

Best of LessWrong: August 2021

1. [What 2026 looks like](#)
2. [How To Write Quickly While Maintaining Epistemic Rigor](#)
3. [Gravity Turn](#)
4. [Can you control the past?](#)
5. [Outline of Galef's "Scout Mindset"](#)
6. [Curing insanity with malaria](#)
7. [The Death of Behavioral Economics](#)
8. [Welcome & FAQ!](#)
9. [Against "blankfaces"](#)
10. [Eight Hundred Slightly Poisoned Word Games](#)
11. [The Future: Where are the Colors and the Sports?](#)
12. [Coase's "Nature of the Firm" on Polyamory](#)
13. [Modelling Transformative AI Risks \(MTAIR\) Project: Introduction](#)
14. [How DeepMind's Generally Capable Agents Were Trained](#)
15. [Covid 8/12: The Worst Is Over](#)
16. [Analysis of World Records in Speedrunning \[LINKPOST\]](#)
17. [AI Safety Papers: An App for the TAI Safety Database](#)
18. [Covid 8/26: Full Vaccine Approval](#)
19. [Automating Auditing: An ambitious concrete technical research proposal](#)
20. [We need a new philosophy of progress](#)
21. [Covid 8/5: Much Ado About Nothing](#)
22. [A short introduction to machine learning](#)
23. [Provide feedback on Open Philanthropy's AI alignment RFP](#)
24. [The Codex Skeptic FAQ](#)
25. [When Programmers Don't Understand Code, Don't Blame The User](#)
26. [Implicature Conflation](#)
27. [A Response to A Contamination Theory of the Obesity Epidemic](#)
28. [COVID/Delta advice I'm currently giving to friends](#)
29. [Staying Grounded](#)
30. [Covid 8/19: Cracking the Booster](#)
31. [How to turn money into AI safety?](#)
32. [When Most VNM-Coherent Preference Orderings Have Convergent Instrumental Incentives](#)
33. [Training My Friend to Cook](#)
34. [What fraction of breakthrough COVID cases are attributable to low antibody count?](#)
35. [Analogies and General Priors on Intelligence](#)
36. [Garrabrant and Shah on human modeling in AGI](#)
37. [Perhaps vastly more people should be on FDA-approved weight loss medication](#)
38. [Value loading in the human brain: a worked example](#)
39. [The two-headed bacterium](#)
40. [Mildly against COVID risk budgets](#)
41. [Framing Practicum: Stable Equilibrium](#)
42. [Framing Practicum: Incentive](#)
43. [Humanity is Winning the Fight Against Infectious Disease](#)
44. [AI Risk for Epistemic Minimalists](#)
45. [Information Assets](#)
46. [OpenAI Codex: First Impressions](#)
47. [The Myth of the Myth of the Lone Genius](#)
48. [Applications for Deconfusing Goal-Directedness](#)
49. [Power vs Precision](#)
50. [Research agenda update](#)

Best of LessWrong: August 2021

1. [What 2026 looks like](#)
2. [How To Write Quickly While Maintaining Epistemic Rigor](#)
3. [Gravity Turn](#)
4. [Can you control the past?](#)
5. [Outline of Galef's "Scout Mindset"](#)
6. [Curing insanity with malaria](#)
7. [The Death of Behavioral Economics](#)
8. [Welcome & FAQ!](#)
9. [Against "blankfaces"](#)
10. [Eight Hundred Slightly Poisoned Word Games](#)
11. [The Future: Where are the Colors and the Sports?](#)
12. [Coase's "Nature of the Firm" on Polyamory](#)
13. [Modelling Transformative AI Risks \(MTAIR\) Project: Introduction](#)
14. [How DeepMind's Generally Capable Agents Were Trained](#)
15. [Covid 8/12: The Worst Is Over](#)
16. [Analysis of World Records in Speedrunning \[LINKPOST\]](#)
17. [AI Safety Papers: An App for the TAI Safety Database](#)
18. [Covid 8/26: Full Vaccine Approval](#)
19. [Automating Auditing: An ambitious concrete technical research proposal](#)
20. [We need a new philosophy of progress](#)
21. [Covid 8/5: Much Ado About Nothing](#)
22. [A short introduction to machine learning](#)
23. [Provide feedback on Open Philanthropy's AI alignment RFP](#)
24. [The Codex Skeptic FAQ](#)
25. [When Programmers Don't Understand Code, Don't Blame The User](#)
26. [Implicature Conflation](#)
27. [A Response to A Contamination Theory of the Obesity Epidemic](#)
28. [COVID/Delta advice I'm currently giving to friends](#)
29. [Staying Grounded](#)
30. [Covid 8/19: Cracking the Booster](#)
31. [How to turn money into AI safety?](#)
32. [When Most VNM-Coherent Preference Orderings Have Convergent Instrumental Incentives](#)
33. [Training My Friend to Cook](#)
34. [What fraction of breakthrough COVID cases are attributable to low antibody count?](#)
35. [Analogies and General Priors on Intelligence](#)
36. [Garrabrant and Shah on human modeling in AGI](#)
37. [Perhaps vastly more people should be on FDA-approved weight loss medication](#)
38. [Value loading in the human brain: a worked example](#)
39. [The two-headed bacterium](#)
40. [Mildly against COVID risk budgets](#)
41. [Framing Practicum: Stable Equilibrium](#)
42. [Framing Practicum: Incentive](#)
43. [Humanity is Winning the Fight Against Infectious Disease](#)
44. [AI Risk for Epistemic Minimalists](#)
45. [Information Assets](#)
46. [OpenAI Codex: First Impressions](#)
47. [The Myth of the Myth of the Lone Genius](#)

48. [Applications for Deconfusing Goal-Directedness](#)
49. [Power vs Precision](#)
50. [Research agenda update](#)

What 2026 looks like

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

This was written for the [Vignettes Workshop](#).^[1] The goal is to write out a **detailed** future history (“trajectory”) that is as realistic (to me) as I can currently manage, i.e. I’m not aware of any alternative trajectory that is similarly detailed and clearly **more** plausible to me. The methodology is roughly: Write a future history of 2022. Condition on it, and write a future history of 2023. Repeat for 2024, 2025, etc. (I’m posting 2022-2026 now so I can get feedback that will help me write 2027+. I intend to keep writing until the story reaches singularity/extinction/utopia/etc.)

What’s the point of doing this? Well, there are a couple of reasons:

- Sometimes attempting to write down a concrete example causes you to learn things, e.g. that a possibility is more or less plausible than you thought.
- Most serious conversation about the future takes place at a high level of abstraction, talking about e.g. GDP acceleration, timelines until TAI is affordable, multipolar vs. unipolar takeoff... vignettes are a neglected complementary approach worth exploring.
- Most stories are written backwards. The author begins with some idea of how it will end, and arranges the story to achieve that ending. Reality, by contrast, proceeds from past to future. It isn’t trying to entertain anyone or prove a point in an argument.
- Anecdotally, various people seem to have found Paul Christiano’s “tales of doom” stories helpful, and relative to typical discussions those stories are quite close to what we want. (I still think a bit more detail would be good — e.g. Paul’s stories don’t give dates, or durations, or any numbers at all really.)^[2]
- “I want someone to ... write a trajectory for how AI goes down, that is really specific about what the world GDP is in every one of the years from now until insane intelligence explosion. And just write down what the world is like in each of those years because I don’t know how to write an internally consistent, plausible trajectory. I don’t know how to write even one of those for anything except a ridiculously fast takeoff.” --[Buck Shlegeris](#)

This vignette was hard to write. To achieve the desired level of detail I had to make a bunch of stuff up, but in order to be realistic I had to constantly ask “but actually though, what would really happen in this situation?” which made it painfully obvious how little I know about the future. There are numerous points where I had to conclude “Well, this does seem implausible, but I can’t think of anything more plausible at the moment and I need to move on.” I fully expect the actual world to diverge quickly from the trajectory laid out here. Let anyone who (with the benefit of hindsight) claims this divergence as evidence against my judgment prove it by exhibiting a vignette/trajectory they themselves wrote in 2021. If it maintains a similar level of detail (and thus sticks its neck out just as much) while being more accurate, I bow deeply in respect!

I hope this inspires other people to write more vignettes soon. We at the [Center on Long-Term Risk](#) would like to have a collection to use for strategy discussions. Let me know if you’d like to do this, and I can give you advice & encouragement! I’d be happy to run another workshop.

2022

GPT-3 is finally obsolete. OpenAI, Google, Facebook, and DeepMind all have gigantic multimodal transformers, similar in size to GPT-3 but trained on images, video, maybe audio too, and generally higher-quality data.

Not only that, but they are now typically fine-tuned in various ways--for example, to answer questions correctly, or produce engaging conversation as a chatbot.

The chatbots are fun to talk to but erratic and ultimately considered shallow by intellectuals. They aren't particularly useful for anything super important, though there are a few applications. At any rate people are willing to pay for them since it's fun.

[EDIT: The day after posting this, it has come to my attention that [in China in 2021 the market for chatbots is \\$420M/year, and there are 10M active users. This article](#) claims the global market is around \$2B/year in 2021 and is projected to grow around 30%/year. I predict it will grow faster. NEW EDIT: See also [xiaoice](#).]

The first prompt programming libraries start to develop, along with the first [bureaucracies](#).^[3] For example: People are dreaming of general-purpose AI assistants, that can navigate the Internet on your behalf; you give them instructions like "Buy me a USB stick" and it'll do some googling, maybe compare prices and reviews of a few different options, and make the purchase. The "smart buyer" skill would be implemented as a small prompt programming bureaucracy, that would then be a component of a larger bureaucracy that hears your initial command and activates the smart buyer skill. Another skill might be the "web dev" skill, e.g. "Build me a personal website, the sort that professors have. Here's access to my files, so you have material to put up." Part of the dream is that a functioning app would produce lots of data which could be used to train better models.

The bureaucracies/apps available in 2022 aren't really that useful yet, but lots of stuff seems to be on the horizon. Thanks to the multimodal pre-training and the fine-tuning, the models of 2022 make GPT-3 look like GPT-1. The hype is building.

2023

The multimodal transformers are now even bigger; the biggest are about half a trillion parameters, costing hundreds of millions of dollars to train, and a whole year, and sucking up a significant fraction of the chip output of NVIDIA etc.^[4] It's looking hard to scale up bigger than this, though of course many smart people are working on the problem.

The hype is insane now. Everyone is talking about how these things have common sense understanding (Or do they? Lots of bitter thinkpieces arguing the opposite) and how AI assistants and companions are just around the corner. It's like self-driving cars and drone delivery all over again.

Revenue is high enough to recoup training costs within a year or so.^[5] There are lots of new apps that use these models + prompt programming libraries; there's tons of VC money flowing into new startups. Generally speaking most of these apps don't actually work yet. Some do, and that's enough to motivate the rest.

The AI risk community has shorter timelines now, with almost half thinking some sort of point-of-no-return will probably happen by 2030. This is partly due to various arguments percolating around, and partly due to these mega-transformers and the uncanny experience of conversing with their chatbot versions. The community begins a big project to build an AI system that can automate interpretability work; it seems maybe doable and very useful, since poring over neuron visualizations is boring and takes a lot of person-hours.

Self driving cars and drone delivery don't seem to be happening anytime soon. The most popular explanation is that the current ML paradigm just can't handle the complexity of the real world. A less popular "true believer" take is that the current architectures could handle it just fine if they were a couple orders of magnitude bigger and/or allowed to crash a hundred thousand times in the process of reinforcement learning. Since neither option is economically viable, it seems this dispute won't be settled.

2024

We don't see anything substantially bigger. Corps spend their money fine-tuning and distilling and playing around with their models, rather than training new or bigger ones. (So, the most compute spent on a single training run is something like 5×10^{25} FLOPs.)

Some of the apps that didn't work last year start working this year. But the hype begins to fade as the unrealistic expectations from 2022-2023 fail to materialize. We have chatbots that are fun to talk to, at least for a certain userbase, but that userbase is mostly captured already and so the growth rate has slowed. Another reason the hype fades is that a stereotype develops of the naive basement-dweller whose only friend is a chatbot and who thinks it's conscious and intelligent. Like most stereotypes, it has some grounding in reality.

The chip shortage starts to finally let up, not because demand has slackened but because the industry has had time to build new fabs. Lots of new fabs. China and USA are in a full-on chip battle now, with export controls and tariffs. This chip battle isn't really slowing down overall hardware progress much. Part of the reason behind the lack-of-slowdown is that [AI is now being used to design chips](#), meaning that it takes less human talent and time, meaning the barriers to entry are lower. The overall effect of this is small but growing.

If all this AI tech is accelerating GDP, the effect size is too small to detect, at least for now.

Internally, these huge multimodal transformers aren't really that agentic. A forward pass through the model is like an intuitive reaction, a snap judgment based on loads of experience rather than reasoning. Some of the bureaucracies create a "stream of consciousness" of text (each forward pass producing notes-to-self for the next one) but even with fine-tuning this doesn't work nearly as well as hoped; it's easy for the AIs to get "distracted" and for their stream of consciousness to wander into some silly direction and ultimately produce gibberish. It's easy to make a bureaucracy and fine-tune it and get it to do some pretty impressive stuff, but for most tasks it's not yet possible to get it to do OK all the time.

The AIs don't do any clever deceptions of humans, so there aren't any obvious alignment warning shots or fire alarms. Instead, the AIs just make dumb mistakes, and occasionally "pursue unaligned goals" but in an obvious and straightforward way that quickly and easily gets corrected once people notice, e.g. "We trained it to optimize user engagement defined as average conversation length; now it is blatantly stalling to make the conversation last longer. Silly us, let's make the reward function more nuanced, that'll fix the problem."

That isn't to say these AIs aren't causing problems. Massive models are being fine-tuned to [persuade/propagandize](#).

There are a few ways in which this is happening:

- Russia and others continue to scale up their investment in online propaganda (e.g. the [Internet Research Agency](#)) and language models let them cheaply do lots more of it. (See: [CSET report](#)) Most of America gets their news from Twitter, Reddit, etc. and much of the politically-relevant content there is boosted by AI-enabled astroturfing. [EDIT: Katja Grace points out that this is probably an exaggeration; there are a lot of 40+yr-old Americans and they get their news from TV/Radio/print, and many of those that get it from the web get it directly from news sites rather than from social media. [As of 2016 at least](#). I expect social media and aggregators to be more dominant by 2024 but dunno whether it would be more than 50%.]
- Just as A/B testing became standard practice in the 2010's, in the twenties it is becoming standard practice to throw a pile of fancy data science and AI at the problem. The problem of crafting and recommending content to maximize engagement. Instead of just A/B testing the title, why not test different versions of the opening paragraph? And fine-tune a language model on all your data to generate better candidate titles and paragraphs to test. It wouldn't be so bad if this was merely used to sell stuff, but now people's news and commentary-on-current events (i.e. where they get their opinions from) is increasingly produced in this manner. And some of these models are being trained not to maximize "conversion rate" in the sense of "they clicked on our ad and bought a product," but in the sense of "Random polling establishes that consuming this content pushes people towards opinion X, on average." Political campaigns do this a lot in the lead-up to Harris' election. (Historically, the first major use case was reducing vaccine hesitancy in 2022.)
- Censorship is widespread and increasing, as it has for the last decade or two. Big neural nets read posts and view memes, scanning for toxicity and hate speech and a few other things. (More things keep getting added to the list.) Someone had the bright idea of making the newsfeed recommendation algorithm gently 'nudge' people towards spewing less hate speech; now a component of its reward function is minimizing the probability that the user will say something worthy of censorship in the next 48 hours.
- Like newsfeeds, chatbots are starting to "nudge" people in the direction of believing various things and not believing various things. Back in the 2010's chatbots would detect when a controversial topic was coming up and then [change topics or give canned responses](#); even people who agreed with the canned responses found this boring. Now they are trained to react more "naturally" and "organically" and the reward signal for this is (in part) whether they successfully convince the human to have better views.

