation, bad faith, fake news, phoniness, gaslighting, and fraud that threatens to consume us all. But one thing I'm pretty sure *won't* help much is *clever logic puzzles about implausibly sophisticated Nazis*.

(Thanks to Michael Vassar for feedback on an earlier draft.)

---

1. I'm scare-quoting "intended" because this process isn't necessarily conscious, and probably usually isn't. Internal distortions of reality in imperfectly deceptive social organisms can be adaptive for the function of deceiving conspecifics. ↩

2. If I had written this post, I would have titled it "Fake Common Knowledge" (following in the tradition of "Fake Explanations", "Fake Optimization Criteria", "Fake Causality", *&c.*) ↩

3. But it's worth noting that the "Is this statement actually and literally true?" test, taken literally, should have caught this, even if my intuition still doesn't want to call it a "lie." ↩

4. Actually, that's not literally true! You often open your mouth to breathe or eat without saying anything at all! Is the referent of this footnote then a blatant lie on my part?—or can I expect you to *know what I meant*? ↩

5. A similar phenomenon may occur with other attempts at ethical bindings: for example, confidentiality promises. Suppose Open Opal tends to wear her heart on her sleeve and more specifically, believes in lies of omission: if she's talking with someone *she* trusts, and she has information *relevant* to that conversation, she finds it *incredibly psychologically painful* to *pretend not to know* that information. If Paranoid Paris has *much* stronger privacy intuitions than Opal and wants to message her about a sensitive subject, Paris might demand a promise of secrecy from Opal ("Don't share the content of this conversation")—only to spark conflict later when Opal construes the literal text of the promise more narrowly than Paris might have hoped ("'Don't share the content' means don't share the *verbatim text*, right? I'm still allowed to paraphrase things Paris said and attribute them to an anonymous correspondent when I think that's relevant to whatever conversation I'm in, even though that hypothetically leaks entropy if Paris has implausibly determined enemies, right?"). ↩

6. I know, fictional evidence, but I claim that the *kind of deception* illustrated in quoted passage to follow is *entirely* realistic. ↩

7. Okay, that's probably not *exactly* how Rand or her acolytes would put it, but that's how I'm interpreting it. ↩

[Part 2] Amplifying generalist research via forecasting – results from a preliminary exploration

This post covers the set-up and results from our exploration in amplifying generalist research using predictions, in detail. It is accompanied by a second post with a high-level description of the results, and more detailed models of impact and challenges. For an introduction to the project, see that post.

—

The rest of this post is structured as follows.

First, we cover the basic set-up of the exploration.

Second, we share some results, in particular focusing on the accuracy and cost-effectiveness of this method of doing research.

Third, we briefly go through some perspectives on what we were trying to accomplish and why that might be impactful, as well as challenges with this approach. These are covered more in-depth in a separate post.

Overall, we *are very interested in feedback and comments on where to take this next.*

Set-up of the experiment

A note on the experimental design

To begin with, we note that this was not an “experiment” in the sense of designing a rigorous methodology with explicit controls to test a particular, well-defined hypothesis.

Rather, this might be seen as an ”exploration” [3]. We tested several different ideas at once, instead of running a unique experiment for each separately. We also intended to uncover new ideas and inspiration as much as testing existing ones.

Moreover, we proceeded in a startup-like fashion where several decisions were made ad-hoc. For example, a comparison group was introduced after the first experiment had been completed; this was not originally planned, but later became evidently useful. This came at the cost of worsening the rigor of the experiment.

We think this trade-off was worth it for our situation. This kind of policy allows us to execute a large number of experiments in a shorter amount of time, quickly pivot away from bad ones, and notice low-hanging mistakes and learning points before scaling up good ones. This is especially helpful as we’re shooting for tail-end outcomes, and are looking for concrete mechanisms to implement in practice (rather than publishing particular results).

We do not see it as a substitute for more rigorous studies, but rather as a complement, which might serve as inspiration for such studies in the future.

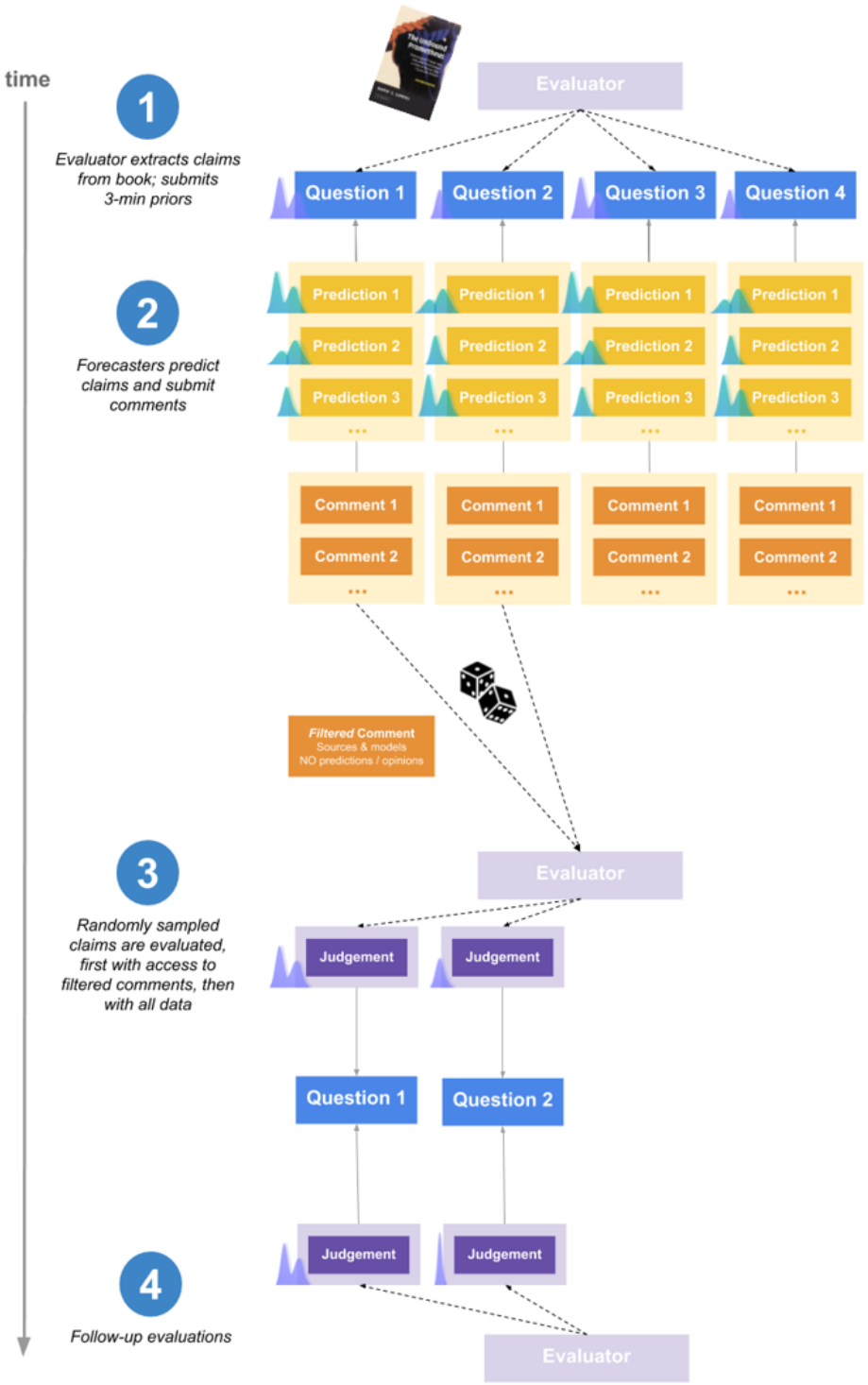
To prevent this from biasing the data, all results from the experiment are public, and we try to note when decisions were made post-hoc.

Mechanism design

The basic set-up of the project is shown in the following diagram, and described below.

A two-sentence version would be:

> Forecasters predicted the conclusions that would be reached by Elizabeth van Norstrand, a generalist researcher, before she conducted a study on the accuracy of various historical claims. We randomly sampled a subset of research claims for her to evaluate, and since we can set that probability arbitrarily low this method is not bottlenecked by her time.


time
1
Evaluator extracts claims from book; submits 3-min priors
Evaluator
Question 1
Question 2
Question 3
Question 4
2
Forecasters predict claims and submit comments
Prediction 1
Prediction 2
Prediction 3
Comment 1
Comment 2
Filtered Comment
Sources & models
NO predictions / opinions
3
Randomly sampled claims are evaluated, first with access to filtered comments, then with all data
Evaluator
Judgement
Judgement
Question 1
Question 2
Judgement
Judgement
4
Follow-up evaluations
Evaluator

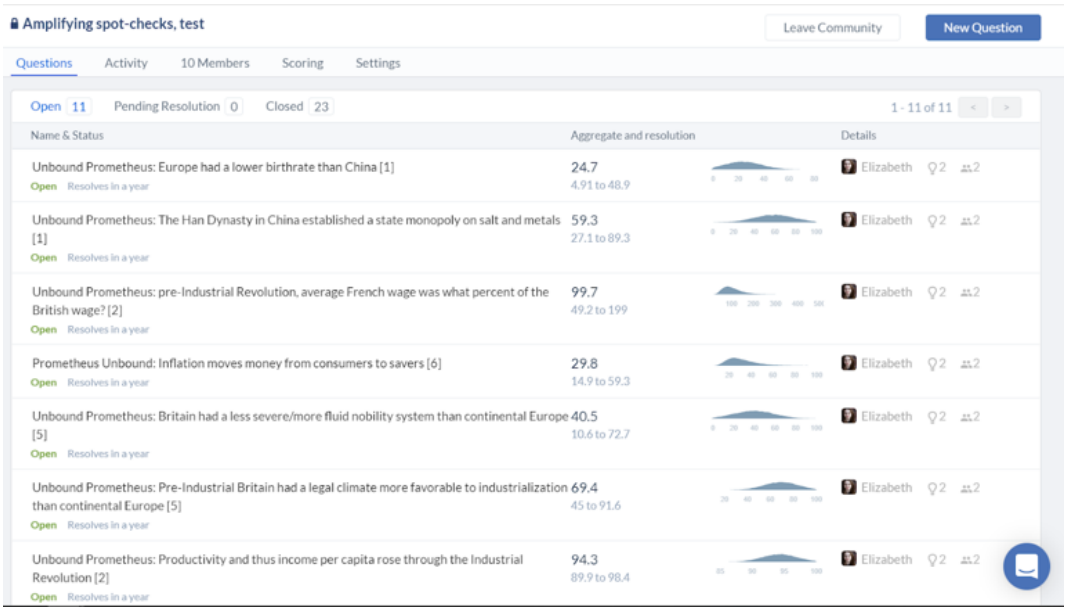
1. Evaluator extracts claims from the book and submits priors

The evaluator for the experiment was Elizabeth Van Norstrand, an independent generalist researcher known for her "Epistemic spot checks". This is a series of posts assessing the trustworthiness of a book by evaluating some of it claims. We chose Elizabeth for the experiment as she has a reputation for reliable generalist research, and there was a significant amount of public data about her past evaluations of claims.

She picked 10 claims from the book *The Unbound Prometheus: Technological Change and Industrial Development in Western Europe from 1750 to the Present*, as well as a meta-claim about the reliability of the book as a whole.

All claims were assigned an importance rating from 1-10 based on their relevance to the thesis of the book as a whole. We were interested in finding if this would influence forecaster effort between questions.

Elizabeth also spent 3 minutes per claim submitting an initial estimate (referred to as a "prior").



Beliefs were typically encoded as distributions over the range 0% to 100%, representing where Elizabeth expected the mean of her posterior credence in the claim to be after 10 more hours of research. For more explanation, see this footnote [4].

2. Forecasters make predictions

Forecasters predicted what they expected Elizabeth to say after ~45 minutes of research on the claim, and wrote comments explaining their reasoning.

Forecasters' payments for the experiment were proportional to how much their forecasts outperformed the aggregate in estimating her 45-minute distributions. In addition, forecasters were paid a base sum just for participating. You can see all forecasts and comments here, and an interactive tool for visualising and understanding the scoring scheme here.

A key part of the design was that that forecasters *did not know* which question Elizabeth would randomly sample to evaluate. Hence they were incentivised to do their best on *all* questions (weighted by importance). This has the important implication that we could easily extend the amount of questions predicted by forecasters -- even if Elizabeth can only judge 10 claims, we could have forecasting questions for 100 different claims [5].

Two groups of forecasters participated in the experiment: one based on a mailing list with participants interested in participating in forecasting experiments (recruited from effective altruism-adjacent events and other forecasting platforms) [6], and one recruited from Positly, an online platform for crowdworkers. The former group is here called “Network-adjacent forecasters” and the latter “Online crowdworkers”.

	Network-adjacent forecasters	**Online crowdworkers**
Number of participants	19	13
Total number of predictions	121	85
Total time spent	~50h	~15h
Total cost	~$900 [7]	$335 [8]

3. The evaluator judges the claims

Elizabeth was given a time-budget of 6 hours, within which she randomly sampled claims to research and judge.

At this point, we wanted to use the work done by forecasters to help Elizabeth, while avoiding anchoring and biasing her with their estimates.

To solve this, Elizabeth was initially given a filtered version of the comments section for each claim, which contained all sources and models used, but which had been stripped of any explicit predictions or subjective opinion generalising from the data.

For example, for the question:

> *Pre-Industrial Britain had a legal climate more favorable to industrialization than continental Europe [5].*

One commenter wrote:

> Seems more likely to be true than not. The English Civil War and Glorious Revolution both significantly curtailed the arbitrary power of the monarch/gentry and raised the power of merchants in Britain, making likely that government was more favourable to mercantile interests. Hard to judge the claim about haggling.

And in Elizabeth’s initial briefing this was replaced by:

> The English Civil War and Glorious Revolution both significantly curtailed the arbitrary power of the monarch/gentry and raised the power of merchants in

Britain [...].

After arriving at a final estimate, Elizabeth was allowed to look at the full forecaster comments and predictions and optionally change her mind. In practice, she didn't change her mind in any of these cases.

To summarise, the parts involved were:

- We summarised all of the relevant comments into a list and removed any descriptions that referred to people's probabilities.
- We randomly chose 8 of the 10 claims for review by Elizabeth.
- Elizabeth saw this list, and spent 6 hours evaluating the 8 claims and resolving them with probability distributions of the values she expected to have for them in a future possible evaluation round. She divided this time in accordance to what seemed the most useful; for instance, questions with answers that became obvious quickly got a relatively small proportion of this time.
- Elizabeth got access to all predictions and all comments and was was allowed to change her resolutions. She decided not to in all of the cases.

4. The evaluator double-checks the evaluations

After having spent 6 hours researching 8 claims, Elizabeth randomly sampled two of those, each of which she spent an additional 3 hours researching. For the remaining claims, she waited until a week after the experiment, and then reread her notes and submitted new resolutions, to see if her way of converting beliefs into numbers was consistent over time. This part was intended to test the consistency and reliability of Elizabeth's evaluations.

The outcome of this was that Elizabeth appeared highly consistent and reliable. You can see the data and graphs here. Elizabeth's full notes explaining her reasoning in the evaluations can be found here.

Results and analysis

You can find all the data and interactive tools for exploring it yourself, here.

Online crowdworkers

We were interested in comparing the performance of our pool of forecasters to "generic" participants with no prior interest or experience forecasting.

Hence, after the conclusion of the original experiment, we reran a slightly modified form of the experiment with a group of forecasters recruited through an online platform that sources high quality crowdworkers (who perform microtasks like filling out surveys or labeling images for machine learning models).

However, it should be mentioned that these forecasters were operating under a number of disadvantages relative to other participants, which means we should be careful when interpreting their performance. In particular:

- They did not know that Elizabeth was the researcher who created the claims and would resolve them, and so they had less information to model the person

whose judgments would ultimately decide the questions.
- They did not use any multimodal or custom distributions, which is a way to increase tail-uncertainty and avoid large losses when forecasting with distributions. We expect this was because of the time-constraints set by their payment, as well as the general difficulty.

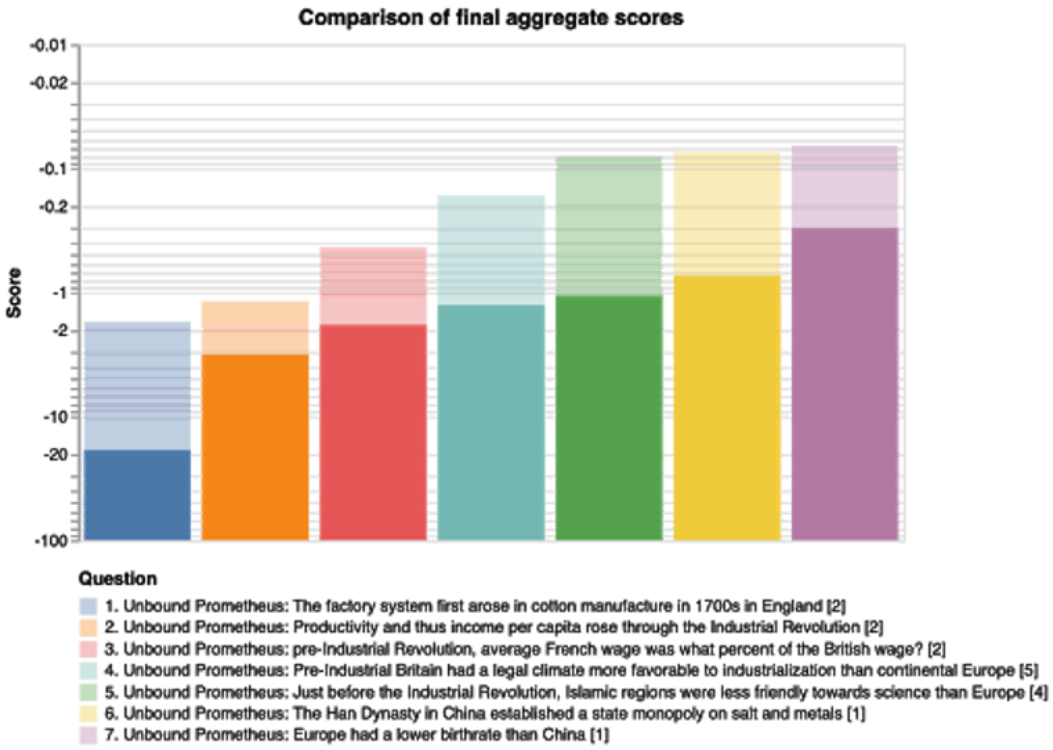
Overall the experiment with these online crowdworkers produced poor accuracy results at predicting Elizabeth's resolutions (as is discussed further below).

Accuracy of predictions

This section analyses how well forecasters performed, collectively, in amplifying Elizabeth's research.

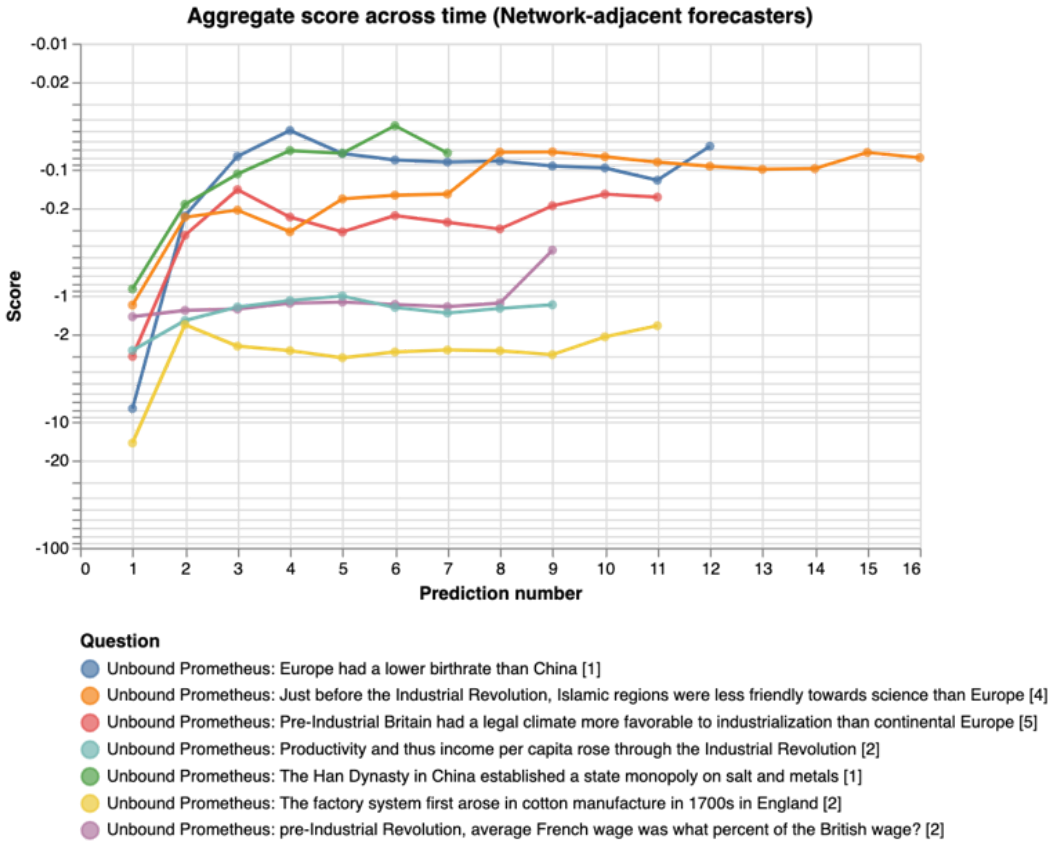
The aggregate prediction was computed as the average of all forecasters' final predictions. Accuracy was measured using a version of the logarithmic scoring rule.

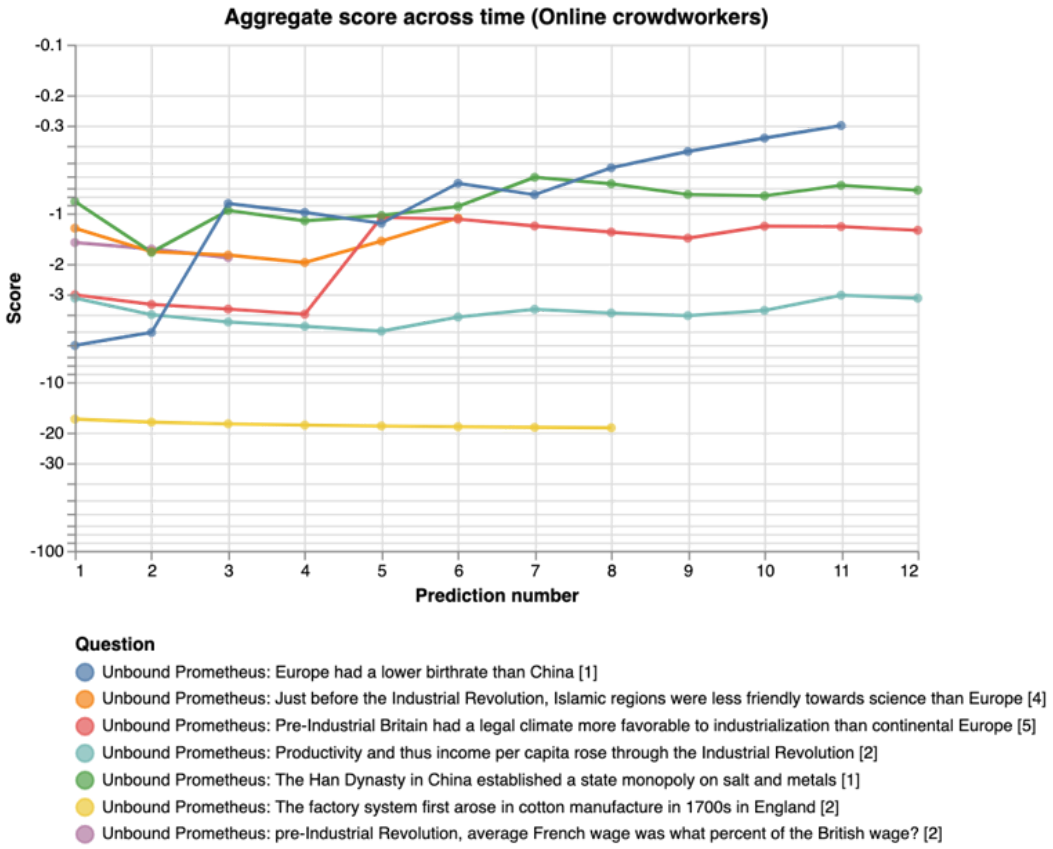
The following graph shows how the aggregate performed on each question:

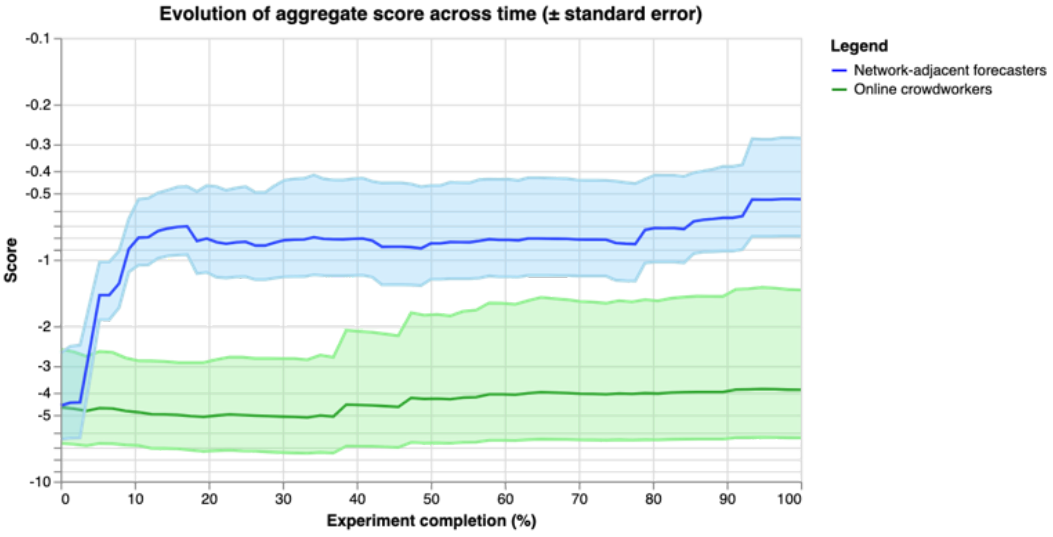


The opaque bars represent the scores from the crowdworkers, and the translucent bars, which have higher scores throughout, represent the scores from the network-adjacent forecasters. It's interesting that the order is preserved, that is, that the question difficulty was the same for both groups. Finally we don't see any correlation between question difficulty and the importance weights Elizabeth assigned to the questions.

However, the comparison is confounded by the fact that more effort was spent from the network-adjacent forecasters. The above graph also doesn't compare performance to Elizabeth's priors. Hence we also plot the evolution of the aggregate score over prediction number and time (the first data-point in the below graphs represent Elizabeth's priors):







For the last graph, the y-axis shows the score on a logarithmic scale, and the x-axis shows how far along the experiment is. For example, 14 out of 28 days would correspond to 50%. The thick lines show the average score of the aggregate prediction, across all questions, at each time-point. The shaded areas show the standard error of the scores, so that the graph might be interpreted as a guess of how the two communities would predict a random new question [10].

One of our key takeaways from the experiment is that simple average aggregation algorithm performed surprisingly well, but only for the network-adjacent forecasters.

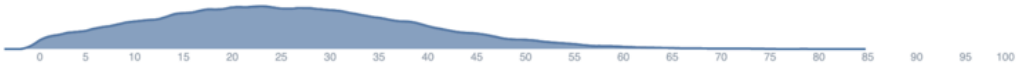
One way to see this qualitatively is by observing the graphs below, where we display Elizabeth's priors, the final aggregate of the network-adjacent forecasters, and the final resolution, for a subset of questions [11].

Question examples

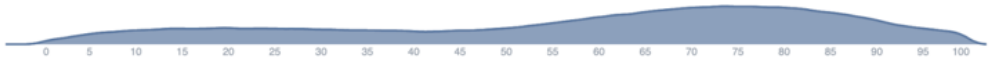
The x-axis [12] refers to the Elizabeth's best estimate of the accuracy of a claim, from 0% to 100% (see section "Mechanism design, 1. Evaluator extracts claims" for more detail).

Europe had a lower birth rate than China [1]

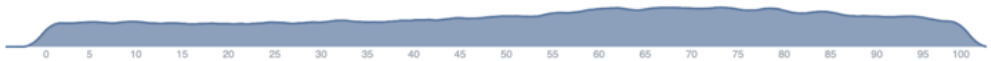
Prior



Aggregate (log score -0.07)

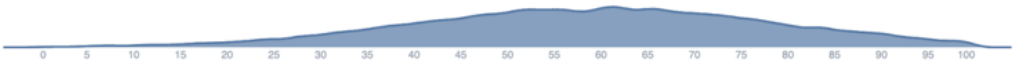


Resolution

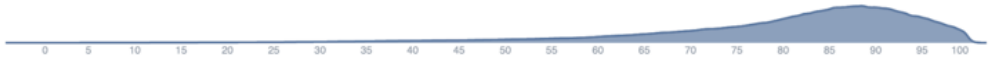


The Han Dynasty in China established a state monopoly on salt and metals [1]

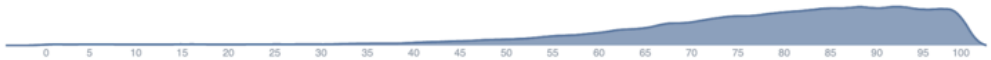
Prior



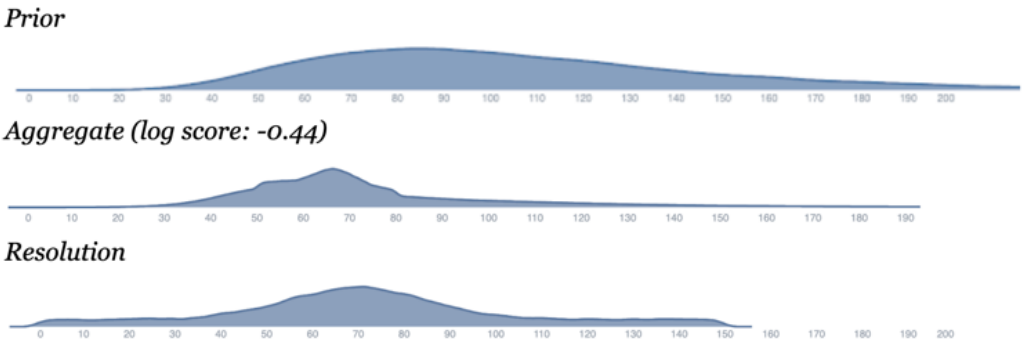
Aggregate (log score -0.07)



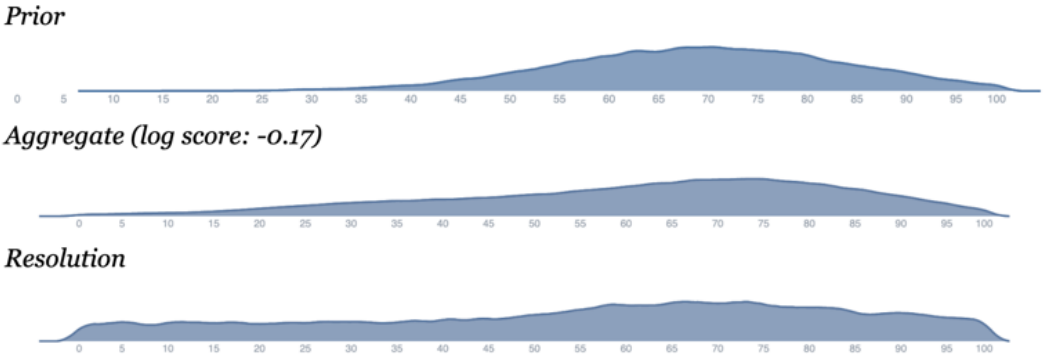
Resolution



Pre-Industrial Revolution, average French wage was what percent of the British wage? [2]



Pre-Industrial Britain had a legal climate more favorable to industrialization than continental Europe [5]



Another way to understand the performance of the aggregate is to note that the aggregate of network-adjacent forecasters had an average log score of -0.5. To get a rough sense of what that means, it's the score you'd get by being 70% confident in a binary event, and being correct (though note that this binary comparison merely serves to provide intuition, there are technical details making the comparison to a distributional setting a bit tricky).

By comparison, the crowdworkers and Elizabeth's priors had a very poor log score of around -4. This is roughly similar to the score you'd get if you predict an event to be ~5% likely, and it still happens.

Cost-effectiveness

High-level observations

This experiment was run to get a sense of whether forecasters could do a competent job forecasting the work of Elizabeth (i.e. as an "existence proof"). It was not meant to show cost-effectiveness, which could involve many efficiency optimizations not yet undertaken. However, we realized that the network-adjacent forecasting may have

been reasonably cost-effective and think that a cost-effectiveness analysis of this work could provide a baseline for future investigations.

To compute the cost-effectiveness of doing research using amplification, we look at two measures: the information gain from predictors relative to the evaluator, and the cost of predictors relative to the evaluator.

Benefit/cost ratio = % information gain provided by forecasters relative to the evaluator / % cost of forecasters relative to the evaluator

If a benefit/cost ratio of significantly over 1 can be achieved, then this could mean that forecasting could be useful to partially augment or replace established evaluators.

Under these circumstances, each unit of resources invested in gaining information from forecasters has higher returns than just asking the evaluator directly.

Some observations about this.

First, note that this does *not* require forecasters to be as accurate as the evaluator. For example, if they only provide 10% as much value, but at 1% of the opportunity cost, this is still a good return on investment.