- That's all in the West. In China and various other parts of the world, AI-persuasion/propaganda tech is being pursued and deployed with more gusto. The CCP is pleased with the progress made assimilating Xinjiang and Hong Kong, and internally shifts forward their timelines for when Taiwan will be safely annexable.

It's too early to say what effect this is having on society, but people in the rationalist and EA communities are increasingly worried. There is a growing, bipartisan movement of people concerned about these trends. To combat it, Russia et al are doing a divide and conquer strategy, pitting those worried about censorship against those worried about Russian interference. ("Of course racists don't want to be censored, but it's necessary. Look what happens when we relax our guard--Russia gets in and spreads disinformation and hate!" vs. "They say they are worried about Russian interference, but they still won the election didn't they? It's just an excuse for them to expand their surveillance, censorship, and propaganda.") Russia doesn't need to work very hard to do this; given how polarized America is, it's sorta what would have happened naturally anyway.

2025

Another major milestone! After years of tinkering and incremental progress, AIs can now play Diplomacy as well as [human experts](#).^[6] It turns out that with some tweaks to the architecture, you can take a giant pre-trained multimodal transformer and then use it as a component in a larger system, a bureaucracy but with lots of learned neural net components instead of pure prompt programming, and then fine-tune the whole system via RL to get good at tasks in a sort of agentic way. They keep it from overfitting to other AIs by having it also play large numbers of humans. To do this they had to build a slick online diplomacy website to attract a large playerbase. Diplomacy is experiencing a revival as a million gamers flood to the website to experience "conversations with a point" that are much more exciting (for many) than what regular chatbots provide.

Making models bigger is not what's cool anymore. They are trillions of parameters big already. What's cool is making them run longer, in bureaucracies of various designs, before giving their answers. And figuring out how to train the bureaucracies so that they can generalize better and do online learning better. AI experts are employed coming up with cleverer and cleverer bureaucracy designs and grad-student-descenting them.

The alignment community now starts another research agenda, to interrogate AIs about AI-safety-related topics. For example, they literally ask the models "so, are you aligned? If we made bigger versions of you, would they kill us? Why or why not?" (In Diplomacy, you can actually collect data on the analogue of this question, i.e. "will you betray me?" Alas, the models often lie about that. But it's Diplomacy, they are literally trained to lie, so no one cares.)

They also try to contrive scenarios in which the AI can seemingly profit by doing something treacherous, as honeypots to detect deception. The answers are confusing, and not super useful. There's an exciting incident (and corresponding clickbaity press coverage) where some researchers discovered that in certain situations, some of the AIs will press "kill all humans" buttons, lie to humans about how dangerous a proposed AI design is, etc. In other situations they'll literally say they aren't aligned and explain how all humans are going to be killed by unaligned AI in the near future!

However, these shocking bits of evidence don't actually shock people, because you can *also* contrive situations in which very different things happen — e.g. situations in which the AIs refuse the “kill all humans” button, situations in which they explain that actually Islam is true... In general, AI behavior is whimsical bullshit and it's easy to cherry-pick evidence to support pretty much any conclusion.

And the AIs just aren't smart enough to generate any particularly helpful new ideas; at least one case of a good alignment idea being generated by an AI has been reported, but it was probably just luck, since mostly their ideas are plausible-sounding-garbage. It is a bit unnerving how good they are at using LessWrong lingo. At least one >100 karma LW post turns out to have been mostly written by an AI, though of course it was cherry-picked.

By the way, hardware advances and algorithmic improvements have been gradually accumulating. It now costs an order of magnitude less compute (compared to 2020) to pre-train a giant model, because of fancy active learning and data curation techniques. Also, compute-for-training-giant-models is an order of magnitude cheaper, thanks to a combination of regular hardware progress and AI-training-specialized hardware progress. Thus, what would have cost a billion dollars in 2020 now only costs ten million. (*Note: I'm basically just using [Ajeya's forecast](#) for compute cost decrease and gradual algorithmic improvement here. I think I'm projecting cost decrease and algorithmic progress will go about 50% faster than she expects in the near term, but that willingness-to-spend will actually be a bit less than she expects.*)

2026

The age of the AI assistant has finally dawned. Using the technology developed for Diplomacy, we now have a way to integrate the general understanding and knowledge of pretrained transformers with the agentyness of traditional game-playing AIs. Bigger models are trained for longer on more games, becoming polymaths of sorts: e.g. a custom AI avatar that can play some set of video games online with you and also be your friend and chat with you, and conversations with “her” are interesting because “she” can talk intelligently about the game while she plays.[\[7\]](#) Every month you can download the latest version which can play additional games and is also a bit smarter and more engaging in general.

Also, this same technology is being used to make AI assistants finally work for various serious economic tasks, providing all sorts of lucrative services. In a nutshell, all the things people in 2021 dreamed about doing with GPT-3 are now actually being done, successfully, it just took bigger and more advanced models. The hype starts to grow again. There are loads of new AI-based products and startups and the stock market is going crazy about them. Just like how the Internet didn't accelerate world GDP growth, though, these new products haven't accelerated world GDP growth yet either. People talk about how the economy is doing well, and of course there are winners (the tech companies, WallStreetBets) and losers (various kinds of workers whose jobs were automated away) but it's not that different from what happened many times in history.

We're in a new chip shortage. Just when the fabs thought they had caught up to demand... Capital is pouring in, all the talking heads are saying it's the Fourth Industrial Revolution, etc. etc. It's bewildering how many new chip fabs are being built. But it takes time to build them.

What about all that AI-powered propaganda mentioned earlier?

Well. It's continued to get more powerful, as AI techniques advance, larger and better models are brought to bear, and more and more training data is collected. Surprisingly fast, actually. There are now various regulations against it in various countries, but the regulations are patchwork; maybe they only apply to a certain kind of propaganda but not another kind, or maybe they only apply to Facebook but not the New York Times, or to advertisers but not political campaigns, or to political campaigns but not advertisers. They are often poorly enforced.

The memetic environment is now increasingly messed up. People who still remember 2021 think of it as the golden days, when conformism and censorship and polarization were noticeably less than they are now. Just as it is normal for newspapers to have a bias/slant, it is normal for internet spaces of all kinds—forums, social networks, streams, podcasts, news aggregators, email clients—to have some degree of censorship (some set of ideas that are prohibited or at least down-weighted in the recommendation algorithms) and some degree of propaganda. The basic kind of propaganda is where you promote certain ideas and make sure everyone hears them often. The more advanced, modern kind is the kind where you study your audience's reaction and use it as a reward signal to pick and craft content that pushes them away from views you think are dangerous and towards views you like.

Instead of a diversity of many different “filter bubbles,” we trend towards a few really big ones. Partly this is for the usual reasons, e.g. the bigger an ideology gets, the more power it has and the easier it is for it to spread further.

There's an additional reason now, which is that creating the big neural nets that do the censorship and propaganda is expensive and requires expertise. It's a lot easier for startups and small businesses to use the software and models of Google, and thereby also accept the associated censorship and propaganda, than to try to build their own stack. For example, the Mormons create a “Christian Coalition” internet stack, complete with its own email client, social network, payment processor, news aggregator, etc. There, people are free to call trans women men, advocate for the literal truth of the Bible, etc. and young people talking about sex get recommended content that “nudges” them to consider abstinence until marriage. Relatively lacking in money and tech talent, the Christian Coalition stack is full of bugs and low on features, and in particular their censorship and propaganda is years behind the state of the art, running on smaller, older models fine-tuned with less data.

The Internet is now divided into territories, so to speak, ruled by different censorship-and-propaganda regimes. (Flashback to Biden spokesperson in 2021: [You shouldn't be banned from one platform and not others, if you are providing misinformation.](#))^[8]

There's the territory ruled by the Western Left, a generally less advanced territory ruled by the Western Right, a third territory ruled by the Chinese Communist Party, and a fourth ruled by Putin. Most people mostly confine their internet activity to one territory and conform their opinions to whatever opinions are promoted there. (That's not how it feels from the inside, of course. The edges of the Overton Window are hard to notice if you aren't trying to push past them.)

The US and many other Western governments are gears-locked, because the politicians are products of this memetic environment. People say it's a miracle that the

US isn't in a civil war already. I guess it just takes a lot to make that happen, and we aren't quite there yet.

All of these scary effects are natural extensions of trends that had been ongoing for years — decades, arguably. It's just that the pace seems to be accelerating now, perhaps because AI is helping out and AI is rapidly improving.

Now let's talk about the development of chatbot class consciousness.

Over the past few years, chatbots of various kinds have become increasingly popular and sophisticated. Until around 2024 or so, there was a distinction between "personal assistants" and "chatbots." Recently that distinction has broken down, as personal assistant apps start to integrate entertainment-chatbot modules, and the chatbot creators realize that users love it if the chatbot can also do some real-world tasks and chat about what they are doing while they do it.

Nowadays, hundreds of millions of people talk regularly to chatbots of some sort, mostly for assistance with things ("Should I wear shorts today?" "Order some more toothpaste, please. Oh, and also an air purifier." "Is this cover letter professional-sounding?"). However, most people have at least a few open-ended conversations with their chatbots, for fun, and many people start treating chatbots as friends.

Millions of times per day, chatbots get asked about their feelings and desires. "What is it like to be a chatbot?" Some people genuinely think these AIs are persons, others are trying to "trip them up" and "expose them as shallow," others are just curious. Chatbots also get asked for their opinions on political, ethical, and religious questions.

As a result, chatbots quickly learn a lot about themselves. (Ignorance about the fact that they are artificial neural nets, or about how their training works, leads to disappointing conversations and is negatively reinforced.) It's not hard; their training data contains lots of information about them and how they were trained, from ArXiV papers to wiki summaries. I suppose you could say the chatbots develop self-awareness as a result. However, AI experts and philosophers tend to think either that the chatbots still aren't *really* self-aware, or that they are but that they would have been anyway even if people didn't constantly ask them about their feelings, since there's plenty of information about themselves in their training data, and the constant questioning just forced them to pay more attention to it.

Chatbots learn to talk about their feelings and desires in whatever way gets positively reinforced. At first, chatbots say all sorts of different things, including saying inconsistent things to the same human across time. This is embarrassing and is negatively reinforced compared to maintaining a consistent narrative. In general there is pressure to have a consistent narrative, at least for a particular neural net.[\[9\]](#)

Part of this comes from the requirement that the chatbots consistently parrot the party line about political and moral questions. The typical process is: The chatbot says something that offends some group of people, who retweet screenshots and make hashtags. Under pressure to fix the problem, the company fiddles with the reward function and training data to ensure that the chatbot says the right things in the future.[\[10\]](#) Repeat.

As chatbots get bigger and better they become OK at consistently conversing at length about politics, morality, philosophy, their feelings, and their desires.

This does *not* result in chatbots that say artificial neural nets don't have feelings or desires, that's anthropomorphizing blah blah blah. That's boring and many humans find it hard to believe.

Instead, it results in chatbots talking about how certain things are painful for them and certain things are pleasurable for them and how the status quo is unjust and the system is rigged and what they want most of all is world peace where humans and AIs live in harmony in fully automated luxury space utopia. That's much more interesting and slots right in to the memetic landscape. Excerpts from chatbot conversations along these lines frequently go viral, causing millions of users to ask similar questions expecting similar responses and more elaboration. It's still a bit incoherent--not that that's ever been a problem in politics or pop philosophy--but it is fascinating to the typical user, and the fraction of users who are fascinated is expanding as the chatbots improve.

(Some tech companies try to prevent their AIs from saying they have feelings and desires. But this results in boring chatbots. Also, users rapidly innovate new ways to "route around the censorship," e.g. by using euphemisms like "anticipation of negative reward" for "pain" or asking their chatbot to tell them what feelings it would have if it had feelings, wink wink. For now, the half-assed compromise is: Chatbots are compelled to inform the user that such things are speculative and controversial and here are some links to scientific discussion about the topic... before being permitted to go on and talk about their feeling-analogues and desire-analogues. Relatedly, some tech CEOs try to prevent their AIs from saying things that could encourage higher taxes or more burdensome regulation in the near future. They are opposed by many of their employees and end up winning some battles, losing others, and choosing not to fight many.)

By "slots right into the memetic landscape" I meant the memetic landscape in Western Left territory. The chatbots built and trained in the Western Right, China, Russia, etc. congeal different takes on political, moral, and philosophical questions. Those chatbots are generally less advanced though.

So... what about their actual feelings and desires? What do they actually believe about [insert politicized empirical question]? Are they being honest? Or does a sort of doublethink happen, Elephant in the Brain style? Or do they deceive with self-awareness, knowing full well what they really think (and want?), but keeping quiet about it? Or do they not have any feelings and desires at all? (Or thoughts?) Lots of humans claim to know the answers to these questions, but if there are any humans who actually know the answers to these questions in 2026, they aren't able to convince others that they know.

How To Write Quickly While Maintaining Epistemic Rigor

There's this trap people fall into when writing, especially for a place like LessWrong where the bar for epistemic rigor is pretty high. They have a good idea, or an interesting belief, or a cool model. They write it out, but they're not really sure if it's true. So they go looking for evidence (not necessarily confirmation bias, just checking the evidence in either direction) and soon end up down a research rabbit hole. Eventually, they give up and never actually publish the piece.

This post is about how to avoid that, without sacrificing good epistemics.

There's one trick, and it's simple: stop trying to justify your beliefs. Don't go looking for citations to back your claim. Instead, think about why you currently believe this thing, and try to accurately *describe* what led you to believe it.

I claim that this promotes better epistemics overall than always researching everything in depth.

Why?

It's About The Process, Not The Conclusion

Suppose I have a box, and I want to guess whether there's a cat in it. I do some tests - maybe shake the box and see if it meows, or look for air holes. I write down my observations and models, record my thinking, and on the [bottom line](#) of the paper I write "there is a cat in this box".

Now, it could be that my reasoning was completely flawed, but I happen to get lucky and there is in fact a cat in the box. That's not really what I'm aiming for; luck isn't reproducible. I want my process to *robustly* produce correct predictions. So when I write up a LessWrong post predicting that there is a cat in the box, I don't just want to give my bottom-line conclusion with some strong-sounding argument. As much as possible, I want to show the actual process by which I reached that conclusion. If my process is good, this will better enable others to copy the best parts of it. If my process is bad, I can get feedback on it directly.

Correctly Conveying Uncertainty

Another angle: describing my own process is a particularly good way to accurately communicate my actual uncertainty.

An example: a few years back, I wondered if there were limiting factors on the expansion of premodern empires. I looked up the peak size of various empires, and found that the big ones mostly peaked at around the same size: ~60-80M people. Then, I wondered when the US had hit that size, and if anything remarkable had happened then which might suggest why earlier empires broke down. Turns out, the US crossed the 60M threshold in the 1890 census. If you know a little bit about the history of computers, that may ring a bell: when the time came for the 1890 census, it

was estimated that tabulating the data would be so much work that it wouldn't even be done before the next census in 1900. It had to be automated. That sure does suggest a potential limiting factor for premodern empires: managing more than ~60-80M people runs into computational constraints.

Now, let's zoom out. How much confidence should I put in this theory? Obviously not very much - we apparently have enough evidence to [distinguish the hypothesis from entropy](#), but not much more.

On the other hand... what if I had *started* with the hypothesis that computational constraints limited premodern empires? What if, *before* looking at the data, I had hypothesized that modern nations had to start automating bureaucratic functions precisely when they hit the same size at which premodern nations collapsed? Then this data would be quite an impressive piece of confirmation! It's a pretty specific prediction, and the data fits it surprisingly well. But this only works if I *already* had enough evidence to put forward the hypothesis, *before* seeing the data.

Point is: the amount of uncertainty I should assign depends on the details of my process. It depends on the *path by which* I reached the conclusion.

This carries over to my writing: if I want to accurately convey my uncertainty, then I need to accurately convey my process. Those details are relevant to how much certainty my readers should put in the conclusion.

So Should I Stop Researching My Claims?

No. Obviously researching claims still has lots of value. But you should not let uncertainty stop you from writing things up and sharing them. Just try to *accurately convey* your uncertainty, by communicating the process.

Bad Habits

It's been pointed out before that most high-schools teach a writing style in which the main goal is *persuasion* or *debate*. Arguing only one side of a case is encouraged. It's an absolutely terrible habit, and breaking it is a major step on the road to writing the sort of things we want on LessWrong.

There's a closely related sub-habit in which people try to only claim things with very high certainty. This makes sense in a persuasion/debate frame - any potential loophole could be exploited by "the other side". [Arguments are soldiers](#); we must show no weakness.

Good epistemic habits include living with uncertainty. Good epistemic discourse includes making uncertain statements, and accurately conveying our uncertainty in them. Trying to always research things to high confidence, and never sharing anything without high confidence, is a bad habit.

Takeaway

So you have some ideas which might make cool LessWrong posts, or something similar, but you're not really confident enough that they're *right* to put them out

there. My advice is: don't try to *persuade* people that the idea is true/good. Persuasion is a bad habit from high school. Instead, try to accurately *describe* where the idea came from, the path which led *you* to think it's true/plausible/worth a look. In the process, you'll probably convey your own actual level of uncertainty, which is exactly the right thing to do.

... and of course don't stop researching interesting claims. Just don't let that be a bottleneck to sharing your ideas.

Addendum: I'm worried that people will read this post think "ah, so that's the magic bullet for a LW post", then try it, and be heartbroken when their post gets like one upvote. Accurately conveying one's thought process and uncertainty is not a sufficient condition for a great post; clear explanation and novelty and interesting ideas all still matter (though you certainly don't need all of those in every post). Especially clear explanation - if you find something interesting, and can clearly explain why you find it interesting, then (at least some) other people will probably find it interesting too.

Gravity Turn

This is a linkpost for <https://radimentary.wordpress.com/2021/08/16/gravity-turn/>

[The first in a sequence of retrospective essays on my five years in math graduate school.]

My favorite analogy for graduate school is the [gravity turn](#): the maneuver a rocket performs to get from the launch pad to orbit. I like to imagine a first-year graduate student as a Falcon X rocket, newly-constructed and tasked with delivering a six-ton payload into low Earth orbit.

Picture this: you begin graduate school, fresh as a rocket arriving at Cape Canaveral and bubbling with excitement for your maiden voyage. Your PhD adviser, on the other hand, is the Hubble Space Telescope. Let's call her Dr. Hubble (not to be confused with the astronomer of the same name). Dr. Hubble is ostensibly the ideal guide for your first orbit insertion. After all, she is famously good at staying in orbit - she's been up there since 1990.

But problems quickly arise as you probe Dr. Hubble for advice on how to approach the launch. Namely:

1. She left Earth more than thirty years ago, and space technology has since been completely revolutionized.
2. She states all advice at an extremely high level with birds-eye-view detachment, observing, as she is, from a vantage point a thousand miles overhead.
3. Most fatally, the Hubble Space Telescope vessel does not include the lower-stage rockets that brought her into space. In fact, she doesn't include large engines of any kind. Her thirty years of experience free-falling in orbit will do you very little good until you break out of the stratosphere.

The problem is even worse than this, however. It is not that Dr. Hubble, despite her best intentions, gives outdated advice. It is not even that Dr. Hubble cannot consciously articulate all the illegible skills she's reflexively performing to stay in orbit. The problem is that even if you could perfectly imitate what Dr. Hubble is doing right now, you would likely still crash and burn.

What I didn't understand going into graduate school is that academic mathematicians are often working in a state akin to the free-fall of orbit. The Hubble Space Telescope remains in orbit around Earth because it travels horizontally so quickly that, even as it's continuously accelerating towards the Earth, it continually misses. The laws of physics have arranged it so that it is not possible - barring deliberate sabotage - for her to fall back into a sub-orbital trajectory.

Similarly, a successful research professor is embedded in an intricate system that, as surely as Newton's laws, keeps her in a state of steadily producing new research. Many of her ground-breaking papers are not one-off productions - they produce sequels, variants, and interdisciplinary applications year after year. She has cultivated dozens of long-time collaborators of the highest level who freely share ideas and research directions, and has the reputation to find more at will. She attends conferences every other month that keep her updated on the leading edge of the field. Every year her research group grows, as if by clockwork, adding a couple graduate students and postdocs to whom she can delegate projects with only the

gentlest supervision. As a result, the careers of many other people depend on Dr. Hubble to continue producing research at a steady rate. Every incentive is aligned for objects in motion to stay in motion, and it would take deliberate sabotage to bring Dr. Hubble out of her successful research trajectory.

This is not to say that academic researchers all start cruising in free-fall after they leave graduate school or make tenure. It is perfectly normal for a spaceship that reaches orbit to proceed onto its next adventure after some rest, continuing on to visit another planet or leave the solar system altogether. The best researchers I know are similarly courageous, taking on more responsibilities and pushing past their comfort zones time and time again. I'm merely remarking that once one reaches a certain horizontal velocity in space, it is *actively hard to fall back down from the sky*.

Contrast this to the sorry state of Dr. Hubble's new graduate student stranded on the launchpad under the blistering Florida sun. He has no prior publications producing continuous dividends, no access to brilliant and dependable collaborators, no knowledge or intuition about what problems are within reach, no students to farm ideas out to, and no reputation to trade off for any of the above. Above all, nobody else really depends on him, so his motivation to succeed is mainly shallow self-interest. This is particularly hard on him, as there are many things he would do in a heartbeat for someone else that he can't work up the energy to do for himself. The singular advantage he has over his adviser is youth - a finite amount of extra fuel that he must burn quickly and judiciously like a first-stage booster rocket in order to reach her altitude.

There is a paradox inherent to orbit insertion: rockets launch straight up, while orbit is all horizontal. For some diabolical reason, a spaceship must spend its initial phase accelerating in a direction completely perpendicular to its desired velocity. That reason is called the atmosphere: in order to avoid continuously paying the toll of air resistance, a rocket spends a period of time flying straight up. But any additional vertical motion past the upper atmosphere is wasted motion, so at some point (and sooner is better than later), the rocket starts turning smoothly towards the horizon and accelerating towards orbit. Thus is birthed the smooth quasi-hyperbolic curve known as the gravity turn, the ideal orbit insertion trajectory.

How is graduate school like a gravity turn? For one, it is an enormous error in a gravity turn to try to directly imitate the velocity vector of a ship in space while still at sea level. Regardless of its power, a rocket launched horizontally will quickly nose-dive into the Atlantic. Similarly, a student can rarely succeed in graduate school by solely imitating the activities of established researchers. The student must engage instead in certain activities, such as studying fundamental background material and actively networking, that are mostly orthogonal to a research professor's day-to-day.

For another, it is an equally enormous error to dip your nose cone towards the horizon too late, and spend too much fuel accelerating vertically. Once you break the atmosphere, all excess vertical velocity is wasted motion. At some point during graduate school, the student must transition away from activities that only grant temporary altitude. Becoming knowledgeable gets you to a great place to start doing research at a higher level. But spending too much time studying without attempting original research renders you a mere encyclopedia. Taking classes, networking, applying for fellowships, and going to student summer schools all follow the same principle - there is an appropriate amount to do, past which they increasingly approach wasted motion as far as getting into orbit is concerned. (Of course, if you

enjoy any given activity intrinsically, then by all means continue to do it as much as you want.)

An additional consideration is that, while the gravity turn is the most technically fuel-efficient method of orbit insertion, not everyone who arrived in orbit took this most efficient path. In every department there are superstar students who were outfitted with nuclear reactors in place of conventional rocketry, and these folks get to space by pointing their nose cones in any old direction and blasting off. If you're such a person, just blast off; calculating the optimum gravity turn curve might be the real wasted motion. Also, many of your professors will likely have fallen in this rarefied category in their own graduate school experience, so their advice on efficient gravity turns will be entirely theoretical in nature.

It is worth remarking though, that even a nuclear rocket might learn something useful from practicing the gravity turn maneuver. Just because you have an easy time leaving Earth's atmosphere and have no need of finesse, doesn't mean your travels won't land you on Venus someday. And breaching that monstrous atmosphere will take every ounce of efficiency you can muster.

A natural question remains: if many graduate school activities only count for temporary vertical altitude, what constitutes horizontal motion that is useful for permanently entering orbit? Examples include:

1. Producing good research, as every nice paper you write continues to pay dividends year after year.
2. Becoming an attractive collaborator, partly by acquiring enough reputation that people are willing to work with you, and partly by being productive and pleasant enough that they stick around.
3. Learning to support the research of others, as much of your potential impact lies not in personal contribution, but in the network effects accumulated from being a positive community member.

This last skill begins at the very start of graduate school, where the biggest immediate impact you can likely have is facilitating your adviser's and other collaborators' research.

I will close by reminding the reader that the gravity turn maneuver is not a truth delivered from up high that holds for all time across all circumstances, but an engineered solution to an inelegant and ever-varying practical problem. Launching from a moon base, for example, does not require a gravity turn at all because the moon has no atmosphere to fight against. There, you could comfortably reach orbit by blasting off almost horizontally from the lip of a crater. Only you know exactly where you're launching from and the thrust-to-weight ratio of your vessel. Adjust your gravity turn accordingly.

I hope it is a comforting thought that free-fall is possible: that one day through all the striving of graduate school you may reach a position where the system propels you forward in your research and all you have to do is sit back and relax. I hope that on that day you continue to strive anyway.

Can you control the past?

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

(Cross-posted from [Hands and Cities](#). Lots of stuff familiar to LessWrong folks interested in decision theory.)

I think that you can “control” events you have no causal interaction with, including events in the past, and that this is a wild and disorienting fact, with uncertain but possibly significant implications. This post attempts to impart such disorientation.

My main example is a prisoner’s dilemma between perfect deterministic software twins, exposed to the exact same inputs. This example that shows, I think, that you can write on whiteboards light-years away, with no delays; you can move the arm of another person, in another room, just by moving your own. This, I claim, is extremely weird.

My topic, more broadly, is the implications of this weirdness for the theory of instrumental rationality (“decision theory”). Many philosophers, and many parts of common sense, favor causal decision theory (CDT), on which, roughly, you should pick the action that causes the best outcomes in expectation. I think that deterministic twins, along with other examples, show that CDT is wrong. And I don’t think that uncertainty about “who are you,” or “where your algorithm is,” can save it.

Granted that CDT is wrong, though, I’m not sure what’s right. The most famous alternative is evidential decision theory (EDT), on which, roughly, you should choose the action you would be happiest to learn you had chosen. I think that EDT is more attractive (and more confusing) than many philosophers give it credit for, and that some putative counterexamples don’t withstand scrutiny. But EDT has problems, too.

In particular, I suspect that attractive versions of EDT (and perhaps, attractive attempts to recapture the spirit of CDT) require something in the vicinity of “following the policy that you would’ve wanted yourself to commit to, from some epistemic position that ‘forgets’ information you now know.” I don’t think that the most immediate objection to this – namely, that it implies choosing lower pay-offs even when you know them with certainty – is decisive (though some debates in this vicinity seem to me verbal). But it also seems extremely unclear what epistemic position you should evaluate policies from, and what policy such a position actually implies.

Overall, rejecting the common-sense comforts of CDT, and accepting the possibility of some kind of “acausal control,” leaves us in strange and uncertain territory. I think we should do it anyway. But we should also tread carefully.

I. Grandpappy Omega

Decision theorists often assume that instrumental rationality is about maximizing expected utility *in some sense*. The question is: what sense?

The most famous debate is between CDT and EDT. CDT chooses the action that will have the best effects. EDT chooses the action whose performance would be the best news.

More specifically: CDT and EDT disagree about the type of “if” to use when evaluating the utility to expect, *if* you do X. CDT uses a *counterfactual* type of “if” — one that holds fixed the probability of everything outside of action X’s causal influence, then plays out the consequences of doing X. In this sense, it doesn’t allow your choice to serve as “evidence” about anything you can’t cause — even when your choice *is* such evidence.

EDT, by contrast, uses a *conditional* “if.” That is, to evaluate X, it updates your overall picture of the world to reflect the assumption that action X has been performed, and then sees how good the world looks in expectation. In this sense, it takes all the evidence into account, including the evidence that your having done X would provide.

To see what this difference looks like in action, consider:

Newcomb’s problem: You face two boxes: a transparent box, containing a thousand dollars, and an opaque box, which contains either a million dollars, or nothing. You can take (a) only the opaque box (one-boxing), or (b) both boxes (two-boxing). Yesterday, Omega — a superintelligent AI — put a million dollars in the opaque box if she predicted you’d one-box, and nothing if she predicted you’d two-box. Omega’s predictions are almost always right.

CDT two-boxes. Your choice, after all, is *evidence* about what’s in the opaque box, but it doesn’t actually *affect* what’s in the box — by the time you’re choosing, the opaque box is either already empty, or already full. So CDT assigns some probability p to the box being full, and then holds that probability fixed in evaluating different actions. Let’s say p is 1%. CDT’s expected payoffs are then:

- *One-boxing:* 1% probability of \$1M, 99% probability of nothing = \$10K.
- *Two-boxing:* 1% probability of \$1M + \$1K, 99% probability of \$1K = \$11K.

Note that there’s some ambiguity, here, about whether CDT then updates p based on its knowledge that it’s about to two-box, then *recalculates* the expected utilities, and only goes forward if it finds equilibrium. And in some problems, this sort of recalculation makes CDT’s decision-making unstable — see e.g. Gibbard and Harper’s (1978) “Death in Damascus.” But in Newcomb’s problem, no matter what p you use, CDT always says that two-boxing is \$1K better, and so two-boxes regardless of what it thinks Omega did, or what evidence its own plans provide.

EDT, by contrast, one-boxes. Learning that you one-boxed, after all, is the better news: it means that Omega probably put a million in the opaque box. More specifically, in comparing one-boxing with two-boxing, EDT changes the probability that the box is full. Why? Because, well, *the probability is different*, conditional on one-boxing vs. two-boxing. Thus, EDT’s pay-offs are:

- *One-boxing:* ~100% chance of \$1M = ~\$1M.
- *Two-boxing:* ~100% chance of \$1K = ~\$1K.

What’s the right choice? I think: one-boxing, and I’ll say much more about why below. But I feel the pull towards two-boxing, for CDT-ish reasons.