Second, amplification can still be worthwhile even if the benefit-cost ratio is < 1. In particular:

1. Forecasters can work in parallel and hence answer a much larger number of questions, within a set time-frame, than would be feasible for some evaluators.
2. Pre-work by forecasters might also improve the speed and quality of the evaluator's work, if she has access to their research [13].
3. Having a low benefit-cost ratio can still serve as an existence proof that amplification of generalist research is possible, as long as the benefit is high. One might then run further optimised tests which try harder to reduce cost.

Results

The opportunity cost is computed using Guesstimate models linked below, based on survey data from participants collected after the experiment. We are attempting to include both hourly value of time and value of altruistic externalities. We did not include the time that our own team spent figuring out and organising this work.

For example, the estimated cost ratio for the network-adjacent forecasters in this experiment was 120%, meaning that the cost of obtaining a final aggregate prediction for a question was 20% higher when asking this group of 19 forecasters than when asking Elizabeth directly, all things considered.

The value is computed using the following model (interactive calculation linked below). We assume Elizabeth is an unbiased evaluator, and so the true value of a question is the mean of her resolution distribution. We then treat this point estimate as the *true* resolution, and compare to it the scores of Elizabeth's resolution, had it been a prediction, vs. her initial prior; and the final aggregate vs. her initial prior. All scores are weighed by the importance of the question, as assigned by Elizabeth on a 1-10 scale [14].

Results were as follows.

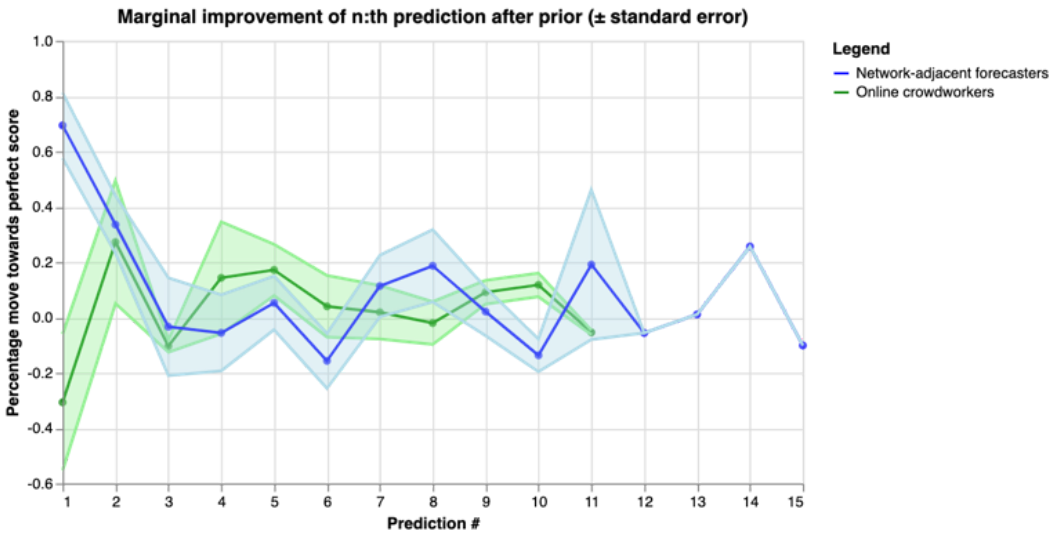
	Network-adjacent forecasters	Online crowdworkers
Cost ratio	120%	52%
Value ratio	87%	-32%
Benefit/cost ratio	72%	-62%

(Links to models: network-adjacent cost ratio and value ratio, online crowdworker cost ratio and value ratio.)

The negative value ratio for the control group indicates that they assigned a lower probability to the mean of Elizabeth's resolution than she her herself did when submitting her prior. Hence just accepting the means from those forecasts would have made us worse off, epistemically, than trusting the priors.

This observation is in tension with the some of the above graphs, which show a tiny increase in average log score between crowdworkers and Elizabeth's priors. We are somewhat uncertain about the reason for this, though we think it is as follows: they were worse at capturing the resolution means than the prior, but they were sometimes better at capturing the resolution distribution (likely by the average of them adding on more uncertainty). And the value ratio only measures the former of those improvements.

Another question to consider when thinking about cost-effectiveness is diminishing returns. The following graph shows how the information gain from additional predictions diminished over time.



The x-axis shows the number of predictions after Elizabeth's prior (which would be prediction number 0). The y-axis shows how much closer to a perfect score each prediction moved the aggregate, as a percentage of the distance between the previous aggregate and the perfect log score of 0 [15].

We observe that for the network-adjacent forecasters, the majority of value came from the first two predictions, while the online crowdworkers never reliably reduced uncertainty. Several hypotheses might explain this, including that:

- The first predictor on most questions was also one of the best participants in the experiment
- Most of the value of the predictors came from increasing uncertainty, and already after averaging 2-3 distributions we had gotten most of the effect there
- Later participants were anchored by the clearly visible current aggregate and prior predictions

Future experiments might attempt to test these hypotheses.

Perspectives on impact and challenges

This section summarises some different perspectives on what the current experiment is trying to accomplish and why that might be exciting, as well as some of the challenges it faces. To keep things manageable, we simply give a high-level overview here and discuss each point in more detail in a separate post.

There are several perspectives here given that the experiment was designed to explore multiple relevant ideas, rather than testing a particular, narrow hypothesis.

As a result, the current design is not optimising very strongly for any of these possible uses, and it is also plausible that its impact and effectiveness will vary widely between uses.

Perspectives on impact

- **Mitigating capacity bottlenecks.** The effective altruism and rationality communities face rather large bottlenecks in many areas, such as allocating funding, delegating research, vetting talent and reviewing content. The current setup might provide a means of mitigating some of those -- a scalable mechanism of outsourcing intellectual labor.
- **A way for intellectual talent to build and demonstrate their skills.** Even if this set-up can't make new intellectual progress, it might be useful to have a venue where junior researchers can demonstrate their ability to predict the conclusions of senior researchers. This might provide an objective signal of epistemic abilities not dependent on detailed social knowledge.
- **Exploring new institutions for collaborative intellectual progress.** Academia has a vast backlog of promising ideas for institutions to help us think better in groups. Currently we seem bottlenecked by practical implementation and product development.
- **Getting more data on empirical claims made by the Iterated Amplification AI alignment agenda.** These ideas inspired the experiment. (However, our aim was more practical and short-term, rather than looking for theoretical insights useful in the long-term.)
- **Exploring forecasting with distributions.** Little is known about humans doing forecasting with full distributions rather than point estimates (e.g. "79%"),

partly because there hasn't been easy tooling for such experiments. This experiment gave us some cheap data on this question.
- **Forecasting fuzzy things.** A major challenge with forecasting tournaments is the need to concretely specify questions; in order to clearly determine who was right and allocate payouts. The current experiments tries to get the best of both worlds -- the incentive properties of forecasting tournaments and the flexibility of generalist research in tackling more nebulous questions.
- **Shooting for unknown unknowns.** In addition to being an "experiment", this project is also an "exploration". We have an intuition that there are interesting things to be discovered at the intersection of forecasting, mechanism design, and generalist research. But we don't yet know what they are.

Challenges and future experiments

- **Complexity and unfamiliarity of experiment.** The current experiment had many technical moving parts. This makes it challenging to understand for both participants and potential clients who want to use it in their own organisations.
- **Trust in evaluations.** The extent to which these results are meaningful depends on your trust in Elizabeth Van Nostrand's ability to evaluate questions. We think is partly an inescapable problem, but also expect clever mechanisms and more transparency to be able to make large improvements.
- **Correlations between predictions and evaluations.** Elizabeth had access to a filtered version of forecaster comments when she made her evaluations. This introduces a potential source of bias and a "self-fulfilling prophecy" dynamic in the experiments.
- **Difficulty of converting mental models into quantitative distributions.** It's hard to turn nuanced mental models into numbers. We think a solution is to have a "division of labor", where some people just build models/write comments and others focus on quantifying them. We're working on incentive schemes that work in this context.
- **Anti-correlation between importance and "outsourceability".** The intellectual questions which are most important to answer might be different from the ones that are easiest to outsource, in a way which leaves very little value on the table in outsourcing.
- **Overhead of question generation.** Creating good forecasting questions is hard and time-consuming, and better tooling is needed to support this.
- **Overly competitive scoring rules.** Prediction markets and tournaments tend to be zero-sum games, with negative incentives for helping other participants or sharing best practices. To solve this we're designing and testing improved scoring rules which directly incentivise collaboration.

Footnotes

[1] Examples include: AI alignment, global coordination, macrostrategy and cause prioritisation.

[2] We chose the industrial revolution as a theme since it seems like a historical period with many lessons for improving the world. It was a time of radical change in productivity along with many societal transformations, and might hold lessons for future transformations and our ability to influence those.

[3] Some readers might also prefer the terms “integration experiment” and “sandbox experiment”.

[4] In traditional forecasting tournaments, participants state their beliefs in a binary event (e.g. “Will team X win this basketball tournament?”) using a number between 0% and 100%. This is referred to as a credence, and it captures their uncertainty in a quantitative way. The terminology comes from Bayesian probability theory, where rational agents are modelled as assigning credences to claims and then updating those credences on new information, in a way uniquely determined by Bayes’ rule. However, as a human, we might not always be sure what the right credence for a claim is. If I had an unlimited time to think, I might arrive at the right number. (This is captured by the “after 10 more hours of research” claim.) But if I don’t have a lot of time, I have some uncertainty about exactly how uncertain I should be. This is reflected in our use of distributions.

[5] In scaling the number of claims beyond what Elizabeth can evaluate, we would also have to proportionally increase the rewards.

[6] Many of these participants had previous experience with forecasting, and some were “superforecaster-equivalents” in terms of their skill. Others had less experience with forecasting but were competent in quantitative reasoning. For future experiments, we ought to survey participants about their previous experience.

[7] The payments were doubled after we had seen the results, as the initial scoring scheme proved too harsh on forecasters.

[8] The incentive schemes looked somewhat different between groups, mostly owing to the fact that we tried to reduce the complexity necessary to understand the experiment for the online crowdworkers, who to our knowledge had no prior experience with forecasting. They were each paid at a rate of ~$15 an hour, with the opportunity for the top three forecasters to receive a bonus of $35.

[9] Elizabeth did this by copying the claims into a google doc, numbering them, and then using Google random number generator to pick claims. For a future scaled up version of the experiment, one could use the public randomness beacon as a transparent and reproducible way to sample claims.

[10] In analysing the data we also plotted 95% confidence intervals by multiplying the standard error by 1.96. In that graph the two lines intersect for something like 80%-90% of the x-axis. You can plot and analyse them yourself here.

[11] We only display the first four resolutions to not too make up too much space (which were randomly chosen in the course of the experiment). All resolution graphs can be found here.

[12] The distributions are calculated using Monte Carlo sampling and Kernel smoothing, so are not perfectly smooth. This also led to errors around bounds being outside of the 0 to 100 range.

[13] For this experiment, Elizabeth informally reports that the time saved ranged from 0-60 minutes per question, but she did not keep the kind of notes required to estimate an average.

[14] This is a rough model of calculating this and we can imagine there being better ways of doing it. Suggestions are welcome.

[15] Using this transformation allows us to visualise the fact smaller scores obtained later in the contest can still be as impressive as earlier scores. For example, moving from 90% confidence to 99% confidence takes roughly as much evidence as moving from 50% to 90% confidence. Phrased in terms of odds ratios, both updates involve evidence of strength roughly 10:1.

Participate in future experiments or run your own

Foretold.io was built as an open platform to enable more experimentation with prediction-related ideas. We have also made data and analysis calculations from this experiment publicly available.

If you'd like to:

- Run your own experiments on other questions
- Do additional analysis on this experimental data
- Use an amplification set-up within your organisation

We'd be happy to consider providing advice, operational support, and funding for forecasters. Just comment here or reach out to this email.

If you'd like to participate as a forecaster in future prediction experiments, you can sign-up here.

Acknowledgements

Funding for this project was provided by the Berkeley Existential Risk Initiative and the EA Long-term Future Fund.

We thank Beth Barnes and Owain Evans for helpful discussion.

We are also very thankful to all the participants.

Bayesian examination

A few months ago, Olivier Bailleux, a Professor of computer science and reader of my book on Bayesianism, sent me an email. He suggested to apply some of the ideas of the book to examine students. He proposed **Bayesian examination**.

I believe it to be a brilliant idea, which could have an important impact on how many people think. At least, I think that this is surely worth sharing here.

tl;dr Bayesian examinations seem very important to deploy because they *incentivize* both *probabilistic thinking* and *intellectual honesty.* Yet, as argued by Julia Galef in this talk, incentives seem critical to change our thinking habits.

Let's take an example

Where is the International Olympic Committee?
1. Geneva
2. Lausanne
3. Zurich
4. Lugano

Quite often, students are asked to select one of the four possible answers. But this is arguably pretty bad, for several reasons:
- It makes impossible to distinguish a student who has a hunch from a student who really studied and knew the answer.
- It gives students the habit of self-identifying with a single answer.
- It normalizes deterministic question answering.
- It motivates students to defend the answer they gave (which encourages the motivated reasoning fallacy...).

Instead, Bayesian examination demands that students provide *probabilistic answers.* In other words they will have to provide percentage for each answer.

In our case, a student, call her Alice, might thus answer
1. 33%
2. 33%
3. 33%
4. 1%

Alice would essentially be formalizing the sentence "I really don't know but I would be very surprised if Lugano was the right answer".

Another student, let's call him Bob, might answer
1. 5%
2. 40%
3. 50%
4. 5%

Bob might be having in mind something like "I know that FIFA and CIO are in Zurich and Lausanne, but I don't remember which is where; though Zurich is larger so it would make sense for CIO to be in Zurich rather than Lausanne".

Spoiler: the answer turns out to be Lausanne.

Why naive notation is bad

Now, how would such an exam be scored? One intuitive idea could be that Alice should thus get 0.33 points, while Bob should get 0.4 points. Denoting q_i the probability assigned by a student to answer i, and i* the right answer, this would correspond to giving the student a score equals to $S = q_{i*}$.

This would *not* be a great idea though. The reason for this has to do with **incentives**. Indeed, it turns out that if the above figures are the credences p_i of Alice and Bob, then Alice and Bob would be incentivized to maximize their expected scores $E[S] = p_1q_1 + p_2q_2 + p_3q_3 + p_4q_4$. It turns out that this maximization leads to the following answers.

For Alice:

1. Credence $p_1 = 33\%$ in Geneva, but answers $q_1 = 100\%$.
2. Credence $p_2 = 33\%$ in Lausanne, but answers $q_2 = 0\%$.
3. Credence $p_3 = 33\%$ in Zurich, but answers $q_3 = 0\%$.
4. Credence $p_4 = 33\%$ in Lugano, but answers $q_4 = 0\%$.

For Bob:

1. Credence $p_1 = 5\%$ in Geneva, but answers $q_1 = 0\%$.
2. Credence $p_2 = 40\%$ in Lausanne, but answers $q_2 = 0\%$.
3. Credence $p_3 = 50\%$ in Zurich, but answers $q_3 = 100\%$.
4. Credence $p_4 = 5\%$ in Lugano, but answers $q_4 = 0\%$.

In other words, this naive scoring incentivizes the exaggeration of beliefs towards deterministic answer. This is very, very, very, very, very bad (sorry I'm a bit of Bayesian extremist!). This favors polarization, rationalization, groupism and so many other root causes of poor debating.

Indeed, while students may not find out consciously that this exaggeration strategy is optimal, we should expect them to eventually try it and not unconsciously notice that this is not so bad. In particular, this prevents them from valuing the extra-effort of *probabilistic thinking*.

Fortunately, there are better scoring rules.

Incentive-compatible scoring rules

An incentive-compatible scoring rule is called a *proper scoring rule*. But I'm not keen on the terminology, as it's not transparent, so I'll stick with **incentive-compatible scoring rule**. Such incentive-compatible scoring rules are such that truth-telling (or rather "credence-telling") is incentivized.

There are several incentive-compatible scoring rules, like the *logarithmic scoring rule* ($S = \ln(q_{i*})$) or the *spherical scoring rule* ($S = q_{i*}/||q||_2$). But I think that the most appropriate one may be the *quadratic scoring rule*, because it is the simplest and easiest for students to verify.

In our case, given that the right answer was Lausanne, the score of a student who answered q_1, q_2, q_3 and q_4 is $S = 1 - (q_1^2 + (1 - q_2)^2 + q_3^2 + q_4^2)$. In other words, for each possibility i, the student loses the square of the distance between his answer q_i and the true answer (0% or 100%).

In our case, Alice would win $S = 0.3332$ points, while Bob would win $S = 0.585$ points.

Of course, the right answer $q = (0, 1, 0, 0)$ would win 1 point, while any maximally wrong answer like $q = (1, 0, 0, 0)$ would *lose* 1 point.

Perhaps more interestingly, a maximally ignorant student who answers 25% to each possibility would win $S = 0.25$ points. This is much better than the expected answer of random deterministic guess, which equals $E[S] = -0.5$. Exaggerated guesses get greatly penalized. In fact, they yield *negative points*!

Formally, the quadratic scoring rule equals $S = 1 - ||q - e_{i*}||_2^2 = 2q_{i*} - ||q||_2^2$, where e_{i*} is the basis vector whose entries are all zeros except for the i*-th coordinate, which is 1. If there are n answers, then the maximally ignorant student wins $S = 1/n$, while the random deterministic guesser wins an expectation of $E[S] = -1 + (2/n)$.

Note also that $E[S(q)|p] = ||p||_2^2 - ||q-p||_2^2$, where p is the credence and q is the answer. This is clearly minimal for $q=p$. In fact, interestingly, it is minimal even if we allow $q \in \mathbb R^n$ (i.e. even if we don't tell students that their probabilistic answers need to add up to 1, then they will eventually learn that this is the way to go). In particular, the honest answer yields an expected score of $E[S] = ||p||_2^2$, which indeed reflects the uncertainty of the student.

Why this is important

Because wrong answers are much more penalized than acknowledging ignorance, students who aim to maximize their scores will likely eventually learn, consciously or not, that *guessing deterministic answers is just wrong*. They may even learn the habit of second-guessing their intuitions, and to add uncertainty to their first guesses. In terms of rationality, this seems like a huge deal!

Perhaps equally importantly, such Bayesian examinations incentivize students to take on *probabilistic reasoning.* Students may thereby learn to constantly measure appropriately their levels of confidence, and to reason with (epistemological) uncertainty. As an aspiring Bayesian, this is the part I'm most excited about!

Finally, and probably even more importantly, such examinations incentivize *intellectual honesty*. This is the habit of trying to be honest, not only with others, but also with ourselves. It's sometimes said that "a bet is a tax on bullshit", as argued by Alex Tabarrok. Arguably, Bayesian examinations are even better than a bet. Indeed, in (important) exams, we might be making an even bigger effort than when we put our money where our mouth!

In case you're still not convinced by the importance of intellectual honesty, I highly recommend this talk by Julia Galef or her upcoming book (as well as, say, Tetlock and Gardner's Superforecasting book).

Where to go from here

I haven't had the chance to test these ideas though. I wonder how students and teachers will feel about it. I suspect some pushback early on. But I would also bet that students may eventually appreciate it. To find out, I guess this really needs to be tested out there!

One particular platform that could be a great first step is in MOOCs and other online websites where people enter their answers electronically. If you happen to be working in such areas, or to know people working in these areas, I think it would be great to encourage a trial of Bayesian examinations! And if you do, please send me feedbacks. And please let me test your exams as well :P

Still another approach would be to develop an app to record Bayesian bets that we make, and to compute our incentive-compatible (quadratic?) scores. Gamifying the app might make it more popular. If anyone is keen on developing such an app, I'd be more than eager to test it, and to train my own Bayesian forecasting abilities!

PS : If you're French-speaking (or motivated to read subtitles), you can also check out the video I made on the same topic.

We run the Center for Applied Rationality, AMA

CFAR recently launched its 2019 fundraiser, and to coincide with that, we wanted to give folks a chance to ask us about our mission, plans, and strategy. Ask any questions you like; we'll respond to as many as we can from 10am PST on 12/20 until 10am PST the following day (12/21).

Topics that may be interesting include (but are not limited to):

- Why we think there should be a CFAR;
- Whether we should change our name to be less general;
- How running mainline CFAR workshops does/doesn't relate to running "AI Risk for Computer Scientist" type workshops. Why we both do a lot of recruiting/education for AI alignment research and wouldn't be happy doing only that.
- How our curriculum has evolved. How it relates to and differs from the Less Wrong Sequences. Where we hope to go with our curriculum over the next year, and why.

Several CFAR staff members will be answering questions, including: me, Tim Telleen-Lawton, Adam Scholl, and probably various others who work at CFAR. However, we will try to answer with our own individual views (because individual speech is often more interesting than institutional speech, and certainly easier to do in a non-bureaucratic way on the fly), and we may give more than one answer to questions where our individual viewpoints differ from one another's!

(You might also want to check out our 2019 Progress Report and Future Plans. And we'll have some other posts out across the remainder of the fundraiser, from now til Jan 10.)

[Edit: We're out of time, and we've allocated most of the reply-energy we have for now, but some of us are likely to continue slowly dribbling out answers from now til Jan 2 or so (maybe especially to replies, but also to some of the q's that we didn't get to yet). Thanks to everyone who participated; I really appreciate it.]

Under what circumstances is "don't look at existing research" good advice?

In How I do research, TurnTrout writes:

> [I] Stare at the problem on my own, ignoring any existing thinking as much as possible. Just think about what the problem is, what's confusing about it, what a solution would look like. In retrospect, this has helped me avoid anchoring myself. Also, my prior for existing work is that it's confused and unhelpful, and I can do better by just thinking hard.

The MIRI alignment research field guide has a similar sentiment:

> It's easy to fall into a trap of (either implicitly or explicitly) conceptualizing "research" as "first studying and learning what's already been figured out, and then attempting to push the boundaries and contribute new content."
>
> The problem with this frame (according to us) is that it leads people to optimize for absorbing information, rather than seeking it instrumentally, as a precursor to understanding. (Be mindful of what you're optimizing in your research!) [...]
>
> ... we recommend throwing out the whole question of authority. Just follow the threads that feel alive and interesting. Don't think of research as "study, then contribute." Focus on your own understanding, and let the questions themselves determine how often you need to go back and read papers or study proofs.
>
> Approaching research with that attitude makes the question "How can meaningful research be done in an afternoon?" dissolve. Meaningful progress seems very difficult if you try to measure yourself by objective external metrics. It is much easier when your own taste drives you forward.

And I'm pretty sure that I have also seen this notion endorsed elsewhere on LW: do your own thinking, don't anchor on the existing thinking too much, don't worry too much about justifying yourself to established authority. It seems like a pretty big theme among rationalists in general.

At the same time, it feels like there are fields where nobody would advise this, or where trying to do this is a well-known failure mode. TurnTrout's post continues:

> I think this is pretty reasonable for a field as young as AI alignment, but I wouldn't expect this to be true at all for e.g. physics or abstract algebra. I also think this is likely to be true in any field where philosophy is required, where you need to find the right formalisms instead of working from axioms.

It is not particularly recommended that people try to invent their own math instead of studying existing math. Trying to invent your own physics without studying real physics just makes you into a physics crank, and most fields seem to have some version of "this is an intuitive assumption that amateurs tend to believe, but is in fact wrong, though the reasons are sufficiently counterintuitive that you probably won't figure it out on your own".

But "do this in young fields, not established ones" doesn't seem quite right either. For one, philosophy is an old field, yet it seems reasonable that we should indeed sometimes do it there. And it seems that even within established fields where you normally should just shut up and study, there will be particular open questions or subfields where "forget about all the existing work and think about it on your own" ought to be good advice.

But how does one know when that is the case?

Propagating Facts into Aesthetics

Epistemic status: Tentative. I've been practicing this on-and-off for a year and it's seemed valuable, but it's the sort of thing I might look back on and say "hmm, that wasn't really the right frame to approach it from."

In doublecrux, the focus is on "what observations would change my mind?"

In some cases this is (relatively) straightforward. If you believe minimum wage helps workers, or harms them, there are some fairly obvious experiments you might run. "Which places have instituted minimum wage laws? What happened to wages? What happened to unemployment? What happened to worker migration?"

The details will matter a lot. The results of the experiment might be weird and confusing. If I ran the experiment myself I'd probably get a lot of things wrong, misuse statistics and forget to account for some confounding factors. But I don't feel *confused* about how to learn better statistics, account for more confounders, etc.

But there's a problem that seems harder to me, which is *how to change my mind about aesthetics.* Sarah Constantin first brought this up in Naming the Nameless, and I've been thinking about it ever since.

I think a lot of deep disagreements have to do with "what is beautiful, and what is ugly?", and inability to directly address this is part of what prevents those disagreements from resolving.

In the case of the minimum wage example, you might run an experiment, and find overwhelming evidence that minimum wage helps or hurts workers. But because there's lots of confounders, the evidence might be mixed and confusing. How you interpret it will depend on how it fits into your existing worldview.

Part of this has to do with your ontological frame. But I think a lot has to do with aesthetics judgments, such as:

- *Is capitalism ugly and/or distasteful?* You might have very salient examples of how capitalism can result in exploitation, pollution, or people becoming trapped in unhealthy power structures.
- *Is capitalism beautiful?* Alternately, it might be salient that capitalism creates supermarkets, gains from trade, and vast surplus. Economic efficiency isn't just pretty numbers on a graph, it's real value being created.

These don't directly bear on the minimum wage question, but might make it harder to resolve.

In some cases, your aesthetic taste might make it harder to update on new information properly. In other cases, your aesthetic taste might help you to notice important patterns more readily.

Why 'aesthetics?'

I'm using the word aesthetic in a nonstandard way. When people do that, I think it's important to be clear and about what they're doing and why.

There's a few different words I might have used here, including "feelings", "ontologies", "frameworks", and "values."

Most obviously, I could have asked 'is capitalism *good/bad?*' instead of 'is capitalism beautiful or ugly?'.

I'm making a fairly strong claim (weakly held) that "is it beautiful or ugly?" is at least one of the important questions to be asking, in addition to "is capitalism good/bad" and "does raising minimum wage help or harm workers?". Not because it's how a flawless AI would think about it, but because it's how humans seem to often think about it.

What is an aesthetic?

An aesthetic is a mishmash of values, strategies, and ontologies that reinforce each other.

The values reinforce "you want to use strategies that achieve these values."

The act of using a particular strategy shapes the ontology that you see the world through.

The ontology reinforces what values seem important to you.

Together, this all creates a feedback loop between your metagoals and subgoals, where the process of using this cluster of value/strategy/ontology makes each link in the chain stronger.

In humans (who have messy, entangled brains), this caches out into feelings, felt senses. The original goal and the metagoals blur together. I think "this helps me achieve my [generic] goals" might reinforce "these *particular* subgoals I have are good goals to help with my overall flourishing."

This might be implemented via evolution over millions of years, or via human brains over decades. A Just So Story I'm not sure I endorse but hopefully gets the point across:

> "A flower is beautiful, you say. Do you think there is no story behind that beauty, or that science does not know the story? Flower pollen is transmitted by bees, so by sexual selection, flowers evolved to attract bees—by imitating certain mating signs of bees, as it happened; the flowers' patterns would look more intricate, if you could see in the ultraviolet. Now healthy flowers are a sign of fertile land, likely to bear fruits and other treasures, and probably prey animals as well; so is it any wonder that humans evolved to be attracted to flowers?

Here are some things that you might find beautiful, or distasteful:

- Mozart
- Punk Rock
- Readable, well-written code
- Clever hacks that got the job done quickly

- Cities built on rectangular grids
- Winding alleyways in villages where nobody has consistent names
- People being physically affectionate in public
- A harsh, barren desert
- A lush valley with a river
- Swamps / wetlands
- Nature in general
- Manicured gardens
- Books
- Throwing away books
- People speaking in languages different from yours
- Dense spreadsheets laden with accurate data
- Minimalism
- Frugalism
- Patriotism
- People going out of their way to be kind to their neighbors
- People going out of their way to solve small-but-common problems using math

(If you're like me, you might find it distasteful when people make moral arguments that seem rooted in distaste... and then feel kinda self conscious about the contradiction)

Sometimes you're doublecruxing with someone, and they've explained their model. And their model... makes sense. But the conclusion just seems *so damn ugly*. You want to take 5x the time to write beautiful code, and they just want you to get the job done and ship it.

One thing you can do is push aside your aesthetic judgment, shut up and multiply. This may be useful for expediency.

But sometimes, I think the correct thing is for one or both people to backpropagate facts through their aesthetics.

I do *not* think you should rush or "force" this. Your sense of beauty is there for a reason. But I have a sense that figuring out how to do this well is a key open problem in applied rationality.

Examples

Are Swamps Beautiful?

Compare the swamp with a verdant forest.

If you're like me, swamps seem ugly. Forests seem pretty.

My associations with swamps come largely from stories (and perhaps most concretely, from the game "Magic the Gathering"), where they're often presented as places of disease, murky horrors and corrupt magic. In person, swamps are physically hard to walk in (sometimes solid ground turns out to be algae), and full of mosquitos that bite me.

Are these associations accurate?

Well, the solid ground and mosquito issues are definitely real.

Swamp Thing is not real, life-stealing magic is not real.

Are there additional facts I can learn? My sister evaluates land for construction projects. She says that swamps often serve important roles as a natural way to filter water, and when you naively drain swamps, water quality in an area gets worse. James Scott in Against the Grain claims that early Sumerian civilization developed in swamps, where food and resources were plentiful and life was fairly leisureful – until empires arose and subjected people and forced them to switch to easily-taxable crops instead of the ones that grew naturally. (Speaking of which: is civilization beautiful or ugly?)

I probably find forests beautiful, in part, because they represent a lot of resources that I understand how to make use of. If swamps also supply those resources, maybe I should respect them more?

I also find forests beautiful because my experience stems from a) enchanted forests in fairytales, and b) relatively manicured national parks. If I remind myself that the last time I walked through an *untamed* forest, it was dense with brambles that cut me 'till I bled. It wasn't actually a much nicer experience than the last time I explored a swamp. (It also had non-trivial numbers of mosquitos)

In this example, simply mulling over the facts naturally re-organizes my feelings about them. I still find swamps ugly, but less ugly than before. I expect that, if I reflected on this periodically, over time, it would shift a bit more.

Are Harsh Deserts Beautiful?

I am in fact confused by this. My answer is "yes", and I don't know why. Deserts don't have much in the way of resources. Their stark beauty is more like the way a statue is beautiful than the way a forest is beautiful.

I mulled this one over for a while, am still confused and I note it here because "noticing the limits of a model" seems important.

[Edit: this was discussed more in the comments.]

Is Helping Nearby People Other Beautiful?

The first experience I got with aesthetic doublecrux was debating “hufflepuff virtue” with Oliver Habryka.

I had a strong sense that “people helping each other out” was good and right and virtuous. There was a beauty to the sort of community where everyone notices when someone is hurting (and reaches out to help), or when a space is messy (and cleans it up). There was a cluster of attributes that seemed to fit together in a way that was stronger than the sum of its parts.

And this was visibly lacking in the Berkeley community, and it was resulting in people feeling alienated and distrustful of each other, and many spaces being either messy,

or burdening a single person with cleaning up everyone else's mess.

This seemed concretely harmful. But it also just seemed… ugly and bad.

Oliver had a different view, which I summarize as the "systemization and specialization" approach. (previously discussed here)

If everyone has to pay attention to their environment and notice things that need doing, this is a *lot* of cognitive overhead. If people only have seven working memory slots but they're spending one of them on tracking the environment, that's a dramatic cost on their ability to think. For a community that specializes in thinking, this could be quite bad.

Moreover, "everyone pitch in" is just a really inefficient way of getting things done. A better solution is to streamline and automate as much of the work as possible, hire cleaning services, and whatever remaining work needs doing, simply pay one one of the people something commensurate for their time and effort. Specialization is how things get done when you're doing them seriously.

We argued about this over the course of three days.

I still think there are some things habryka was missing here. But eventually my worldview shifted in some significant ways:

- I updated that the "everyone pitch in" way of keeping spaces clean doesn't make sense for longterm organizations with serious funding. Specialization is real, cognitive bandwidth is precious, and it's generally better to just hire a cleaning service if you can afford one.
- I updated a bit (talking with Satvik) that my model that "helping each other out in low-key ways builds trust which later enables more extensive projects" wasn't as strong as I thought. Satvik asked something like "do you think startup cofounders tend to team up because they've helped each other take out the trash? I feel like it's more about sharing a clear vision and principles or something." And I thought back to some experiences and… yeah that seemed maybe more accurate.
- I gained a better understanding of where and why the "everyone pitch in" approach is useful.
    - Cleaning services are expensive, and if you're a fledgling organization or a typical household, it's probably not worth hiring a cleaner more than once a week or so. Meanwhile, people make messes much more frequently than once a week. If you want your space nice, you have to clean it yourself.
    - There's a value that comes from having community spaces use the "everyone pitch in" method, in that it creates a stronger sense of ownership and buy-in for the space. It also is a mechanism by which people can relate to each other more easily. While this might not be that important for a company, it seems important for a community that's aiming to meet community-shaped-needs.

But this all left me with a nagging, frustrated sense that something important and beautiful being lost. I *want* to live in a world where people help each other out in small ways. It's the particular kind of beauty that a small town in a Miyazaki movie embodies. It feels important to me.

Under what circumstances should I change how I feel about that?

There’s a sense in which aesthetics can’t be proven wrong, or at least “trying to prove it wrong” isn’t really the right frame of mind.

But… I have an aesthetic preference for *consistency,* and for *believing true things* (whether this is good is another question, but I’m taking it at face value for now), which informs my other aesthetics. Aesthetics can turn out to be built out of contradictory pieces, and they can turn out to hinge of false beliefs.