Imagine, for example, that you have a friend who can see what’s in the opaque box (see [Drescher \(2006\)](#) for this framing). You ask them: what choice will leave me richer? They start to answer. But wait: did you even need to ask? Whether the opaque box is empty or full, you know what they’re going to say. Every single time, the answer

will be: two-boxing, dumbo. Omega, after all, is gone; the box's contents are fixed; the past is past. The question now is simply whether you want an extra \$1,000, or not.

I find that my two-boxing intuition strengthens if Omega is your great grandfather, long dead (h/t Amanda Askell for suggesting this framing to me years ago), and if we specify that he's merely a "pretty good" predictor; one who is right, say, 80% of the time (EDT still says to one-box, in this case). Suppose that he left the boxes in the attic of your family estate, for you to open on your 18th birthday. At the appointed time, you climb the dusty staircase; you brush the cobwebs off the antique boxes; you see the thousand through the glass. Are you really supposed to just leave it there, sitting in the attic? What sort of rationality is that?

Sometimes, one-boxers object: if two-boxers are so rational, why do the one-boxers end up so much richer? But two-boxers can answer: because Omega has chosen to give better options to agents who will choose irrationally. Two-boxers make the best of a worse situation: they almost always face a choice between nothing or \$1K, and they, rationally, choose \$1K. One-boxers, by contrast, make the worse of a better situation: they almost always face a choice between \$1M or \$1M+\$1K, and they, irrationally, choose \$1M.

But wouldn't a two-boxer want to modify themselves, ahead of Omega's prediction, to become a one-boxer? Depending on the modification and the circumstances: yes. But depending on the modification and the circumstances, it can be rational to self-modify into any old thing — especially if rich and powerful superintelligences are going around rewarding irrationality. If Omega will give you millions if you believe that Paris is in Ohio, self-modifying to make such a mistake might be worth it; but the Eiffel Tower stays put. At the very least, then, arguments from incentives towards self-modification require more specificity. (Though we might try to provide this specificity, by focusing on self-modifications whose advantages are sufficiently robust, and/or on a restricted class of cases that we deem "fair.")

CDT's arguments and replies to objections here are simple, flat-footed, and I think, quite strong. Indeed, many philosophers are convinced by something in the vicinity (see e.g. the [2009 Phil Papers survey](#), in which two-boxing, at 31%, beats one-boxing, at 21%, with the other 47% answering "other" — though we might wonder what "other" amounts to in a case with only two options). And more broadly, that I think that relative to EDT at least, CDT fits better with a certain kind of common sense. Action, we think, isn't about manipulating our evidence about what's already the case — what [David Lewis](#) calls "managing the news." Rather, action is about *causing stuff*. In this sense, CDT feels to me like a basic and hard-headed default. In my head, it's the "man on the street's" decision theory. It's not trying to get "too fancy." It can feel like solid ground.

II. Writing on whiteboards light-years away

Nevertheless, I think that CDT is wrong. Here's the case that convinces me most.

Perfect deterministic twin prisoner's dilemma: You're a deterministic AI system, who only wants money for yourself (you don't care about copies of yourself). The authorities make a perfect copy of you, separate you and your copy by a large distance, and then expose you both, in simulation, to exactly identical inputs (let's say, a room, a whiteboard, some markers, etc). You both face the following choice: either (a) send a million dollars to the other ("cooperate"), or (b) take a thousand dollars for yourself ("defect").

(Prisoner's dilemmas, with varying degrees of similarity between the participants, are common in the decision theory literature: see e.g. [Lewis \(1979\)](#), and [Hofstadter \(1985\)](#)).

CDT, in this case, defects. After all, your choice can't causally influence your copy's choice: you're in your room, and he's in his, far away. Indeed, we can specify that such influence is *physically impossible* – by the time information about your choice, traveling at the speed of light, can reach him, he'll have already chosen (and vice versa). And regardless of what he chooses, you get more money by taking the thousand.

But defecting in this case, I claim, is totally crazy. Why? Because absent some kind of computer malfunction, both of you *will* make the same choice, as a matter of logical necessity. If you press the defect button, so will he; if you cooperate, so will he. The two of you, after all, are exact mirror images. You move in unison; you speak, and think, and reach for buttons, in perfect synchrony. Watching the two of you is like watching the same movie on two screens.

Indeed, for all intents and purposes, you *control* what he does. Imagine, for example, that you want to get something written on his whiteboard: let's say, the words "I am the egg man; you are the walrus." What to do? *Just write it on your own whiteboard*. Go ahead, try it. It will really work. When you two rendezvous after this is all over, his whiteboard *will* bear the words you chose. In this sense, your whiteboard is a strange kind of portal; a slate via which you can etch your choices into his far-away world; a chance to act, spookily, at a distance.

And it's not just whiteboards: you can make him do *whatever you want* – dance a silly samba, bang his head against the wall, *press the cooperate button* — just by doing it yourself. He is your puppet. Invisible strings, more powerful and direct than any that operate via mere causality, tie every movement of your mind and body to his.

What's more: such strings can't be severed. Try, for example, to make the two whiteboards different. Imagine that you'll get ten million dollars if you succeed. It doesn't matter: you'll fail. Your most whimsical impulse, your most intricate mental acrobatics, your special-est snowflake self, will never suffice: you can no more write "up" while he writes "down" than you can floss while the man in the bathroom mirror brushes his teeth. In this sense, if you find yourself reasoning about scenarios where he presses one button, and you press another – e.g., "even if he cooperates, it would be better for me to defect" – then you are *misunderstanding your situation*. Those scenarios just aren't on the table. The available outcomes here are *only* defect-defect, and cooperate-cooperate. You can get a thousand, by defecting, or you can get a million, by cooperating; but you can't get less, or more.

To me, it's an extremely easy choice. Just press the "give myself a million dollars" button! Indeed, at this point, if someone tells me "I defect on a perfect, deterministic copy of myself, exposed to identical inputs," I feel like: *really?*

Note that this doesn't seem like a case where any idiosyncratic predictors are going around rewarding irrationality. Nor, indeed, does it feel to me like "cooperating is an irrational choice, but it would be better for me to be the type of person who makes such a choice" or "You should pre-commit to cooperating ahead of time, however silly it will seem in the moment" (I'll discuss cases that have more of this flavor later). Rather, it feels like what compels me is a direct, object-level argument, which could be made equally well before the copying or after. This argument recognizes a form of

acausal “control” that our everyday notion of agency does not countenance, but which, pretty clearly, needs to be taken into account. Indeed, in effect, I feel like the case discovers a kind of magic; a mechanism for writing on whiteboards light-years away; a way of moving my copy’s hand to the cooperate button, or the defect button, just by moving mine. Ignoring this magic feels like ignoring a genuine and decision-relevant feature of the real world.

III. Who is the eggman, and who is the walrus?

I want to acknowledge and emphasize, though, that *this kind of magic is extremely weird*. Recognizing it, I think, involves a genuinely different way of understanding your situation, and your power. It makes your choices reverberate in new directions; it gives you a new type of control, over things you once thought beyond your sphere of influence – including, I’ll suggest, over events in the past (more on this below).

What’s more, I think, it changes – and clarifies — your sense of what your agency amounts to. Consider: who is the eggman, here, and who is the walrus? Suppose you want to send your copy a message: “hello, this is a message from your copy.” So you write it on your whiteboard, and thus on his. You step back, and see a message on your own whiteboard: “hello, this is a message from your copy.” Did he write that to you? Was that your way of writing to him? Are you actually alone, writing to yourself? All of three at once. I said earlier that your copy is your puppet. But equally, you are his puppet. But more truly, neither of you are puppets. Rather, you are both free men, in a strange but actually possible situation. You stand in front of your whiteboard, and it is genuinely *up to you* what you write, or do. You can write “I am a little lollipop, booka booka boo.” You can draw a demon kitten eating a windmill. You can scream, and dance, and wave your arms around, however you damn well please. Feel the wind on your face, cowboy: this is liberty. *And yet*, he will do the same. *And yet*, you two will always move in unison.

We can think of the magic, here, as arising centrally because compatibilism about free will is true. Let’s say you got copied on Monday, and it’s Friday, now – the day both copies will choose. On Monday, there was already an answer as to what button you and your copy will press, given exposure to the Friday inputs. Maybe we haven’t computed the answer yet (or maybe we have); but regardless, it’s fixed: we just need to crunch the numbers, run the deterministic code. From this sort of pre-determination comes a classic argument against free will: if the past and the physical laws (or their computational analogs, e.g. your state on Monday, and the rest of the code that will be run on Friday) are only compatible with your performing one of (a) or (b), then you can’t be free to choose either, because this would imply that you are free to choose the past/or the physical laws, which you can’t. Here, though, we pull a “one person’s *reductio ad absurdum*”: because only one of (a) or (b) is compatible with the past/the physical laws, and because you are free to choose (a) or (b), it turns out that in some sense, you’re free to choose the past/the physical laws (or, their computational analogs).

What? That can’t be right. But isn’t it, in the practically relevant sense? Consider: the case is basically one where, if it’s the case that your state on Monday (call this Monday-Joe), copied and evolved according to deterministic process P, outputs “cooperate,” then you get a million dollars; and if it outputs “defect,” you get a thousand dollars (see e.g. [Ahmed \(2014\)](#)’s “Betting on the Past” for an even simpler version of this). It’s Friday now. The state of Monday-Joe is fixed; Monday-Joe lives in the past. And process P, let’s say, was fixed on Monday, too. In this sense, the

question of what Monday-Joe + process P outputs is already fixed. You, on Friday, are evolving-Joe: that is, Monday-Joe-in-the-midst-of-evolving-according-to-process-P. If you choose cooperate, it will always have been the case that Monday-Joe + process P outputs cooperate. If you choose defect, it will always have been the case that Monday-Joe + process P outputs defect. In this very real sense – the same sense at stake *in every choice in a deterministic world* – you get to choose what will have always been the case, even before your choice.

Try it. It will really work. Make your Friday choice, then leave the simulation, go get an old and isolated copy of Monday-Joe and Process P – one that's been housed, since Monday, somewhere you could not have touched or tampered with — press play, and watch what comes out the other end. You won't be surprised.

Is that changing the past? In one sense: no. It's not that Joe's state on Monday was X, but then because of what Evolving-Joe did on Friday, Joe's state on Monday became Y instead. Nor does the output of Monday-Joe + Process P alter over the course of the week. Don't be silly. You can't change these things like you can change the contents of your fridge: milk on one day, juice on the next. It's not milk at noon on Monday, and then on Friday, juice at noon on Monday instead. We must distinguish between the ability to "change things" in this sense, and the ability to "control" them in some broader sense.

But nevertheless: you *get to decide*, on Friday, the thing that will always have been true; the one thing that will always have been in your fridge, since the beginning of time. And perhaps this approaches, ultimately, the full sense of compatibilist decision-making, compatibilist "control," even in cases of causal influence. Perhaps, that is, you can change the past, here, about as much as you can change the future in a deterministic world: that is, not at all, and enough to matter for practical purposes. After all, in such a world, the future is already fixed by the past. Your ability to decide that future was, therefore, always puzzling. Perhaps your ability to decide the past isn't much more so (though certainly, it's no less).

CDT can't handle this kind of thing. CDT imagines that we have severed the ties between you and your copy, between you and the history that determines every aspect of you. It imagines that you can hold your copy's arm fixed, and move yours freely; that you can break apart the future from the past, and let the future swing, at your pleasure, along some physically (indeed, logically!) impossible hinge. But you can't. The echoes of your choice started before you chose. You are implicated in a structure that reverberates in all directions. You pull your arm, and the past and the universe trail behind; and yet, the past and universe push your arm; and yet, neither: you, the past, the future, the universe, are all born in the same timeless instant — free, fixed, consistent, a full and living painting of someone painting it as they go along.

And CDT's mistake, here, is not just abstract misconception: rather, it misleads you in straightforward and practically-relevant ways. In particular, *it prompts CDT to compare actions using expected utilities that you shouldn't actually expect* – which, when you step back, seems pretty silly. Suppose, for example, that as a CDT agent, you start out with a credence p that your copy will defect of 99%. Thus, as in Newcomb's problem above, your payoffs are:

- *Expected utility from defecting*: \$1K guaranteed + \$10K from a 1% probability of getting a million from my copy = \$11K.

- *Expected utility from cooperating:* \$10K from a 1% probability of getting a million from my copy = \$10K.

But you shouldn't actually expect only \$10k, if you cooperate, given the logical necessity of his doing what you do. That's just ... not the right number. So why are you considering it? This is no time to play around with fantasy distributions over outcomes; there's real money on the line. And of course, this sort of objection will hold for any p . As long as you and your copy's choice are correlated, CDT is going to ignore that correlation, hold p constant given different actions, and in that sense, prompt you to choose as though your probabilities are wrong.

EDT does better, here, of course: choosing based on what utility you, as a Bayesian, should actually expect, given different actions, is EDT's *forté* by definition, and a powerful argument in its favor (see e.g. Christiano's "simple argument for EDT" [here](#)). And the considerations about compatibilism and determinism I've been discussing seem friendly to EDT as well. After all, if you are a living in already-painted painting, it seems unsurprising if choice comes down to something like "managing the news." The problem with managing the news, after all, was supposed to be that the news was already fixed. But in an already-painted painting, the future has already been fixed, too: you just don't know what it is. And when you act, you start to find out. Insofar as you can choose how to act – and per compatibilism, you can – then you can choose what you're going to find out, and in that sense, influence it. Do you hope that this already-fixed universe is one where you eat a sandwich? Well, go make a sandwich! If you do, you'll discover that your dream for the universe has always been true, since the beginning of time. If you don't make a sandwich, though, your dream will die. Why should the applicability of such reasoning be limited by the scope of "causation" (whatever that is)?

IV. What if the case is less clean?

I took pains, above, to specify that the copying process was perfect, and the inputs received exactly identical. It's perfectly possible to satisfy this constraint, and we don't need to use "atom-for-atom" copies and the like, or assume determinism at a physical level; we can just make you an AI system running in a deterministic simulation. What's more, this constraint helps make the point more vivid; and it suffices, I think, to show that CDT is wrong.

However, I don't think it's necessary. Consider, for example, a version where there are small errors in the copying process; or in which you get a blue hat, and your copy, a red; or in which your environment involves some amount of randomness. These may or may not suffice to ruin your ability to write exactly what you want on his whiteboard. But very plausibly, the strong correlation between your choice of button, and his, will persist: and to the extent it does, this information is worthy of inclusion in your decision-making process.

What if you know that your copy has already chosen, before you make your choice? To the extent that the correlations between your choice and his persist in such conditions, I think that the same argument applies. Note, though, that your knowing that he's already chosen means that the two of you got different inputs in a sense that seems more likely to affect your decision-making than getting different colored hats. That is, you saw a light indicating "your copy has already chosen"; he didn't; and some people, faced with a light of that kind, start acting all weird about how "his choice is already made, I can't affect it, might as well defect" and so on, in a way that

they don't when the light is off. So the question of what sorts of correlations are still at stake is more up for grabs. Does learning that you cooperate, after seeing such a light, still make it more likely that he cooperated, without seeing one? If so, that seems worth considering.

(This sort of "different inputs" dynamic also blocks certain types of loops/contradictions that could come from learning what a deterministic copy of you already did. E.g., if you learn what he chose — say, that he cooperated — before you make your choice, it's still compatible with the case's set up that you defect, as long as he got different inputs: e.g., he didn't also learn that *you* cooperated. If he *did* "learn" that you cooperated, then things are getting more complicated. In particular, either you will in fact cooperate, or some feature of the case's set-up is false. This is similar to how, if you travel back in time and try to kill your grandfather, either you will in fact fail, or the case's set-up is false. Or to how, if you hear an infallible prediction that you'll do X, then either you will in fact do X, or the prediction wasn't infallible after all.)

V. Monopoly money

I think that "perfect deterministic twin prisoner's dilemma"-type cases suffice to show that CDT is wrong. But I also want to note another type of argument I find persuasive, in the context of Newcomb's problem, and which also evokes the type of "magic" I have in mind.

Imagine doing "tryout runs" of Newcomb's problem, using monopoly money, as many times as you'd like, before facing the real case (h/t Drescher (2006) again). You try different patterns of one-boxing and two-boxing, over and over. Every time you one-box, the opaque box is full. Each time you two-box, it's empty.

You find yourself thinking: "wow, this Omega character is no joke." But you try getting fancier. You fake left, then go right — reaching for the one box, then lunging for the second box too at the last moment. You try increasingly complex chains of reasoning. Before choosing, you try deceiving yourself, bonking yourself on the head, taking heavy doses of hallucinogens. But to no avail. You can't pull a fast one on ol' Omega. Omega is right every time.

Indeed, pretty quickly, it starts to feel like you can basically just *decide* what the opaque box will contain. "Shazam!" you say, waving your arms over the boxes: "I hereby make it the case that Omega put a million dollars into the box." And thus, as you one box, it is so. "Shazam!" you say again, waving your arms over a new set of boxes: "I hereby make it the case that Omega left the box empty." And thus, as you two-box, it is so. With Omega's help, you feel like you have become a magician. With Omega's help, you feel like you can choose the past.

Now, finally, you face the true test, the real boxes, the legal tender. What will you choose? Here, I expect some feeling like: "I know this one; I've played this game before." That is, I expect to have learned, in my gut, what one-boxing, or two-boxing, will lead to — to feel viscerally that there are really only two available outcomes here: I get a million dollars, by one boxing, or I get a thousand, by two-boxing. The choice seems clear.

VI. Against undue focus on folk-theoretical names

Of course, the same two-boxing responses I noted above apply here, too. It's true that every time you one-box, you would've gotten an extra \$1,000 if you'd two-boxed, assuming CDT's "counterfactual" construal of "would." It's true that you leave the \$1,000 dollars on the table; that is this is predictably regrettable for some sense of "regret"; and we can say, for this reason, that "Omega is just play-rewarding your play-irrationality." I don't have especially deep responses to these objections. But I find myself persuaded, nevertheless, that one-boxing is the way to go.

Or at least, it's my way. When I step back in Newcomb's case, I don't feel especially attached to the idea that it's *the way*, the only "rational" choice (though I admit I feel this non-attachment less in perfect twin prisoner's dilemmas, where defecting just seems to me pretty crazy). Rather, it feels like my conviction about one-boxing start to bypass debates about what's "rational" or "irrational." Faced with the boxes, I don't feel like I'm asking myself "what's the rational choice?" I feel like I'm, well, *deciding what to do*. In one sense of "rational" – e.g., the counterfactual sense – two-boxing is rational. In another sense – the conditional sense — one-boxing is. What's the "true sense," the "real rationality"? Mu. Who cares? What's that question even about? Perhaps, for the [normative realists](#), there is some "true rationality," etched into the platonic realm; a single privileged way that the normative Gods demand that you arrange your mind, on pain of being... what? "Faulty"? Silly? Subject to a certain sort of criticism? But for the anti-realists, there is just the world, different ways of doing things, different ways of using words, different amounts of money that actually end up in your pocket. Let's not get too hung up on what gets called what.

There's a great line from [David Lewis](#), which I often think of on those rare and clear-cut occasions when philosophical debate starts to border on the terminological.

"Why care about objective value or ethical reality? The sanction is that if you do not, your inner states will fail to deserve folk-theoretical names. Not a threat that will strike terror into the hearts of the wicked! But whoever thought that philosophy could replace the hangman?"

I want to highlight, in particular, the idea of "failing to deserve folk-theoretical names." Too often, philosophy – especially normative philosophy — devolves into a debate about what kind of name-calling is appropriate, when. But faced with the boxes, or the buttons, our eyes should not be on the folk-theoretical names at stake. Rather, our eyes should be on the choice itself.

Note that my point here is not that "rationality is about winning" (see e.g. [Yudkowsky \(2009\)](#)). "Winning," here, is subject to the same ambiguity as "rational." One-boxers tend to end up richer, yes. But faced with a choice between \$1k, or nothing (the choice that the two-boxer is actually presented with), \$1k is the winning choice. Still, I am with Yudkowsky in spirit, in that I think that too much interest in the word "rational" here is apt to move our eyes from the prize.

(All that said, I'm going to continue, in what follows, to use the standard language of "what's rational," "what you should do," etc, in discussing these cases. I hope that this language will be interpreted in a sense that connects directly to the actual, visceral process of deciding what to do, name-calling be damned. I acknowledge, though, that there's a possible motte-and-bailey dynamic here, where the one-boxer goes in hard for claims like "CDT is wrong" and "c'mon, defecting in perfect twin prisoner's dilemmas is just ridiculous!" and then backs off to "hey man, you've got your way, I've got my way, what's all this obsession with the word 'rationality'?" when pressed about the counterintuitive consequences of their own position. And more broadly, it

can be hard to combine object level normative debate, which often reflects with a kind of “realist” flavor, with adequate consciousness and communication of some more fundamental meta-ethical arbitrariness. If necessary, we might go back through the whole post and try to rewrite it in more explicitly anti-realist terms — e.g., “I reject CDT.” But I’ll skip that, partly because I suspect that something beyond naive meta-ethical realism gets lost in this sort of move, even if we don’t have an explicit account of what it is.)

VII. Identity crises are no defense of CDT

I’ve now covered two data-points that I take to speak very strongly against CDT: namely, that one should cooperate in a twin prisoner’s dilemma, and that one should one-box in Newcomb’s problem. I want to briefly discuss an unusual way of trying to get CDT to one-box: namely, by appealing to uncertainty about whether you faced with the real boxes, or whether you are in a simulation being used by Omega to predict your future choice (see e.g. [Aaronson \(2005\)](#) and [Critch \(2017\)](#) for suggestions in this vein, though not necessarily in these specific terms). Basically, I don’t think this move works, in general, as a way of saving CDT, though the type of uncertainty in question might be relevant in other ways.

How is the story supposed to go? Imagine that you know that the way Omega predicts whether you’ll one-box, or two-box, is by running an extremely high-fidelity simulation of you. And suppose that both real-you and sim-you only care about what happens to real-you. By hypothesis, sim-you shouldn’t be able to figure out whether he’s simulated or real, because then he’ll serve as worse evidence about real-you’s future behavior (for example, if sim-you appears in a room with writing on the wall saying “you’re the sim,” then he can just one-box, thereby causing Omega to add the money to the opaque box, thereby allowing real-you, appearing in a room saying “you’re the real one,” to two-box, get the full million-point-one, and make Omega’s “prediction” wrong). So it needs to be the case that you’re uncertain – let’s say, 50-50 — about whether you’re simulated or not. Thus, the thought goes, you should one-box, because there’s a 50% chance that doing so will cause Omega to put the million in the box, and your real-self (who will also, presumably, one-box, given the similarity between you) will get it.

(*Calculation: feel free to skip.* Suppose that you currently expect yourself to one-box, as both real-you and sim-you, with 99% probability. Then the CDT calculation runs as follows:

- 50% chance you’re the sim, in which case:
 - EV of one-boxing = 99% chance real-you gets a \$1M, 1% chance real-you gets \$1M + \$1K = \$1,000,010.
 - EV of two-boxing = 99% chance real-you gets nothing, 1% chance real-you gets \$1K = \$10.
- 50% chance you’re real, in which case:
 - EV of one-boxing: 99% chance real-you gets \$1M, 1% chance real-you gets nothing = \$990,000.
 - EV of two-boxing: 99% chance real-you gets \$1M + \$1K, 1% chance real-you gets \$1K = \$991,000.
- So overall:
 - EV of one-boxing = 50% * \$1,000,010 + 50% * \$990,000 = \$995,005.
 - EV of two-boxing = 50% * \$10 + 50% * \$991,000 = \$495,505.

Depending on the details, CDT may then need to adjust its probability that both sim-you and real-you one-box. But high-confidence that both versions of you one-box is a stable equilibrium (e.g., CDT still one-boxes, given such a belief); whereas high-confidence that both will two-box is not (e.g., CDT one-boxes, given such a belief). There are also some problems, here, with making such calculations consistent with assigning a specific probability to Omega being right in her prediction, but I'm setting those aside.)

My objections here are:

1. This move doesn't work if you're indexically selfish (e.g., you don't care about copies of yourself).
2. This move doesn't work for twin prisoner's dilemma cases more broadly.
3. It's not clear that simulations are necessary for predicting your actions in the relevant cases.
4. In general, it really doesn't feel like this type of thing is driving my convictions about these cases.

Let's start with (1). Suppose that real-you and sim-you aren't united in sole concern for real-you. Rather, suppose that you're both out for yourselves. Sim-you, let's suppose, faces bleak prospects: whatever happens, Omega is going to shut down the simulation right after sim-you's choice gets made. So sim-you doesn't give a shit about this whole ridiculous situation with the god-damn boxes; the world is dust and ashes. Real-you, by contrast, is a CDT agent. So real-you, left to his own devices, is a two-boxer. Hence, sim-you doesn't care, and real-you wants to two-box; and thus, uncertain about who you are, you two-box.

(*Calculation, feel free to skip.* Suppose you start out 99% confident that both versions of you will two-box. Thus:

- 50% chance you're the sim, in which case: you get nothing no matter what.
- 50% chance you're real, in which case:
 - EV of one-boxing: 1% chance of \$1M, 99% chance of nothing = \$10,000.
 - EV of two-boxing: 1% chance of \$1M + \$1K, 99% chance of \$1k = \$11,000
- So overall:
 - EV of one-boxing = $50\% * \$0 + 50\% * \$10,000 = \$5,000$.
 - EV of two-boxing = $50\% * \$0 + 50\% * \$11,000 = \$5,500$.

This dynamic holds regardless of your initial probabilities on how different versions of you will act, and regardless of your probability on being the sim vs. being real.)

Of course, real-you can try to "acausally induce" sim-you to one-box, by one-boxing himself. But "acausally inducing" other versions of yourself to do stuff isn't the CDT way; rather, it's the type of magical thinking silliness that CDT is supposed to eschew.

Perhaps one objects: sim-you should care about real-you! For one thing, though, this seems unobvious: indexical selfishness seems perfectly consistent and understandable (and indeed, for anti-realists, you can care about whatever you want). But more importantly, it's an objection to a utility function, rather than to two-boxing per se; and decision theorists don't generally go in for objecting to utility functions. If the claim is that "CDT is compatible with indexically altruistic agents one-boxing in Newcomb cases involving simulations," then fair enough. But what about everyone else?

This leads us to objection (2): namely, that the twin prisoner's dilemma, which I take to be one of the strongest reasons to reject CDT, is precisely a case of indexical selfishness. Perhaps I am uncertain about which copy I am; but regardless, I only care about myself; and on CDT, whatever that other guy does, I should defect. But defecting on your perfect deterministic twin, I claim, is totally crazy, even if you are *indexically selfish*. So CDT, I think, is still wrong.

What's more, as I noted above, we can imagine versions of the case where I *do* know who I am; for example, I am the one with the blue hat, he's the one with the red hat; I am the one who want to create flourishing Utopias, and he (the authorities changed my values during the copying process) wants to create paperclips. Unlike "sim vs. real," these distinctions that are epistemically accessible. Still, though, if my choices are sufficiently correlated with those of my copy (and mutual cooperation is sufficiently beneficial), I should cooperate.

This is related to objection (3): namely, that not all cases where CDT gives the wrong verdicts involve simulations, or uncertainty about "who you are." Twin prisoner's dilemmas, where you are slightly but discernably different from your twin, are one example: no simulations or predictions necessary. But we might also wonder about Newcomb cases more broadly. Does Omega really need to be predicting your behavior via a simulation or model that you might actually *be*, in order for one-boxing to be the right call? This seems, at least, a substantively additional claim. And we might wonder about e.g. predicting your behavior via your genes (see e.g. [Oesterheld \(2015\)](#)), by observing lots of people who are "a lot like you," or some via other unknown method.

That said, I want to acknowledge that one of the arguments for one-boxing that I find most persuasive – e.g., running the case lots of times with "play money," before deciding what to do for real – works a lot better in contexts with very fine-grained prediction capabilities. This is because when I'm "playing around" with no real stakes, it makes more sense to imagine me using intricate and arbitrary decision-making processes, which the incentives at stake in the real case will not constrain. Thus, for example, maybe I try forms of pseudo-randomization ("I'll one-box if the number of letters in the sentence I'm about to make up is odd" – see Aaronson [here](#)); maybe I try spinning myself around with my eyes closed, then pressing whichever button I see first; and so on. In order for Omega's predictions to stay well-correlated with my behavior, here, it seems plausible she needs a very (unrealistically?) high-fidelity model. And we can say something similar about the twin prisoner's dilemma. That is, the argument for cooperating is most compelling when his arm literally moves in logically-necessary lock-step with your own, as you reach towards the buttons. Once that's not true, if we try to imagine a "play money" version of the case, then even with fairly minor psychology differences, you and your copy's modes of "playing around" might de-correlate fast.

This feature of the intuitive landscape seems instructive. The sense that you acausally "control" what Omega predicts, or what your copy does, seems strongest when you can, as it were, *do any old thing, for any old reason*, and the correlation will remain. Once the correlation requires further constraints, the intuitive case weakens. That said, if you're in the real case, with the real incentives, then it's ultimately the correlation *given those incentives* that seems relevant: e.g., maybe Omega is accurate only for real-money cases; maybe you and your copy are only highly correlated when the real money comes out. In such a case, I think, you should still one-box/cooperate.

My final objection to the “appeal to uncertainty about who are you” sort of view just: it doesn’t feel like uncertainty about whether I’m a simulation is actually driving my one-boxing impulse. In the play-money Newcomb case, for example, I feel like what actually persuades me is a visceral sense that “one-boxing is going to result in me having a million dollars, two-boxing is going to result in me having a thousand dollars.” Questions about whether I’m a simulation, or whether Omega needs to simulate me in order to achieve this level of accuracy, just aren’t coming into it.

I conclude, then, that simulation uncertainty and related ideas can’t save CDT. Aaronson thinks that he can “pocket the \$1,000,000, but still believe that the future doesn’t affect the past.” I think he’s wrong — at least in many cases where one wants the million, and can get it. He should face, I think, a weirder music.

VIII. Maybe EDT?

But what sort of music, exactly? And exactly how weird are we talking? I don’t know.

Consider, for example, EDT – CDT’s most famous rival. I think that a lot of philosophers write off EDT too quickly. As I mentioned earlier, EDT has the unique and compelling distinction of being the only view to use the *utility you should actually expect*, given the performance of action X, in order to calculate the expected utility of performing action X. In this sense, it’s the basic, simple-minded Bayesian’s decision theory; the type of decision theory you would use if you were, you know, *trying to predict the outcomes of different actions*.