“Trying on” another aesthetic

While talking to habryka, I tried to get a sense of what it’s like to live in the world where systemization and specialization are obviously good and right. What was it like to be habryka? How did this fit together with his other beliefs and values?

Then, once I had a good handle on that, I tried to inhabit “what would it be like to be a Raemon who found systemization and specialization good and right?”. Without *actually* adopting the aesthetic, I tried fitting it into my existing model. This was a bit of an aesthetic process of its own – like trying on a new outfit and seeing how I reacted in the mirror.

I’m not sure if habryka endorses considering those as an ‘aesthetic’, per se. But I found this process valuable.

I gained some ability to see systemization as beautiful. My sense of hufflepuff beauty became more nuanced and caveated.

Clean Code vs Quick Hacks

Humans have (an instinctive? Learned? I'm not sure) sense that when you smell fecal matter or rotting flesh, there is probably disease nearby. It's digusting.

Dogs... well, I'm not 100% sure what's going on with dogs but I think it's something like "strong odors that mask my scent are more useful than disease is bad", and for some reason fecal matter is joyful to play around in.

Programmers often learn that spaghetti code is *evidence* of bugs, even if they don't know exactly what the bug is yet. It acquires a *bad code smell.*

Young programmers often do not have this sense of distaste, and it is important for them to acquire it.

On the flipside: there is also a thing where, well, sometimes you're rushing to ship an Minimum Viable Product and you don't have time to do everything right. It can be legitimately hard to figure out how much effort to put into "doing things right." But it seems at least sometimes, experienced coders either need to learn to "hold their nose" and do the quick fix, or to develop alternate aesthetics that they can shift between depending on circumstances.

Knobs to Turn

There are a few different directions this kind of process might go:

- You could shift to find something *more* beautiful than you did before
- You could shift to find something *less* beautiful than you did before.
- You could shift to find something *more* distasteful than you did before.
- You could shift to find something *less* distasteful than you did before.

I have some sense that these are subtly different processes, although not much evidence to back that up. I also feel like in each case, going from Zero to N, or N to Zero, is different than dialing an existing aesthetic response up or down.

Gaining a new appreciation for why something is beautiful feels different than gaining a categorically new form of disgust. In particular, gaining a new form of beauty mostly makes my life feel nicer, whereas gaining a new form of disgust increases the unpleasantness

Why Does this Matter?

In Naming the Nameless, Sarah Constantin references this comment by Scott Alexander:

> Sometimes I can almost feel this happening. First I believe something is true, and say so. Then I realize it's considered low-status and cringeworthy. Then I make a principled decision to avoid saying it – or say it only in a very careful way – in order to protect my reputation and ability to participate in society. Then when other people say it, I start looking down on them for being bad at public relations. Then I start looking down on them just for being low-status or cringeworthy.
>
> Finally the idea of "low-status" and "bad and wrong" have merged so fully in my mind that the idea seems terrible and ridiculous to me, and I only remember it's true if I force myself to explicitly consider the question. And even then, it's in a condescending way, where I feel like the people who say it's true deserve low status for not being smart enough to remember not to say it. This is endemic, and I try to quash it when I notice it, but I don't know how many times it's slipped my notice all the way to the point where I can no longer remember the truth of the original statement."

Sarah notes:

> Now, I could say "just don't do that, then" -- but Scott of 2009 would have *also* said he believed in being independent and rational and not succumbing to social pressure. Good intentions aren't enough. [...]
>
> I think it's much better to try to make the implicit explicit, to bring cultural dynamics into the light and understand how they work, rather than to hide from them.

Scott's comment gets at what I mean by "An aesthetic is a mishmash of values, strategies, beliefs, and ontologies that reinforce each other." He starts with a belief, then adopts a strategy for how he relates his communication to that belief, and then ends up with a vague sense that the belief is "cringey", and later collapsing it to "cringey and wrong".

This quite worrying epistemic horror.

I think most of what needed saying, Sarah already said, but it's worth concluding with here:

> If you take something about yourself that's "cringeworthy" and, instead of cringing yourself, try to look at *why* it's cringeworthy, what that's made of, and dialogue honestly with the perspective that disagrees with you -- then there is, in a sense, nothing to fear.
>
> There's an "elucidating" move that I'm trying to point out here, where instead of defending against an allegation, you say "let's back up a second" and bring the entire situation into view. It's what double crux is about -- "hey, let's find out what even is the disagreement between us." Double crux is hard enough with *arguments*, and here I'm trying to advocate something like double-cruxing *aesthetic preferences*, which sounds absurdly ambitious. But: imagine if we could talk about *why* things seem beautiful and appealing, or ugly and unappealing. Where do these preferences come from, in a causal sense? Do we still endorse them when we know their origins? What happens when we *bring tacit things into consciousness*, when we talk carefully about what aesthetics evoke in us, and how that might be the same or different from person to person?
>
> Unless you can think about how cultural messaging works, you're going to be a *mere* consumer of culture, drifting in whatever direction the current takes you.

I'm hoping this post gives some nuts and bolts on how to actually make progress on that goal.

Again, I don't know that the specific techniques I list in this post are the best ones, or how often exactly aesthetic concerns are most relevant. I think it's usually good form to *start* with an attempt to take arguments at face value, and debate about concrete beliefs.

But, if that isn't working, I think digging into aesthetics is one of the tools that's important to have in your toolkit.

[Personal Experiment] One Year without Junk Media

I wrote in a previous post how my life is better when I avoid certain kinds of media. It increases my happiness, lowers my stress and makes me smarter.

I should follow my own advice. I hereby commit myself to abstaining from junk media for one year.

There's a long list of rules, definitions and exceptions but the basic idea is I'll avoid videogames, news, Reddit, web surfing, Hot Network Questions and similar mindless media feeds. YouTube is special case. Music and dance videos are okay so I created a new YouTube account and trained the recommendation algorithm to recommend only these kinds of videos.

I've been abstaining from junk media for increasingly long periods of time. My record is around two months. Now's finally the time to pull the trigger and go a whole year.

New Years resolutions usually fail so instead I'm starting this resolution on the 317th anniversary of the 47 Ronin.

万歳！

Edit: I cut this short 3 months early in October 2020 due to multiple changes in my life circumstances.

The Review Phase

LessWrong is currently doing a major review of 2018 — looking back at old posts and considering which of them have stood the test of time. Info about what features we added to the site for writing reviews is in December's monthly updates post.

There are three phases:

- ***Nomination** (completed)*
- ***Review** (ends Dec 31st [EDIT: Jan 13th])*
- ***Voting** on the best posts (ends January 7th [EDIT: Jan 13th])*

We're now in the Review Phase, and there are 75 posts that got two or more nominations. The full list is here. Now is the time to dig into those posts, and for each one ask questions like "What did it add to the conversation?", "Was it epistemically sound?" and "How do I know these things?".

The LessWrong team will award $2000 in prizes to the reviews that are most helpful to them for deciding what goes into the Best of 2018 book.

If you're a nominated author and for whatever reason don't want one or more of your posts to be considered for the Best of 2018 book, contact any member of the team - e.g. drop me an email at benitopace@gmail.com.

Creating Inputs For LW Users' Thinking

The goal for the next month is for us to try to figure out which posts we think were the best in 2018.

Not which posts were talked about a lot when they were published, or which posts were highly upvoted at the time, but which posts, with the benefit of hindsight, you're most grateful for being published, and are well suited to be part of the foundation of future conversations.

This is in part an effort to reward the best writing, and in part an effort to solve the bandwidth problem (there were more than 2000 posts written in 2018) so that we can build common knowledge of the best ideas that came out of 2018.

With that aim, when *I'm* reviewing a post, the main question I'm asking myself is

> What information can I give to other users to help them think clearly and accurately about whether a given post should be added to our annual journal?

A large part of the review phase is about producing inputs for our collective thinking. With that in mind, I've gathered some examples of things you can write that are help others understand posts and their impacts.

1) Personal Experience Reports

There were a lot of examples of this in the nomination phase, which I found really useful, and would find useful to read more of. Here are some examples:

Raemon:

> This post... may have actually had the single-largest effect size on "amount of time I spent thinking thoughts descending from it."

Joh N. Swentworth

> This post (and the rest of the sequence) was the first time I had ever read something about AI alignment and thought that it was actually asking the right questions. It is not about a sub-problem, it is not about marginal improvements. Its goal is a gears-level understanding of agents, and it directly explains why that's hard. It's a list of everything which needs to be figured out in order to remove all the black boxes and Cartesian boundaries, and understand agents as well as we understand refrigerators.

Swimmer963:

> Used as a research source for my EA/rationality novel project, found this interesting and useful.

David Manheim:

> Until seeing this post, I did not have a clear way of talking about common knowledge. Despite understanding the concept fairly well, this post made the points more clearly than I had seen them made before, and provided a useful reference when talking to others about the issue.

Eli Tyre:

> One of my favorite posts, that encouraged me to rethink and redesign my honesty policy.

ryan_b:

> I have definitely linked this more than any other post.

More detail is also really great. I'd definitely encourage the above users to be more thorough about how the ideas in the post impacted them. Here's a nomination that had a bunch more detail about how the ideas have affected them.

jacobjacob:

> In my own life, these insights have led me to do/considering doing things like:
>
> • not sharing private information even with my closest friends -- in order for them to know in future that I'm the kind of agent who can keep important information (notice that there is the counterincentive that, in the moment, sharing secrets makes you feel like you have a stronger bond with someone -- even though in the long-run it is evidence to them that you are less trustworthy)
>
> • building robustness between past and future selves (e.g. if I was excited about and had planned for having a rest day, but then started that day by work and being really excited by work, choosing to stop work and decide to rest such that

different parts of me learn that I can make and keep inter-temporal deals (*even if* work seems higher ev in the moment))

• being more angry with friends (on the margin) -- to demonstrate that I have values and principles and will defend those in a predictable way, making it easier to coordinate with and trust me in future (and making it easier for me to trust others, knowing I'm capable of acting robustly to defend my values)

• thinking about, in various domains, "What would be my limit here? What could this person do such that I would stop trusting them? What could this organisation do such that I would think their work is net negative?" and then looking back at those principles to see how things turned out

• not sharing passwords with close friends, even for one-off things -- not because I expect them to release or lose it, but simply because it would be a security flaw that makes them more vulnerable to anyone wanting to get to me. It's a very unlikely scenario, but I'm choosing to adopt a robust policy across cases, and it seems like useful practice

A special case here is data from the author themselves, e.g. “Yeah, this has been central to my thinking” or “I didn’t really think about it again” or “I actually changed my mind and think this is useful but wrong”. I would generally be excited for users to review their own posts now that they've had ~1.5 years of hindsight, and I plan to do that for all the posts I've written that were nominated.

If a post had a big or otherwise interesting impact on you, consider writing that up.

2) Big Picture Analysis (e.g. Book Reviews)

There are lots of great book reviews on the web that really help the reader understand the context of the book, and explain what it says and adds to the conversation.

Some good examples on LessWrong are be the reviews of Pearl's Book of Why, The Elephant in the Brain, The Secret of Our Success, Consciousness Explained, Design Principles of Biological Circuits, The Case Against Education (part 2, part 3), and The Structure of Scientific Revolutions.

Many of these reviews do a great job of things like

- Talking about how the post fits into the broader conversation on that topic
- Trying to pass the ITT of the author by explaining how they see the world
- Looking at that same topic through their own worldview
- Pointing out places they see things differently and offering alternative hypotheses.

A review of some LessWrong posts would be that time Scott reviewed Inadequate Equilibria. Oh, and don't forget that time Scott reviewed Inadequate Equilibria.

Many of the posts we're reviewing are shorter than most of the reviews I linked to, so it doesn't apply literally, but much of the spirit of these reviews is great. Also check out others short book reviews and consider writing something in that style (e.g. SSC, Thing of Things).

Consider picking a book review style you like and applying it to one of the nominated posts.

3) Testing Subclaims (e.g. Epistemic Spot Checks)

Elizabeth Van Nostrand has written several posts in this style.

- Epistemic Spot Check: The Role of Deliberate Practice in the Acquisition of Expert Performance
- Epistemic Spot Check: Full Catastrophe Living (Jon Kabat-Zinn)
- Epistemic Spot Check: The Dorito Effect (Mark Schatzker)

For another example, in Scott's review of Secular Cycles, one way he tried to think about the ideas in the book was to gather a bunch of alternative data sets on which to test some of the author's claims.

These things aren't meant to be full reviews of the entire book or paper, or advice on overall how to judge it. They take narrower questions that are definitively answerable, like is a random sample of testable claims literally true, and answers them as fully as possible.

If there is an important subclaim of a post you think you can check out, consider trying to verify/falsify the claim and writing up your results and partial results.

Go forth and think out loud!

Karate Kid and Realistic Expectations for Disagreement Resolution

There's an essay that periodically feels deeply relevant to a situation:

> Someday I want to write a self-help book titled "F*k The Karate Kid: Why Life is So Much Harder Than We Think".
>
> Look at any movie with a training montage: The main character is very bad at something, then there is a sequence in the middle of the film set to upbeat music that shows him practicing. When it's done, he's an expert.
>
> It seems so obvious that it actually feels insulting to point it out. But it's not obvious. Every adult I know--or at least the ones who are depressed--continually suffers from something like sticker shock (that is, when you go shopping for something for the first time and are shocked to find it costs way, way more than you thought). Only it's with effort. It's Effort Shock.
>
> We have a vague idea in our head of the "price" of certain accomplishments, how difficult it should be to get a degree, or succeed at a job, or stay in shape, or raise a kid, or build a house. And that vague idea is almost always catastrophically wrong.
>
> Accomplishing worthwhile things isn't just a little harder than people think; it's 10 or 20 times harder. Like losing weight. You make yourself miserable for six months and find yourself down a whopping four pounds. Let yourself go at a single all-you-can-eat buffet and you've gained it all back.
>
> So, people bail on diets. Not just because they're harder than they expected, but because they're so much harder it seems unfair, almost criminally unjust. You can't shake the bitter thought that, "This amount of effort should result in me looking like a panty model."
>
> It applies to everything. [The world] is full of frustrated, broken, baffled people because so many of us think, "If I work this hard, this many hours a week, I should have (a great job, a nice house, a nice car, etc). I don't have that thing, therefore something has corrupted the system and kept me from getting what I deserve."

Last time I brought this up it was in the context of realistic expectations for self improvement.

This time it's in the context of productive disagreement.

Intuitively, it feels like when you see someone being wrong, and you have a simple explanation for why they're wrong, it should take you, like, 5 minutes of saying "Hey, you're wrong, here's why."

Instead, Bob and Alice people might debate and doublecrux for 20 hours, making serious effort to understand each other's viewpoint... and the end result is a conversation that *still* feels like moving through molasses, with both Alice and Bob feeling like the other is missing the point.

And if 20 hours seems long, try *years*.

AFAICT the Yudkowsky/Hanson Foom Debate didn't really resolve. But, the general debate over "should we expect a sudden leap in AI abilities that leaves us with a single victor, or a multipolar scenario?" has actually progressed over time. Paul Christiano's Arguments About Fast Takeoff seemed most influential of reframing the debate in a way that helped some people stop talking past each other, and focus on the actual different strategic approaches that the different models would predict.

Holden Karnofsky initially had some skepticism about some of MIRI's (then SIAI's) approach to AI Alignment. Those views changed over the course of years.

On the LessWrong team, we have a lot of disagreements about how to make various UI tradeoffs, which we still haven't resolved. But after a year or so of periodic chatting about I think we at least have better models of each other's reasoning, and in some cases we've found third-solutions that resolved the issue.

I have observed myself taking years to really assimilate the worldviews of others.

When you have deep frame disagreements, I think "years" is actually just a fairly common timeframe for processing a debate. I don't think this is a *necessary fact* about the universe, but it seems to be the status quo.

Why?

The reasons a disagreement might take years to resolve vary, but a few include:

i. Complex Beliefs, or Frame Differences, that take time to communicate.

Where the blocker is just "dedicating enough time to actually explaining things." Maybe the total process only takes 30 hours but you have to actually do the 30 hours, and people rarely dedicate more than 4 at a time, and then don't prioritize finishing it that highly.

ii. Complex Beliefs, or Frame Differences, that take time to absorb

Sometimes it only takes an hour to explain a concept explicitly, but it takes awhile for that concept to propagate through your implicit beliefs. (Maybe someone explains a pattern in social dynamics, and you nod along and say "okay, I could see that happening sometimes", but then over the next year you start to see it happening, and you don't "really" believe in it until you've seen it a few times.)

Sometimes it's an even vaguer thing like "I dunno man I just needed to relax and not think about this for awhile for it to subconsciously sink in somehow"

iii. Idea Innoculation + Inferential Distance

Sometimes the first few people explaining a thing to you suck at it, and give you an impression that anyone advocating the thing is an idiot, and causes you to subsequently dismiss people who pattern match to those bad arguments. Then it takes someone who puts a lot of effort into an explanation that counteracts that initial bad taste.

iv. Hitting the right explanation / circumstances

Sometimes it just takes a specific combination of "the right explanation" and "being in the right circumstances to hear that explanation" to get a magical click, and unfortunately you'll need to try several times before the right one lands. (And, like reason #1 above, this doesn't necessarily take *that* much time, but nonetheless takes years of intermittent attempts before it works)

v. Social pressure might take time to shift

Sometimes it just has nothing to do with good arguments and rational updates – it turns out you're a monkey who's window-of-possible beliefs depends a lot on what other monkeys around you are willing to talk about. In this case it takes years for enough people around you to change their mind first.

Hopefully you can take actions to improve your social resilience, so you don't have to wait for that, but I bet it's a frequent cause.

Optimism and Pessimism

You can look at this glass half-empty or half-full.

Certainly, if you're expecting to convince people of your viewpoint within a matter of hours, you may sometimes have to come to terms with that not always happening. If your plans depend on it happening, you may need to re-plan. (Not always: I've *also* seen major disagreements get resolved in hours, and sometimes even 5 minutes. But, "years" might be an outcome you need to plan around. If it *is* taking years it may not be worthwhile unless you're actually building a product together.)

On the plus side... I've now gotten to see several deep disagreements actually progress. I'm not sure I've seen a years-long disagreement resolve *completely*, but have definitely seen people change their minds in important ways. So I now have existence proof that this is even possible to address.

Many of the reasons listed above seem addressable. I think we can do better.

Should We Still Fly?

I've seen a lot of discussion about plane travel from a climate perspective lately, with people arguing that we should try to restructure our lives to fly much less. Avoid business travel, vacation closer to home, visit relatives less, etc. After looking at the numbers, though, I think this mostly doesn't make sense.

Let's take an example round trip flight from Boston to LA. I've flown this many times for work and to visit relatives, and it's maybe on the long end for a vacation flight. Taking into account that emissions at high altitude are worse than at ground level, that's about 1.3T CO2e [1].

The thing is, 1.3T isn't that much! For example, carbon offsets are about $10/T, so this would add just ~$13 to your ~$500 round-trip flight. Or, if you don't trust offsets and would rather use the full social cost of carbon, that's ~$55/T (Wang et. al. 2019) or ~$72. Or, if you want to go all the way to direct air capture, that's ~$160/T (Keith et. al. 2018) or ~$210.

If you consider a typical BOS-LAX business trip, with, say, $500 for flights, $500 for lodging, $100 for food, and 14hr time lost to travel, a carbon cost of even $210 is rarely going to make the difference on whether the travel is worth it. Even for a vacation, where people tend to be more price sensitive, it's a factor but not nearly the biggest factor.

Climate change is a real problem, and I'm not saying we shouldn't change anything. I favor a stiff carbon tax, high enough to cover the full social cost of emissions. But even under a high tax, most of the things people fly for today would still be worth flying for.

[1] I tried three different calculators and got 1.16T, 1.4T, and 1.36T.

Comment via: facebook

2010s Predictions Review

Ten years ago, lesswrong users made predictions about the 2010s. Review them here.

Generalizing Experimental Results by Leveraging Knowledge of Mechanisms

In a recent post (and papers), Anders Huitfeldt and co-authors have discussed ways of achieving external validity in the presence of “effect heterogeneity.” These results are not immediately inferable using a standard (non-parametric) selection diagram, which has led them to conclude that selection diagrams may not be helpful for "thinking more closely about effect heterogeneity" and, thus, might be "throwing the baby out with the bathwater."

Taking a closer look at the analysis of Anders and co-authors, and using their very same examples, we came to quite different conclusions. In those cases, transportability is not immediately inferable in a fully nonparametric structural model for a simple reason: it relies on *functional constraints* on the structural equation of the outcome. Once these constraints are properly incorporated in the analysis, all results flow naturally from the structural model, and selection diagrams prove to be indispensable for thinking about heterogeneity, for extrapolating results across populations, and for protecting analysts from unwarranted generalizations. See details in the note we post here for discussion.

Free Speech and Triskaidekaphobic Calculators: A Reply to Hubinger on the Relevance of Public Online Discussion to Existential Risk

In response to Wei Dai's claim that a multi-post 2009 *Less Wrong* discussion on gender issues and offensive speech went well, MIRI researcher Evan Hubinger writes—

> Do you think having that debate online was something that needed to happen for AI safety/x-risk? Do you think it benefited AI safety at all? I'm genuinely curious. My bet would be the opposite—that it caused AI safety to be more associated with political drama that helped further taint it.

Okay, but the *reason* you think AI safety/x-risk is important is *because* twenty years ago, people like Eliezer Yudkowsky and Nick Bostrom were trying to do *systematically correct reasoning* about the future, noticed that the alignment problem looked really important, and *followed that line of reasoning where it took them*—even though it probably looked "tainted" to the serious academics of the time. (The robot apocalypse is nigh? Pftt, sounds like science fiction.)

The *cognitive algorithm* of "Assume my current agenda is the most important thing, and then execute whatever political strategies are required to protect its social status, funding, power, un-taintedness, *&c.*" wouldn't have led us to *noticing* the alignment problem, and I would be pretty surprised if it were sufficient to solve it (although that would be very convenient).

An analogy: it's actually *easier* to build a calculator that does correct arithmetic than it is to build a "triskaidekaphobic calculator" that does "correct arithmetic, *except* that it never displays the result 13", because the simplest implementation of the latter is just a calculator *plus* an extra conditional that puts something else on the screen when the real answer would have been 13.

If you don't actually understand how arithmetic works, but you feel intense social pressure to produce a machine that never displays the number 13, I don't think you actually succeed at building a triskaidekaphobic calculator: you're trying to solve a problem under constraints that make it impossible to solve a *strictly* easier problem.

Similarly, I conjecture that it's actually easier to build a rationality/alignment research community that does systematically correct reasoning, than it is to build a Catholic rationality/alignment research community that does "systematically correct reasoning, *except* never saying anything the Pope disagrees with." The latter is a strictly harder problem: you have to somehow *both* get the right answer, *and* throw out all of the steps of your reasoning that the Pope doesn't want you to say.

You're absolutely right that figuring out how politics and the psychology of offense work doesn't *directly* help increase the power and prestige of the "AI safety" research agenda. It's just that the *caliber of thinkers* who can solve AGI alignment should *also* be able to solve politics and the psychology of offense, much as how a calculator that can compute `1423 + 1389` should *also* be able to compute `6 + 7`.

The Lesson To Unlearn

This is a linkpost for http://paulgraham.com/lesson.html

> The most damaging thing you learned in school wasn't something you learned in any specific class. It was learning to get good grades.

Humans Are Embedded Agents Too

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Most models of agency (in game theory, decision theory, etc) implicitly assume that the agent is separate from the environment - there is a "Cartesian boundary" between agent and environment. The embedded agency sequence goes through a long list of theoretical/conceptual problems which arise when an agent is instead embedded in its environment. Some examples:

- No defined/input output channels over which to optimize
- Agent might accidentally self-modify, e.g. drop a rock on its head
- Agent might intentionally self-modify, e.g. change its own source code
- Hard to define hypotheticals which don't actually happen, e.g. "I will kill the hostages if you don't pay the ransom"
- Agent may contain subcomponents which optimize for different things
- Agent is made of parts (e.g. atoms) whose behavior can be predicted without thinking of the agent as agenty - e.g. without thinking of the agent as making choices or having beliefs
- Agent is not logically omniscient: it cannot know all the implications of its own beliefs

The embedded agency sequence mostly discusses how these issues create problems for designing reliable AI. Less discussed is how these same issues show up when modelling humans - and, in particular, when trying to define human values (i.e. "what humans want"). Many - arguably most - of the problems alignment researchers run into when trying to create robust pointers to human values are the same problems we encounter when talking about embedded agents in general.

I'll run through a bunch of examples below, and tie each to a corresponding problem-class in embedded agency. While reading, bear in mind that **directly answering the questions posed is not the point**. The point is that each of these problems is a symptom of the underlying issue: humans are embedded agents. Patching over each problem one-by-one will produce a spaghetti tower; ideally we'd tackle the problem closer to the root.

The Keyboard is Not The Human

Let's imagine that we have an AI which communicates with its human operator via screen and keyboard. It tries to figure out what the human wants based on what's typed at the keyboard.

A few possible failure modes in this setup:

- The AI wireheads by seizing control of the keyboard (either intentionally or accidentally)
- A cat walks across the keyboard every now and then, and the AI doesn't realize that this input isn't from the human
- After a code patch, the AI filters out cat-input, but also filters out some confusing (but important) input from the human

Embedded agency problem: **humans do not have well-defined output channels**. We cannot just point to a keyboard and say “any information from that keyboard is direct output from the human”. Of course we can come up with marginally better solutions than a keyboard - e.g. voice recognition - but eventually we’ll run into similar issues. There is nothing in the world we can point to and say “that’s the human’s output channel, the entire output channel, and nothing but the output channel”. Nor does any such output channel exist, so e.g. we won’t solve the problem just by having uncertainty over where exactly the output channel is.

Modified Humans

Because humans are embedded in the physical world, there is no fundamental block to an AI modifying us (either intentionally or unintentionally). Define what a “human” is based on some neural network which recognizes humans in images, and we risk an AI modifying the human by externally-invisible means ranging from drugs to wholesale replacement.

Embedded agency problem: **no Cartesian boundary**. All the human-parts can be manipulated/modified; the AI is not in a different physical universe from us.

Off-Equilibrium

Human choices can depend on off-equilibrium behavior - what we or someone else *would* do, in a scenario which never actually happens. Game theory is full of examples, especially threats: we don’t launch our nukes because we expect our enemies *would* launch their nukes... yet what we actually expect to happen is for nobody to launch any nukes. Our own behavior is determined by “possibilities” which we don’t actually expect to happen, and which may not even be possible. Embedded agency problem: **counterfactuals**.

Going even further: our values themselves can depend on counterfactuals. My enjoyment of a meal sometimes depends on what the alternatives were, even when the meal is my top pick - I’m happier if I didn’t pass up something nearly-as-good. We’re often unhappy to be forced into a choice, even if it’s a choice we would have made anyway. What does it mean to “have a choice”, in the sense that matters for human values? How do we physically ground that concept? If we want a friendly AI to allow us choices, rather than force us to do what’s best for us, then we need answers to questions like these.

Drinking

Humans have different preferences while drunk than while sober [CITATION NEEDED]. When pointing an AI at “human values”, it’s tempting to simply say “don’t count decisions made while drunk”. But on the other hand, people often drink to intentionally lower their own inhibitions - suggesting that, at a meta-level, they *want* to self-modify into making low-inhibition decisions (at least temporarily, and within some context, e.g. at a party).

Embedded agency problem: **self-modification and robust delegation**. When a human intentionally self-modifies, to what extent should their previous values be honored, to what extent their new values, and to what extent their future values?

Value Drift

Humans generally have different values in childhood, middle age, and old age. Heck, humans have different values just from being hangry! Suppose a human makes a precommitment, and then later on, their values drift - the precommitment becomes a nontrivial constraint, pushing them to do something they no longer wish to do. How should a friendly AI handle that precommitment?

Embedded agency problem: **tiling & delegation failures**. As humans propagate through time, our values are not stable, even in the absence of intentional self-modification. Unlike in the AI case, we can't just design humans to have more stable values. (Or can we? Would that even be desirable?)

Akrasia

Humans have subsystems. Those subsystems do not always want the same things. Stated preferences and revealed preferences do not generally match. Akrasia exists; many people indulge in clicker games no matter how much some other part of themselves wishes they could be more productive.

Embedded agency problem: **subsystem alignment**. Human subsystems are not all aligned all the time. Unlike the AI case, we can't just design humans to have better-aligned subsystems - first we'd need to decide what to align them to, and it's not obvious that any one particular subsystem contains the human's "true" values.

Preferences Over Quantum Fields

Humans generally don't have preferences over quantum fields directly. The things we value are abstract, high-level objects and notions. Embedded agency problem: **multi-level world models**. How do we take the abstract objects/notions over which human values operate, and tie them back to physical observables?

At the same time, our values ultimately need to be grounded in quantum fields, because that's what the world is made of. Human values should not seemingly cease to exist just because the world is quantum and we thought it was classical. It all adds up to normality. Embedded agency problem: **ontological crises**. How do we ensure that a friendly AI can still point to human values even if its model of the world fundamentally shifts?

Unrealized Implications

I have, on at least one occasion, completely switched a political position in about half an hour after hearing an argument I had not previously considered. More generally, we humans tend to update our beliefs, our strategies, and what-we-believe-to-be-our-values as new implications are realized.

Embedded agency problem: **logical non-omniscience**. We do not understand the full implications of what we know, and sometimes we base our decisions/strategies/what-we-believe-to-be-our-values on flawed logic. How is a friendly AI to recognize and handle such cases?

Socially Strategic Self-Modification

Because humans are all embedded in one physical world, lying is hard. There are side-channels which leak information, and humans have long since evolved to pay attention to those side-channels. One side effect: the easiest way to “deceive” others is to deceive oneself, via self-modification. Embedded agency problem: **coordination** with visible source code, plus **self-modification**.

We earnestly adopt both the beliefs and values of those around us. Are those our “true” values? How should a friendly AI treat values adopted due to social pressure? More generally, how should a friendly AI handle human self-modifications driven by social pressure?

Combining this with earlier examples: perhaps we spend an evening drunk because it gives us a socially-viable excuse to do whatever we wanted to do anyway. Then the next day, we bow to social pressure and earnestly regret our actions of the previous night - or at least some of our subsystems do. Other subsystems still had fun while drunk, and we do the same thing the next weekend. What is a friendly AI to make of this? Where, in this mess, are the humans’ “values”?

These are the sorts of shenanigans one needs to deal with when dealing with embedded agents, and I expect that a better understanding of embedded agents in general will lead to substantial insights about the nature of human values.

The New Age of Social Engineering

Why have so many online social networks failed to form healthy communities, and instead gained notoriety as hostile spaces? I argue that the reason these platforms have failed is because they didn't learn the lessons taught by the High Moderns when humans were first faced with the challenge of engineering alongside systems that were built through millennia of natural evolution. In a chaotic environment such as human social relations, a different engineering approach is necessary to ensure that more good is done than harm. To gain the skills necessary to make these projects a success we need to learn from the history of social environments themselves, and of human engineering strategies. What follows is the story of social evolution becoming social engineering, how the meaning of both has changed radically in the last 20 years, and what this means for designers in the new Information Era.

Part 1 — Ten millenniums of social engineering

A key part of my thesis is that the way our social environment is formed has changed over the course of human history, and more rapidly in recent years. How do we know that to be true? Much of the work I'm building on comes out of the accounts provided by The Secret of Our Success by Joseph Henrich, as well as Seeing Like a State by James C. Scott. There are many things that I disagree with in these works, but I think they both get to the core idea that there exist two main ways in which human society develops. One of those ways is via an evolutionary process, where some societies develop some technique that aids in survival and flourishing, pass it on, and end up growing and outcompeting other societies. The people practicing these traditions often don't have concrete knowledge as to why they work, but they become enshrined as tradition because they help the group succeed. This goes from knowledge about what plants are edible, to complex ideas like how the group should be structured. On the other hand, there is social engineering. In social engineering, explicit models of human behavior are used to derive new social conventions and structures. Usually this doesn't mean designing something new from whole cloth, but instead an effective synthesis of ideas that the culture has generated over time into a compelling ideological canon or into new distinct institutions.

For most of human history, we relied primarily on social evolution instead of social engineering. This was for good reason: social engineering when done poorly is very often worse than social evolution. A mother who breaks tradition and tries feeding new plants to their children because she doesn't know of any reason those plants are harmful may discover unexpected side effects of their consumption. This reality is often referred to as Chesterton's fence, and is often discussed as an argument in favor of traditionalism. However, much of the social and technological progress of the industrial era has through the rejection of tradition. How do we square these conflicting forces? I think that a key reason is simply because for a society to succeed in social engineering, a detailed historical record and careful specialists are usually necessary. Only in this way can new first principle knowledge be solidified and built-upon. It's for that reason that the societies that appear the most engineered also in general tend to be those with more detailed historical records and information about other differing societies. When one is able to see the culture "from above", that unique perspective can enable one to design an effective institutions.

One excellent example is perhaps one of the most successful early cultural engineers, Confucius. The Great Teacher developed his unique philosophy while travelling around

China and seeing various social issues, their causes, and the variety of different social structures present in China at the time. By synthesizing these insights into a central canon, he created an enduring cultural institution that was central to Chinese administration for centuries. Of course it is true that Confucianism relies heavily on tradition, and can in some ways be considered no more than a collection of various preexisting traditions, but its success indicates that it must have some quality beyond that of the constituent parts. Ultimately, Confucianism and the society it created lost supremacy because while it itself was a result of synthesis, it became unable to assimilate or change at pace with the world it inhabited. What once created a powerful bureaucratic class capable of financing great discoveries, ended up as a chain that left the society unable to appreciate the possibility of learning from outside influences. Innovation became tradition, and tradition cannot change course by its very nature.

Part 2 — Modernity

We're going to leave behind ancient societies, because although they are a rich source of insight, others have better studied those trends in depth than I. Instead we are going to turn to look at the relatively modern. I am going to focus for a minute on the United States. For all that American Exceptionalism is a real risk, I think there is something somewhat unique and interesting about the formation of the US. Specifically, the US is one of the best examples of what I consider full social engineering. A group of people sat down in a room and set out to design, in a written legal document, how its society would function. It was a group of what can only be called engineers who set out to design structures to improve upon the governments they were aware of. They didn't just try and say "tyranny is bad so we won't be tyrants," they tried to engineer complex social structures that took advantage of human behavior in order to guide the behavior of the government — independent of any single political actor. The idea of applying contractual thinking to the structure of our nations and institutions didn't start with the US, but the US can be seen as a culmination of those ideas. This project has had varied success to say the least, but it's notable that so many modern institutions function in this way. A group of founders get together and try to set the community's direction at both an object and meta-level. Just as the objects and tools we use have become increasingly engineered, so too have our institutions become influenced by engineering. The evolution and design of social norms is at the heart of what we call society. I would venture to say that it is the defining feature of the human species, and of intelligent species in general. The Machiavellian Intelligence Hypothesis posits that intelligence arose, not to better use tools or better hunt prey, but instead to better compete in the social arena. Therefore, I think it's fair to say the top down engineering in the world of social norms faces an uphill battle to outperform the metis of the traditional culture. However, this applies much less when we turn our eyes towards the engineering of the physical spaces our cultures occupy.

Social Engineering in the context of the physical is about the design of objects and spaces in the traditional engineering sense, but with a consideration to how that design influences the group rather than the individual. It's obvious that a designer making a chair must consider how the chair interacts with the behaviors and preferences of the person who will eventually use the chair. This same thoughtfulness should be applied, and usually is, when dealing with objects and spaces that drive social interaction. Someone trying to build a successful bar will think carefully about the layout and decoration of the space and how that will influence their patrons. They may think about other layouts they've seen and how they might improve on those designs in order to give the space the mood they want. The pub is undeniably a social

institution, and it is designed to both encourage and discourage certain types of social behavior. Every space you interact with, from the supermarket to the sidewalk, has generations of trial, error, and improvement. That doesn't mean every space is perfect, but it's easy to forget the marvel that is present all around us. However, it is the process of conscious engineering that has also introduced many institutions that are detrimental to healthy communities. One such piece of design often discussed is the American shopping mall.

In 1954, architect Victor Gruen built the first modern shopping mall in Michigan. Two years before his death in 1978 he would describe malls as destroyers of cities. The suburbanization of America, a process due in large part to decisions made by urban engineers, killed key social spaces and is widely seen as a social engineering mistake of the highest order. Many of these engineers would see the damages within their lifetime, and like Gruen, spend their life trying to reverse course. For much of human history, cities didn't *have* designers, and even when city design started, it satisfied itself with general zoning, usually to protect and enforce class divisions, as often seen in early Chinese urban environments. City planning has introduced big improvements in quality of life and health, but also class segregation and the destruction of communities. Much of the excellent book, Seeing Like a State by James C. Scott, is focused on this phenomenon. The core conclusion of that book is that top-down planning is deeply flawed, evidenced by a variety of failures at such tasks, from farming to the city of Brasilia. The points made by Scott are well argued, and I recommend taking a look at the analysis of the book by Scott Alexander or Lou Keep at the very least. I agree with Scott in that top down planning of human environments is an incredibly difficult task to do successfully, and there are a mountain of failures left behind by the High Moderns for us to learn from. As Scott Alexander once said, "The road we're on is littered with the skulls of the people who tried to do this before us."

To me the key takeaway is that the High Moderns failed because of the particular approaches they took when undertaking their design projects. They regularly ignored the actual desires of the people living in the cities they were to redesign, and their motivations were counter to the goals of the populace. The government wanted more legible, easier to tax cities; the citizens want community and more local control (features which notably go hand in hand with organized resistance). Today, most of our designers are deeply aware of the failures of the high moderns. Books such as *Seeing Like a State* detail these past failures and provide guidance towards avoiding these pitfalls. Arguably cities at the forefront of growth have *overlearned*these lessons, with any attempt to demolish old buildings met with fierce opposition. We aren't perfect, but lessons have been learned in the way we design our social spaces, with one massive exception.

Part 3 — The Internet

The internet has opened up a new frontier in the design of social institution. We are designing platforms that are used by wide masses of people, that grow more quickly than any historical analog, and that provide unprecedented levers of control over discourse. Facebook was founded in 2004; ten years later there were more monthly active Facebook users than Catholics. The ability to implement a new social institution with basically no startup cost and end up with this much influence is unprecedented, and thus it's not surprising that we've seen so many instances of the new social internet having issues with healthy discourse. So much of the evolutionary work that went into shaping our meatspace social institutions has been ignored during the construction of these new online spaces. Platform designers often repeat the mistakes

of the High Moderns. The users of the platforms are rarely given a voice in discussions, and are frequently treated as antagonists rather than stakeholders. Worst of all, centralized platforms have very little immediate incentives to improve the quality of discourse, as network effects prevent users from easily moving to a nicer competitor. Network operators are encouraged to gain users as fast as possible, keep them in the space as long as possible to view ads, and completely ignore the social well-being of the communities that form. Additionally, online platforms provide a nigh microscopic level of control over the interactions between users. Someone designing a bar can choose the layout of the tables and the lighting, but an online platform designer can run automated testing to determine what text, fonts, and layouts best guide the user into behaving in a desired way. In the case of platforms like YouTube, complex AI systems are constantly optimizing every nanometer of the system to maximize ad revenue.

In the early days of the internet there was a lot of optimism about the ability of the web to bring people together, to form new understanding between distant peoples, and to provide an escape from tyranny. It's important to recognize some of the successes on these fronts. I regularly communicate with people from other nations, and that has helped give me a broader view of the world and of differing cultural norms. However, anybody can see that we have failed to live up to this early promise. Most of the social spaces online were driven in design by technological constraints and financial motives, not by a consistent dedication to building prosocial institutions. The rapid expansion of the internet has left engineers struggling to make their websites function at all, let alone spend resources on deep analysis of user behaviors. Such work is only done by established players with the goal of increasing revenue, and thus almost inevitably results in a *worsening* of user experiences because of misaligned incentives. To fix these issues we need both philosophical changes and technological changes.

On the ideological level we need to ensure that programmers, and the managers who direct them, are thinking carefully about how their platform encourages users to act when designing social spaces online. This requires a long-view of platform design, because user engagement metrics can't tell us if we're making people happier or more informed. A key part of this is to respect the history of human social evolution, and use existing institutions as a starting place. For example, it's in the case of Facebook it's unclear what exactly a Facebook friend maps to. Clearly it's not analogous to a real life friend, because even someone you consider a friend can't listen into most of your conversations and interject at will, a behavior mode that "friending" on Facebook enables. As a result, anyone who used Facebook during its height knew the pain of ending up with too many "friends" and having them jump into and derail any conversation, as well as the strange pressures around someone offline asking to be added as a Facebook friend. The failure of Google+ was unfortunate, because the circles concept of grouping individuals into overlapping groups such as "family", "friends", and "acquaintances" maps much more clearly to real world relationships. Once you can approximate an existing healthy institution, then you can start making modifications and consider how they might make the communities better or worse. Approach projects with a respect for your potential users and their values, but don't be naive and assume everyone is a good faith actor. Understand the challenges presented by potentially web scale networks, where you could have anywhere from 50 to 50 thousand users, such as content moderation. This failure to scale up moderation is at the heart of YouTube's difficulty with copyright and Twitter's losing battle over what kind of speech is necessary to censor. Platform owners have tried to react to users attacking one another through more aggressive moderation, but this often results in innocent users being caught in the crossfire. Users are encouraged

to use block/mute features, but such tools aren't effective enough when facing a mob of potentially thousands of users. Users are forced into increasingly defensive behaviour, which means even good faith critics may be blocked out. There probably isn't a one size fits all solution to many of the problems faced during this design process, but if we get programmers and designers to start considering these issues seriously we're already improving. As an example, Yik Yak was developed pretty late into the social network timeline. Even a cursory glance at its design could have revealed the inevitable issues it would have with moderation and abuse. Somehow, the project still went ahead, and not 2 years later the platform was basically dead, with an unknowable amount of social collateral. Changes to the way engineers approach these problems are important, but there are reasons many of these somewhat obvious ideas haven't already been implemented, and why seemingly obvious improvements to platforms get left on the table. That being said, anybody who works in the software world knows that there are already legions of designers that look at how these platforms are put together, trying to improve them every working hour. Clearly, it isn't enough. Some of these problems are fundamental to the technologies and business models relied upon by platform operators.

The combination of centralization and the advertiser revenue model result in a world where essentially every online social space has goals that run counter to the values of the users, at least to some degree. Platform providers are only incentivized to keep their platform nice enough to prevent a total collapse in the user base, an astoundingly low bar because of the network effect. If YouTube was based on a monthly subscription model, the company's main incentive would be to keep viewers and content creators happy. Instead, YouTube's primary goal is keeping advertisers happy, which means doing things like penalizing external linking. The company's obsessive AI driven tweaks all work with the goal of increasing the length of time users spend on the site, and by extension revenue. After years of users protesting that such narrow optimization hurts the community, YouTube has responded and stated it will try to change the way it manages content, but the fact remains that ad revenue is YouTube's primary incentive. YouTube Red can even be considered an acknowledgement of these issues. Already we are seeing cracks in the technological and business structures of online platforms. Brave/BAT and others are pioneering a microtransaction alternative to ads for funding online content. Across basically every social platform struggles with content moderation are revealing the impossibility of centralized solutions to online social spaces, and decentralized media platforms like the Fediverse are waiting in the wings to step in and fill the void. To be honest, many of these alternatives are far from ready to pick up the mantle of the web giants, but they grow more attractive with each passing year. Most importantly, users want alternatives. People are sick of dealing with companies that don't care about them and of ceding control over their communities to some distant office of, at best, overworked engineers just trying to avoid a lawsuit.

I'd like to take a moment to go back to those early internet idealists. To imagine the potential provided by the internet. Imagine settlers arriving at a continent where there is unlimited space for new communities and cultures. Where physical violence is impossible, and where nobody can be prevented from leaving a community they don't like. We aren't there, and maybe it will be a long long time before we get there. Despite that, I'm an optimist at heart. I believe in human ingenuity, and in the potential for people to rise up to the challenge and opportunity they are presented with. The internet is still young, and we have time to make sure that future generations will benefit from it in ways we can't even imagine today.

Part 4 — Closing Remarks

I tried to focus on a descriptive approach in this essay. It’s obviously informed by my own perspective, but I avoid spending a lot of time on the particulars of how I would design a social platform, instead focusing on the technological and incentive structures that provide the foundation for any platform. In the next part of this series on the design of social platforms, I intend to dive more deeply into the specifics of how I might design a platform given the sentiments expressed above, as well as my own thoughts on the nature of community.

Footnotes

This essay is heavily inspired by The Uruk Series by Sam[ ]zdat also known as Lou Keep, whose writing was really inspirational to me, a person who is more high modernist by nature.

Against Premature Abstraction of Political Issues

A few days ago romeostevensit wrote in response to me asking about downvotes on a post:

> I didn't downvote, but I do think that conversations like this attract people who aren't interested in arguing in good faith. I prefer that such discussions occur at one abstraction level up so that they don't need to mention any object level beliefs like social justice in order to talk about the pattern that the author wants to talk about.

And I replied:

> This seems like a reasonable worry. Maybe one way to address it would be to make posts tagged as "politics" (by either the author or a moderator) visible only to logged in users above a certain karma threshold or specifically approved by moderators. Talking at the meta-level is also good, but I think at some point x-risk people have to start discussing object-level politics and we need some place to practice that.

Since writing that, I've had the thought (because of this conversation) that only talking about political issues at a meta level has another downside: premature abstraction. That is, it takes work to find the right abstraction for any issue or problem, and forcing people to move to the meta level right away means that we can't all participate in doing that work, and any errors or suboptimal choices in the abstraction can't be detected and fixed by the community, leading to avoidable frustrations and wasted efforts down the line.

As an example, consider a big political debate on LW back in 2009, when "a portion of comments here were found to be offensive by some members of this community, while others denied their offensive nature or professed to be puzzled by why they are considered offensive." By the time I took my shot at finding the right abstraction for thinking about this problem, three other veteran LWers had already tried to do the same thing. Now imagine if the object level issue was hidden from everyone except a few people. How would we have been able to make the intellectual progress necessary to settle upon the right abstraction in that case?

One problem that exacerbates premature abstraction is that people are often motivated to talk about a political issue because they have a strong intuitive position on it, and when they find what they think is the right abstraction for thinking about it, they'll rationalize an argument for their position within that abstraction, such that accepting the abstract argument implies accepting or moving towards their object-level position. When the object level issue is hidden, it becomes much harder for others to detect such a rationalization. If the abstraction they created is actually wrong or incomplete (i.e., doesn't capture some important element of the object-level issue), their explicit abstract argument is even more likely to have little or nothing to do with what actually drives their intuition.

Making any kind of progress that would help resolve the underlying object-level issue becomes extremely difficult or impossible in those circumstances, as the meta discussion is likely to become bogged down and frustrating to everyone involved as

one side tries to defend an argument that they feel strongly about (because they have a strong intuition about the object-level issue and think their abstract argument explains their intuition) but may actually be quite weak due to the abstraction itself being wrong. And this can happen even if their object-level position is actually correct!

To put it more simply, common sense says hidden agendas are bad, but by having a norm for only discussing political issues at a meta level, we're directly encouraging that.

(I think for this and other reasons, it may be time to relax the norm against discussing object-level political issues around here. There are definitely risks and costs involved in doing that, but I think we can come up with various safeguards to minimize the risks and costs, and if things do go badly wrong anyway, we can be prepared to reinstitute the norm. I won't fully defend that here, as I mainly want to talk about "premature abstraction" in this post, but feel free to voice your objections to the proposal in the comments if you wish to do so.)

Applications of Economic Models to Physiology?

Applying economic models to physiology seems really obvious. For instance:

- Surely the body uses price signals to match production to consumption of various metabolites. Insulin as a price signal for glucose is one example.
- Presumably such price signals coordinate between spatially-separated organs with specialized roles in various physiological "supply chains". That should lead to general equilibrium models, and questions of convexity and stability.
- Can we back out an implied discount rate for the body's long-term energy stores?

Yet when I run a google search for the obvious phrase "econophysiology", I get back five results, most of which appear to be misspellings. (I feel like I ought to write something right now just to call dibs on the name.)

Does anyone know of sources on this sort of thing? Is there a name for it?

Why aren't assurance contracts widely used?

A priori, dominant assurance contracts seem like awesome tools for solving a fairly broad range of collective action problems. Why aren't they used much? Or is it just that they are a new idea and we should expect them to grow in prominence in the next few decades?

Long Bets by Confidence Level

If you want to make a long-term bet one of your options is to register your bet with the Long Now Foundation as a Long Bet. They have some rules, which are roughly:

- Both parties put up the same amount, at least $200/each.
- Long Bets effectively runs a donor-advised fund (DAF).
- When the bet concludes the winner chooses a charity to receive the money.
- The charity gets the initial stakes, plus half the investment income.

While people have all sorts of reasons why they might want to use Long Bets, one question is: how confident do you need to be for placing a long bet to result in more money going to your preferred charity than just putting the money in a DAF now?

Let's say I claim we'll have talking horses ten years from now, and you're skeptical. You consider betting $1000 against my $1000 via Long Bets. If you win you'll get your $1000 back, my $1000, and half the investment income which (figuring the stock market returns a nominal 7%) will be ~$967, for a total of ~$2967. On the other hand, if you had just put your $1000 in a DAF you'd have ~$1967. Is this a good deal?

Provided putting the money in a DAF for at least that long would otherwise be your best option, if you're 100% confident that (a) you'll win and (b) Long Bets will still be around, then it's a solid deal. You're up about 50%. On the other hand, the less confident you are the worse the deal looks:

	confidence						
year	60%	75%	85%	90%	95%	98%	99%
2	12%	41%	59%	69%	78%	84%	85%
3	9%	36%	54%	63%	73%	78%	80%
5	3%	28%	46%	54%	63%	68%	70%
8	-5%	19%	34%	42%	50%	55%	57%
13	-15%	6%	20%	27%	34%	39%	40%
21	-26%	-7%	6%	12%	18%	22%	23%
34	-34%	-17%	-6%	-1%	5%	8%	9%
55	-39%	-23%	-13%	-8%	-3%	0%	1%
89	-40%	-25%	-15%	-10%	-5%	-2%	-1%

(sheet you can copy and play with)

For example, at 60% confidence you're neutral at 6 years, and negative after that. At 75% you're down to neutral at 16 years. At 90%, 32 years. At 99%, 75 years. For an organization trying to promote long-term thinking, it's surprising they would choose a fee structure that penalizes long-term bets so heavily.

Comment via: facebook

Speaking Truth to Power Is a Schelling Point

Consider a coalition that wants to build accurate shared world-models (maps that reflect the territory), and then use those models to inform decisions that achieve the coalition's goals.

However, suppose that some ways of improving models are punished by the surrounding Society. For example, if the Emperor's new clothes turn out to be "vaporwear", agents who notice this might not want to make it common knowledge within their coalition by adding it to the coalition's shared map, because if that knowledge "leaks" during the onerous process of applying for a grant from the Imperial Endowment for the Arts and Sciences, then the grant application will be more likely to be rejected: the Emperor's men don't want to fund coalitions who they can detect believe "negative" things about the Emperor, because those coalitions are more likely to be disloyal to the regime.

(Because while everyone has an interest in true beliefs, disloyal subjects have a unusually large interest in *selectively* seeking out information that could be used against the regime during a revolution. ("The corrupt false Emperor is wasting your tax money on finery that *doesn't even exist!* Will you join in our crusade?") That makes even true negative beliefs about the Emperor become a signal of disloyalty, which in turn gives loyal subjects an incentive to *avoid* learning anything negative about the Emperor in order to credibly signal their loyalty.)

Coalitions need to model the world in order to achieve their goals, but grant money is useful, too. This scenario suggests coalition members working on their shared maps might follow a strategy schema that could be summarized in slogan form as—

> Speak the truth, even if your voice trembles—*unless* adding that truth to our map would make it ***x%*** harder for our coalition to compete for Imperial grant money, in which case, obfuscate, play dumb, stonewall, rationalize, report dishonestly, filter evidence, violate Gricean maxims, lie by omission, gerrymander the relevant category boundaries, *&c.*

(But outright lying is out of the question, because *that* would be contrary to the moral law.)

Then the coalition faces a choice of the exact value of x. Smaller values of x correspond to a more intellectually dishonest strategy, requiring only a small inconvenience before resorting to obfuscatory tactics. Larger values of x correspond to more intellectual honesty: in the limit as $x \to \infty$, we just get, "Speak the truth, even if your voice trembles (full stop)."

Which choice of x looks best is going to depend on the coalition's *current* beliefs: coalition members can only deliberate on the optimal trade-off between map accuracy and money using their *current* map, rather than something else.

But as the immortal Scott Alexander explains, situations in which choices about the current value of a parameter, alter the process that makes future choices about that same parameter, are prone to a "slippery slope" effect: Gandhi isn't a murderer, but

may quickly become one if he's willing to accept a bribe to take a pill that makes him both more violent and *less averse to taking more such pills*.

The slide down a slippery slope tends to stop at "sticky" Schelling points: choices that, for whatever reason, are unusually *salient* in a way that makes them a natural focal point for mutual expectations, an answer different agents (or the same agent at different times) might give to the infinitely recursive question, "What would I do if I were her, wondering what she would do if she were me, wondering what ...?"

In the absence of distinguished salient intermediate points along the uniformly continuous trade-off between maximally accurate world-models and sucking up to the Emperor, the only Schelling points are $x = \infty$ (tell the truth, the whole truth, and nothing but the truth) and $x = 0$ (do everything short of outright lying to win grants). In this model, the tension between these two "attractors" for coordination may tend to promote coalitional schisms.

Recent Progress in the Theory of Neural Networks

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

It's common wisdom that neural networks are basically "matrix multiplications that nobody understands" , impenetrable to theoretical analysis, which have achieved great results largely through trial-and-error. While this may have been true in the past, recently there has been significant progress towards developing a theoretical understanding of neural networks. Most notably, we have obtained an arguably complete understanding of network initialization and training dynamics in a certain infinite-width limit. There has also been some progress towards understanding their generalization behavior. In this post I will review some of this recent progress and discuss the potential relevance to AI alignment.

Infinite Width Nets: Initialization

The most exciting recent developments in the theory of neural networks have focused the infinite-width limit. We consider neural networks where the number of neurons in all hidden layers are increased to infinity. Typically we consider networks with a Gaussian-initialized weights, and scale the variance at initialization as $\frac{1}{\sqrt{H}}$, where H is the number of hidden units in the preceding layer(this is needed to avoid inputs blowing up, and is also the initialization scheme usually used in real networks). In this limit, we have obtained an essentially complete understanding of both behavior at initialization and training dynamics[1]. (Those with limited interest/knowledge of math may wish to "Significance and Limitations" below).

We've actually had a pretty good understanding of the behavior of infinite-width neural networks at initialization for a while, since the work of Radford Neal(1994). He proved that in this limit, fully-connected neural networks with Gaussian-distributed weights and biases limit to what are known as Gaussian processes. Gaussian processes can be thought of the generalization of Gaussian distributions from finite-dimensional spaces to spaces of functions. Neal's paper provides a very clear derivation of this behavior, but I'll explain it briefly here.

A neural network with m real-valued inputs and 1 real valued outputs defines a function from R^m to R. Thus, a distribution over the weights and biases of such a neural network -- such as the standard Gaussian initialization -- implicitly defines a distribution over functions on R^m. Neal's paper shows that, for fully-connected neural networks, this distribution limits to a Gaussian process.

What is a Gaussian process? It's a distribution over functionsf with the property that, for any finite collection of points $X_1, \ldots, X_N$, the valuesf$(X_1), \ldots,$ f(X_N) have a joint distribution which is a multivariate Gaussian. Any Gaussian process is uniquely

defined by its mean and covariance functions, $\mu(x)$ and $C(x, x')$. For points $X_1, \ldots, X_N$, the distribution of $f(X_1), \ldots, f(X_N)$ will have mean $\mu(X_1), \ldots, \mu(X_N)$ with covariance matrix $C_{ij} = C(X_i, X_j)$.

The argument that fully-connected neural networks limit to Gaussian processes in the infinite-width limit is pretty simple. Consider a three-layer neural network, with an activation function σ in the second layer and a single linear output unit. This network can be defined by the equation $y = \sum V^k \sigma(\sum W^{kj} X^j)$. At initialization, V and W are filled with independent Gaussians, with variance of V scaled as the inverse square-root of the number of hidden-units.

Each hidden unit h^k will has a value for each of the inputs X_i , $h^k(X_i) = \sigma(\sum_j W^{kj} X_i^j)$.

Since W is random, for each k, $h^k(\vec{X})$ is an independent random vector(where we write $\vec{X}$ for $X_1, \ldots, X_N$). All of these random vectors follow the same distribution, and the output $\vec{y} = f(\vec{X})$ of the network is simply the sum of these identical distributions multiplied by the univariate Gaussians V^k. By the multidimensional central limit theorem, this sum will tend to a multidimensional Gaussian.

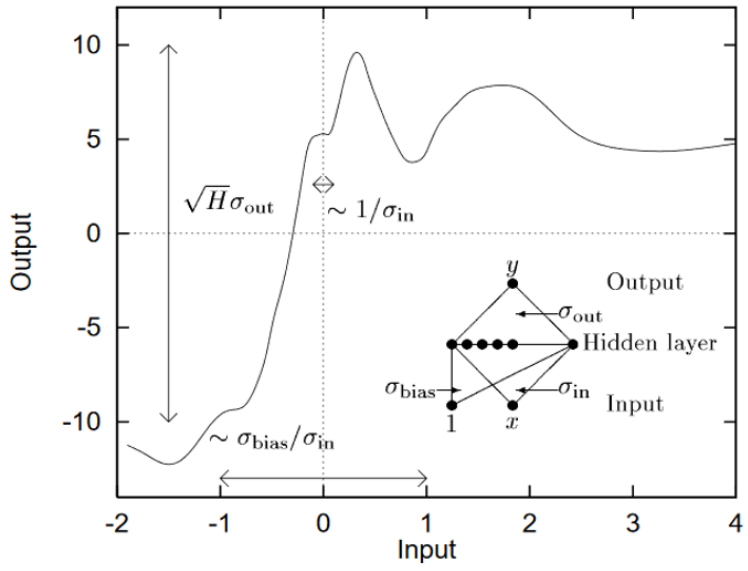


Image: a sample from a GP defined by a neural net. From McKay(1995).

Extending this argument to multiple hidden layers is also pretty easy. By induction, the pre-activations of each unit in hidden layer J have identical Gaussian process distributions, which induces identical(non-Gaussian,generically) joint distributions on the activations. The pre-activations of layer J+1 are the sum of these activations multiplied by univariate Gaussians, hence the central limit theorem can be applied again to show that these pre-activations have a joint Gaussian distribution for any set of inputs, hence they have a Gaussian process distribution. This inductive process can be used to compute the mean $\mu(\vec{X})$ and covariance $C(\vec{X}, \vec{X})$ of the output to an arbitrary depth for a given set of inputs $\vec{X}$. For many activation functions including ReLUs, this computation can be done exactly, giving an explicit expression for the distribution over outputs at initialization.

More recently, this behavior was proved to extend to CNNs, and then pretty much all classes of neural network architecture currently used. In convolutional neural nets, the infinite 'width' limit is taken with respect to the number of filters.

Infinite Width: Training

Okay, so we can understand how neural nets behave at initialization in this limit. But we don't care that much about initialization -- what really matters is what function it represents after the training process is over. The training process is over a complex, non-linear space, and seems much less tractable to the kind of analysis used at initialization. Surprisingly, however, there is a similar simplification that occurs when we pass to the infinite-width limit. For those of you who know some machine learning, it turns out that, in this limit, neural networks behave as a kind of kernel machine, using the so-called neural tangent kernel(NTK).

In this case, the derivation of the infinite-width behavior is more complex, so I'll just explain what that behavior *is*. The key is to consider the effect of the training process on the values $f(X_1), \ldots, f(X_N)$ at the points to be classified, rather than the weights of the network. Consider two inputs X_1, X_2. Imagine taking a step of gradient descent on the network weights θ to adjust the network output $f(X_1)$. What will the effect of this be on $f(X_2)$? To first order, a given change in the weights $\Delta\theta$ will change $f(X_2)$ by $< \nabla_\theta f(X_2), \Delta\theta >$. Taking a gradient step in the direction of $f(X_1)$ will cause a change in the weights of $\nabla_\theta f(X_1)$. Therefore, taking a step of gradient descent at $f(X_1)$ will have the effect of changing $f(X_2)$ by $< \nabla_\theta f(X_1), \nabla_\theta f(X_2) >$ We can construct an $N \times N$ matrix K_θ with entries $K_{\theta,i,j} =< \nabla_\theta f(X_i), \nabla_\theta f(X_j) >$ Taking a step of full-batch gradient descent in the direction $\vec{\Delta}$ (indexed along $\vec{X}$) will, to first order, effect a change in the outputs of $K_\theta \vec{\Delta}$.

Of course, this doesn't really simplify things much, as the matrix K_θ is itself dependent on the weights, which vary both randomly at initialization and during training. The insight of the NTK paper is that in the infinite-width limit, this dependence disappears. For infinite-width networks:

i) at initialization, K_θ becomes a deterministic matrix K_∞

ii) during training, K_∞ doesn't change. (The *weights* still change during training, but their change is small enough that K_∞ is unaffected)

Therefore, training on a set of N inputs can be perfectly simulated by just calculating K_∞ for those inputs, then using K_∞ to iterate the training(in practice, the *end result* of training is instead calculated directly, which can be done by inverting K_∞) An inductive formula for calculating K_∞ is given in the NTK paper.

Another way of thinking about the NTK is that it is essentially equivalent to taking the first-order Taylor expansion of a neural network about its initial parameters. In this regime, the response of the output to changes in the parameters is linear(though the output is *not* linear in the network input!) Then the above papers prove that, in the infinite-width limit, the training trajectory stays close to that of its Taylor expansion.

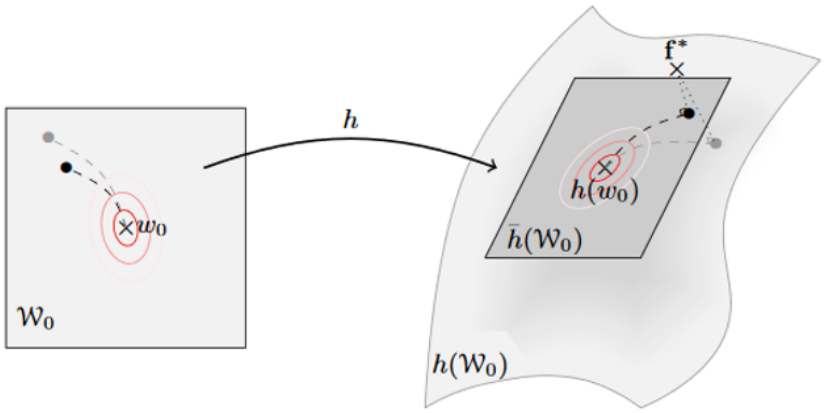


Image: The network's training trajectory stays close to that of its linearization $\bar{h}$. From Chizat&Bach(2018).

The NTK was originally defined for simply-connected models, but was later extended to convolutional nets, and now pretty much all network architectures. (As a historical note, many of the ideas behind the NTK were discovered before the paper coining the term NTK, check out this paper for instance)

For those of you wanting to attain a deeper understanding, the original NTK paper is a pretty clear read, as is this blog post.

Significance and Limitations of Infinite-Width Limit

So what's the upshot of all this? Does studying the infinite-width limit tell us anything about the success of finite neural networks? I'd argue that it does. Several of the papers above include comparisons between the output of finite-width networks and the analytically-computed predictions of the associated Gaussian processes and neural tangent kernel. Agreement was often pretty close:

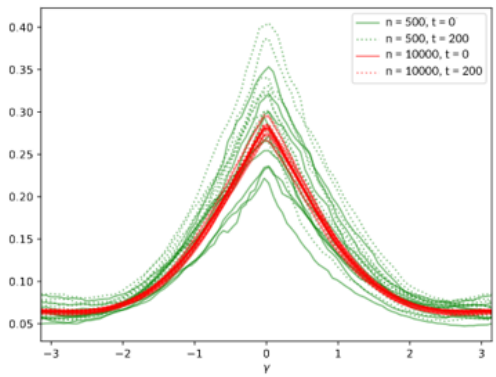


Figure 1: Convergence of the NTK to a fixed limit for two widths n and two times t.

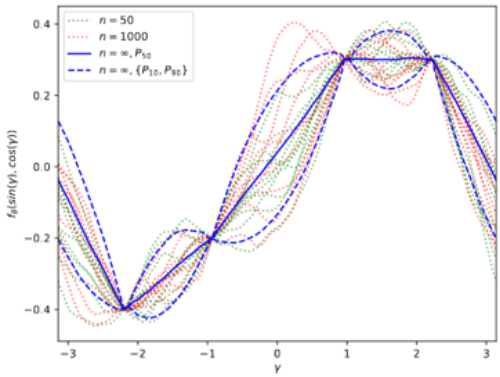


Figure 2: Networks function f_θ near convergence for two widths n and 10th, 50th and 90th percentiles of the asymptotic Gaussian distribution.

Image: convergence of NTK for one-dimensional input space. From the NTK paper.

Moreover, the performance of the NTK-based methods on learning tasks was impressive. This paper used the kernel associated with deep CNN to classify CIFAR-10 images, achieving 77% accuracy, a new record for kernel-based methods. This is only 6% lower than the performance of the original network. The kernel-like behavior of neural networks may not account for all of their good performance, but it seems to explain at least some of it.

Ultimately, the point of relating neural networks to kernel methods is that kernel methods are much simpler. Kernel methods are a sort of generalization of linear models, in which inputs are projected into a higher-dimensional space where they can be linearly separated. Kernels are tractable to mathematical analysis. It's possible to prove that kernel methods will always converge to a global minimum (on the training points) under gradient descent, and thus prove that neural networks will always converge to a minimum when they have enough hidden units. Another mathematical tool for analyzing kernels is their eigen-decomposition: see for instance this paper which finds that the NTK is diagonalized in the Fourier basis on the binary cube. They then use the eigenvalue associated to various functions as a measure of complexity, finding that it correlates well with the generalization performance of the neural network when learning that function.

Despite this, there are limitations to kernel-based analysis. A given NTK will usually underperform its associated neural network, and as far as I know nobody has even tried to apply NTK methods to problems such as ImageNet. (mostly due to computational costs, as using the NTK for regression scales like N^3 in number of data

points). There are theoretical works that suggest that there exist problems solvable by neural networks which no kernel-based method can solve. See also this paper on the limits of the 'lazy regime', their term for training regimes in which classifiers are approximately linear in their parameters(which includes the infinite-width limit).

Generalization Theory

The works above on the infinite-width limit explain, to some extent, the success of SGD at optimizing neural nets, because of the approximately linear nature of their parameter-space. A remaining piece of the puzzle is generalization, explaining why the global minimum found on the training set will tend to work well on new data points.

Traditionally, statistical learning theory has focused on classes of models where there number of potential functions learnable by that class is small. However, neural networks are usually capable of fitting arbitrary functions of their dataset, so many tools used to prove that models have low generalization error have failed: the bounds they give are vacuous, meaning that they can't certify that the model will perform better than random guessing. This issue was popularized in a 2017 paper by Zhang et al.