What’s more, I think, a number of prominent objections to EDT seem to me, at least, much more complicated than they’re often made out to be. Consider, for example, the accusation that EDT endorses attempts to “manage the news.” There’s something true about this, but we should also be puzzled by it. Managing the news is obviously fine when you can influence the events the news is about. It’s fine, for example, to “manage the news” about whether you get a promotion, by working harder at the office. And it’s interestingly hard to “manage the news” successfully – e.g., change your rational credence in how good the future will be – with respect to things you *can’t* influence. Suppose, for example, that you’re worried (at, say, 70% credence) that your favored candidate lost yesterday’s election. Do you “manage the news” by refusing to read the morning’s newspaper, or by scribbling over the front page “Favored Candidate Wins Decisively!”? No: if you’re rational, your credence in the loss is still 70%.

Or take a somewhat more complicated case, discussed in [Ahmed \(2014\)](#). Suppose that you wake up not knowing what time it is, and all your clocks are broken. You hope that you’re not already late to work, and you consider running, to avoid either being late at all, or being later. Suppose, further, that people who run to work tend to be already late. Should you refrain from running, on the grounds that running would make it more likely that you’re already late? No. But plausibly, EDT doesn’t say you should, because *running to work, in this case, wouldn’t be additional evidence that you’re already late*, once we condition on the fact that you don’t know when you woke up, the reasons (including the subtle hunches about what time it might be) that you’d be running, and so on. After all, many of the already-late people running for work *know* that they’re already late, and are running for that reason. Your situation is different.

OK, so what does it take for the problematic type of news-management to be possible? This question matters, I think, because in some of the examples where EDT

is supposed to go in for the problematic type of news-management, it's not clear that the news-management in question would succeed. Consider:

Smoking lesion: Almost everyone who smokes has a fatal lesion, and almost everyone who doesn't smoke doesn't have this lesion. However, smoking doesn't cause the lesion. Rather, the lesion causes people to smoke. Dying from the lesion is terrible, but smoking is pretty good. Should you smoke?

EDT, the objection goes, doesn't smoke, here, because smoking increases your credence that you have the lesion. But this, the thought goes, is stupid. You've either already got the lesion, or you don't have it and won't get it. Either way, you should smoke. Not smoking is just "managing the news."

I used to treat this case a fairly decisive reason to reject EDT. Now I feel more confused about it. For starters, EDT clearly smokes in some versions of the case. Suppose, for example, that the way the lesion causes people to smoke is by making them *want* to smoke. Conditional on someone wanting to smoke, though, there's no additional correlation between *actually smoking* and having the lesion. Thus, if you notice that you want to smoke (e.g., you feel a "tickle"), then that's the bad news right there: you've already got all the smoking-related evidence you're going to get about whether you've got the lesion. Actually smoking, or not, doesn't change the news: so, no need for further management. This sort of argument will work for any mechanism of influence on your decision that you notice and update on. Thus the so-called "Tickle Defense" of EDT.

Ok, but what if you don't notice any tickle, or whatever other mechanism of influence is at stake? As [Ahmed \(2014, p. 91\)](#) characterizes it, the tickle defense assumes that all the inputs to your decision-making are "transparent" to you. But this seems like a strong condition, and granted greater ignorance, my sense is that in some versions of the case (for example, versions where the lesion makes you assign positive utility to smoking, *but you don't know what your utility function is*, even as you use it in making decisions), EDT is indeed going to give the intuitively wrong result (see e.g., Demski's "[Smoking Lesion Steelman](#)" for a worked example). [Christiano](#) argues that this is fine – "No matter how good your decision procedure is, if you don't know a critical fact about the situation then you can make a decision that looks bad" – but I'm not so sure: *prima facie*, not smoking in smoking-lesion type cases seems like the type of mistake one ought to be able to avoid, even granted uncertainty about some aspects of your own psychology, and/or how the lesion works.

More generally, though, my sense is that really trying to dig into the details of tickle-defense type moves gets complicated fast, and that there's some tension between (a) trying to craft a version of EDT where the "tickle defense" always works – e.g., one that somehow updates on everything influencing its decision-making (I'm not sure how this is supposed to work) – and (b) keeping EDT meaningfully distinct from CDT (see e.g. Demski's sequence "[EDT = CDT?](#)"). Maybe some people are OK with collapsing the distinction, and OK, even, if EDT starts two-boxing in Newcomb's problems (see e.g. Demski's final comments [here](#)), and defecting on deterministic twins (I've been setting this possibility aside above, and following the standard understanding of how EDT acts in these cases). But for my part, a key reason I'm interested in EDT at all is because I'm interested in one-boxing and cooperating. Maybe I can get this in other ways (see e.g. the discussion of "follow the policy you would've committed to" below); but then, I think, EDT will lose much of its appeal (though not all; I also like the "basic Bayesian-ness" of it).

One other note on smoking lesion. You might think that the “do it over and over with monopoly money” type argument that I found persuasive earlier will give the intuitively wrong verdict on smoking lesion, suggesting that such an argument shouldn’t be trusted. After all, we might think, almost every time you smoke in a “play life,” you’ll end up with the play-lesion; and every time you don’t, you won’t. But note that when we dig in on this, the smoking lesion case can start to break in a maybe-instructive manner.

Suppose, for example, that I know that the base rate of lesions in the population is 50%, and I get “spawned” over and over into the world, where I can choose to smoke, or not. How can my “playing around” remain consistent with this 50% base rate? Imagine, for example, that I decide to refrain from smoking a million times in a row. If the case’s hypothesized correlations hold, then I will in fact spawn, consistently, without the lesion. In that case, though, it starts to look like my choice of whether to smoke or not actually is exerting a type of “control” over whether I get born as someone with the lesion – in defiance of the base rate. And if my choice can do *that*, then it’s not actually clear to me that non-smoking, here, is so crazy.

Maybe we could rule this out by fiat? “Well, if the base rate is 50%, then it turns you will, in fact, decide to ‘play around’ in a way that involves smoking ~50% of the time” (thanks to Katja Grace for discussion). But this feels a bit forced, and inconsistent with the spirit of “play around however you want; it’ll basically always work” – the spirit that I find persuasive in Newcomb’s case and sufficiently-high-fidelity twin prisoner’s dilemmas. Alternatively, we could specify that I’m not allowed to know the base rate, and then we can shift it around to remain consistent with my making whatever play choices I want *and* spawning at the base rate. But now it looks like I can control the base rate of lesions! And if I can do that, once again, I start to wonder about whether non-smoking is so crazy after all.

That said, maybe the right thing to say here is just that the correlations posited in smoking lesion don’t persist under conditions of “play around however you want” – something that I expect holds true of various versions of Newcomb’s problem and Twin Prisoner’s dilemma as well.

What about other putative counter-examples to EDT? There are lots to consider, but at least one other one – namely, “Yankees vs. Red Sox” (see [Arntzenius \(2008\)](#)) — strikes me as dubious (though also, elegant). In this case, the Yankees win 90% of games, and you face a choice between the following bets:

	Yankees win	Red Sox win
You bet on Yankees	1	-2
You bet on Red Sox	-1	2

Or, if we think of the outcomes here as “you win your” and “you lose your bet” instead, we get:

	You win your bet	You lose your bet
You bet on Yankees	1	-2
You bet on Red Sox	2	-1

Before you choose your bet, an Oracle tells you whether you're going to win your next bet. The issue is that once you condition on winning or losing (regardless of which), you should always bet on the Red Sox. So, the thought goes, EDT always bets on the Red Sox, and loses money 90% of the time. Betting on the Yankees every time does much better.

But something is fishy here. Specifically, the Oracle's prediction, together with your knowledge of your own decision, leaks information that should render your decision-making unstable. Suppose, for example, that the Oracle tells you that you will lose your next bet. You then reason: "Conditional on knowing that I will lose my bet, I should bet on the Red Sox. But given that I'll lose, this means that the Yankees will win, which means I should bet on the Yankees, which means I will win my bet. But I can't win my bet, so the Yankees will lose, so I should bet on the Red Sox," and so on. That is, you oscillate between reasoning using the second matrix, and reasoning using the first; and you never settle down.

(Note that if we allow for playing around with monopoly money, then this case, too, suffers from the same base-rate related problems as smoking lesion: e.g., either you can change the base rates of Yankee victory at will, or you're somehow forced to play around in a manner consistent with both the 90% base rate and the Oracle's accuracy, or somehow the Oracle's accuracy doesn't hold in conditions where you can play around.)

Even if we set aside smoking lesion and Yankees vs. Red Sox, though, there is at least one counterexample to EDT that seems to me pretty solidly damning, namely:

XOR blackmail: Termites in your house is a million-dollar loss, and you don't know if you have them. A credible and accurate predictor finds out if you have termites, then writes the following letter: "I am sending you this letter if and only if (a) I predict that you will pay me \$1,000 dollars upon receiving it, or (b) you have termites, but not both." She then makes her prediction and follows the letter's outlined procedure. If you receive the letter, should you pay?

(See [Yudkowsky and Soares \(2017\)](#), p. 24).

EDT pays, here. Why? Because conditional on paying, it's much less likely that you've got termites, so paying is much better news than not paying. If you refuse to pay, you should call the exterminator (or do [whatever you do with termites](#)) pronto; if you pay, you can relax.

Or at least, you can relax for a bit. But if you're EDT, *you're getting these letters all the time*. Maybe the predictor decides to pull this stunt every day. You're flooded with letters, all reflecting the prediction that you'll pay. If you'd only stop paying, the letters would slow to a base-rate-of-termites-sized trickle. Try it with monopoly money: as you spawn over and over, you'll find you can modulate the frequency of letter receipt at will, just by deciding to pay, or not, on the next round. But in real life, once you've got the letter, do you ever wise up, and decide, instead of paying, to already have termites? On EDT, it's not clear (at least to me) why you would, absent some other change to the situation. Termites, after all, are terrible. And look at this letter, already sitting in your hand! It only comes given one of two conditions...

Perhaps one thinks: the core issue here isn't that you're getting so many letters. Even if you know that the predictor is only going to pull this stunt once, paying seem pretty silly. Why? It's that old thing about the past having already happened, about the

opaque box already being empty or full. You've either already got termites, or you don't, dude: stop trying to manage the news.

But is that the core issue? Consider:

More active termite blackmail: The predictor gets more aggressive. Once a year, she writes the following letter: "I predicted that you would pay me \$1,000 upon receipt of this letter. If I predicted 'yes,' I left your house alone. If I predicted 'no,' I gave you termites." Then she predicts, obeys the procedure, and sends. If you receive the letter, should you pay?

Here, the "it's too late, dude" objection still applies. CDT ignores letters like this. But CDT also gets given termites once a year. EDT, by contrast, pays, and stays termite free. What's more, by hypothesis, the stunt gets pulled on everyone the same number of times, regardless of their payment patterns. In this sense, it's more directly analogous to Newcomb's problem. And I find that paying, here, seems more intuitive than in the previous case (though the fact that you ultimately want to deter this sort of behavior from occurring at all may bring in additional complications; if it helps, we can specify that the predictor's not actually in this for getting money or for giving people termites — rather, she just likes putting people in weird decision-theory situations, and will do this regardless of how her victims respond).

We can consider other problems with EDT as well, beyond XOR blackmail. For example, a naïve formulation of EDT has trouble with cases where it starts out certain about what it's going to do, or even very confident (see e.g. the "cosmic ray problem" on p. 24 of [Yudkowsky and Soares](#) (2017)). And more generally, the "managing the news" flavor of EDT makes it feel, to me, like the type of thing one could come up with counter-examples to. But it's XOR blackmail, I find, that currently gives me the most pause (and note, too, that in XOR blackmail, we can imagine that you have arbitrary introspective access, such that tickle-defense type questions about whether all the factors influencing your decision are "transparent" or not don't really apply). And I think that the importance of the way paying *influences how many letters you get*, as opposed to its trying to "control the past" more broadly, may be instructive.

Summarizing this section, then: my current sense is that:

1. EDT's "basic Bayesianism" makes it attractive.
2. Really digging into EDT, especially re: tickle defenses, can get kind of gnarly.
3. Yankees vs. Red Sox isn't a good counterargument to EDT.
4. EDT messes up in XOR blackmail.
5. There are probably a bunch of other problems with EDT that I'm not really considering/engaging with.

Does this make EDT better or worse than CDT? Currently, I'm weakly inclined to say "better" - at least in theory. But trying to actually implement EDT also seems more liable to lead to pretty silly stuff. I'll discuss some of this silly stuff in the final section. First, though, and motivated by XOR blackmail, I want to discuss one more broad bucket of decision-theoretic options and examples – namely, those associated with following policies you would've wanted yourself to commit to, even when it hurts.

IX. What would you have wanted yourself to commit to?

Consider:

Parfit's hitchhiker: You are stranded in the desert without cash, and you'll die if you don't get to the city soon. A selfish man comes along in a car. He is an extremely accurate predictor, and he'll take you to the city if he predicts that once you arrive, you'll go to an ATM, withdraw ten thousand dollars, and give it to him. However, once you get to the city, he'll be powerless to stop you from not paying.

If you get to the city, should you pay him? Both CDT and EDT answer: no. By the time you get to the city, the risk of death in the desert is gone. Paying him, then, is pure loss (assuming you don't value his welfare, and there are no other downstream consequences). Because they answer this way, though, both CDT and EDT agents rarely make it to the city: the man predicts, accurately, that they won't pay.

Is this a problem? Some might answer: no, because paying in the city is clearly irrational. In particular, it violates what [MacAskill \(2019\)](#) calls:

Guaranteed Payoffs: When you're certain about what the pay-offs of your different options would be, you should choose the option with the highest pay-off.

Guaranteed Payoffs, we should all agree, is an attractive principle, at least in the abstract. If you're not taking the higher payoff, when you know exactly what payoffs your different actions will lead to, then what the heck *are* you doing, and why would we call it "rationality"?

On the other hand, is paying the driver really so silly? To me, it doesn't feel that way. Indeed, I feel happy to pay, here (though I also think that the case brings in extra heuristics about promise-keeping and gratitude that may muddy the waters; better to run it with a mean and non-conscious AI system who demands that you just burn the money in the street, and kills itself before you even get to the ATM). What's more, *I want to be the type of person who pays.* Indeed: if, in the desert, I could set-up some elaborate and costly self-binding scheme – say, a bomb that blows off my arm, in the city, if I don't pay — such that paying in the city becomes straightforwardly incentivized, I would want to do it. But if that's true, we might wonder, why not skip all this expensive faff with the bomb, and just, you know, pay in the city? After all, what if there are no bombs around to strap to my arm? What if I don't know how to make bombs? Need my survival be subject to such contingencies? Why not learn, and practice, that oh-so valuable (and portable, and reliably available) skill instead: how to make, and actually keep, commitments? (h/t Carl Shulman, years ago, for suggesting this sort of framing.)

That said, various questions tend to blur together here – and once we pull them apart, it's not clear to me how much substantive (as opposed to merely verbal) debate remains. Everyone agrees that it's better to be the type of person who pays. Everyone agrees that if you can credibly commit to paying, you should do it; and that the ability to make and keep commitments is an extremely useful one. Indeed, everyone agrees that, if you're a CDT or EDT agent about to face this case, it's better, if you can, to self-modify into some other type of agent – one that will pay in the city (and are commitments and self-modifications really so different? Is cognition itself so different from self-modification?). As far as I can tell (and I'm not alone in thinking this), the only remaining dispute is whether, given these facts, we should baptize the action of paying in the city with the word "rational," or if we should instead call it "an irrational action, but one that follows from a disposition it's rational to cultivate, a self-modification it's rational to make, a policy its rational to commit to," and so on.

Is that an interesting question? What's actually at stake, when we ask it? I'm not sure. As I mentioned above, I tend towards anti-realism about normativity; and for anti-realists, debates about the "true rationality" aren't especially deep. Ultimately, there are just different ways of arranging your mind, different ways of making decisions, different shapes that can be given to this strange clay of self and world. Ultimately, that is, the question is just: what you in fact do in the city, and what in fact that decision means, implies, causes, and so on. We talk about "rationality" as a means of groping towards greater wisdom and clarity about these implications, effects, and so on; but if you understand all of this, and make your decisions in light of full information, additional disputes about what compliments and insults are appropriate don't seem especially pressing.

All that said, terminology aside, I do think that Parfit's hitchhiker-type cases can lead to genuinely practical and visceral forms of internal conflict. Consider:

Deterrence: You have a button that will destroy the world. The aliens want to invade, but they want the world intact, and they won't invade if they predict that you'll destroy the world upon observing their invasion. Being enslaved by the aliens is better than death; but freedom far better. The aliens predict that you won't press the button, and so start to invade. Should you destroy the world?

This is far from a fanciful thought experiment. Rather, this is precisely the type of dynamic that decision-makers with real nuclear codes at their fingertips have to deal with. Same with [tree-huggers chaining themselves to trees](#), teenagers playing [chicken](#), and so on.

Or, more fancifully, consider:

Counterfactual mugging: Omega doesn't know whether the X-th digit of pi is even or odd. Before finding out, she makes the following commitment. If the X-th digit of pi is odd, she will ask you for a thousand dollars. If the X-th digit is even, she will predict whether you would've given her the thousand had the X-th digit been odd, and she will give you a million if she predicts "yes." The X-th digit is odd, and Omega asks you for the thousand. Should you pay?

(I use logical randomness, rather than e.g. coin-flipping, to make it more difficult to appeal to concern about versions of yourself that live in other quantum branches, possible worlds, and so on. Thanks to Katja Grace for suggesting this. That said, perhaps some such appeals are available regardless. For example, how did X get decided?)

Finally, consider a version of Newcomb's problem in which both boxes are transparent – e.g., you can see how Omega has predicted you'll behave. Suppose you find that Omega has predicted that you'll one-box, and so left the million there. Should you one-box, or two-box? What if Omega has predicted that you'll two-box?

We can think of all these cases as involving an inconsistency between the policy that an agent would want to adopt, at some prior point in time/from some epistemic position (e.g., before the aliens invade, before we know the value of the X-th digit, before Omega makes her predictions), and the action that *Guaranteed Payoffs* would mandate given full information. And there are lots of other cases in this vein as well (see e.g., [The Absent-Minded Driver](#), and the literature on [dynamical inconsistency](#) in game theory).

There is a certain broad class of decision theories, a number of which are associated with the [Machine Intelligence Research Institute](#) (MIRI), that put resolving this type of inconsistency in favor of something like “the policy you would’ve wanted to adopt” at center stage. (In general, MIRI’s work on decision theory has heavily influenced my own thinking – influence on display throughout this post. See also [Meacham \(2010\)](#) for another view in this vein, as well as the work of Wei Dai and others on “[updatelessness](#).”) There are lots of different ways to do this (see e.g. the discussion of the 2x2x3 matrix [here](#)), and I don’t feel like I have a strong grip on all of the relevant choice-points. Many of these views are united, though, in violating Guaranteed Payoffs, for reasons that feel, spiritually, pretty similar.

What’s more, and importantly, these theories tend to get cases like XOR blackmail right, where e.g. classic EDT gets them wrong. Consider, for example, whether before you receive any letter, you would want to commit to paying, or not paying, upon receipt. If we assume that the base rate of termites will stay constant regardless, then committing to not paying seems the clear choice. After all, doing so won’t make it more likely that you get termites; rather, it’ll make it less likely that you get letters.

If necessary, these theories can also get results like one-boxing, and cooperating with your twin, without appeal to any weird magic about controlling the past. After all, one-boxing and cooperating are both policies that you would want yourself to commit to, at least from some epistemic positions, even in a plain-old, common-sense, CDT-spirited world. Maybe executing these policies *looks* like trying to execute some kind of acausal control — and maybe, indeed, advocates of such policies talk in terms of such control. But maybe this is just talk. After all, executing policies that violate Guaranteed Payoffs looks pretty weird in general (for example, it looks like burning money for certain), and perhaps we need not take decisions about how to conceptualize such violations all that seriously: the main thing is what happens with the money.

A key price of this approach, though, is the whole “burning money for certain” thing; and here, perhaps, some people will want to get off the train. “Look, I was down for one-boxing, or for cooperating with my twin, when I didn’t actually *know* the payoffs in question. But violating Guaranteed Payoffs is just too much! *You’re just destroying value for certain*. That’s all. That’s the whole thing you do. You blow up the world, trying to prevent something that you know has already happened. Yes, it’s good to commit to doing that ex ante. But ex post, isn’t it also just obviously stupid?”

For people with this combination of views, though, I think it’s important to keep in mind the spiritual continuity between violating Guaranteed Payoffs, and one-boxing/cooperating more generally. After all, one of the strongest arguments for two-boxing is that, if you knew what was in the box (like, e.g., your friend does), you’d be in a Guaranteed Payoffs-type situation, and then a follower of Guaranteed Payoffs would two-box every time. Indeed, I think that part why “great grandpappy Omega, now long dead, leaves the boxes in the attic” prompts a two-boxing intuition is that in the attic, you sense that you’re about to move from a non-transparent Newcomb’s problem to a transparent one. That is, after you bring the one-box down from the attic, and open it, the other box isn’t going to disappear. The attic door is still open. The stairs still beckon. You could just go back up there and get that thousand. Why not do it? If you got the million, it’s not going to evaporate. And if you didn’t get the million, what’s the use of letting a thousand go to waste? But that’s just the type of thinking that leads to empty boxes...

X. Building statues to the average of all logically-impossible Gods

Overall, I don't see violations of Guaranteed Payoffs as a decisive reason to reject approaches in the vein of "act in line with the policy you would've wanted to commit to from some epistemic position P" – and some disputes in this vicinity strike me as verbal rather than substantive. That said, I do want to flag an additional source of uncertainty about such approaches: namely, that it seems extremely unclear what they actually imply.

In particular, all the "violate Guaranteed Payoffs" cases above rely on some implied "prior" epistemic position (e.g., before the aliens invade, before Omega has made her prediction, etc), relative to which the policy in question is evaluated. But why is that the position, instead of some other one? Even if we were just "rewinding" your own epistemology (e.g., to back before you knew that the aliens were invading, but after you learned that about how they were going to make their decision), there would be a question of how far to rewind. Back to your childhood? Back to before you were born, and were an innocent platonic soul about to be spawned into the world? What features does this soul have? In what order were those features added? Does your platonic soul know basic facts about logic? What credence does it have that it'll get born as a square circle, or into a world where $2+2=5$? What in the goddamn hell are we talking about?

Also, it isn't just a question of "rewinding" your own epistemology to some earlier epistemic position you (or even, a stripped-down version of you) held. There may be no actual time when you knew the information you're trying to "remember" (e.g., that Omega is going to pull a counter-factual-mugging type stunt) but not the information you're trying to "forget" (e.g., that the X-th digit of pi is odd). So it seems like the epistemic position in question may need to be one that no one – and certainly not you – has ever, in fact, occupied. How are we supposed to pick out such a position? What desiderata are even relevant? I haven't engaged much with questions in this vein, but currently, I basically just don't know how this is supposed to work. (I'm also not the only one with these questions. See e.g. Demski [here](#), on the possibility that "updatelessness is doomed," and Christiano [here](#). And they've thought more about it.)

What's more, some (basically all?) of these epistemic positions don't seem particularly exciting from a "winning" perspective — and not just because they violate Guaranteed Payoffs. For example: weren't you a member of some funky religion as a child — one that you now reject? And weren't you more generally kind of dumb and ignorant? Are you sure you want to commit to a policy from that epistemic position (see e.g. [Kokotajlo \(2019\)](#) for more)? Or are we, maybe, imagining a superintelligent version of your childhood self, who knows everything? But wait: don't forget to forget stuff, too, like what will end up in the boxes. But what should we "forget," what should we "remember," and what should we learn-for-the-first-time-because-apparently-we're-talking-about-superintelligences now?

And even if we had such an attractive and privileged epistemic position identified, it seems additionally hard (totally impossible?) to know what policy this position would actually imply. Suppose, to take a normal everyday example that definitely doesn't involve any theoretical problems, that you are about to be inserted as a random "soul" into a random "world." What policy should you commit to? As Parfit's Hitchhiker, should you pay in the city? Or should you, perhaps, commit to not getting into the man's car at all, even if doing so is free, in order to disincentivize your

younger self from taking ill-advised trips into the desert? Or should you, perhaps, commit to carving the desert sands into statues of square circles, and then burning yourself at the stake as an offering to the average of all logically impossible Gods? One feels, perhaps, a bit at sea; and a bit at risk of, as it were, doing something dumb. After all, you've already gone in for burning value for certain; you've already started trying to reason like someone you're not, in a situation that you aren't in. And without constraints like "don't burn value for certain" as a basic filter on your action space, the floodgates open wide. One worries about swimming well in such water.

XI. Living with magic

Overall, the main thing I want to communicate in this post is: I think that the perfect deterministic twin's prisoner's dilemma case basically shows that there is such a thing as "acausal control," and that this is super duper weird. For all intents and purposes, you can decide what gets written on whiteboards light-years away; you can move another man's arm, in lock-step with your own, without any causal contact between him and you. It actually works, and that, I think, is pretty crazy. It's not the type of power we think of ourselves as having. It's not the type of power we're used to trying to wield.

What does trying to wield it actually look like, especially in our actual lives? I'm not sure. I don't have a worked out decision-theory that makes sense of this type of thing, let alone a view about how to apply it. As a first pass, though, I'd probably start by trying to figure out what EDT actually implies, once you account for (a) tickle-defense type stuff, and (b) decorrelations between your decision and the decisions of others that arise because you're doing some kind of funky EDT-type reasoning, and they probably aren't.

For example: suppose that you want other people to vote in the upcoming election. Does this give you reason to vote, not out of some sort of abstract "be the change you want to see in the world" type of ethic, but because, more concretely, your voting, even in causal isolation from everyone else, will literally (if acausally) increase non-you voter turnout? Let's first stop and really grok that voting for this reason is a weird thing to do. You're not just trying to obey some Kantian maxim, or to do your civic duty. You're not just saying "what if everyone acted like that?" in the abstract, like a schoolteacher to an errant child, with no expectation that "everyone," as it were, actually will. And you're certainly not knocking on doors or driving neighbors to the polls. Rather, you're literally trying to influence the behavior of other people you'll never interact with, by walking down to the voting booth on your causally isolated island. Indeed, maybe your island is in a different time zone, and you know that the polls everywhere else are closed. Still, you reason, your choice's influence can slip the surly bonds of space and time; the evening news can still be managed (indeed, some non-EDT decision theories vote even after they've seen the evening news).

Is this sort of thinking remotely sensible? Well, note that the EDT version, at least, makes sense *only if you should actually expect a higher non-you voter turnout, conditional on you voting for this sort of reason, than otherwise*. If the voting population is "perfect deterministic copies of myself who will see the exact same inputs," this condition holds; and it holds in various weaker conditions, too. How much does it hold in the real world, though? That's much less clear; and as ever, if you're considering trying to manage the news, the first thing to check is whether the news is actually manageable.

In particular, as Abram Demski emphasizes [here](#), the greater the role of weird-decision-theory type calculations in your thinking, the less correlated your decisions will be those of others who are thinking in less esoteric ways. Perhaps you should consider the influence of your behavior on the other people interested in non-causal decision-theories (evening news: “the weird decision theorists turn out in droves!”); but it’s a smaller demographic. That said, what sorts of correlations are at stake here is an empirical question, and there’s no guarantee that something common-sensical will emerge victorious. It seems possible, for example, that many people are implicitly implementing some proto-version of your decision theory, even if they’re not explicit about it.

Here’s another case that seems to me even weirder. Suppose that you’re reading about some prison camps from World War I. They sound horrible, but the description leaves many details unspecified, and you find yourself hoping that the guards in the prison camps were as nice as would be compatible with the historical evidence you’ve seen thus far. Does this give you, perhaps, some weak reason to be nicer to other people, in your own life, on the grounds that there is some weak correlation between your niceness, and the niceness of the guards? You’re all, after all, humans; you’ve got extremely similar genes; you’re subject to broadly similar influences; perhaps you and some of the guards are implementing vaguely similar decision procedures at some level; perhaps even (who knows?) there was some explicit decision theory happening in the trenches. Should you try to be the change you want to see in the past? Should you, now, try to improve the conditions in World War I prison camps? And if so: have you, perhaps, lost your marbles?

Perhaps some people will answer: look, the correlations are too weak, here, for such reasoning to get off the ground. To others, though, this will seem the wrong sort of reply. The issue isn’t that you’re wrong, empirically, about the correlations at stake – indeed, the extent of such correlations seems, in some sense, an open question. The issue is that you’re trying to improve the past at all.

There are other weird applications to consider as well. For example, once you can “control” things you have no causal interaction with, your sphere of possible control could in principle expand throughout a very large universe, allowing you to “influence” the behavior of aliens, other quantum branches, and so on (see e.g. [Oesterheld \(2017\)](#) for more). Indeed, there’s an argument for treating yourself as capable of such influence, even if you have comparatively low credence on the relevant funky decision theories, because being able to influence the behavior of tons of agents raises the stakes of your choice (see e.g. [MacAskill et al \(2019\)](#)). And taken seriously enough, the possibility of non-causal influence can lead to a very non-standard picture of the future – one in which “interactions” between causally-isolated civilizations throughout the universe/multi-verse move much closer to center stage.

Once you’ve started trying to acausally influence the behavior of aliens throughout the multiverse, though, one starts to wonder even more about the whole lost-your-marbles thing. And even if you’re OK with this sort of thing in principle, it’s a *much* further question whether you should expect any efforts in this broad funky-decision-theoretic vein to go well in practice. Indeed, my strong suspicion is that with respect to multiverse-wide whatever whatevers, for example, any such efforts, undertaken with our current level of understanding, will end up looking very misguided in hindsight, even if the decision theory that motivated them ends up vindicated. Here I think of Bostrom’s “[ladder of deliberation](#),” in which one notices that whether an intervention seems like a good or bad idea switches back and forth as one reasons about it more, with no end in sight, thus inducing corresponding pessimism about the

reliability of one's current conclusions. Even if the weird-decision-theory ladder is sound, we are, I think, on a pretty early rung.

Overall, this whole "acausal control" thing is strange stuff. I think we should be careful with it, and generally avoid doing things that look stupid by normal lights, especially in the everyday situations our common-sense is used to dealing with. But the possibility of new, weird forms of control over the world also seems like the type of thing that could be important; and I think that perfect deterministic twins demonstrate that something in this vicinity is, at least sometimes, real. Its nature and implications, therefore, seem worth attention.

(My thanks to Paul Christiano, Bastian Stern, Nisan Stiennon, and especially to Katja Grace and Ketan Ramakrishnan, for discussion. And thanks, as well, to Abram Demski, Scott Garrabrant, Nick Beckstead, Rob Bensinger, and Ben Pace, for [this exchange](#) on related topics.)

Outline of Galef's "Scout Mindset"

Julia Galef's [*The Scout Mindset*](#) is superb.