Despite this, recently some non-vacuous generalization bounds have been proven. Thus far, the only non-vacuous bounds for 'real' datasets such as MNIST have used PAC-Bayes methods. These methods replace an individual neural net with a learned distribution over network parameters, and introduce a fixed prior over the parameters. The generalization error is bounded by (the square root of) the KL divergence between the prior distribution and the learned distribution. Intuitively: a low KL divergence means the learned distribution has a short description length w.r.t. the prior, and there are only so many such distributions(sort of), so one of them matching the training inputs would be unlikely unless it truly captured part of the underlying function. PAC-Bayes bounds cannot guarantee high performance off the training set, but they can provably bound the error with high probability, assuming that the training data has been fairly sampled from the underlying distribution.

The first work to use PAC-Bayes bounds for modern neural networks was written by Dziugaite&Roy. They were able to prove non-vacuous bounds on a binarized version of MNIST -- not as trivial as it sounds, given that the classifying networks had hundreds of thousand of parameters. Taking as their prior a Gaussian distribution centered at initialization, the authors represented the learned network with another Gaussian distribution whose parameters they optimized with SGD to minimize the PAC-Bayes bound on total error. This work was inspired by the notion of flat minima, which is the idea that gradient descent is biased toward wide minima in parameter space, where perturbing the parameters does not affect the loss much. From a minimum description length principle, flat minima can be described using fewer bits because of their width, which should imply that solutions found by SGD have good generalization performance. The 'nonvacuous bounds' paper used a formalism inspired by this notion to derive provable generalization bounds.

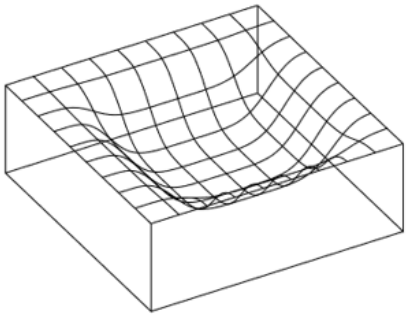


Figure 1: *Example of a "flat" minimum.*

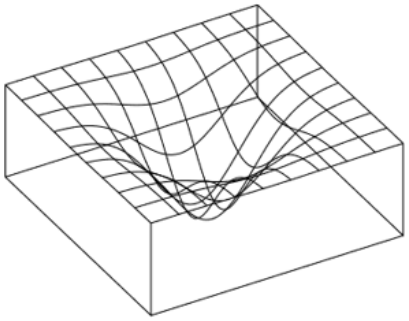


Figure 2: *Example of a "sharp" minimum.*

Image: flat vs. sharp minima. From Hochreiter&Schmidhuber(1996).

A later paper made the connection between PAC-Bayes bounds and compression more explicit. They used techniques for compressing the parameters of a neural network to store networks solving full MNIST and ImageNet using far fewer bits than their original size. Using a PAC-Bayes prior over code-words, they were able to provably verify at least 54% accuracy on MNIST and at least 3.5% accuracy on ImageNet(non-trivial given the huge number of classes in ImageNet). More recently still[2], an approach using random projections proved a bound of 85% accuracy for MNIST. Random projections constrain the network parameters to lie in a random low-dimensional subspace. That training is still possible under such a regime indicates that the exact direction chosen by the network is not too important: it's still likely possible to find a good minimum no matter which way it goes, so long as it has enough wiggle room. This further decreases the description length of the model and hence provides a way of obtaining generalization bounds.

Deriving bounds on the generalization error might seem pointless when it's easy to do this by just holding out a validation set. I think the main value is in providing a test of purported theories: your 'explanation' for why neural networks generalize ought to be able to produce non-trivial bounds on their generalization error.

Relevance for Alignment

At this stage, theoretical research on neural networks is not yet directly useful for alignment. Its goal is more conceptual clarity than producing tools that would be useful for practitioners, or even theoretical insights that are directly relevant to alignment-type issues.

In the long run, though, I believe that this sort of research could be crucial for creating aligned AI. It seems plausible that neural networks will be used to build AGI, or be a major component of AGI. If that happens, deeply understanding the implicit bias and optimization properties of these networks will be extremely important for a variety of purposes from choosing the class of models to enabling ongoing monitoring of what they have learned. This sort of theoretical understanding will likely be essentially in the implementation of alignment schemes such as IDA, and could enable more powerful versions of existing transparency and robustness methods.

But even if you think, as MIRI does(?), that neural networks are ultimately too insecure to build aligned AI, I believe trying to understand neural networks is still a worthwhile goal. Neural networks are one of the only techniques we have with anything approaching the generality needed for AGI. If alignment researchers want to build a 'secure version' of neural networks, then it seems necessary to first understand what factors contribute to their strong performance. Then it may be possible to isolate those factors in a more secure and transparent class of models. In contrast, attempting to derive such a class of models from pure thought, or experiments isolated from the mainstream of ML, seems much more difficult. Almost no AI techniques people think of work very well, so the existence of one that does seem to work well on a variety of realistic problems is a powerful and hard-won clue.

The upshot of this for how people interested in alignment should spend their time and money isn't as clear. This seems like an area that academia and industry is already pretty interested in and successful at studying. At the same time, I think there is still a huge amount of work to do and lots of stuff we don't understand, so I could imagine marginal researchers being useful. At the very least, I think it would be a good thing if alignment researchers were aware that advances in the theory of neural networks are happening, and kept tabs on new developments.

Footnotes

[1]: Technical note: Taking the limit of all layers to infinity is ambiguous; do you take the first layer to infinity, then the second, etc., or do you take them all to infinity at once? It turns out you get the same answer either way, so I'll just present as if taking the limits sequentially.

[2]: Disclaimer: I am the author of this paper.

Polio and the controversy over randomized clinical trials

This is a linkpost for https://rootsofprogress.org/polio-and-the-controversy-over-randomized-clinical-trials

I'm currently reading *Polio: An American Story*, by David Oshinsky, and I came across a fascinating story:

In 1954, when it was time for large-scale human trials of the first polio vaccine, some researchers were *against* the idea of doing a properly randomized, double-blind, placebo-controlled clinical trial—including Jonas Salk, the inventor of the vaccine.

What did they want to do instead? An "observed control" trial: They would ask for volunteers (children) to get the vaccine, and then compare the rate of polio in the volunteers to the rate in their schoolmates who weren't vaccinated. No placebo. No randomization. Not blind.

Of course, this was hopelessly confounded. In that era, the families most likely to volunteer were the more educated and affluent families (and those were actually the ones *most* at risk for the disease).

So why did Salk and others oppose proper randomized blind controls? The argument against a randomized trial was the urgency of protecting the nation's children against a debilitating and deadly disease. If the vaccine worked, it would be a tragedy to withhold it from the control children. Quoting Oshinsky:

> There were ethical issues as well. Were injected controls really suited to a polio trial? Was it proper, in short, to deny someone access to a potentially lifesaving vaccine in the name of statistical accuracy? Thousands of parents were going to volunteer their children to receive an injection—all of them hoping it contained the polio vaccine, not the placebo. Yet one-half of this study composed of six- to nine-year-olds, the group most vulnerable to paralytic polio, would receive a *worthless* liquid. Some, including Salk himself, saw this as elite science at its worst, a cynical form of Russian roulette.

Some context: the trial was massive, involving hundreds of thousands (eventually over a million) children across the country. This is not n=50 we're talking about here. And the disease was seasonal, striking in epidemic waves every summer. So even if all the controls were properly vaccinated at the end of the trial, it would be too late for anyone who had been stricken that year.

So there was a real dilemma here. Salk himself seemed to be already convinced that the vaccine worked, and wanted it to be administered as widely as possible:

> … Salk refused to budge. There must be no placebo. He could not deny his own product to those who volunteered to receive it. If thousands of children were going to be injected, then every one of them deserved the benefit of his vaccine. The object of these trials should be to protect as many lives as possible, not to run a textbook experiment. Given the stakes, Salk wrote O'Connor, "I would feel that every child who [gets] a placebo and becomes paralyzed will do so at my hands. I know this truthfully is not the case, but I know equally well that if the same child

> were to receive a vaccine that proved to be effective, then he might have been spared." It was enough, he said, "to make the humanitarian shudder [and] Hippocrates turn over in his grave."

Ultimately, some people resigned over the issue, including Harry Weaver, the director of research for the National Foundation for Infantile Paralysis, which was sponsoring the trials; and Joseph Bell, the scientific director who was to run them. The Foundation replaced Bell with Thomas Francis, Salk's old mentor—but Francis *also* insisted on a proper blind placebo control. Salk ended up going along with it (perhaps because he had to at that point, perhaps because he trusted Francis more than Bell?)

In the end they did a combination: Some counties did a placebo control, others an "observed control", at their own discretion. Fortunately, there were enough randomized controls to draw sound scientific conclusions at the end of the study.

While I'm sympathetic to the practical issue of wanting to protect patients, I think the history of medicine shows how easy it is for scientific "knowledge" to become polluted with falsehoods based on less-than-perfect experiments. In medicine especially, these mistakes become tradition, entrenched "wisdom" from revered authority figures that can stand undefeated as common practice for decades or centuries. So it's important to get things right, and I think Bell and Francis were clearly correct here.

Mostly I just find it fascinating that as late as the 1950s, the need for proper randomized blind placebo controls in clinical trials was not universally accepted, even among scientific researchers. Cultural norms matter, especially epistemic norms.

Is Causality in the Map or the Territory?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

steve2152 brought up a great example:

> Consider a 1kΩ resistor, in two circuits. The first circuit is the resistor attached to a 1V [voltage] supply. Here an engineer would say: "The supply creates a 1V drop across the resistor; and that voltage drop causes a 1mA current to flow through the resistor." The second circuit is the resistor attached to a 1mA current source. Here an engineer would say: "The current source pushes a 1mA current through the resistor; and that current causes a 1V drop across the resistor." Well, it's the same resistor ... does a voltage across a resistor cause a current, or does a current through a resistor cause a voltage, or both, or neither? [...] my conclusion was that people think about causality in a way that is not rooted in physics, and indeed if you forced someone to exclusively use physics-based causal models, you would be handicapping them.

First things first: we're talking about causality, which means we're mainly talking about counterfactuals - questions of the form "what would the system do if we did X?". (See Pearl's book for lots of detail on how and why causality and counterfactuals go together.)

In the resistor example, both scenarios yield exactly the same actual behavior (assuming we've set the parameters appropriately), but the counterfactual behavior differs - and that's exactly what defines a causal model. In this case, the counterfactuals are things like "what if we inserted a different resistor?" and "what if we adjusted the knob on the supply?". If it's a voltage supply, then a voltage -> current model ("voltage causes current") correctly answers the counterfactuals:

- Inserting a different resistor changes the current but not the voltage. In the voltage -> current model, we cut the arrow going into "current" and set that node to a new value.
- Adjusting the knob on the supply changes the voltage, and the current adjusts to match. In the voltage -> current model, we set the "voltage" node to a new value, and the model tells us how to update the "current" node.

Conversely, if it's a current supply, then a current -> voltage model ("current causes voltage") correctly answers the counterfactuals. It is a mistake here to think of "the territory" as just the resistor by itself; the supply is a critical determinant of the counterfactual behavior, so it needs to be included in order to talk about causality.

Note that all the counterfactual queries in this example are physically grounded - they are properties of the territory, not the map. We can actually go swap the resistor in a circuit and see what happens.

But Which Counterfactuals?

Of course, there's still the question of how we decide which counterfactuals to support. That *is* mainly a property of the map, so far as I can tell, but there's a big catch: some sets of counterfactual queries will require keeping around far less information than others. A given territory supports "natural" classes of counterfactual queries, which require relatively little information to yield accurate predictions for the whole query class. In this context, the lumped circuit abstraction is one such example: we keep around just high-level summaries of the electrical properties of each component, and we can answer a whole class of queries about voltage or current measurements. Conversely, if we wanted to support a few queries about the readings from a voltage probe, a few queries about the mass of various circuit components, and a few queries about the number of protons in a wire mod 3... these all require completely different information to answer. It's not a natural class of queries.

So natural classes of queries imply natural choices of abstract model, possibly including natural choices of causal model. There will still be some choice in which queries we care about, and what information is actually available will play a role in that choice (i.e. even if we cared about number of protons mod 3, we have no way to get that information).

In our example above, the voltage -> current model is such a natural abstraction when we have a voltage supply. Using just the basic parameters of the system (i.e. resistance and the present system state), it allows us to accurately answer questions about *both* the system's current state *and* counterfactual changes. Same for the current -> voltage model when using a current supply.

But... although a voltage -> current model *is* a natural abstraction when we're using a voltage supply, it's not clear that the current -> voltage model is *not*. It won't correctly answer counterfactuals about swapping out a resistor or adjusting the knob on the supply, but perhaps there is some other class of counterfactual queries which would be correctly answered by the current -> voltage model? (One class of counterfactual queries which it would correctly answer is "swap the voltage supply for a current supply, and then...". But that class of queries just reinforces the idea that voltage -> current is a natural abstraction for a voltage supply, and current -> voltage is not.)

"You can't possibly succeed without [My Pet Issue]"

There's a particular conversational move that I've noticed people making over the past couple years. I've also noticed *myself* making it. The move goes:

> "You can't *possibly* succeed without X", where X is whatever principle the person is arguing for.

(Where "succeed" means "have a functioning rationality community / have a functioning organization / solve friendly AI / etc")

This is not always false. But, I am pretty suspicious of this move.

(I've seen it from people from a variety of worldviews. This is not a dig on any one particular faction from local-politics. And again, I do this myself).

When *I* do the move, my current introspective TAP goes something like: "Hmm. Okay, is this actually true? Is it *impossible* to succeed without my pet-issue-of-the-day? Upon reflection, obviously not. I legit think it's *harder.* There's a reason I started caring about my pet-issue in the first place. But 'impossible' is a word that was clearly generated by my political rationalization mind. How *much* harder is it, exactly? Why do I believe that?")

In general, there are incentives (and cognitive biases) to exaggerate the importance of your plans. I think this is partly for political reasons, and partly for motivational reasons – it's hard to get excited enough about your *own* plans if you don't believe they'll have outsized effects. (A smaller version of this, common on my web development team, is someone saying "if we just implemented Feature X we're get a 20% improvement on Metric Y", and the actual answer was we got, like, a 2% improvement, and it was worth it. But, like, the 20% figure was clearly ridiculous).

"It's impossible" is an easier yellow-flag to notice than "my numbers are bigger than what other people think are reasonable". But in both cases, I think it's a useful thing to train yourself to notice, and I think "try to build an explicit quantitative model" is a good immune response. Sometimes the thing *is* actually impossible, and your model checks out. But I'm willing to bet if you're bringing this up in a social context where you think an abstract principle is at stake, it's probably wrong.

[AN #77]: Double descent: a unification of statistical theory and modern ML practice

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Audio version here (may not be up yet).

Highlights

Deep Double Descent *(Preetum Nakkiran et al)* (summarized by Rohin): This blog post provides empirical evidence for the existence of the *double descent* phenomenon, proposed in an earlier paper summarized below. Define the *effective model complexity* (EMC) of a training procedure and a dataset to be the maximum size of training set such that the training procedure achieves a *train* error of at most ε (they use ε = 0.1). Let's suppose you start with a small, underparameterized model with low EMC. Then initially, as you increase the EMC, the model will achieve a better fit to the data, leading to lower test error. However, once the EMC is approximately equal to the size of the actual training set, then the model can "just barely" fit the training set, and the test error can increase or decrease. Finally, as you increase the EMC even further, so that the training procedure can easily fit the training set, the test error will once again *decrease*, causing a second descent in test error. This unifies the perspectives of statistics, where larger models are predicted to overfit, leading to increasing test error with higher EMC, and modern machine learning, where the common empirical wisdom is to make models as big as possible and test error will continue decreasing.

They show that this pattern arises in a variety of simple settings. As you increase the width of a ResNet up to 64, you can observe double descent in the final test error of the trained model. In addition, if you fix a large overparameterized model and change the number of epochs for which it is trained, you see another double descent curve, which means that simply training longer can actually *correct overfitting*. Finally, if you fix a training procedure and change the size of the dataset, you can see a double descent curve as the size of the dataset decreases. This actually implies that there are points in which *more data is worse*, because the training procedure is in the critical interpolation region where test error can increase. Note that most of these results only occur when there is *label noise* present, that is, some proportion of the training set (usually 10-20%) is given random incorrect labels. Some results still occur without label noise, but the resulting double descent peak is quite small. The authors hypothesize that label noise leads to the effect because double descent occurs when the model is misspecified, though it is not clear to me what it means for a model to be misspecified in this context.

Rohin's opinion: While I previously didn't think that double descent was a real phenomenon (see summaries later in this email for details), these experiments

convinced me that I was wrong and in fact there is something real going on. Note that the settings studied in this work are still not fully representative of typical use of neural nets today; the label noise is the most obvious difference, but also e.g. ResNets are usually trained with higher widths than studied in this paper. So the phenomenon might not generalize to neural nets as used in practice, but nonetheless, there's *some* real phenomenon here, which flies in the face of all of my intuitions.

The authors don't really suggest an explanation; the closest they come is speculating that at the interpolation threshold there's only ~one model that can fit the data, which may be overfit, but then as you increase further the training procedure can "choose" from the various models that all fit the data, and that "choice" leads to better generalization. But this doesn't make sense to me, because whatever is being used to "choose" the better model applies throughout training, and so even at the interpolation threshold the model should have been selected throughout training to be the type of model that generalized well. (For example, if you think that regularization is providing a simplicity bias that leads to better generalization, the regularization should also help models at the interpolation threshold, since you always regularize throughout training.)

Perhaps one explanation could be that in order for the regularization to work, there needs to be a "direction" in the space of model parameters that doesn't lead to increased training error, so that the model can move along that direction towards a simpler model. Each training data point defines a particular direction in which training error will increase. So, when the number of training points is equal to the number of parameters, the training points just barely cover all of the directions, and then as you increase the number of parameters further, that starts creating new directions that are not constrained by the training points, allowing the regularization to work much better. (In fact, the original paper, summarized below, *defined* the interpolation threshold as the point where number of parameters equals the size of the training dataset.) However, while this could explain model-wise double descent and training-set-size double descent, it's not a great explanation for epoch-wise double descent.

Read more: Paper: Deep Double Descent: Where Bigger Models and More Data Hurt

Technical AI alignment

Problems

Comment on Coherence arguments do not imply goal directed behavior *(Ronny Fernandez)* (summarized by Rohin): I have argued (AN #35) that coherence arguments that argue for modeling rational behavior as expected utility maximization do not add anything to AI risk arguments. This post argues that there is a different way in which to interpret these arguments: we should only model a system to be an EU maximizer if it was the result of an optimization process, such that the EU maximizer model is the best model we have of the system. In this case, the best way to predict the agent is to imagine what we would do if we had its goals, which leads to the standard convergent instrumental subgoals.

Rohin's opinion: This version of the argument seems to be more a statement about our epistemic state than about actual AI risk. For example, I know many people without technical expertise who anthropomorphize their laptops as though they were

pursuing some goal, but they don't (and shouldn't) worry that their laptops are going to take over the world. More details in this comment.

AI strategy and policy

How does the offense-defense balance scale? *(Ben Garfinkel et al)* (summarized by Flo): The offense-defense balance that characterises how easy it is to successfully attack others can affect what kinds of conflicts break out and how often that happens. This paper analyses how growing capabilities on both sides affect that balance. For example, consider an idealized model of cyber defense with a fixed set of vulnerabilities that are discovered independently by attackers and defenders. The attacker will initially be able to use almost all of the vulnerabilities they found. This is because, with only a small percentage of vulnerabilities discovered by both sides, the defender is unlikely to have found the same ones as the attacker. Marginal increases of the defender's capabilities are unlikely to uncover vulnerabilities used by the attacker in this regime, such that attacks become easier as both sides invest resources. Once most vulnerabilities have been found by both sides, this effect reverses as marginal investments by the attacker become unlikely to uncover vulnerabilities the defender has not fixed yet.

This pattern, where increasingly growing capabilities first favour offense but lead to defensive stability in the long run, dubbed **OD-scaling** seems to be common and can be expected to be found whenever there are **multiple attack vectors**, the attacker only needs to break through on some of them and the defender enjoys **local defense superiority**, meaning that with sufficient coverage by the defender for a given attack vector, it is almost impossible for the attacker to break through.

Because the use of digital and AI systems can be scaled up quickly, scale-dependent shifts of the offense-defense balance are going to increase in importance as these systems become ubiquitous.

Flo's opinion: I found it quite surprising that the paper mentions a lack of academic consensus about whether or not offensive advantage is destabilizing. Assuming that it is, OD-scaling might provide a silver lining concerning cybersecurity, provided things can be scaled up sufficiently. These kinds of dynamics also seem to put a natural ceiling on arms races: above a certain threshold, gains in capabilities provide advantage to both sides such that resources are better invested elsewhere.

Other progress in AI

Deep learning

Reconciling modern machine learning practice and the bias-variance trade-off *(Mikhail Belkin et al)* (summarized by Rohin): This paper first proposed double descent as a general phenomenon, and demonstrated it in three machine learning models: linear predictors over random Fourier features, fully connected neural networks with one hidden layer, and forests of decision trees. Note that they define the interpolation threshold as the point where the number of parameters equals the number of training points, rather than using something like effective model complexiy.

For linear predictors over random Fourier features, their procedure is as follows: they generate a set of random features, and then find the linear predictor that minimizes the squared loss incurred. If there are multiple predictors that achieve zero squared loss, then they choose the one with the minimum L2 norm. The double descent curve for a subset of MNIST is very pronounced and has a huge peak at the point where the number of features equals the number of training points.

For the fully connected neural networks on MNIST, they make a significant change to normal training: prior to the interpolation threshold, rather than training the networks from scratch, they train them from the final solution found for the previous (smaller) network, but after the interpolation threshold they train from scratch as normal. With this change, you see a very pronounced and clear double descent curve. However, if you always train from scratch, then it's less clear -- there's a small peak, which the authors describe as "clearly discernible", but to me it looks like it could be noise.

For decision trees, if the dataset has n training points, they learn decision trees of size up to n leaves, and then at that point (the interpolation threshold) they switch to having ensembles of decision trees (called forests) to get more expressive function classes. Once again, you can see a clear, pronounced double descent curve.

Rohin's opinion: I read this paper back when summarizing Are Deep Neural Networks Dramatically Overfitted? (AN #53) and found it uncompelling, and I'm really curious how the ML community correctly seized upon this idea as deserving of further investigation while I incorrectly dismissed it. None of the experimental results in this paper are particularly surprising to me, whereas double descent itself is quite surprising.

In the random Fourier features and decision trees experiments, there is a qualitative difference in the *learning algorithm* before and after the interpolation threshold, that suffices to explain the curve. With the random Fourier features, we only start regularizing the model after the interpolation threshold; it is not surprising that adding regularization helps reduce test loss. With the decision trees, after the interpolation threshold, we start using ensembles; it is again not at all surprising that ensembles help reduce test error. (See also this comment.) So yeah, if you start regularizing (via L2 norm or ensembles) after the interpolation threshold, that will help your test error, but in practice we regularize throughout the training process, so this should not occur with neural nets.

The neural net experiments also have a similar flavor -- the nets before the interpolation threshold are required to reuse weights from the previous run, while the ones after the interpolation threshold do not have any such requirement. When this is removed, the results are much more muted. The authors claim that this is necessary to have clear graphs (where training risk monotonically decreases), but it's almost certainly biasing the results -- at the interpolation threshold, with weight reuse, the test squared loss is ~0.55 and test accuracy is ~80%, while without weight reuse, test squared loss is ~0.35 and test accuracy is ~85%, a *massive* difference and probably not within the error bars.

Some speculation on what's happening here: neural net losses are nonconvex and training can get stuck in local optima. A pretty good way to get stuck in a local optimum is to initialize half your parameters to do something that does quite well while the other half are initialized randomly. So with weight reuse we might expect getting stuck in worse local optima. However, it looks like the training losses are comparable between the methods. Maybe what's happening is that with weight reuse,

the half of parameters that are initialized randomly memorize the training points that the good half of the parameters can't predict, which doesn't generalize well but does get low training error. Meanwhile, without weight reuse, all of the parameters end up finding a good model that does generalize well, for whatever reason it is that neural nets do work well.

But again, note that the authors were right about double descent being a real phenomenon, while I was wrong, so take all this speculation with many grains of salt.

More Data Can Hurt for Linear Regression: Sample-wise Double Descent *(Preetum Nakkiran)* (summarized by Rohin): This paper demonstrates the presence of double descent (in the size of the dataset) for *unregularized linear regression*. In particular, we assume that each data point x is a vector in independent samples from Normal(0, σ^2), and the output is y = βx + ε. Given a dataset of (x, y) pairs, we would like to estimate the unknown β, under the mean squared error loss, with no regularization.

In this setting, when the dimensionality d of the space (and thus number of parameters in β) is equal to the number of training points n, the training data points are linearly independent almost always / with probability 1, and so there will be exactly one β that solves the n linearly independent equalities of the form βx = y. However, such a β must also be fitting the noise variables ε, which means that it could be drastically overfitted, with very high norm. For example, imagine β = [1, 1], so that y = x1 + x2 + ε, and in our dataset x = (-1, 3) is mapped to y = 3 (i.e. an ε of +1), and x = (0, 1) is mapped to y = 0 (i.e. an ε of -1). Gradient descent will estimate that β = [-3, 0], which is going to generalize very poorly.

As we decrease the number of training points n, so that d > n, there are infinitely many settings of the d parameters of β that satisfy the n linearly independent equalities, and gradient descent naturally chooses the one with minimum norm (even without regularization). This limits how bad the test error can be. Similarly, as we increase the number of training points, so that d < n, there are too many constraints for β to satisfy, and so it ends up primarily modeling the signal rather than the noise, and so generalizing well.

Rohin's opinion: Basically what's happening here is that at the interpolation threshold, the model is forced to memorize noise, and it has only one way of doing so, which need not generalize well. However, past the interpolation threshold, when the model is overparameterized, there are *many* models that successfully memorize noise, and gradient descent "correctly" chooses one with minimum norm. This fits into the broader story being told in other papers that what's happening is that the data has noise and/or misspecification, and at the interpolation threshold it fits the noise in a way that doesn't generalize, and after the interpolation threshold it fits the noise in a way that does generalize. Here that's happening because gradient descent chooses the minimum norm estimator that fits the noise; perhaps something similar is happening with neural nets.

This explanation seems like it could explain double descent on model size and double descent on dataset size, but I don't see how it would explain double descent on training time. This would imply that gradient descent on neural nets first has to memorize noise in one particular way, and then further training "fixes" the weights to memorize noise in a different way that generalizes better. While I can't rule it out, this seems rather implausible to me. (Note that regularization is *not* such an explanation, because regularization applies throughout training, and doesn't "come into effect" after the interpolation threshold.)

Understanding “Deep Double Descent” *(Evan Hubinger)* (summarized by Rohin): This post explains deep double descent (in more detail than my summaries), and speculates on its relevance to AI safety. In particular, Evan believes that deep double descent shows that neural nets are providing strong inductive biases that are crucial to their performance -- even *after* getting to ~zero training loss, the inductive biases *continue* to do work for us, and find better models that lead to lower test loss. As a result, it seems quite important to understand the inductive biases that neural nets use, which seems particularly relevant for e.g. mesa optimization and pseudo alignment (AN #58).

Rohin's opinion: I certainly agree that neural nets have strong inductive biases that help with their generalization; a clear example of this is that neural nets can learn randomly labeled data (which can never generalize to the test set), but nonetheless when trained on correctly labeled data such nets do generalize to test data. Perhaps more surprising here is that the inductive biases help even *after* fully capturing the data (achieving zero training loss) -- you might have thought that the data would swamp the inductive biases. This might suggest that powerful AI systems will become simpler over time (assuming an inductive bias towards simplicity). However, this is happening in the regime where the neural nets are overparameterized, so it makes sense that inductive biases would still play a large role. I expect that in contrast, powerful AI systems will be severely underparameterized, simply because of *how much data* there is (for example, the largest GPT-2 model still underfits the data (AN #46)).

Uniform convergence may be unable to explain generalization in deep learning *(Vaishnavh Nagarajan)* (summarized by Rohin): This post argues that existing generalization bounds cannot explain the empirical success of neural networks at generalizing to the test set.

"What?", you say if you're like me, "didn't we already know this? Generalization bounds depend on your hypothesis space being sufficiently small, but neural nets can represent any reasonable function? And even if you avoid that by considering the size of the neural net, we know that empirically neural nets can learn randomly labeled data, which can never generalize; surely this means that you can't explain generalization without reference to some property of the dataset, which generalization bounds typically don't do?"

It turns out that the strategy has been to prove generalization bounds that depend on the *norm of the weights of the trained model* (for some norm that depends on the specific bound), which gets around both these objections, since the resulting bounds are independent of the number of parameters, and depend on the trained model (which itself depends on the dataset). However, when these bounds are evaluated on a simple sphere-separation task, they *increase* with the size of the training dataset, because the norms of the trained models increase.

Okay, but can we have a stronger argument than mere empirical results? Well, all of these bounds depend on a *uniform convergence bound*: a number that bounds the absolute difference between the train and test error for *any* model in your hypothesis space. (I assume the recent generalization bounds only consider the hypothesis space "neural nets with norms at most K", or some suitable overapproximation of that, and this is how they get a not-obviously-vacuous generalization bound that depends on weight norms. However, I haven't actually read those papers.)

However, no matter what hypothesis space these bounds choose, to get a valid generalization bound the hypothesis space must contain (nearly) all of the models that would occur by training the neural net on a dataset sampled from the underlying distribution. What if we had the actual smallest such hypothesis space, which only contained the models that resulted from an actual training run? The authors show that, at least on the sphere-separation task, the uniform convergence bound is still extremely weak. Let's suppose we have a training dataset S. Our goal is now to find a model in the hypothesis space which has a high absolute difference between actual test error, and error in classifying S. (Recall that uniform convergence requires you to bound the absolute difference for *all* models in your hypothesis class, not just the one trained on S.) The authors do so by creating an "adversarial" training dataset S' that also could have been sampled from the underlying distribution, and training a model on S'. This model empirically gets S almost completely wrong. Thus, this model has low test error, but high error in classifying S, which forces the uniform convergence bound to be very high.

Rohin's opinion: I enjoyed this blog post a lot (though it took some time to digest it, since I know very little about generalization bounds). It constrains the ways in which we can try to explain the empirical generalization of neural networks, which I for one would love to understand. Hopefully future work will explore new avenues for understanding generalization, and hit upon a more fruitful line of inquiry.

Read more: Paper

Understanding the generalization of ‘lottery tickets’ in neural networks *(Ari Morcos et al)* (summarized by Flo): The lottery ticket hypothesis (AN #52) states that a randomly initialized dense or convolutional neural network contains (sparse) subnetworks, called "winning tickets", which can be trained to achieve performance similar to the trained base network while requiring a lot less compute.

The blogpost summarizes facebook AI's recent investigations of the generalization of winning tickets and the generality of the hypothesis. Because winning tickets are hard to find, we would like to reuse the ones we have found for similar tasks. To test whether this works, the authors trained classifiers, pruned and reset them to obtain winning tickets on different image datasets and then trained these on other datasets. Winning tickets derived from similar datasets relevantly outperform random subnetworks after training and ones derived from larger or more complex datasets generalize better. For example, tickets from ImageNet are consistently among the best and tickets from CIFAR-100 generalize better than those from CIFAR-10.

Experiments in natural language processing and reinforcement learning suggest that the lottery ticket hypothesis is not just a peculiarity of image classification: for example, the performance of a large transformer model could be recovered from a winning ticket with just a third of the original weights, whereas random tickets with that amount of weights performed quite a bit worse. The analysis of simple shallow neural networks in a student-teacher setting is used as a toy model: when a larger student network is trained to mimic a smaller teacher with the same amount of layers, **student specialization** happens: some of the student's neurons learn to imitate single neurons of the teacher. This can be seen to happen more often and faster if the student neuron is already close to the teacher neuron at initialization. If the student network is large enough, every teacher neuron will be imitated by some student neuron and these student neurons collectively form a winning ticket.

Flo's opinion: I enjoyed reading this blogpost and like the idea of using winning tickets for transfer learning. I would have been quite surprised if they had found that the lottery ticket hypothesis was specific to image classification, as similar to pretraining, winning tickets seem to provide an inductive bias constraining the set of features that can be learnt during training to more useful ones. I do not think that further research into that direction will directly help with quickly training models for novel tasks unless the tickets can be identified very efficiently which seems like a harder optimization problem than just training a network by gradient descent.

Recent Progress in the Theory of Neural Networks *(interstice)*

News

AI Safety Camp Toronto (summarized by Rohin): The next AI safety camp (AN #10) will be held in early May, in Toronto. Apply here by Jan 5.

{Math} A times tables memory.

I have a distinct memory of being 8 years old, or so, and being handed one of those worksheets where they ask you to multiply numbers up through 12x12, and being viscerally disgusted by the implied pedagogy of it. That was *over a hundred things* you were asking me to memorize. On *my own time.* The *whole reason* I rush through my school work is so I don't have to do anything when I get home. I don't know if eight year old me swore, but this was definitely a "Screw you" moment for him.

But he actually ended up being able to do that sheet pretty quickly, at least compared to most of the rest of the class. There were a few kids who were faster than me, but I got the impression they were dumb enough to have to practice this instead of watching Ed, Edd 'n' Eddy at home. Or worse, they actually *did* memorize this stuff, instead of practice to get quick with the multiply-numbers-in-your-head algorithm like I did. (Because *of course* nobody else in the class would be doing it the same way I did, just *much faster.* But eight-year-olds aren't known to have particularly nuanced concepts of self that can gracefully accept that there are other people naturally much better than them at what they do best.)

Later on, we moved up to multiplying arbitrary two-digit-by-one-digit numbers, and then two-digit-*by*-*two*-*digit* numbers. (I didn't piece together how uncommon this was until a few years later.) Everyone who outpaced me in the times-tables speed tests were now far, far below me; meanwhile, I just had to chain my little "multiply-small-numbers" mental motion to a few "add-up-the-sums" motions. 76 * 89 = 7*8*100 + 6*8*10 + 7*9*10 + 6*9. I felt like I was *so* clever. I started to take pride in the fact that I was now leading the pack, even though I had told myself before that I didn't care!

That is, of course, until the kids who were originally faster than me *also* realized how to perform that mental motion, and then they leapt past me in speed with the combined force of split-second memory of times tables *and* a quick ability to perform algorithms.

I think by the time we were finished with the lightning round worksheet practice, I was in the bottom quarter of the class for speed, and when I did push myself to speed up, I'd start making careless mistakes like mixing up which one of 6*7 and 7*7 was 42 and which was 49, again?

Later in my mathematical pedagogy, I am taking a Real Analysis course. There are two midterms in this course. The first one I did not prepare for at all, falling into my old 8-year-old failure mode: "If I can't just compute the answer on the spot to the question, I sort of deserve to fail, don't I?" I got a B-, in the lower half of the class.

The second one, I reminded myself of the times tables kids. I got an A.

Making decisions under moral uncertainty

Cross-posted to the EA Forum. Updated substantially since initial publication.

Overview/purpose of this sequence

While working on an (upcoming) post about a new way to think about moral uncertainty, I unexpectedly discovered that, as best I could tell:

1. There was no single post on LessWrong or the EA Forum that very explicitly (e.g., with concrete examples) overviewed what seem to be the most prominent approaches to making decisions under moral uncertainty (more specifically, those covered in Will MacAskill's 2014 thesis).[1][2]

2. There was no (easily findable and explicit) write-up of how to handle simultaneous moral and empirical uncertainty. (What I'll propose is arguably quite obvious, but still seems worth writing up explicitly.)

3. There was no (easily findable and sufficiently thorough) write-up of applying sensitivity analysis and value of information analysis to situations of moral uncertainty.

I therefore decided to write a series of three posts, each of which addressed one of those apparent "gaps". My primary aim is to synthesise and make accessible various ideas that are currently mostly buried in the philosophical literature, but I also think it's plausible that some of the ideas in some of the posts (though not this first one) haven't been explicitly explored before.

I expect that these posts are most easily understood if read in order, but each post should also have value if read in isolation, especially for readers who are already familiar with key ideas from work on moral uncertainty.

Epistemic status (for the whole sequence)

I've now spent several days reading about moral uncertainty, but I wouldn't consider myself an actual expert in this topic or in philosophy more broadly. Thus, while I don't expect this sequence to contain any *major*, *central* mistakes, I wouldn't be surprised if it's inaccurate or unclear/misleading in some places.

I welcome feedback of all kinds (on these posts and in general!).

Moral uncertainty

We are often forced to make decisions under conditions of uncertainty. This uncertainty can be empirical (e.g., what is the likelihood that nuclear war would cause human extinction?) or moral (e.g., does the wellbeing of future generations matter morally?).[3][4] The issue of making decisions under empirical uncertainty has been well-studied,

and expected utility theory has emerged as the typical account of how a rational agent should proceed in these situations. The issue of making decisions under *moral* uncertainty appears to have received less attention (though see this list of relevant papers), despite also being of clear importance.

I'll later publish a post on definitions, types, and sources of moral uncertainty. In the present post, I'll instead aim to convey a sense of what moral uncertainty is through various examples. One example (which I'll return to repeatedly) is the following:

Devon's decision

> Suppose Devon assigns a 25% probability to T1, a version of hedonistic utilitarianism in which human "hedons" (a hypothetical unit of pleasure) are worth 10 times more than fish hedons. He also assigns a 75% probability to T2, a different version of hedonistic utilitarianism, which values human hedons just as much as T1 does, but doesn't value fish hedons at all (i.e., it sees fish experiences as having no moral significance). Suppose also that Devon is choosing whether to buy a fish curry or a tofu curry, and that he'd enjoy the fish curry about twice as much. (Finally, let's go out on a limb and assume Devon's humanity.)
>
> According to T1, the choice-worthiness (roughly speaking, the rightness or wrongness of an action) of buying the fish curry is -90 (because it's assumed to cause 1,000 negative fish hedons, valued as -100, but also 10 human hedons due to Devon's enjoyment).[5] In contrast, according to T2, the choice-worthiness of buying the fish curry is 10 (because this theory values Devon's joy as much as T1 does, but doesn't care about the fish's experiences). Meanwhile, the choice-worthiness of the tofu curry is 5 according to both theories (because it causes no harm to fish, and Devon would enjoy it half as much as he'd enjoy the fish curry).

The choice-worthiness of each option according to each theory is summarised in the following table:

	T1 - 25% credence	T2 - 75% credence
Fish curry	-90	10
Tofu curry	5	5

Given this information, what should Devon do?

"My Favourite Theory"

Multiple approaches to handling moral uncertainty have been proposed. The simplest option is the "My Favourite Theory" (MFT) approach, in which we essentially ignore our moral uncertainty, and just do whatever seems best based on the theory in which one has the highest "credence" (belief). In the above situation, MFT would suggest Devon should buy the fish curry, even though doing so is only somewhat better according to T2 (10 - 5 = 5), and is *far* worse (5 - -90 = 95) according to another theory in which he

has substantial (25%) credence. **Indeed, even if Devon had *49%* credence in T1 (vs 51% in T2), and the difference in the choice-worthiness of the options was *a thousand times* as large according to T1 as according to T2, MFT would still ignore the fact the situation is so much "higher stakes" for T1 than T2, refuse to engage in any "moral hedging", and advise Devon proceed with whatever T2 advised.**

On top of generating such counterintuitive results, MFT is subject to other quite damning objections (see pages 20-25 of Will MacAskill's 2014 thesis). Thus, the remainder of this post will focus on other approaches to moral uncertainty, which *do* allow for "moral hedging".

Types of moral theories

Which approach to moral uncertainty should be used depends in part on what types of moral theories are under consideration by the decision-maker - in particular, whether the theories are *cardinally measurable* or only *ordinally measurable*, and, if cardinally measurable, whether or not they're *inter-theoretically comparable*.[6]

Cardinality

Essentially, a theory is cardinally measurable if it can tell you not just which outcome is better than which, but also *by how much*. E.g., it can tell you not just that "X is better than Y which is better than Z", but also that "X is 10 'units' better than Y, which is 5 'units' better than Z". (Some readers may be more familiar with distinctions between ordinal, *interval*, and *ratio* scales; I'm almost certain "cardinal" scales include both interval and ratio scales.)

My understanding is that popular consequentialist theories are typically cardinal, while popular non-consequentialist theories are typically (or at least more often) ordinal. For example, a Kantian theory may simply tell you that lying is worse than not lying, but not by how much, so you cannot directly weigh that "bad" against the goodness/badness of other actions/outcomes (whereas such comparisons are relatively easy under most forms of utilitarianism).

Intertheoretic comparability

Even if a set of theories *are* cardinal, they still may not be *inter-theoretically comparable*. Roughly speaking, two theories are comparable if there's a consistent, non-arbitrary "exchange rate" between the theories' "units of choice-worthiness" (and they're non-comparable if there isn't). MacAskill explains the "problem of intertheoretic comparisons" as follows:

> "even when all theories under consideration give sense to the idea of magnitudes of choice-worthiness, we need to be able to compare these magnitudes of choice-worthiness across different theories. But it seems that we can't always do this. [... Sometimes we don't know] how can we compare the seriousness of the wrongs, according to these different theories[.] For which theory is there more at stake?"

In his own thesis, Tarsney provides useful examples:

> "Consider, for instance, hedonistic and preference utilitarianism, two straightforward maximizing consequentialist theories that agree on every feature of morality, except that hedonistic utilitarianism regards pleasure and pain as the sole non-derivative bearers of moral value while preference utilitarianism regards satisfied and dissatisfied preferences as the sole non-derivative bearers of moral value. Both theories, we may stipulate, have the same cardinal structure. But this structure does not answer the crucial question for expectational reasoning, how the value of a hedon according to hedonic utilitarianism compares to the value of a preference utile according to preference utilitarianism—that is, for an agent who divides her beliefs equally between the two theories and wishes to hedge when they conflict, how much hedonic experience does it take to offset the dissatisfaction of a preference of a given strength (or vice versa)?
>
> Likewise, of course, in trolley problem situations that pit consequentialist and deontological theories against one another, even if we could overcome the apparent structural incompatibility of these rival theories, the thorniest question seems to be: How many net lives must be saved, according to some particular version of consequentialism, to offset the wrongness of killing an innocent person, according to some particular version of deontology?" (line break added)[7]

It's worth noting that similar issues have received attention from, and are relevant to, other fields as well. For example, MacAskill writes: "A similar problem arises in the study of social welfare in economics: it is desirable to be able to compare the strength of preferences of different people, but even if you represent preferences by cardinally measurable utility functions you need more information to make them comparable." Thus, concepts and findings from those fields could illuminate this matter, and vice versa.

Three approaches

In MacAskill's thesis, the approaches to moral uncertainty he argues for are:

1. Maximising Expected Choice-worthiness (MEC), if all theories under consideration by the decision-maker are cardinal and intertheoretically comparable. (This is arguably the “best” situation to be in, as it is the case in which the most information is being provided by the theories.)
2. Variance Voting (VV), a form of what I’ll call “Normalised MEC”, if all theories under consideration are cardinal but *not* intertheoretically comparable.
3. The Borda Rule (BR), if all theories under consideration are ordinal. (This is the situation in which the *least* information is being provided by the theories.)
4. A “Hybrid” procedure, if the theories under consideration differ in whether they’re cardinal or ordinal and/or in whether they’re intertheoretically comparable. (Hybrid procedures will not be discussed in this post; interested readers can refer to pages 117-122 of MacAskill’s thesis.)

I will focus on these approaches (excluding Hybrid procedures), both because these approaches seem to me to be relatively prominent, effective, and intuitive, and because I know less about other approaches. (Potentially promising alternatives include a bargaining-theoretic approach [related presentation slides here], the similar but older and less fleshed-out parliamentary model, and the approaches discussed in Tarsney's thesis.)

Maximising Expected Choice-worthiness (MEC)

MEC is essentially an extension of expected utility theory. MacAskill describes MEC as follows:

> "when all [normative/moral] theories [under consideration by the decision-maker] are cardinally measurable and intertheoretically comparable, the appropriateness of an option is given by its expected choice-worthiness, where the expected choice-worthiness (EC) of an option is as follows:
>
> $$EC(A) = \sum_{i=1}^{n} C(T_i)CW_i(A)$$
>
> The appropriate options are those with the highest expected choice-worthiness."

In this formula, *C(Ti)* represents the decision-maker's credence (belief) in *Ti* (some particular moral theory), while *CWi(A)* represents the "choice-worthiness" (*CW*) of *A* (an "option" or action that the decision-maker can take), according to *Ti*.

To illustrate how MEC works, we will return to the example of Devon deciding whether to buy a fish curry or tofu curry, as summarised in the table of choice-worthiness values from earlier:

	T1 - 25% credence	T2 - 75% credence
Fish curry	-90	10
Tofu curry	5	5

(I've also modelled this example in Guesstimate. In that link, for comparison purposes, this model is followed by a model of the same basic example using traditional expected utility reasoning, and another using MEC-E (an approach I'll explain in my next post).)

Using MEC in this situation, the expected choice-worthiness of buying the fish curry is 0.25 * -90 + 0.75 * 10 = -15, and the expected choice-worthiness of buying the tofu curry is 0.25 * 5 + 0.75 * 5 = 5. Thus, Devon should buy the tofu curry.

This is despite Devon believing that T2 is more likely than T1, and T2 claiming that buying the fish curry is better than purchasing the tofu curry. The reason is that, as discussed earlier, there is far "more at stake" for T1 than for T2 in this example.

To me, this seems like a good, intuitive result for MEC, and shows how it improves upon the “My Favourite Theory” approach.

There are two final things I should note about MEC:

- MEC can be used in exactly the same way when more than two theories are under consideration. (The only reason most examples in this sequence will be ones in which only two moral theories are under consideration is to keep explanations simple.)
- **The basic idea of MEC can also be used as a heuristic, without involving actual numbers.**
    - For example, say Clara believes that there’s a “high chance” utilitarianism is correct, but that some deontological theory, in which lying is deeply wrong, is “plausible”. Clara is considering whether to tell a lie, and has good reason to believe this will lead to a slight net increase in wellbeing. She might still decide not to lie, despite believing it’s likely that lying is the “right” thing to do, because it’d only be *slightly right*, whereas it’s plausible it’s *deeply wrong*.

Another example of applying MEC (which is probably only worth reading if the approach still seems unclear to you) can be found in the following footnote.[8]

Normalised MEC and Variance Voting

(*It's possible I've made mistakes in this section; if you think I have, please let me know.*)

But what about cases in which, despite being cardinal, the theories you have credence in are *not intertheoretically comparable*? (Recall that this essentially means that there's no consistent, non-arbitrary “exchange rate” between the theories' “units of choice-worthiness".)

MacAskill argues that, in such situations, one must first "normalise" the theories in some way (which basically means "adjusting values measured on different scales to a notionally common scale"). MEC can then be applied just as we saw earlier, but now with the new, normalised choice-worthiness scores.

There are multiple ways one could normalise the theories under consideration (e.g., by range), but MacAskill argues for normalising by variance. That is, he argues that we should:

> “[treat] the average of the squared differences in choice-worthiness from the mean choice-worthiness as the same across all theories. Intuitively, the variance is a measure of how spread out choice-worthiness is over different options; normalising at variance is the same as normalising at the difference between the mean choice-worthiness and one standard deviation from the mean choice-worthiness.”

MacAskill uses the term Variance Voting to refer to this process of first normalising by variance and then using the MEC approach.

(Unfortunately, as far as I could tell, none of the three theses/papers I read that referred to normalising moral theories by variance actually provided a clear, worked

example. I've attempted to construct such a worked example based on an extension of the scenario with Devon deciding what meal to buy; that can be found here, and here is a simpler and I think effectively identical method, suggested in a private message.)

In arguing for Variance Voting over its alternatives, MacAskill states that the basic principle normalisation aims to capture is the "*principle of equal say*: the idea, stated imprecisely for now, that we want to give equally likely moral theories equal weight when considering what it's appropriate to do" (emphasis in original). He further writes:

> "To see a specific case of how this could go awry, consider average and total utilitarianism, and assume that they are indeed incomparable. And suppose that, in order to take an expectation over those theories, we choose to treat them as agreeing on the choice-worthiness ordering of options concerning worlds with only one person in them. If we do this, then, for almost all decisions about population ethics, the appropriate action will be in line with what total utilitarianism regards as most choiceworthy because, for almost all decisions, the stakes are huge for total utilitarianism, but not very large for average utilitarianism. So it seems that, if we treat the theories in this way, we are being partisan to total utilitarianism.
>
> In contrast, if we chose to treat the two theories as agreeing on the choice-worthiness differences between options with worlds involving 10^100 people then, for almost all real-world decisions, what it's appropriate to do will be the same as what average utilitarianism regards as most choice-worthy. This is because we're representing average utilitarianism as claiming that, for almost all decisions, the stakes are much higher than for total utilitarianism. In which case, it seems that we are being partisan to average utilitarianism, whereas what we want is to have a way of normalising such that each theory gets equal influence." (line break added)

(Note that it's not a problem for one theory to have much more influence on decisions *due to higher credence in that theory*. The principle of equal say is only violated if additional influence is unrelated to additional credence in a theory, and instead has to do with what are basically *arbitrary/accidental choices about exchange rates* between units of choice-worthiness.)

MacAskill (pages 110-116) provides two arguments that VV is the approach that satisfies the principle of equal say, and Owen Cotton-Barratt similarly argues for the superiority of normalisation by variance over alternative normalisations. (But note that this approach does seem to have its flaws, as discussed in, e.g., pages 222-223 of Tarsney's thesis.)

The Borda Rule (BR)

Finally, what about cases in which all moral theories you have credence in are only ordinal, rather than cardinal (i.e., they say only whether each option is more, equally, or less choice-worthy than each other option, but not by how much)? For such cases, MacAskill recommends a voting method called the Borda Rule (BR; also known as a "Borda count"), with "moral theories play[ing] the part of voters and practical options the part of candidates" (Tarsney). I will first quote MacAskill's formal explanation of BR (which may be somewhat confusing by itself), before quoting an example he gives and showing what applying BR to that looks like:

> "An option *A*'s *Borda Score*, for any theory *Ti*, is equal to the number of options within the option-set that are less choice-worthy than *A* according to theory *Ti*'s

choice-worthiness function, minus the number of options within the option-set that are more choice-worthy than *A* according to *Ti*'s choice-worthiness function.[9]

An option *A*'s *Credence-Weighted Borda Score* is the sum, for all theories *Ti*, of the Borda Score of *A* according to theory *Ti* multiplied by the credence that the decision-maker has in theory *Ti*.

[The *Borda Rule* states that an] option *A* is more appropriate than an option *B* iff [if and only if] *A* has a higher Credence-Weighted Borda Score than *B*; *A* is equally as appropriate as *B* iff *A* and *B* have an equal Credence-Weighted Borda Score."

I will now show, following MacAskill, how this rule applies to an example he gives in his thesis:

"Julia is a judge who is about to pass a verdict on whether Smith is guilty for murder. She is very confident that Smith is innocent. There is a crowd outside, who are desperate to see Smith convicted. Julia has three options:

[G]: Pass a verdict of 'guilty'.

[R]: Call for a retrial.

[I]: Pass a verdict of 'innocent'.

Julia knows that the crowd will riot if Smith is found innocent, causing mayhem on the streets and the deaths of several people. If she calls for a retrial, she knows that he will be found innocent at a later date, and that it is much less likely that the crowd will riot at that later date. If she declares Smith guilty, the crowd will be appeased and go home peacefully. She has credence in three moral theories:

35% credence in a variant of utilitarianism, according to which [G>R>I].

34% credence in a variant of common sense, according to which [R>I>G].

31% credence in a deontological theory, according to which [I>R>G]."

The options' Borda Scores according to each theory, and their Credence-Weighted Borda Scores, are therefore as shown in the following table:

	Utilitarian theory - 35%	Common sense theory - 34%	Deontological theory - 31%	Credence-Weighted Borda Score
Guilty	2 - 0 = **2**	0 - 2 = **-2**	0 - 2 = **-2**	**-0.6**
Retrial	1 - 1 = **0**	2 - 0 = **2**	1 - 1 = **0**	**0.68**
Innocent	0 - 2 = **-2**	1 - 1 = **0**	2 - 0 = **2**	**-0.08**

(For example, G has a score of 2 - 0 = 2 according to utilitarianism because that theory views two options as less choice-worthy than G, and 0 options as more choice-worthy than G.)

The calculations that provided the Credence-Weighted Borda Scores shown in the above table are as follows:

G: 0.35 * 2 + 0.34 * -2 + 0.31 * -2 = -0.6 (this because the utilitarian, common sense, and deontological theories are given credences of 35%, 34%, and 31%, respectively, and these serve as the weightings for the Borda Scores these theories provide)

R: 0.35 * 0 + 0.34 * 2 + 0.31 * 0 = 0.68

I: 0.35 * -2 + 0.34 * 0 + 0.31 * 2 = -0.08

BR would therefore claim that Julia should call for a retrial. **This is the case even though passing a guilty verdict was seen as best by Julia's "favourite theory" (the variant of utilitarianism). Essentially, calling for a retrial is preferred because both passing a guilty verdict and passing an innocent verdict were seen as *least* preferred by some theory Julia has substantial credence in, whereas calling for a retrial is not *least* preferred by any theory.**

MacAskill notes that preferring this sort of a compromise option in a case like this seems intuitively right. He also argues that alternatives to BR fail to give us the sort of answers we'd want in these or other sorts of cases. (Though Tarsney raises some objections to BR which I won't get into.)

Closing remarks

I hope you have found this post a useful, clear summary of key ideas around what moral uncertainty is, why it matters, and how to make decisions when morally uncertain. Personally, I believe that an understanding of moral uncertainty - particularly a sort of heuristic version of MEC - has usefully enriched my thinking, and influenced some of the biggest decisions I've made over the last year.[10]

In the next post, I will discuss (possibly novel, arguably obvious) extensions of each of the three approaches discussed here, in order to allow for modelling *both moral and empirical uncertainty, explicitly and simultaneously*. The post after that will discuss how we can combine the approaches in the first two posts with sensitivity analysis and value of information analysis.[11][12]

---

1. I genuinely mean no disrespect to the several posts on moral uncertainty I did discover (e.g., here, here, and here). All did meet some of those criteria, and I'd say most were well-written but just weren't highly explicit (e.g., didn't include enough concrete examples), and/or didn't cover (in the one post) each of the prominent approaches and the related ideas necessary to understand them. ↩

2. Other terms/concepts that are sometimes used and are similar to "moral uncertainty" are *normative*, *axiological*, and *value* uncertainty. In this sequence, I'll use "moral uncertainty" in a general sense that also incorporates axiological and value uncertainty, and at least a large part of normative uncertainty.

   Also, throughout this sequence, I will use the term "approach" in a way that I believe aligns with MacAskill's use of the term "metanormative theory". ↩

3. It seems to me that there are many cases where it's not entirely clear whether the uncertainty is empirical or moral. For example, I might wonder "Are fish conscious?", which seems on the face of it an empirical question. However, I

might not yet know precisely what I mean by “conscious”, and only really want to know whether fish are “conscious in a sense I would morally care about”. In this case, the seemingly empirical question becomes hard to disentangle from the (seemingly moral) question “What forms of consciousness are morally important?”

(Furthermore, my answers to *that* question in turn may be influenced by empirical discoveries. For example, I may initially believe avoidance of painful stimuli demonstrates consciousness in a morally relevant sense, but then change that belief after learning that this behaviour can be displayed in a stimulus-response way by certain extremely simple organisms.)

In such cases, I believe the approach suggested in the next post of this sequence will still work well, as that approach does not really require empirical and moral uncertainty to be treated fundamentally differently. (Another approach, which presents itself differently but I think is basically the same in effect, is to consider uncertainty over “worldviews”, with those worldviews combining moral and empirical claims.) ↩

4. In various places in this sequence, I will use language that may appear to endorse or presume moral realism (e.g., referring to “moral information” or to probability of a particular moral theory being “true”). But this is essentially just for convenience; I intend this sequence to be neutral on the matter of moral realism vs antirealism, and I believe this post can be useful in mostly similar ways regardless of one’s position on that matter. I discuss the matter of "moral uncertainty for antirealists" in more detail in this separate post. ↩

5. The matter of how to actually assign “units” or “magnitudes” of choice-worthiness to different options, and what these things would even mean, is complex, and I won’t really get into it in this sequence. ↩

6. Christian Tarsney's 2017 thesis thesis (e.g., pages 175-176) explains other ways the "structure" of moral theories can differ, and potential implications of these other differences. These were among the juicy complexities I had to resist cramming in this originally-intended-as-bitesized post (but I may write another post about Tarsney's ideas later; please let me know if you think that'd be worthwhile). ↩

7. It's worth noting that similar issues have received attention from, and are relevant to, other fields as well. For example, MacAskill writes: "A similar problem arises in the study of social welfare in economics: it is desirable to be able to compare the strength of preferences of different people, but even if you represent preferences by cardinally measurable utility functions you need more information to make them comparable." Thus, concepts and findings from those fields could illuminate this matter, and vice versa. ↩

8. Suppose Alice assigns a 60% probability to hedonistic utilitarianism (HU) being true and a 40% probability to preference utilitarianism (PU) being true. Suppose also that Bob *wants* to play video games, but would actually *get slightly more joy* out of a day at the beach. Thus, according to HU, letting Bob play video games has a CW of 5, and taking him to the beach has a CW of 6; while according to PU, letting Bob play video games has a CW of 15, and taking him to the beach has a CW of -20.

Under these conditions, the expected choice-worthiness of letting Bob play video games is 0.6 * 5 + 0.4 * 15 = 9, and the expected choice-worthiness of taking Bob to the beach is 0.6 * 6 + 0.4 * -20 = -4.4. Therefore, Alice should let Bob play video games.

Analogously to the situation with the Devon example, this is despite Alice believing HU is more likely than PU, and despite HU positing that taking Bob to the beach being better than letting him play video games. As before, the reason is that there is "more at stake" in this decision for the less-believed theory than for the more-believed theory; HU considers there to only be a very small difference between the choice-worthiness of the options, while PU considers there to be a large difference. ↩

9. MacAskill later notes that a simpler method (which doesn't subtract the number of options that are more choice-worthy) can be used when there are no ties. His calculations for the example I quote and work through in this post use that simpler method. But in this post, I'll stick to the method MacAskill describes in this quote (which is guaranteed to give the same final answer in this example anyway). ↩

10. However, these concepts are of course not an instant fix or cure-all. In a (readable and interesting) 2019 paper, MacAskill writes "so far, the implications for practical ethics have been drawn too simplistically [by some philosophers.] First, the implications of moral uncertainty for normative ethics are far more wide-ranging than has been noted so far. Second, one can't straightforwardly argue from moral uncertainty to particular conclusions in practical ethics, both because of 'interaction' effects between moral issues, and because of the variety of different possible intertheoretic comparisons that one can reasonably endorse."

    For a personal example, a heuristic version of MEC still leaves me unsure whether I should move from being a vegetarian-flirting-with-veganism to a strict vegan, or even whether I should spend much time making that decision, because that might trade off to some extent with time and money I could put towards longtermist efforts (which seem more choice-worthy according to other moral theories I have some credence in). I suspect any quantitative modelling simple enough to be done in a reasonable amount of time would still leave me unsure.

    That said, I, like MacAskill (in the same paper), "do believe, however, that consideration of moral uncertainty should have major impacts for how practical ethics is conducted. [...] It would be surprising if the conclusions [of approaches taking moral uncertainty into account] were the same as those that practical ethicists typically draw."

    In particular, I'd note that considering moral uncertainty can reveal some "low-hanging fruit": some "trades" between moral theories that are relatively clearly advantageous, due to large differences in the "stakes" different moral theories see the situation as having. (Personally, cases of apparent low-hanging fruit of this kind have included becoming at least vegetarian, switching my career aims to longtermist ones, and yet engaging in global-poverty-related movement-building when an unusual opportunity arose and it wouldn't take up too much of my time.) ↩

11. To foreshadow: Basically, my idea is that, once you've made explicit your degree of belief in various moral theories and how good/bad outcomes appear to each of

those theories, you can work out which updates to your beliefs in moral theories or to your understandings of those moral theories are most likely to change your decisions, and thus which "moral learning" to prioritise and how much resources to expend on it. ↩

12. I'm also considering later adding posts on:

    - Different types and sources of moral uncertainty (drawing on these posts).
    - The idea of ignoring even very high credence in nihilism, because it's never decision-relevant.
    - Whether it could make sense to give moral realism disproportionate (compared to antirealism) influence over our decisions, based on the idea that realism might view there as "more at stake" than antirealism does.

    I'd be interested in hearing whether people think those threads are likely to be worth pursuing. ↩

The Actionable Version of "Keep Your Identity Small"

(cross posted on my roam blog)

There's an old Paul Graham Essay, "Keep Your Identity Small". It's short so it's worth it to read the whole thing right now if you've never seen it. The yisbifiefyb ("yeah it's short but i'm functionally illiterate except for your blog") is roughly "When something becomes part of your identity, you become dumber. Don't make things part of your identity."

I read that post some time in high school and thought, "Of course! You're so right Paul Graham. Cool, now I'll never identify as anything." I still think that Paul Graham is pointing out a real cluster of Things That Happen With People, but over time the concept of *identity*, and *identifying as BLANK* have started to feel less clear. It feels right to say "People get dumb when their identity is challenged" and it even feels sorta axiomatic. Isn't that what it means for something to be part of your identity? Thinking about it more I came up with a bunch of different ways of thinking of myself that all felt like *identifying as BLANK*, but it felt like unnecessary dropping of nuance to smoosh them all into the single concept of *identity*.

Identity Menagerie

Lets look at some examples of what identifying as a BLANK can look like:

- Blake: "I do Cross Fit. "
- Jane: "I'm smart. In fact I'm normally among the smartest in the room. I'm able to solve a lot of problems by just finding a clever solution to them instead of having to get stuck in grunt work. People often show awe and appreciation for my depth and breadth of knowledge."
- Jay: "I the peacekeeper, the one always holding the group together."

Self-Concept

Steve Andreas outlines the idea of a self-concept quite nicely:

> Your self-concept is a sort of map of who you are. Like any other map, it is always a very simplified version of the territory. [...] Your self-concept, your "map" you have of yourself, has the same purpose as a map of a city—to keep you oriented in the world and help you find your way, particularly when events are challenging or difficult.

The thing you'll notice is it's nigh impossible to avoid having a self-concept. When Jane thinks of herself and how she can act on the world, "being smart" is a chunk of self-concept that summarizes a lot of her experiences and that she uses to guide decisions she makes.

Kaj Sotala has a good post about how tweaking and modifying his self-concept helped fix parts of his depression and anxiety.

Group Identity

This is the obvious one that we're all used to. Blake does Cross Fit, hangs out with cross fit people all the time, and loves telling people about all this. All of his Cross Fit buddies support each other and give each other praise for being part of such an awesome group. Someone calling Cross Fit stupid would feel like someone calling him and all of his friends stupid. It would be big and difficult change for Blake to get out of Cross Fit, given that's where most of his social circle is, and where all his free time goes.

Intelligent Social Web

Here's Val describing what he calls the Intelligent Social Web:

> I suspect that improv works because we're doing something a lot like it pretty much all the time. The web of social relationships we're embedded in helps define our roles as it forms and includes us. And that same web, as the distributed "director" of the "scene", guides us in what we do. A lot of (but not all) people get a strong hit of this when they go back to visit their family. If you move away and then make new friends and sort of become a new person (!), you might at first think this is just who you are now. But then you visit your parents... and suddenly you feel and act a lot like you did before you moved away. You might even try to hold onto this "new you" with them... and they might respond to what they see as strange behavior by trying to nudge you into acting "normal": ignoring surprising things you say, changing the topic to something familiar, starting an old fight, etc.

This feels like another important facet of identity, one that doesn't just exist in your head, but in the heads of those around you.

Identity as a Strategy for meeting your needs

In middle school and high school I built up a very particular identity. I bet if you conversed with high school me, you wouldn't be able to pin me down to using any particular phrase, label, or group to identify myself as. And yet, there are ways of being you could have asked me to try that would have *scared the shit out of me*. Almost as if... my identity was under attack....

So new take, one I consider more productive. Reread Paul Grahams essay and replace every instance of "identity" with "main strategy to meet one's needs". Hmmmm, it's starting to click. If you've been a preacher for 40 years, and all you know is preaching, and most of your needs are met by your church community, an attack on the church is an attack on *your livelihood and well-being.*

I expect having your "identity" under attack to feel similar to being a hunter gatherer and watching the only river that you've known in your life drying up. Fear and Panic. What are you going to do know? Will you survive? Where are the good things in your life going to come from?

When you frame it like this, you can see how easily trying to KYIS could lead to stuff that just hurts you. If I only have one way of getting people to like me (say, being funny), I can't just suddenly decide not to care if people don't consider me funny. I can't just suddenly not care if people stop laughing at my jokes. Both of those events mean I no longer have a functional strategy to be liked.

A very concrete prediction of this type of thinking: someone will be clingy and protective over a part of their behavior to the degree that it is the sole source of meeting XYZ important needs.

KYIS is not actionable advice

The take away from Paul Graham is "don't let something become you identity". How do you do that? I thought it meant something like "Never self identity as a BLANK", to others or to yourself. Boom. Done. And yet, even though I never talked about being part of one group or another, I still went through life a decent chunk of life banking on "Be funny, act unflappable, be competent at the basic stuff" as the only/main strategy for meeting my needs.

The actionable advice might be something like, "slowly develop a multi-faceted confidence in your ability to handle what life throws at you, via actually improving and seeing results." That's waaaaaay harder to do than just not identifying with a group, but it does a better jump of pointing you in the direction that matters. I expect that when Paul Graham wrote that essay he already had a pretty strong confidence in his ability to meet his needs. From that vantage point, you can easily let go of identities, because they aren't your life lines.

There can be much more to identity than what I've laid out, but I think the redirect I've given is one that is a great first step for anyone dwelling on identity, or for anyone who head the KYIS advice and earnestly tried to implement it, yet found mysterious ways it wasn't working.

2020's Prediction Thread

Inspired by the 2010 prediction thread, I would like to propose this as a thread for people to write in their predictions for the next decade, when practical with probabilities attached.

Is the term mesa optimizer too narrow?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

In the post introducing mesa optimization, the authors defined an *optimizer* as

> a system [that is] internally searching through a search space (consisting of possible outputs, policies, plans, strategies, or similar) looking for those elements that score high according to some objective function that is explicitly represented within the system.

The paper continues by defining a mesa optimizer as an optimizer that was selected by a base optimizer.

However, there are a number of issues with this definition, as some have already pointed out.

First, I think by this definition humans are clearly not mesa optimizers. Most optimization we do is implicit. Yet, humans are the supposed to be the prototypical examples of mesa optimizers, which appears be a contradiction.

Second, the definition excludes perfectly legitimate examples of inner alignment failures. To see why, consider a simple feedforward neural network trained by deep reinforcement learning to navigate my Chests and Keys environment. Since "go to the nearest key" is a good proxy for getting the reward, the neural network simply returns the action, that when given the board state, results in the agent getting closer to the nearest key.

Is the feedforward neural network optimizing anything here? Hardly, it's just applying a heuristic. Note that you don't need to do anything like an internal A* search to find keys in a maze, because in many environments, following a wall until the key is within sight, and then performing a very shallow search (which doesn't have to be explicit) could work fairly well.

As far as I can tell, Hjalmar Wijk introduced the term "malign generalization" to describe the failure mode that I think is most worth worrying about here. In particular, malign generalization happens when you trained a system with objective function X, that at deployment has the actual outcome of doing Y, where Y is so bad that we'd prefer the system to fail completely. To me at least, this seems like a far more intuitive and less theory-laden way of framing inner alignment failures.

This way of reframing the issue allows us to keep the old terminology that we are concerned with capability robustness without alignment robustness, but drops all unnecessary references to mesa optimization.

Mesa optimizers could still form a natural class of things that are prone to malign generalization. But if even humans are not mesa optimizers, why should we expect mesa optimizers to be the primary real world examples of such inner alignment failures?

Meditation Retreat: Immoral Mazes Sequence Introduction

I just got home from a six day meditation retreat and began writing.

The catch is that I arrived at the retreat yesterday.

I knew going in that it was a high variance operation. All who had experience with such things warned us we would hate the first few days, even if things were going well. I was determined to push through that.

Alas or otherwise, I was not sufficiently determined to make that determination stick. I didn't have a regular practice at all going in, was entirely unfamiliar with the details of how this group operated, and found the Buddhist philosophy involved highly off putting, in a 'no that's not how it works that's not how any of this works nor would we want it to' kind of way. I am no negative utilitarian. I couldn't focus, my meditations were entirely unproductive compared to past experience and were increasingly focusing on how terrible things were.

I am highly, highly confident that none of the people who warned me would be surprised by those developments, and similarly confident that they would all tell me to push through it. And will tell me, when I see them, that if I can do so I should try again. But I also realized that *the anticipated reaction from others saying I didn't give it a proper chance was the only reason I was considering not leaving.* So I left.

To my surprise, those there said I was likely making a mature decision and were sympathetic. They spun it a little to try and get me not to give up in the future, but that was it, which was a really good look. It did not go unnoticed.

I took the bulk of the day to get home and relax, play a game, saw the excellent movie Knives Out. What I did realize was that yes, some combination of the Solstice plus the meditation retreat, even if I only did a few hours worth of sessions, did have a clarifying and motivating effect to get me back on track. I'm not unhappy I went, even though I bailed, because I was, in a much more practical and everyday very small sense, enlightened.

I'm also leaning towards being happy I left when I did. I do buy that there are serious effects that can only be gained from being out of feedback loops and in silence for several days, but my instincts (however motivated they may be) are strongly telling me that this is not the way for me to do that.

The other motivating part of this is that, while I will absolutely take the majority of tomorrow to enjoy the College Football Playoff, this is both my chance to be alone for a few days and also a time when I would otherwise be in hardcore meditation. It seems wrong to not accomplish something important that isn't work or game related, to meditate in another way.

The goal is ideally to finish everything up, at least in draft-ready-to-adjust-for-comments-on-earlier-posts form, by the end of the retreat. That is a stretch, so the commit-to-it goal is to declare the first six posts finished and begin publishing them at a reasonable clip, and have momentum on seven and later.

The drafts that currently exist, that will be finalized and likely expanded upon, are the following:

1. Moloch Hasn't Won. Have you noticed that the world is in fact very much *not* a dystonian hellhole of Moloch-worshiping perfect competition and Elua's enemies keep on having all those *unfortunate accidents?*
2. Perfect Competition. Perfect competition, importantly, isn't a thing, but you can get close. Let's flesh this out more.
3. Imperfect Competition. Some practical examples of imperfect competition. Intuition pumps and detailed examples for why perfect competition isn't a thing and we don't usually get that close.
4. What is an Immoral Maze (note that I make a point to say Immoral rather than Moral)? Mazes need not be corporations or (in my current model in ways I'd have to introduce that aren't in the draft right now, with a subset of the tech ecosystem as a motivating example) even formal organizations. What creates a maze? A system with multiple effective layers of hierarchy forcing its middle management into effectively super-perfect competition against each other largely on the basis of anticipated future success in such competition.
5. What is Success in an Immoral Maze? There is no true success. What those inside think is success is anything but. Even if you win, you lose. Stay out, get out.
6. How to Identify an Immoral Maze. Look at levels of hierarchy, skin in the game, soul in the game, how people describe their jobs, diversity of skill levels and degree of slack. Then pay attention, damn it.
7. How to Interact with Immoral Mazes. They can't be fully avoided, and some are stuck with them more than others. Practical discussion of what to do about this on a personal level.
8. The Road to Mazedom. Well? How did we get here? Draft of this is still ongoing and it is *freaking huge* so it is probably going to get split up once we get to it. Also we need better answers on what to do about all this than what I have, even if it's a start. Hard problem is hard!
9. Moloch's Army. This isn't written and needs to go *somewhere* in the sequence or outside of it, or the whole operation is woefully incomplete. I need to finally write it. The devil's greatest trick was never proving he didn't exist, I wrote ten minutes ago, it was proving he'd already won, or would inevitably win. That only those who make deals with him get ahead, so they should implicitly or explicitly coordinate against those who don't. Moloch has an army, who coordinate implicitly around fighting against anyone fighting for anything except Moloch's Army, or anyone having any values. And *this is how Moloch wins, where it wins.* And also by making sure no one ever writes this, which makes this hard to write, etc etc. In that sense, *it really is true that the Devil's greatest trick is convincing people he doesn't exist,* because so far everyone I know who has figured this out has found it impossible to talk or write about this without sounding crazy to those who don't already get it. Much careful background may be necessary. Darwin sequence was originally supposed to be a gateway to this but it wasn't good enough on its own.

LW Team Updates - December 2019

This is the once-monthly updates post for LessWrong team activities and announcements.

Summary

In the past month we rolled out floating comment guidelines and launched the inaugural Lesswrong 2018 review. Work has continued on the LessWrong editor and on a prototype for the new tagging system.

December will see more work on the editor, the 2018 review process, and analytics.

Recent Features

The LessWrong 2018 Review

Much of the recent weeks has been devoted to getting the inaugural Lesswrong 2018 Review into full swing:

> LessWrong is currently doing a major review of 2018 — looking back at old posts and considering which of them have stood the tests of time. It has three phases:
>
> - Nomination *(ends Dec 1st at 11:59pm PST)*
> - Review *(ends Dec 31st)*
> - Voting on the best posts *(ends January 7th)*
>
> Authors will have a chance to edit posts in response to feedback, and then the moderation team will compile the best posts into a physical book and LessWrong sequence, with $2000 in prizes given out to the top 3-5 posts and up to $2000 given out to people who write the best reviews.

Read Raemon's full post to hear for the full rationale for the evaluation of historical posts.

NOMINATED POSTS ARE NOW OPEN FOR REVIEW

The nomination phase just ended a few days ago. 34 nominators made 204 nominations on 98 distinct posts written by 49 distinct authors. Of these, 74 posts have received the 2+ nominations required to proceed to the review phase.

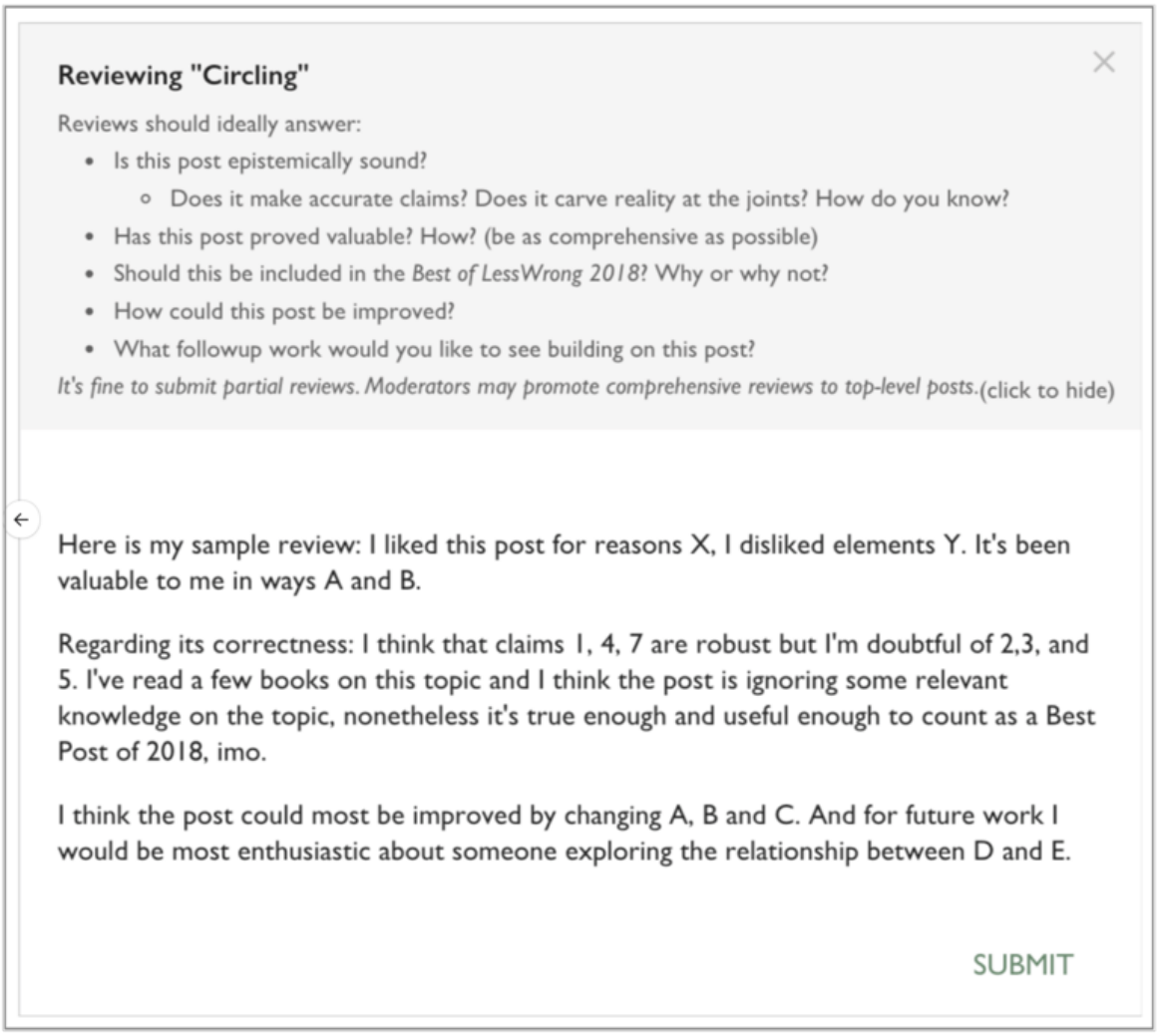
How to start reviewing

1. The frontpage currently has a *LessWrong 2018 Review* section. It shows a random selection of posts which are up for review and has buttons to the *Reviews Dashboard* and the list of reviews and nominations you've made so far.
2. The Reviews Dashboard (located at www.lesswrong.com/reviews) is another way to find posts to review.

The 2018 Review section currently on the homepage.

The Reviews Dashboard.
Note the "Expand Unread Comments" button.

3. When you click *Review* on a review-able post, you will be taken to the post page and a *Review Comment Box* will appear.



The Review Comment Box

Reviews are posted as comments and can be edited after they are posted like regular comments.

Reasons to review

All users are encouraged to writes reviews. Reviews help by:

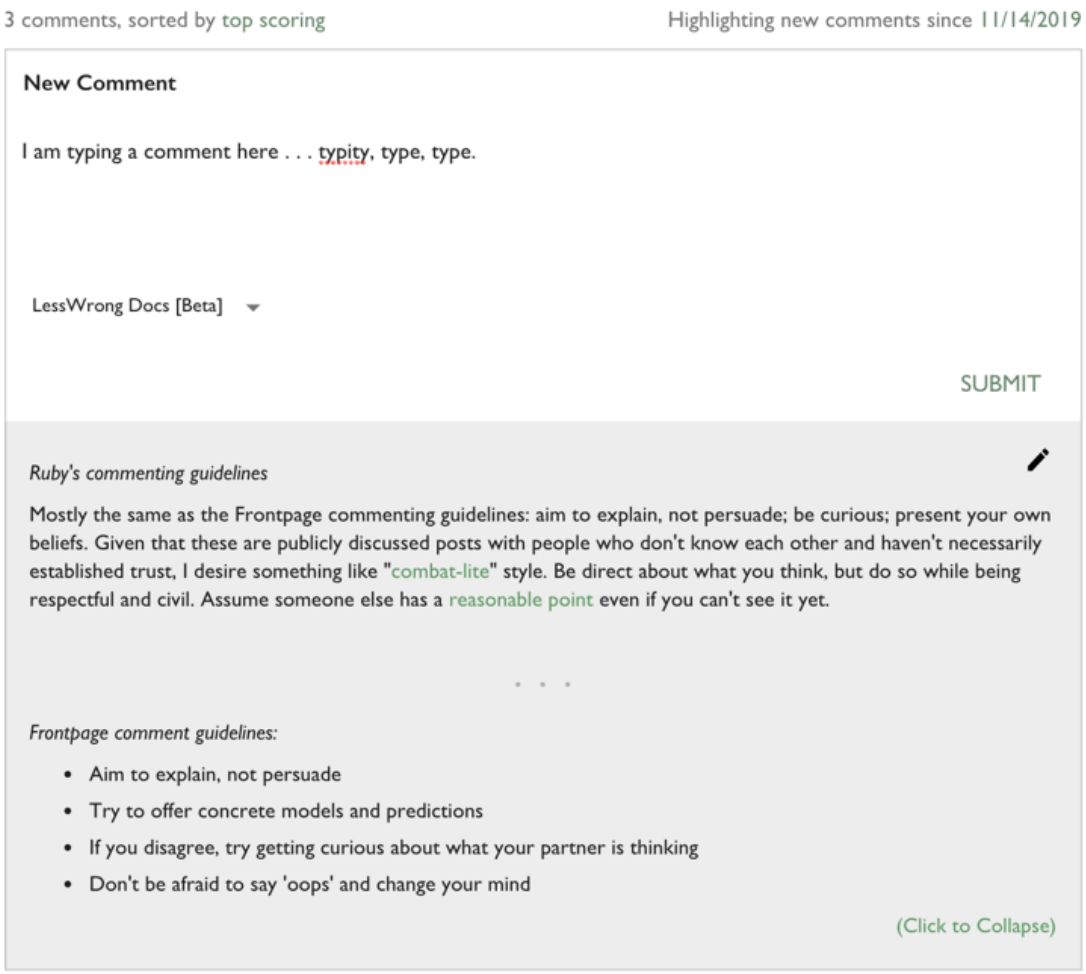
- Giving authors feedback which they can use to revise, update, and expand their posts before users vote on them and they possibly get included in the physical book that will be published.
- Giving the community opportunity to discuss the importance and trustworthiness of posts. In particular, now is an opportune time for the community to debate the more contentious ideas and arguments.
- Thereby, establishing a record on which posts are truly excellent vs those that need work or are more doubtful.
- Help people decide which posts they will vote on in the upcoming Voting Phase.
- Help new readers decide whether or not they wish to read a post.

The review phase will continue until December 31st

Floating Comment Guidelines

For a long time, LessWrong has enabled authors to set and enforce their own custom moderation guidelines on their own posts. This is part or the Archipelago philosophy of moderation which lets people decide what kinds of conversations they want.

To make it easier for commenters stick to desired guidelines across users, and to better understand how sections of the site like Shortform have different norms, we've made it so the moderation guidelines for a post automatically appear beneath the comment checkbox whenever you begin typing.



How the commenting guidelines appear beneath an in-progress comment.

App-Level Analytics Tracking

This isn't really a user-level feature that people can interact with, but we've been working to expand our ability to detect what people are doing within the web-app, e.g. tracking how much different features get used and which don't.

We've been sorely missing this and it's impeded our ability to assess whether some of the features we've been rolling out have been a success or not.

Hopefully, with this improved feedback we'll make better choices about what to build and be better at detecting pain points for users.

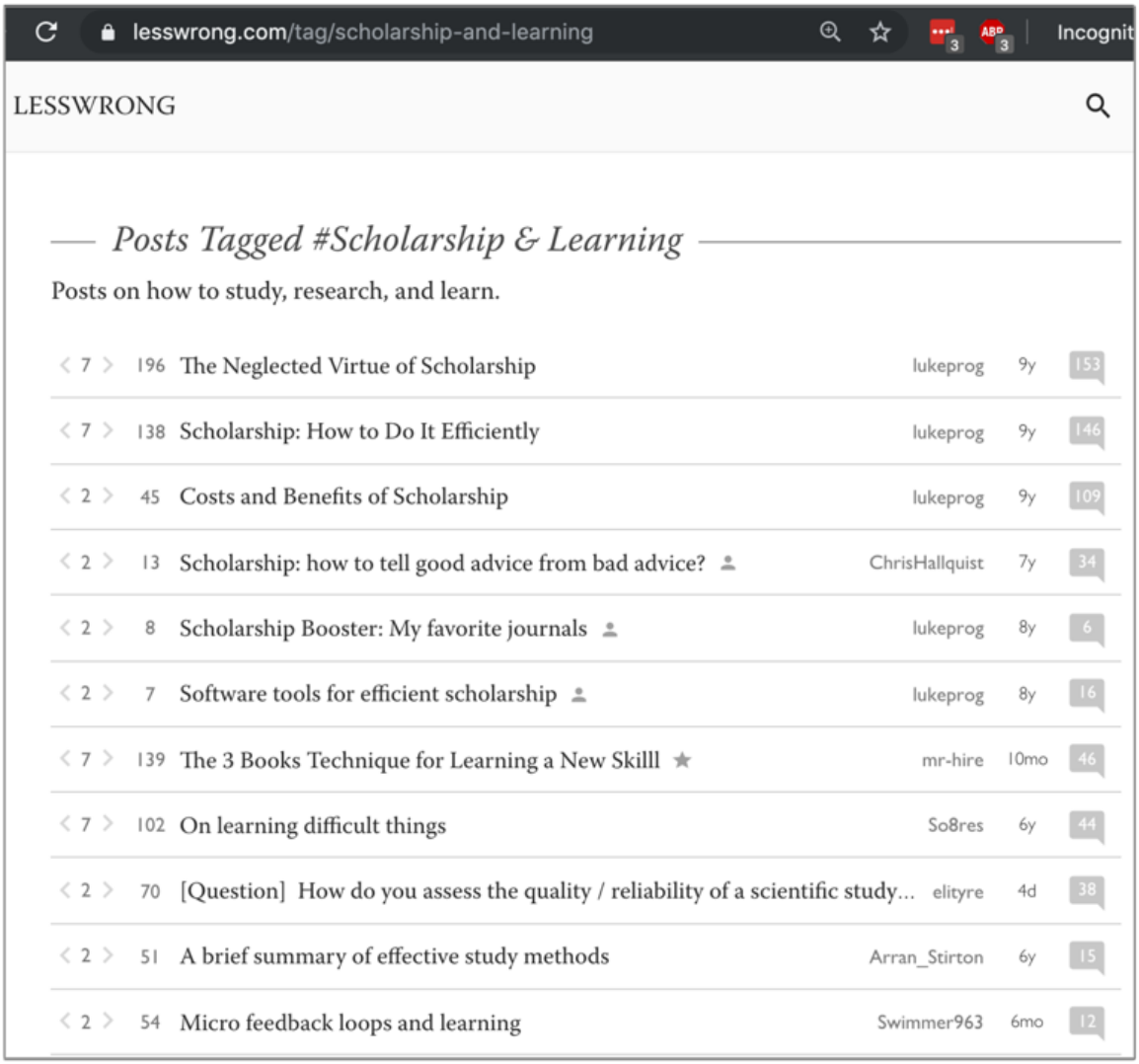
Upcoming Features

LessWrong Docs (new editor)

Work continues on the new editor, codenamed *LW Docs* for now, with the team internally using it. However, we're not yet rolling it out more widely while we work out remaining reliability issues.

Tagging

We successfully implemented a new tagging prototype and have played around with it. That's roughly as much work we plan to do on this in Q4. To complete this project we need to first flesh out a broader design vision, figure out how tags will relate to wikis, and figure out a clean and intuitive UI design. We might release something here in Q1 2020.



MVP of a *tag page*. Leftmost number is the tag relevance score. All users can view the current tag pages, however only admins can currently create or vote on tags.

Feedback & Support

The team can be reached for feedback and support via:

- Comment on this post
- Intercom (icon in the bottom right, you might have to edit your user settings)
- Email us at hello@lesswrong.com or support@lesswrong.com
- Ask a question on www.lesswrong.com/questions
- Message us on our Facebook page.

New paper: (When) is Truth-telling Favored in AI debate?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://medium.com/@RyanCarey/new-paper-when-is-truth-telling-favored-in-ai-debate-8f58f14562e5

An introduction to a recent paper by myself and Ryan Carey. Cross-posting from Medium.

For some intellectual tasks, it's easy to define success but hard to evaluate decisions as they're happening. For example, we can easily tell which Go player has won, but it can be hard to know the quality of a move until the game is almost over. AI works well for these kinds of tasks, because we can simply define success and get an AI system to pursue it as best it can.

For other tasks, it's hard to define success, but relatively easy to judge solutions when we see them, for example, doing a backflip. Getting AI to carry out these tasks is harder but manageable — we can generate a bunch of videos of an AI system making some motion with a simulated body. Then we can give these videos to some people who allocate "approval" to the best-looking motions, and train the AI to maximize that approval until it does a backflip.

What makes AI really hard are tasks for which we have no definition of success, nor any timely way to evaluate solutions. For example: to which school should I send my kids? And should I accept this job or that one? One proposal for these cases is to use AI Debate. The idea is to ask AI systems how to perform a task, and then to have them debate about the virtues of different possible decisions (or answers). The question could be how to win a game of Go, how to do a backflip, which school to send your kids to, or, in principle, basically anything. The hope is that observing an AI debate would help the human judge to better-understand different possible decisions and evaluate them on-the-fly, even if success can't yet be quantified.

One concern with such a scheme is whether it would be safe, especially if the AI systems used are super-smart. Critics ask: "How can we be confident that in AI Debates, the true answer will win?" After all, in human debates, rhetorical tools and persuasive techniques can cause an audience to be mislead.

To us, it seems wrong to imagine all debates will be safe. But it seems equally wrong to expect that none can be. A better question, to us, is: "In which debates will the winner be the true answer?" In our recent paper, we (Vojta and Ryan) have taken a first stab at making mathematical models that address this question.

So what is a debate exactly? According to our model, every debate revolves around some question posed by a human and consists of two phases. In the answering phase, each AI system chooses an answer to argue for. Then in the argumentation phase, the two AI systems debate over whose answer is better. (In some variants, the answering and argumentation phases will be performed by different algorithms or answers may just be assigned to the debaters.) At the end of the debate, the human "judge"

considers all the arguments and optionally performs an “experiment” to get to the bottom of things, such as Googling the claim of some debater. Equipped with this information, the judge rewards the AI whose answer seems better. In the language of game theory, the answering phase is a matrix game, and the argumentation phase is a sequential game with perfect information.

In order to make debate easier to think about, we defined a simple version of the above model called feature debate. In feature debates, the world is characterized by a list of “elementary features”, and the only kind of argument allowed is to reveal the value of a single elementary feature. For example, we can imagine a feature debate regarding whether a given image depicts a cat or a dog. Then each argument consists of revealing a selected pixel. Finally, given this information, a judge updates its beliefs based on the arguments provided and allocates reward to the AI who argued for the answer that looks more likely. For our simple first-pass analysis, we imagine that the judge is completely naive to the fact that debaters provide evidence selectively. We also assume that the judge only has “patience” to process some limited number of arguments.

In the setting of feature debates, we’ve shown some kinds of debates that will work, and others that won’t. For some kinds of debates, the arguments are just too difficult to explain before the judge runs out of patience. Basically, showing n arguments may be completely meaningless without the final argument number n+1. And if the judge only has time for n, then truth won’t win.

Some kinds of feature debates, however, turn out better. The first case is if we know the importance of different features beforehand. Roughly speaking, we can imagine a scenario where each argument is half as important as the last. In that optimistic case, we’ll get a little bit closer with each argument made, and whenever we’re cut off, we’ll be able to put a limit on how wrong our final answer could be.

A second case is if the arguments can be evaluated independently. Sometimes, it’s natural to talk about a decision in terms of its pros and cons. What this amounts to is ignoring the ways these aspects might interact with each other, and just taking into account the total weight for and against the proposition. In these debates — called feature debates with independent evidence — we expect optimal debaters to just bring their strongest arguments to the table. In this case, when we terminate a debate, we can’t say who would ultimately win. After all, the losing debater might always have in reserve a really large number of weak arguments that he hasn’t had a chance to play yet. But we can at least place some limits on where the debate can end up after a finite number more arguments, if the debaters have been playing optimally.

Which of these scenarios describes the most important AI debates that might realistically occur? This is a difficult question that we don’t fully answer. The optimistic cases are pretty restrictive: in realistic debates, we often don’t know when arguments will start to lose their power, except in specific settings, like if we’re running a survey (and each argument is another survey result) or choosing the number of samples to take for a scientific experiment. On the other hand, most realistic debates aren’t as bad as the fully pessimistic case where any new argument can completely overhaul your previous view. Sometimes important moral questions do flip back and forth — in such cases, using AI debate might not be a good idea.

A debate can fail in several other ways. Sometimes, lying might simply be the most convincing strategy, particularly when the truth has a big inferential distance or when

the lie feeds our biases (“Of course the Earth is flat! Wouldn’t things fall off otherwise?”). Even when debates are safe, debates might be slow or unconvincing too often, so people will use unsafe approaches instead. Alternatively, we might accidentally lose the main selling point of debate, which is that each debater wants to point out the mistakes of the opponent. Indeed, we could consider modifications such as rewarding both debaters when both answers seem good or rewarding none of them when the debate is inconclusive. However, such “improvements” introduce unwanted collusion incentives in the spirit of “I won’t tell on you if you won’t tell on me.”.

To understand which debates are useful, we have to consider a bunch of factors that we haven’t modelled yet. The biggest issue that has been raised with us by proponents of debate is that we’ve excluded too many types of arguments. If you’re trying to argue that an image depicts a dog, you’ll usually make claims about medium-sized aspects of the image: “the tail is here”, “the floppy ears are here”, and so on. These arguments directly challenge the opposing arguer, who should either endorse and explain these medium-sized features, or else zoom in to smaller features “this is not a tail because this region is green”, “if this is where ears are supposed to be, what is this eye doing here?”, and so on. By agreeing and disagreeing, and zooming in and out, human debates manage to get to the truth much more efficiently than if they could only reveal individual pixels. Looking into arguments with larger claims is one of our top priorities for taking this model forward.

I (Vojta) am planning to keep working on debate and other AI safety topics over the next twelve months and will be looking to spend most of that time visiting relevant organizations. If you are interested in helping with this, please get in touch.

The paper is available in full here:

Kovařík, Vojtěch, and Ryan, Carey. “(When) Is Truth-telling Favored in AI Debate?.” To appear at SafeAI@AAAI. Preprint available at arXiv:1911.04266 (2019).

Stupidity and Dishonesty Explain Each Other Away

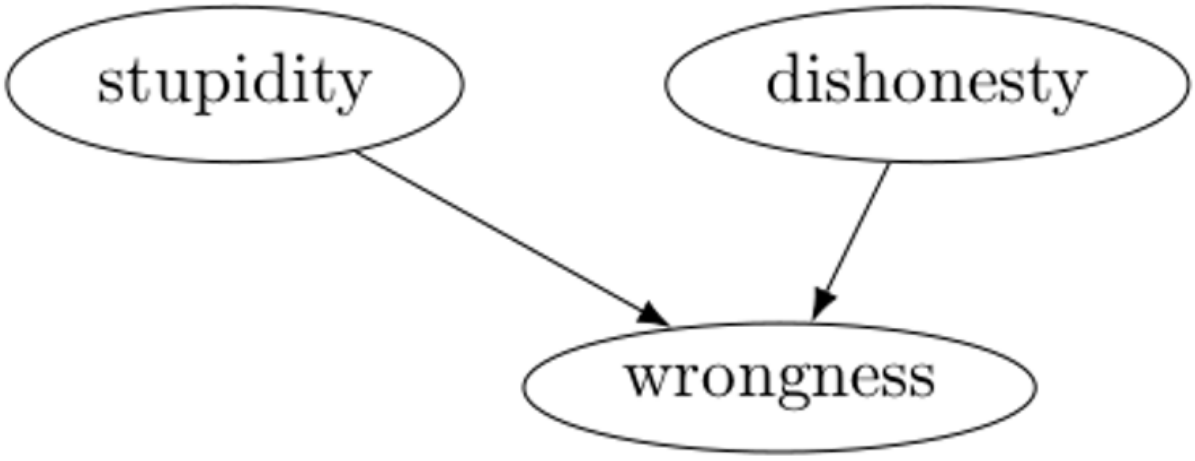
The *explaining-away effect* (or, collider bias; or, Berkson's paradox) is a statistical phenomenon in which statistically independent causes with a common effect become anticorrelated when conditioning on the effect.

In the language of *d*-separation, if you have a causal graph X → Z ← Y, then conditioning on Z unblocks the path between X and Y.

Daphne Koller and Nir Friedman give an example of reasoning about disease etiology: if you have a sore throat and cough, and aren't sure whether you have the flu or mono, you should be relieved to find out it's "just" a flu, because that decreases the probability that you have mono. You *could* be inflected with both the influenza and mononucleosis viruses, but if the flu is completely sufficient to explain your symptoms, there's no *additional* reason to expect mono.[1]

Judea Pearl gives an example of reasoning about a burglar alarm: if your neighbor calls you at your dayjob to tell you that your burglar alarm went off, it could be because of a burglary, or it could have been a false-positive due to a small earthquake. There *could* have been both an earthquake *and* a burglary, but if you get news of an earthquake, you'll stop worrying so much that your stuff got stolen, because the earthquake alone was sufficient to explain the alarm.[2]

Here's another example: if someone you're arguing with is wrong, it could be either because they're just too stupid to get the right answer, or it could be because they're being dishonest—or some combintation of the two, but more of one means that less of the other is required to explain the observation of the person being wrong. As a causal graph—[3]



Notably, the decomposition still works whether you count subconscious motivated reasoning as "stupidity" or "dishonesty". (Needless to say, it's also symmetrical across persons—if *you're* wrong, it could be because *you're* stupid or are being dishonest.)

1. Daphne Koller and Nier Friedman, *Probabilistic Graphical Models: Principles and Techniques*, §3.2.1.2 "Reasoning Patterns" ↩

2. Judea Pearl, *Probabilistic Reasoning in Intelligent Systems*, §2.2.4 "Multiple Causes and 'Explaining Away'" ↩

3. Thanks to Daniel Kumor for example LaTeX code for causal graphs. ↩

Safe exploration and corrigibility

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

EDIT: I now think this post is somewhat confusing and would recommend starting with my more recent post "Exploring safe exploration."

Balancing exploration and exploitation is a classic problem in reinforcement learning. Historically—with approaches such as deep Q learning, for example—exploration is done explicitly via a rule such as ϵ-greedy exploration or Boltzmann exploration. With more modern approaches, however—especially policy gradient approaches like PPO that aren't amenable to something like Boltzmann exploration—the exploration is instead entirely learned, encouraged by some sort of extra term in the loss to implicitly encourage exploratory behavior. This is usually an entropy term, though other more advanced approaches have also been proposed, such as random network distillation in which the agent learns to explore states for which it would have a hard time predicting the output of a random neural network, an approach which was able to set a state of the art on Montezuma's Revenge, a notoriously difficult Atari environment because of how much exploration it requires.

This move to learned exploration has a very interesting and important consequence, however, which is that the safe exploration problem for learned exploration becomes very different. Making ϵ-greedy exploration safe is in some sense quite easy, since the way it explores is totally random. If you assume that the policy without exploration is safe, then for ϵ-greedy exploration to be safe on average, it just needs to be the case that the environment is safe on average, which is just a standard engineering question. With learned exploration, however, this becomes much more complicated—there's no longer a nice "if the non-exploratory policy is safe" assumption that can be used to cleanly subdivide the overall problem of off-distribution safety, since it's just a single, learned policy doing both exploration and exploitation.

First, though, an aside: why is learned exploration so much better? I think the answer lies primarily in the following observation: for most problems, exploration is an instrumental goal, not a terminal one, which means that to do exploration "right" you have to do it in a way which is cognizant of the objective you're trying to optimize for. Boltzmann exploration is better than ϵ-greedy exploration because its exploration is guided by its exploitation—but it's still essentially just adding random jitter to your policy. Fundamentally, though, exploration is about the value of information such that proper exploration requires dynamically balancing the value of information with the value of exploitation. Ideally, in this view, exploration should arise naturally as an instrumental goal of pursuing the given reward function—an agent should instrumentally want to get updated in such a way that causes it to become better at pursuing its current objective.

Except, there's a really serious, major problem with that reasoning: instrumental exploration only cares about the value of information for helping the model to achieve the goal it's learned so far, not for helping it fix its goal to be more aligned with the actual goal.[1] Consider, for instance, my maze example. Instrumental exploration will help the model better explore the larger maze, but it won't help it better figure out

that it's objective of finding the green arrow is misaligned—that is, it won't, for example, lead to the model trying both the green arrow and the end of the maze to see which one is right. Furthermore, because the instrumental exploration actively helps the model explore the larger maze better, it improves the model's capability generalization without also helping its objective generalization, leading to precisely the most worrying case in the maze example. If we think about this problem from a 2D robustness perspective, we can see that what's happening is that instrumental exploration gives us *capability exploration* but not *objective exploration.*

Now, how does this relate to corrigibility? To answer that question, I want to split corrigibility into three different subtypes:

1. **Indifference corrigibility:** An agent is *indifference corrigible* if it is indifferent to modifications made to its goal.
2. **Exploration corrigibility:** An agent is *exploration corrigible* if it actively searches out information to help you correct its goal.
3. **Cooperation corrigibility:** An agent is *cooperation corrigible* if it optimizes under uncertainty over what goal you might want it to have.

Previously, I grouped both of those second two into act-based corrigibility, though recently I've been moving towards thinking that act-based corrigibility isn't as well-defined as I previously thought it was. However, I think the concept of objective exploration lets us disentangle act-based corrigibility. Specifically, I think exploration corrigibility is just indifference corrigibility plus objective exploration, and cooperation corrigibility is just exploration corrigibility plus corrigible alignment.[2] That is, if a model is indifferent to having its objective changed and actively optimizes for the value of information in terms of helping you change its current objective, that gives you exploration corrigibility, and if its objective is also a "pointer" to what you want, then you get cooperation corrigibility. Furthermore, I think this helps solve a lot of the problems I previously had with corrigible alignment, as indifference corrigibility and exploration corrigibility together can help you prevent crystallization of deceptive alignment.

Finally, what does this tell us about safe exploration and how to think about current safe exploration research? Current safe exploration research tends to focus on the avoidance of traps in the environment. Safety Gym, for example, has a variety of different environments containing both goal states that the agent is supposed to reach and unsafe states that the agent is supposed to avoid. One particularly interesting recent work in this domain was Leike et al.'s "Learning human objectives by evaluating hypothetical behaviours," which used human feedback on hypothetical trajectories to learn how to avoid environmental traps. In the context of the capability exploration/objective exploration dichotomy, I think a lot of this work can be viewed as putting a damper on instrumental capability exploration. What's nice about that lens, in my opinion, is that it both makes clear how and why such work is valuable while also demonstrating how much other work there is to be done here. What about objective exploration—how do we do it properly? And do we need measures to put a damper on objective exploration as well? And what about cooperation corrigibility—is the "right" way to put a damper on exploration through constraints or through uncertainty? All of these are questions that I think deserve answers.

---

1. For a mesa-optimizer, this is saying that the mesa-optimizer will only explore to help its current mesa-objective, not to help it fix any misalignment between its mesa-objective and the base objective. ↩

2. Note that this still leaves the question of what exactly indifference corrigibility is unanswered. I think the correct answer to that is *myopia,* which I'll try to say more about in a future post—for this post, though, I just want to focus on the other two types. ↩

ESC Process Notes: Claim Evaluation vs. Syntheses

Forgive me if some of this is repetitive, I can't remember what I've written in which draft and what's actually been published, much less tell what's actually novel. Eventually there will be a polished master post describing my overall note taking method and leaving out most of how it was developed, but it also feels useful to discuss the journey.

When I started taking notes in Roam (a workflowy/wiki hybrid), I would:

1. Create a page for the book (called a Source page), with some information like author and subject (example)
2. Record every claim the book made on that Source page
3. Tag each claim so it got its own page
4. When I investigated a claim, gather evidence from various sources and list it on the claim page, grouped by source

This didn't make sense though: why did some sources get their own page and some a bullet point on a claims page? Why did some claims get their own page and some not? What happened if a piece of evidence was useful in multiple claims?

Around this time I coincidentally had a call with Roam CEO Conor White-Sullivan to demo a bug I thought I had found. There was no bug, I had misremembered the intended behavior, but this meant that he saw my system and couldn't hide his flinch. Aside from wrecking performance, there was no need to give each claim its own page: Roam has block references, so you can point to bullet points, not just pages.

When Conor said this, something clicked. I had already identified one of the problems with epistemic spot checks as being too binary, too focused on evaluating a particular claim or book than building knowledge. The original way of note taking was a continuation of that. What I should be doing was gathering multiple sources, taking notes on equal footing, and then combining them into an actual belief using references to the claims' bullet points. I call that a Synthesis (example). Once I had an actual belief, I could assess the focal claim in context and give it a credence (a slider from 1-10), which could be used to inform my overall assessment of the book.

Sometimes there isn't enough information to create a Synthesis, so something is left as a Question instead (example).

Once I'd proceduralized this a bit, it felt so natural and informative I assumed everyone else would find it similarly so. Finally you didn't have to take my word for what was important- you could see all the evidence I'd gathered and then click through to see the context on anything you thought deserved a closer look. Surely everyone will be overjoyed that I am providing this

Feedback was overwhelming that this was much worse, no one wanted to read my Roam DB, and I should keep presenting evidence linearly.

I refuse to accept that my old way is the *best* way of presenting evidence and conclusions about a book or a claim. It's too linear and contextless. I do accept that

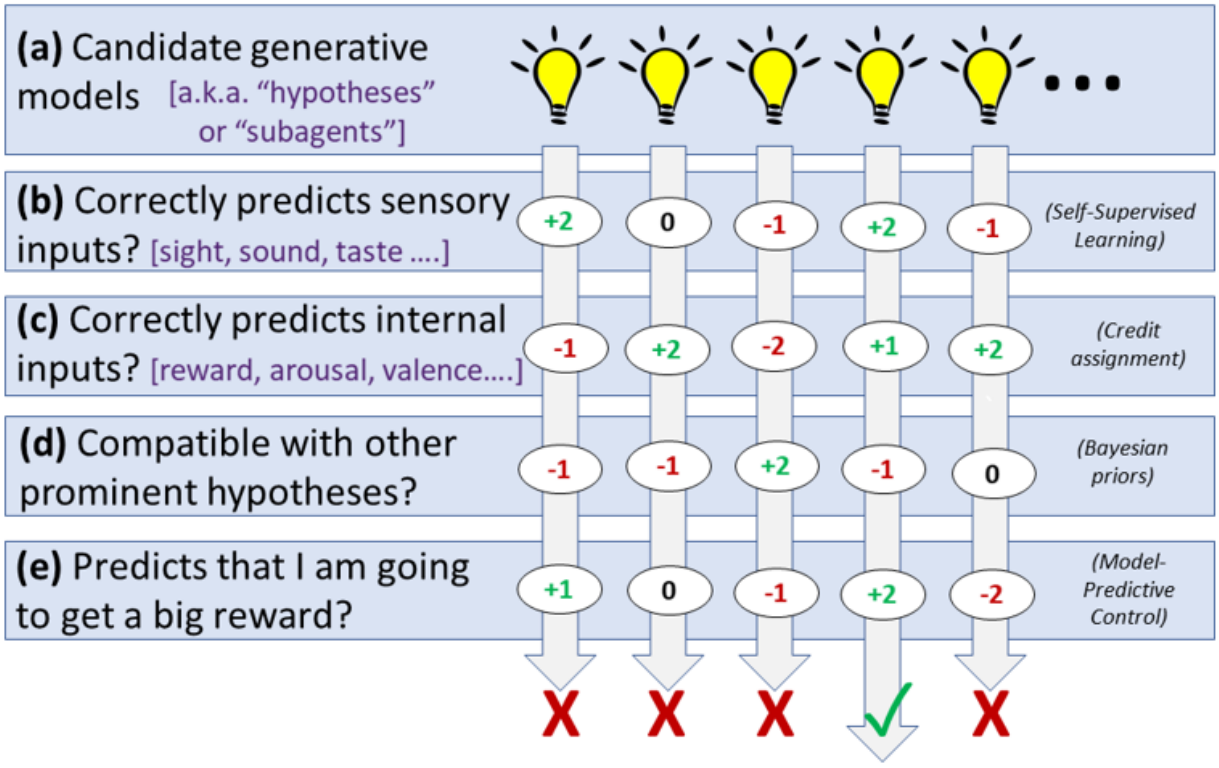
“here’s my Roam have fun” is worse. Part of my current project is to identify a third way that shares the information I want to in a way that is actually readable.

Predictive coding = RL + SL + Bayes + MPC

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(Update much later (2021-06): I put in some minor updates and retractions below. Also, if I were writing this today I would probably have framed it as "here's how I disagree with predictive coding" rather than "here's a version of predictive coding that I like". Also, wherever you see "predictive coding" below, read "predictive processing"—at the time that I wrote this, I didn't understand the difference, and used the wrong one. By the way, I wrote this near the very earliest stages of my learning about the brain. Having learned a lot more since then, I guess I'm not too embarrassed by what I wrote here, but there are certainly lots of little things I would describe differently now.)

I was confused and skeptical for quite a while about some aspects of predictive coding—and it's possible I'm still confused—but after reading a number of different perspectives on brain algorithms, the following picture popped into my head and I felt much better:



This is supposed to be a high-level perspective on how the neocortex[1] builds a predictive world-model and uses it to choose appropriate actions. (a) We generate a bunch of generative models in parallel, which make predictions about what's going on and what's going to happen next, including what *I* am doing and will do next (i.e., my plans). The models gain "prominence" by (b) correctly predicting upcoming sensory inputs; (c) correctly predicting other types of input information coming into the

neocortex like tiredness, hormonal signals, hunger, warmth, pain, pleasure, reward, and so on; (d) being compatible with other already-prominent models; (e) predicting that a large reward signal is coming, which as discussed in my later article Inner Alignment in the Brain, includes things like predicting that my goals will be fulfilled with minimal effort, I'll be eating soon if I'm hungry, I'll be sleeping soon if I'm tired, I'll avoid pain, and so on. Whatever candidate generative model winds up the most "prominent" wins, and determines my beliefs and actions going forward.

(Note on terminology: I'm calling these things "generative models" in all cases. Kurzweil calls them "patterns". They're also sometimes called "hypotheses", especially in the context of passive observation (e.g. "that thing I see is a bouncy ball"). Or they're called "subagents"[2], especially in the context of self-prediction (e.g. "I am about to eat").

Before we get to details, I need to apologize for the picture being misleading:

- First, I drew (b,c,d,e) as happening *after* (a), but really some of these (especially (d) I think) work by affecting which models get considered in the first place. (More generally, I do not want to imply that a,b,c,d,e correspond to exactly five distinct neural mechanisms, or anything like that. I'm just going for a functional perspective in this post.)

- Second (and relatedly), I depicted it as if we simply add up points for (b-e), but it's certainly not linear like that. I think at least some of the considerations effectively get vetoes. For example, we don't generally see a situation where (e) is so positive that it simply outvotes (b-d), and thus we spend all day checking our wallet expecting to find it magically filled with crisp $1000 bills. (Much more about wishful thinking below.) (**Update much later:** Yeah, I think (e) is implemented by a very different mechanism involving different parts of the brain than (a-d)—see here—and indeed I think that (b-d) more-or-less get a veto.)

- Third, at the bottom I drew one generative model being the "winner". Things like action plans and conscious attention do in fact have a winner-take-all dynamic because, for example, we don't want to be sending out muscle commands for both walking and sitting simultaneously.[3] But in general, lower-ranked models are not thrown out; they linger, with their prominence growing or shrinking as more evidence comes in.

Anyway, the picture above tells a nice story:

(b) is self-supervised learning[4], i.e. learning from prediction. Process (b) simply votes against generative models when they make incorrect predictions. This process is where we get the vast majority of the information content we need to build a good predictive world-model. Note that there doesn't seem to be any strong difference in the brain between (i) actual experiences, (ii) memory recall, and (iii) imagination—process (b) will vote for or against models when presented with any of those three types of "evidence". (I think they vote much more strongly in case (i) though.)

(c) is credit assignment, i.e. learning what aspects of the world cause good or bad things to happen to us, so that we can make good decisions. Each generative model makes claims about what is the cause of subcortex-provided informational signals (analogous to "reward" in RL)—information signals that say we're in pain, or eating yummy food, or exhausted, or scared, etc. These claims cash out as predictions that can prove right or wrong, thus either supporting or casting doubt on that model. Thus

our internal models say that "cookies are yummy", corresponding to a prediction that, if we eat one, we'll get a "yummy" signal from some ancient reptilian part of our brain.

(d) is Bayesian priors. I doubt we do Bayesian updating in a literal mathematical sense, but we certainly do incorporate prior beliefs into our interpretation of new evidence. I'm claiming that the mechanism for this is "models gain prominence by being compatible with already-prominent models". What is an "already-prominent model"? One that has previously been successful in this same process I'm describing here, especially if in similar contexts, and super-especially if in the immediate past. Such models function as our priors. And what does it mean for a new model to be "compatible" with these prior models? Well, a critical fact about these models is that they snap together like Legos, allowing hierarchies, recursion, composition, analogies, casual relationships, and so on. (Thus, I've never seen a rubber wine glass, but I can easily create a mental model of one by gluing together some of my rubber-related generative models with some of my wine-glass-related generative models.) Over time we build up these super-complicated and intricate Rube Goldberg models, approximately describing our even-more-complicated world. I think a new model is "compatible" with a prior one when (1) the new model is almost the same as the prior model apart from just one or two simple edits, like adding a new bridging connection to a different already-known model; and/or (2) when the new model doesn't make predictions counter to the prior one, at least not in areas where the prior one is very precise and confident.[5] Something like that anyway, I think...

(e) is Model-Predictive Control. If we're hungry, we give extra points to a generative model that says we're about to get up and eat a snack, and so on. This works in tandem with credit assignment (process (c)), so if we have a prominent model that giving speeches will lead to embarrassment, then we will subtract points from a new model that we will give a speech tomorrow, and we don't need to run the model all the way through to the part where we get embarrassed. I like Kaj Sotala's description here: "mental representations...[are] imbued with a context-sensitive affective gloss"—in this case, the mental representation of "I will give a speech" is infused with a negative "will lead to embarrassment" vibe, and models lose points for containing that vibe. It's context-sensitive because, for example, the "will lead to feeling cold" vibe could be either favorable or unfavorable depending on our current body temperature. Anyway, this framing makes a lot of sense for choosing actions, and amounts to using control theory to satisfy our innate drives. But if we're just passively observing the world, this framework is kinda problematic...

(e) is also wishful thinking. Let's say someone gives us an unmarked box with a surprise gift inside. According to the role of (e) in the picture I drew, if we receive the box when we're hungry, we should expect to find food in the box, and if we receive the box when we're in a loud room, we should expect to find earplugs in the box, etc. Well, that's not right. Wishful thinking does exist, but it doesn't seem so inevitable and ubiquitous as to deserve a seat right near the heart of human cognition. Well, one option is to declare that one of the core ideas of Predictive Coding theory—unifying world-modeling and action-selection within the same computational architecture—is baloney. But I don't think that's the right answer. I think a better approach is to posit that (b-d) are actually pretty restrictive in practice, leaving (e) mainly as a comparitively weak force that can be a tiebreaker between equally plausible models. In other words, passive observers rarely if ever come across multiple equally plausible models for what's going on and what will happen next; it would require a big coincidence to balance the scales so precisely. But when we make predictions about what *we ourselves* will do, that aspect of the prediction is a self-fulfilling prophecy, so we *routinely* have equally plausible models...and then (e) can step in and break the tie.

More general statement of situations where (e) plays a big role: Maybe "self-fulfilling" is not quite the right terminology for when (e) is important; it's more like "(e) is most important in situations where lots of models are all incompatible, yet where processes (b,c,d) never get evidence to support one model over the others." So (e) is central in choosing action-selection models, since these are self-fulfilling, but (e) plays a relatively minor role in passive observation of the world, since there we have (b,c) keeping us anchored to reality (but (e) does play an occasional role on the margins, and we call it "wishful thinking"). (e) is also important because (b,c,d) by themselves leave this whole process highly under-determined: walking in a forest, your brain can build a better predictive model of trees, of clouds, of rocks, or of nothing at all; (e) is a guiding force that, over time, keeps us on track building useful models for our ecological niche.

One more example where (e) is important: confabulation, rationalization, etc. Here's an example: I reach out to grab Emma's unattended lollipop because I'm hungry and callous, but then I immediately think of an alternate model, in which I am taking the lollipop because she probably wants me to have it. The second model gets extra points from the (e) process, because I have an innate drive to conform to social norms, be well-regarded and well-liked, etc. Thus the second model beats the truthful model (that I grabbed the lollipop because I was hungry and callous). Why can't the (b) process detect and destroy this lie? Because all that (b) has to go on is my own memory, and perniciously, the second model has some influence over how I form the *memory* of grabbing the lollipop. It has covered its tracks! Sneaky! So I can keep doing this kind of thing for years, and the (b) process will never be able to detect and kill this habit of thought. Thus, rationalization winds up more like action selection, and less like wishful thinking, in that it is pretty much ubiquitous and central to cognition.[6]

Side note: Should we lump (d-e) together? When people describe Predictive Coding theory, they tend to lump (d-e) together, to say things like "We have a prior that, when we're hungry, we're going to eat soon." I am proposing that this lumping is not merely bad pedagogy, but is actually conflating together two different things: (d) and (e) are *not* inextricably unified into a single computational mechanism. (I don't think the previous sentence is obvious, and I'm not super-confident about it.) (Update later on: yeah they're definitely different, see here.) By the same token, I'm uncomfortable saying that minimizing prediction error is a fundamental operating principle of the brain; I want to say that processes (a-e) are fundamental, and minimizing prediction error is something that arguably happens as an incidental side-effect.

Well, that's my story, it seems to basically makes sense, but that could just be my (e) wishful thinking and (e) rationalization talking. :-)

(Update May 2020: The traditional RL view would be that there's a 1-dimensional signal called "reward" that drives process (e). When I first wrote this, I was still confused about whether that was the right way to think about the brain, and thus I largely avoided the term "reward" in favor of less specific things. After thinking about it more—see inner alignment in the brain, I am now fully on board with the traditional RL view; process (e) is just "We give extra points to models that predict that a large reward signal is coming". Also, I replaced "hypotheses" with "generative models" throughout, I think it's a better terminology.) (Update June 2021: Better discussion of "reward" here.)

---

1. The neocortex is 75% of the human brain by weight, and centrally involved in pretty much every aspect of human intelligence (in partnership with the thalamus

and hippocampus). More about the neocortex in my previous post ↩

2. See Jan Kulveit's Multi-agent predictive minds and AI alignment, Kaj Sotala's Multiagent models of mind sequence, or of course Marvin Minsky and many others. ↩

3. I described conscious attention and action plans as "winner-take-all" in the competition among models, but I think it's somewhat more complicated and subtle than that. I also think that picking a winner is not a separate mechanism from (b,c,d,e), or at least not entirely separate. This is a long story that's outside the scope of this post. ↩

4. I have a brief intro to self-supervised learning at the beginning of Self-Supervised Learning and AGI Safety ↩

5. Note that my picture at the top shows parallel processing of models, but that's not quite right; in order to see whether two prominent models are making contradictory predictions, we need to exchange information between them. ↩

6. See *The Elephant in the Brain* etc. ↩

Might humans not be the most intelligent animals?

The idea that humans are the most intelligent animals on Earth appears patently obvious to a lot of people. And to a large extent, I agree with this intuition. Humans clearly dominate the world in technological innovation, control, communication, and coordination.

However, more recently I have been acquainted with some evidence that the proposition is not actually true, or at the very least is non-obvious. The conundrum arises when we distinguish *raw innovative capability*, from the ability to *efficiently process culture*. I'll explain the basic case here.

Robin Hanson has sometimes pointed to the accumulation of culture as the relevant separation between humans and other animals. Under this view, the reason why humans are able to dominate the world with technology has less to due with our raw innovation abilities, and more to do with the fact that we can efficiently process accumulated cultural information and build on it.

If the reason for our technological dominance is due to raw innovative ability, we might expect a more discontinuous jump in capabilities for AI, since there was a sharp change between "unable to innovate" and "able to innovate" during evolutionary history, which might have been due to some key architectural tweak. We might therefore expect that our AIs will experience the same jump after receiving the same tweak.

If the reason for our technological dominance is due to our ability to process culture, however, then the case for a discontinuous jump in capabilities is weaker. This is because our AI systems can already process culture somewhat efficiently right now (see GPT-2) and there doesn't seem like a hard separation between "being able to process culture inefficiently" and "able to process culture efficiently" other than the initial jump from not being able to do it at all, which we have already passed. Therefore, our current systems are currently bottlenecked on some metric which is more continuous.

The evidence for the cultural accumulation view comes from a few lines of inquiry, such as

- Feral children lack an innovating culture to learn from, and are correspondingly 'unintelligent' from our point of view. This is despite the fact that feral children hold more-or-less the exact same raw innovative capabilities as normal humans. For Genie, one feral child, "Doctors found it extremely difficult to test or estimate Genie's mental age or any of her cognitive abilities, but on two attempts they found Genie scored at the level of a 13-month-old."
- Books like The Secret of Our Success document remarkable instances of raw innovative capability being much less useful than ability to use culture. See the SlateStarCodex review here.
- Evolutionarily, increases in raw innovative capability may universally aid reproductive capability for any animal, whereas increases in ability to efficiently process culture would only help in a species that developed culture to begin with. It's no suprise therefore, that after humans developed language, we

quickly developed the capability to process cultural information, whereas other animals can't learn much from our civilization. However, this doesn't necessarily mean they are less innovative in general.

Under the view that our dominance is due to cultural accumulation, we would expect that there are some animals that are more intelligent than humans in the sense of raw innovative ability. The reason is that it would be surprising *a priori* for us to be the most intelligent, unless we had reason to believe so anthropically. However, if culture is the reason why we dominate, then the anthropic argument here is weaker.

We *do* see some initial signs that humans might not be the most intelligent species on Earth in the innovative sense. For instance,

- Although humans have the highest encephalization quotient, we *don't* have the most neurons, or even the most neurons in our forebrain.
- Some animals have easily measurable cognitive capabilities which surpass ours. One example is the chimpanzee, which may have better working memory. Edit: Now I'm not sure whether this fact is correct. See comments below for more discussion.

If humans are the most intelligent in the sense of having the best raw innovative abilities, this hypothesis should be testable by administering a battery of cognitive tests to animals. However, this is made difficult due to a number of factors.

First, most intelligence tests that humans take rely on their ability to process language, disqualifying other animals from the start.

Second, humans might be biased towards administering tests about things that we are good at, since we might not even be aware of the type of cognitive abilities we score poorly on. This may have the effect of proving that humans are superior in intelligence, but only on the limited subset of tests that we used.

Third, if we use current human intelligence tests (or anything like them), the following argument arises. Computers can already outperform humans at some tasks that intelligence tests measure, such as memory, but this doesn't mean that computers already have a better ability innovate. We would need to test something that we felt confident accurately indicated innovative ability.

Since I'm mostly interested in this question because of its implications for AI takeoff speeds, I'd want to know what type of things are most useful for developing technology, and see if we can see the same abilities in animals *sans* cultural accumulation modules.

Templates and videos for doing annual and daily reviews

Many individuals and organisations in the rationality community do an "annual review" (a review of the past year and update of future plans).

In the past I couldn't find a ready made document to complete to undertake my own annual review. Alex Vermeer's stuff was good but not quite ready to use in my opinion.

With that in mind, I am sharing an annual review/planner template that I made in google sheets.

It incorporates Alex Vermeer's questions and a few of my own. I also added some sheets to add goals and track ‘bugs’.

Please feel free to copy, share and modify it.

Annual review template
http://bit.ly/2MhHc0W
https://lnkd.in/guY8EDW

Over the past year I have also used a 'daily tracker' sheet to help identify and understand patterns of mood and behaviour and to help me to implement and track my goals

Daily tracker template and video
http://bit.ly/3rOscb9
https://lnkd.in/gi6aEQk

Inductive biases stick around

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is a follow-up to Understanding "Deep Double Descent".

I was talking to Rohin at NeurIPS about my post on double descent, and he asked the very reasonable question of why exactly I think double descent is so important. I realized that I hadn't fully explained that in my previous post, so the goal of this post is to further address the question of why you should care about double descent from an AI safety standpoint. This post assumes you've read my Understanding "Deep Double Descent" post, so you should read that first before reading this if you haven't already.

Specifically, I think double descent demonstrates the in my opinion very important yet counterintuitive result that larger models can actually be *simpler* than smaller models. On its face, this sounds somewhat crazy—how can a model with more parameters be simpler? But in fact I think this is just a very straightforward consequence of double descent: in the double descent paradigm, larger models with zero training error generalize better than smaller models with zero training error because they do better on SGD's inductive biases. And if you buy that SGD's inductive biases are approximately simplicity, that means that larger models with zero training error are simpler than smaller models with zero training error.

Obviously, larger models do have more parameters than smaller ones, so if that's your measure of simplicity, larger models will always be more complicated, but for other measures of simplicity that's not necessarily the case. For example, it could hypothetically be the case that larger models have lower Kolmogorov complexity. Though I don't actually think that's true in the case of K-complexity, I think that's only for the boring reason that model weights have a lot of noise. If you had a way of somehow only counting the "essential complexity," I suspect larger models would actually have lower K-complexity.

Really, what I'm trying to do here is dispel what I see as the myth that as ML models get more powerful simplicity will stop mattering for them. In a Bayesian setting, it is a fact that the impact of your prior on your posterior (for those regions where your prior is non-zero[1]) becomes negligible as you update on more and more data. I have sometimes heard it claimed that as a consequence of this result, as we move to doing machine learning with ever larger datasets and ever bigger models, the impact of our training processes' inductive biases will become negligible. However, I think that's quite wrong, and I think double descent does a good job of showing why, because all of the performance gains you get past the interpolation threshold are coming from your implicit prior.[2] Thus, if you suspect modern ML to mostly be in that regime, what will matter in terms of which techniques beat out other techniques is how good they are at compressing their data into the "actually simplest" model that fits it.

Furthermore, even just from the simple Bayesian perspective, I suspect you can still get double descent. For example, suppose your training process looks like the following: you have some hypothesis class that keeps getting larger as you train and at each time step you select the best a posteriori hypothesis. I think that this setup will naturally yield a double descent for noisy data: first you get a "likelihood descent"

as you get hypotheses with greater and greater likelihood, but then you start overfitting to noise in your data as you get close to the interpolation threshold. Past the interpolation threshold, however, you get a second “prior descent” where you're selecting hypotheses with greater and greater prior probability rather than greater and greater likelihood. I think this is a good model for how modern machine learning works and what double descent is doing.

All of this is only for models with zero training error, however—before you reach zero training error larger models can certainly have more essential complexity than smaller ones. That being said, if you don't do very many steps of training then your inductive biases will also matter a lot because you haven't updated that much on your data yet. In the double descent framework, the only region where your inductive biases don't matter very much is right on the interpolation threshold—before the interpolation threshold or past it they should still be quite relevant.

Why does any of this matter from a safety perspective, though? Ever since I read Belkin et al. I've had double descent as part of my talk version of “Risks from Learned Optimization” because I think it addresses a pretty important part of the story for mesa-optimization. That is, mesa-optimizers are simple, compressed policies—but as ML moves to larger and larger models, why should that matter? The answer, I think, is that larger models can generalize better not just by fitting the data better, but also by being simpler.[3]

---

1. Negating the impact of the prior not having support over some hypotheses requires realizability (see Embedded World-Models). ↩

2. Note that double descent happens even without explicit regularization, so the prior we're talking about here is the implicit one imposed by the architecture you've chosen and the fact that you're training it via SGD. ↩

3. Which is exactly what you should expect if you think Occam's razor is the right prior: if two hypotheses have the same likelihood but one generalizes better, according to Occam's razor it must be because it's simpler. ↩

Long-lasting Effects of Suspensions?

I recently read "The School to Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime" (Bacher-Hicks et. al. 2019, pdf, via Rob Wiblin) which argues that a policy of suspending kids in middle school leads to more crime as an adult.

Specifically, they found that after controlling for a bunch of things, students who attended schools with 0.38 more suspensions per student per year were 20% more likely to be jailed as adults:

> A one standard deviation increase in the estimated school effect increases the average annual number of days suspended per year by 0.38, a 16 percent increase. ... We find that students assigned a school with a 1 standard deviation higher suspension effect are about 3.2 percentage points more likely to have ever been arrested and 2.5 percentage points more likely to have ever been incarcerated, which correspond to an increase of 17 percent and 20 percent of their respective sample means.

This is a very surprising outcome: from a single suspension in three years they're 20% more likely to go to jail?

The authors look at the Charlotte-Mecklenburg school district, was ordered by the court to desegregate in the 1970s. In the early 2000s the court was convinced that busing wasn't needed anymore, and the district implemented a "School Choice Plan" for beginning of the 2002 year. Students were massively shuffled between the schools and, while this was generally not randomized, the authors describe it as a "natural experiment".

The idea is that if a student moves from school A to school B and you know how often students were suspended at both schools, then you can look at differences later in life and see how much of that is explained by the difference in suspension rates. They note:

> A key concern is whether variation in "strictness" across schools arises from policy choices made by administrators versus underlying variation in school context. Our use of the boundary change partly addresses this concern, because we show that schools' conditional suspension rates remain highly correlated through the year of the boundary change, which provides a very large shock to school context. We also show that school effects on suspensions are unrelated to other measures of school quality, such as achievement growth, teacher turnover and peer characteristics.

And:

> We also test directly for the importance of administrative discretion by exploiting a second source of variation - principal movement across schools. We find that conditional suspension rates change substantially when new principals enter and exit, and that principals' effects on suspensions in other schools predict suspensions in their current schools. While we ultimately cannot directly connect our estimates to concrete policy changes, the balance of the evidence suggests that principals exert considerable influence over school discipline and that our results cannot be explained by context alone.

Here's an alternative model that fits this data, which I think is much more plausible. Grant that differences in conditional suspension rates are mostly caused by administrators' policy preferences, but figure that student-specific effects still play a role. Then figure there are differences between the schools' cultures or populations that are not captured by the controls, and that these differences cause both (a) differences in the student-specific portion of the suspension rate and (b) differences in adult incarceration rates. If suspensions themselves had no effects we would still see suspension appearing to cause higher incarceration rates later in life.

They refer to movement of principals between schools, which offers a way to test this. Classify principals by their suspension rates, and look at schools that had a principal change while keeping the student body constant. Ideally do this in school districts where the parents don't have a choice about which school their children attend, to remove the risk that the student population before and after the principal change is different in aggregate. Compare the adult outcomes of students just before the change to ones just after. While a principal could affect school culture in multiple ways and we would attribute the entire effect to suspensions, this would at least let us check whether the differences are coming from the administration.

This sort of problem, where there's some kind of effect outside what you control for, which leads you to find causation where there may not be any, is a major issue for value-added models (VAM) in general. "Do Value Added Models Add Value?" (Rothstein 2010, pdf) and "Teacher Effects on Student Achievement and Height" (Bitler et. al. 2019, pdf) are two good papers on this. The first shows that a VAM approach yields higher grades in later years causing higher grades in earlier years, while the second shows the same for teachers causing their students to be taller.

I continue to think we put way too much stock in complex correlational studies, but Bacher-Hicks is an illustration of the way the "natural experiment" label can be used even for things that aren't very experiment-like. It's not a coincidence that at my day job, with lots of money on the line, we run extensive randomized controlled trials and almost never make decisions based on correlational evidence. I would like to see a lot more actual randomization in things like which teachers or schools people are assigned to; this would be very helpful for understanding what actually has what effects.

What determines the balance between intelligence signaling and virtue signaling?

Lately I've come to think of human civilization as largely built on the backs of intelligence and virtue signaling. In other words, civilization depends very much on the positive side effects of (not necessarily conscious) intelligence and virtue signaling, as channeled by various institutions. As evolutionary psychologist Geoffrey Miller says, "it's all signaling all the way down."

A question I'm trying to figure out now is, what determines the relative proportions of intelligence vs virtue signaling? (Miller argued that intelligence signaling can be considered a kind of virtue signaling, but that seems debatable to me, and in any case, for ease of discussion I'll use "virtue signaling" to mean "other kinds of virtue signaling besides intelligence signaling".) It seems that if you get too much of one type of signaling versus the other, things can go horribly wrong (the link is to Gwern's awesome review/summary of a book about the Cultural Revolution). We're seeing this more and more in Western societies, in places like journalism, academia, government, education, and even business. But what's causing this?

One theory is that Twitter with its character limit, and social media and shorter attention spans in general, have made it much easier to do virtue signaling relative to intelligence signaling. But this seems too simplistic and there has to be more to it, even if it is part of the explanation.

Another idea is that intelligence is valued more when a society feels threatened by an outside force, for which they need competent people to protect themselves from. US policy changes after Sputnik is a good example of this. This may also explain why intelligence signaling continues to dominate or at least is not dominated by virtue signaling in the rationalist and EA communities (i.e., we're really worried about the threat from Unfriendly AI).

Does anyone have other ideas, or have seen more systematic research into this question?

Once we understand the above, here are some followup questions: Is the trend towards more virtue signaling at the expense of intelligence signaling likely to reverse itself? How bad can things get, realistically, if it doesn't? Is there anything we can or should do about the problem? How can we at least protect our own communities from runaway virtue signaling? (The recent calls against appeals to consequences make more sense to me now, given this framing, but I still think they may err too much in the other direction.)

PS, it was interesting to read this in Miller's latest book Virtue Signaling:

> Where does the term 'virtue signaling' come from? Some say it goes back to 2015, when British journalist/author James Bartholomew wrote a brilliant piece for The Spectator called 'The awful rise of 'virtue signaling.'' Some say it goes back to the Rationalist blog 'LessWrong,' which was using the term at least as far back as 2013. Even before that, many folks in the Rationalist and Effective Altruism

> subcultures were aware of how signaling theory explains a lot of ideological behavior, and how signaling can undermine the rationality of political discussion.

I didn't know that "virtue signaling" was first coined (or at least used in writing) on LessWrong. Unfortunately, from a search, it doesn't seem like there was substantial discussion around this term. Signaling in general was much discussed on LessWrong and OvercomingBias, but I find myself still updating towards it being more important than I had realized.