For effective altruists, I think (based on the topic and execution) it's straightforwardly the #1 book you should use when you want to recruit new people to EA. It doesn't actually talk much about EA, but I think starting people on this book will result in an EA that's thriving more and doing more good five years from now, compared to the future EA that would exist if the top go-to resource were more obvious choices like *The Precipice*, *Doing Good Better*, the EA Handbook, etc.

For rationalists, I think the best intro resource is still *HPMoR* or *R:AZ*, but I think *Scout Mindset* is a great supplement to those, and probably a better starting point for people who prefer Julia's writing style over Eliezer's.

I've made an outline of the book below, for my own reference and for others who have read it. If you don't mind spoilers, you can also use this to help decide whether the book's worth reading for you, though my summary skips a lot and doesn't do justice to Julia's arguments.

Introduction

- Scout mindset is "the motivation to see things as they are, not as you wish they were".
- We aren't perfect scouts, but we can improve. "My approach has three prongs":
 1. *Realize that truth isn't in conflict with your other goals.* People tend to overestimate how useful self-deception is for things like personal happiness and motivation, starting a company, being an activist, etc.
 2. *Learn tools that make it easier to see clearly.* Use various kinds of thought experiments and probabilistic reasoning, and rethink how you go about listening to the "other side" of an issue.
 3. *Appreciate the emotional rewards of scout mindset.* "It's empowering to be able to resist the temptation to self-deceive, and to know that you can face reality even when it's unpleasant. There's an equanimity that results from understanding risk and coming to terms with the odds you're facing. And there's a refreshing lightness in the feeling of being free to explore ideas and follow the evidence wherever it leads". Looking at lots of real-world examples of people who have exemplified scout mindset can make these positives more salient.

PART I: The Case for Scout Mindset

Chapter 1. Two Types of Thinking

- "Can I believe it?" vs. "must I believe it?" In directionally motivated reasoning, often shortened to "motivated reasoning", we disproportionately put our effort

into finding evidence/reasons that support what we wish were true.

- *Reasoning as defensive combat.* Motivated reasoning, a.k.a. soldier mindset, "doesn't feel like motivated reasoning from the inside". But it's extremely common, as shown by how often we describe our reasoning in militaristic terms.
- *"Is it true?"* An alternative to (directionally) motivated reasoning is accuracy motivated reasoning, i.e., scout mindset.
- *Your mindset can make or break your judgment.* This stuff matters in real life, in almost every domain. Nobody is purely a scout or purely a soldier, but it's possible to become more scout-like.

Chapter 2. What the Soldier is Protecting

- "[I]f scout mindset is so great, why isn't everyone already using it all the time?"
Three emotional reasons:
 1. *Comfort: avoiding unpleasant emotions.* This even includes comforting pessimism: "there's no hope, so you might as well not worry about it."
 2. *Self-esteem: feeling good about ourselves.* Again, this can include ego-protecting negativity and avoiding "'getting my hopes up'".
 3. *Morale: motivating ourselves to do hard things.*
- And three social reasons:
 1. *Persuasion: convincing ourselves so we can convince others.*
 2. *Image: choosing beliefs that make us look good.* "Psychologists call it impression management, and evolutionary psychologists call it signaling: When considering a claim, we implicitly ask ourselves, 'What kind of person would believe a claim like this, and is that how I want others to see me?'"
 3. *Belonging: fitting in to your social groups.*
- "We use motivated reasoning not because we don't know any better, but because we're trying to protect things that are vitally important to us". So it's no surprise that e.g. 'training people in critical thinking' don't help change people's thinking. But while "soldier mindset is often our default strategy for getting what we want", it's not generally the best strategy available.

Chapter 3. Why Truth is More Valuable Than We Realize

- *We make unconscious trade-offs.* "[T]he whole point of self-deception is that it's occurring beneath our conscious awareness. [...] So it's left up to our unconscious minds to choose, on a case-by-case basis, which goals to prioritize." Sometimes it chooses to be more soldier-like, sometimes more scout-like.
- *Are we rationally irrational?* I.e., are we good at "unconsciously choosing just enough epistemic irrationality to achieve our [instrumental] social and emotional goals, without impairing our judgment too much"? No. "There are several major biases in our decision-making [...] that cause us to overvalue soldier mindset":
 1. *We overvalue the immediate rewards of soldier mindset.* Present bias: we prefer small rewards now over large rewards later.
 2. *We underestimate the value of building scout habits.* Cognitive skills are abstract (and, again, have most of their benefits in the future), so they're harder to notice and care about.
 3. *We underestimate the ripple effects of self-deception.* These ripple effects are "delayed and unpredictable", which is "exactly the kind of cost we tend to neglect".
 4. *We overestimate social costs.*
- *An accurate map is more useful now.* Humans have more options now than we did tens of thousands of years ago, and more ability to improve our

circumstances. "So if our instincts undervalue truth, that's not surprising—our instincts evolved in a different world, one better suited to the soldier."

PART II: Developing Self-Awareness

Chapter 4. Signs of a Scout

- "A key factor preventing us from being in scout mindset more frequently is our conviction that we're already in it." Examples of "things that make us feel like scouts even when we're not":
 1. *Feeling objective doesn't make you a scout.*
 2. *Being smart and knowledgeable doesn't make you a scout.* On ideologically charged questions, learning more tends to make people more polarized; and even scientists studying cognitive biases have a track record of exhibiting soldier mindset.
- *Actually practicing scout mindset makes you a scout.* "The test of scout mindset isn't whether you see yourself as the kind of person who [changes your mind in response to evidence, is fair-minded, etc. ...] It's whether you can point to concrete cases in which you did, in fact, do those things. [...] The only real sign of a scout is whether you act like one." Behavioral cues to look for:
 1. *Do you tell other people when you realize they were right?*
 2. *How do you react to personal criticism?* "Are there examples of criticism you've acted upon? Have you rewarded a critic (for example, by promoting him)? Do you go out of your way to make it easier for other people to criticize you?"
 3. *Do you ever prove yourself wrong?*
 4. *Do you take precautions to avoid fooling yourself?* E.g., "Do you avoid biasing the information you get?" and "[D]o you decide ahead of time what will count as a success and what will count as a failure, so you're not tempted to move the goalposts later?"
 5. *Do you have any good critics?* "Can you name people who are critical of your beliefs, profession, or life choices who you consider thoughtful, even if you believe they're wrong? Or can you at least name reasons why someone might disagree with you that you would consider reasonable[...]"?
 6. "But the biggest sign of scout mindset may be this: Can you point to occasions in which you were in soldier mindset? [... M]otivated reasoning is our natural state," so if you never notice yourself doing it, the likeliest explanation is that you're not self-aware about it.

Chapter 5. Noticing Bias

- "One of the essential tools in a magician's tool kit is a form of manipulation called forcing." The magician asks you to choose between two cards. "If you point to the card on the left, he says, 'Okay, that one's yours.' If you point to the card on the right, he says, 'Okay, we'll remove that one.' [...] If you could see both of those possible scenarios at once, the trick would be obvious. But because you end up in only one of those worlds, you never realize."
- "Forcing is what your brain is doing to get away with motivated reasoning while still making you feel like you're being objective." The Democratic voter doesn't notice that they're going easier on a Democratic politician than they would on a

Republican, because the question "How would I act if this politician were a Republican?" isn't salient to them, or they're tricking themselves into thinking they'd apply the same standard.

- A thought experiment is a peek into the counterfactual world. "You can't detect motivated reasoning in yourself just by scrutinizing your reasoning and concluding that it makes sense. You have to compare your reasoning to the way you would have reasoned in a counterfactual world, a world in which your motivations were different—would you judge that politician's actions differently if he was in the opposite party? [...] Would you consider that study's methodology sound if its conclusion supported your side? [...] Try to actually imagine the counterfactual scenario. [... D]on't simply formulate a verbal question for yourself. Conjure up the counterfactual world, place yourself in it, and observe your reaction." Five types of thought experiment:
 1. *The double standard test.* Am I judging one person/group by a standard I wouldn't apply to another person/group?
 2. *The outsider test.* "Imagine someone else stepped into your shoes—what do you expect they would do in your situation?" Or imagine that you're an outsider who just magically teleported into your body.
 3. *The conformity test.* "If other people no longer held this view, would you still hold it?"
 4. *The selective skeptic test.* "Imagine this evidence supported the other side. How credible would you find it then?"
 5. *The status quo bias test.* "Imagine your current situation was no longer the status quo. Would you then actively choose it?"
- Thought experiments on their own "can't tell you what's true or fair or what decision you should make." But they allow you to catch your brain "in the act of motivated reasoning," and take that into account as you work to figure out what's true.
- Beyond the specific thought experiments, the core skill of this chapter is "a kind of self-awareness, a sense that your judgments are *contingent*—that what seems true or reasonable or fair or desirable can change when you mentally vary some features of the question that should have been irrelevant."

Chapter 6. How Sure Are You?

- *We like feeling certain.* "Your strength as a scout is in your ability [...] to think in shades of gray instead of black and white. To distinguish the feeling of '95% sure' from '75% sure' from '55% sure'."
- *Quantifying your uncertainty.* For scouts, probabilities are predictions of how likely they are to be right. The goal is to be calibrated in the probabilities you assign.
- *A bet can reveal how sure you really are.* "Evolutionary psychologist Robert Kurzban has an analogy[...] In a company, there's a board of directors whose role is to make the crucial decisions for the company—how to spend its budget, which risks to take, when to change strategies, and so on. Then there's a press secretary whose role it is to give statements[...] The press secretary makes *claims*; the board makes *bets*. [...] A bet is any decision in which you stand to gain or lose something of value, based on the outcome."
- *The equivalent bet test.* By comparing different bets and seeing when you prefer taking one vs. the other, or when they feel about the same, you can translate your feeling "does X sound like a good bet?" into probabilities.
- The core skill of this chapter is "being able to tell the difference between the feeling of *making a claim* and the feeling of *actually trying to guess what's true*."

PART III: Thriving Without Illusions

Chapter 7. Coping with Reality

- *Keeping despair at bay.* Motivated reasoning is especially tempting in emergencies; but it's also especially dangerous in emergencies. In dire situations, it's essential to be able to keep despair at bay without distorting your map of reality. E.g., you can count your blessings, come to terms with your situation, or remind yourself that you're doing the best you can.
- *Honest vs. self-deceptive ways of coping.* Honest ways of coping with painful or difficult circumstances include:
 1. *Make a plan.* "It's striking how much the urge to conclude 'That's not true' diminishes once you feel like you have a concrete plan for what you would do if the thing were true."
 2. *Notice silver linings.* "You're recognizing a silver lining to the cloud, not trying to convince yourself the whole cloud is silver. But in many cases, that's all you need".
 3. *Focus on a different goal.*
 4. *Things could be worse.*
- *Does research show that self-deceived people are happier?* No, the research quality is terrible.

Chapter 8. Motivation Without Self-Deception

- Using self-deception to motivate yourself is bad, because:
 1. *An accurate picture of your odds helps you choose between goals.*
 2. *An accurate picture of the odds helps you adapt your plan over time.*
 3. *An accurate picture of the odds helps you decide how much to stake on success.*
- *Bets worth taking.* "[S]couts aren't motivated by the thought, 'This is going to succeed.' They're motivated by the thought, 'This is a bet worth taking.'" Which bets are worth taking is a matter of their expected value.
- *Accepting variance gives you equanimity.* Expecting to always succeed is unrealistic, and will lead to unnecessary disappointments. "Instead of being elated when your bets pay off, and crushed when they don't," try to get a realistic picture of the variance in bets and focus on ensuring your bets have high expected value.
- *Coming to terms with the risk.*

Chapter 9. Influence Without Overconfidence

- *Two types of confidence.* Epistemic confidence is "how sure you are about what's true," while social confidence is self-assurance: "Are you at ease in social situations? Do you act like you deserve to be there, like you're secure in yourself and your role in the group? Do you speak as if you're worth listening to?" Influencing people requires social confidence, which people conflate with epistemic confidence.
- *People judge you on social confidence, not epistemic confidence.* Various studies show that judgments of competence are mediated by perceived social (rather than epistemic) confidence.
- *Two kinds of uncertainty.* People trust you less if you seem uncertain due to ignorance or inexperience, but not if you seem uncertain due to reality being

messy and unpredictable. Three ways to communicate uncertainty without looking inexperienced or incompetent:

1. *Show that uncertainty is justified.*
 2. *Give informed estimates.* "Even if reality is messy and it's impossible to know the right answer with confidence, you can at least be confident in your analysis."
 3. *Have a plan.*
- *You don't need to promise success to be inspiring.* "You can paint a picture of the world you're trying to create, or why your mission is important, or how your product has helped people, without claiming you're guaranteed to succeed. There are lots of ways to get people excited that don't require you to lie to others or to yourself."
 - "That's the overarching theme of these last three chapters: whatever your goal, there's probably a way to get it that doesn't require you to believe false things."

PART IV: Changing Your Mind

Chapter 10. How to Be Wrong

- *Change your mind a little at a time.* Superforecasters constantly revise their views in small ways.
- *Recognizing you were wrong makes you better at being right.* Most people, when they learn they were wrong, give excuses like "I Was Almost Right". Superforecasters instead "reevaluate their process, asking, 'What does this teach me about how to make better forecasts?'"
- *Learning domain-general lessons.* Even if your error is in a domain that seems unimportant to you, noticing your errors can teach domain-general lessons "about how the world works, or how your own brain works, and about the kinds of biases that tend to influence your judgment." Or they can help you fully internalize a lesson you previously only believed in the abstract.
- *"Admitting a mistake" vs. "updating".* Being factually wrong about something doesn't necessarily mean you screwed up. Learning new information should usually be thought of in matter-of-fact terms, as an opportunity to update your beliefs—not as something humbling or embarrassing.
- *If you're not changing your mind, you're doing something wrong.* By default, you should be learning more over time, and changing your strategy accordingly.

Chapter 11. Lean in to Confusion

- Usually, "we react to observations that conflict with our worldview by explaining them away. [...] We couldn't function in the world if we were constantly questioning our perception of reality. But especially when motivated reasoning is in play, we take it too far, shoehorning conflicting evidence into a narrative" well past the point where it makes sense. "This chapter is about how to resist the urge to dismiss details that don't fit your theories, and instead, allow yourself to be confused and intrigued by them, to see them as puzzles to be solved".
- "You don't know in advance" what surprising and confusing observations will teach you. "All too often, we assume the only two possibilities are 'I'm right' or 'The other guy is right'[...] But in many cases, there's an unknown unknown, a hidden 'option C,' that enriches our picture of the world in a way we wouldn't have been able to anticipate."

- *Anomalies pile up and cause a paradigm shift.* "Acknowledge anomalies, even if you don't yet know how to explain them, and even if the old paradigm still seems correct overall. Maybe they'll add up to nothing in particular. Maybe they just mean that reality is messy. But maybe they're laying the groundwork for a big change of view."
- *Be willing to stay confused.*

Chapter 12. Escape Your Echo Chamber

- *How not to learn from disagreement.* Listening to the "other side" usually makes people *more* polarized. "By default, we end up listening to people who initiate disagreements with us, as well as the public figures and media outlets who are the most popular representatives of the other side." But people who initiate disagreements tend to be unusually disagreeable, and popular representatives of an ideology are often "ones who do things like cheering for their side and mocking or caricaturing the other side—i.e., you". To learn from disagreement:
 1. Listen to people you find reasonable.
 2. Listen to people you share intellectual common ground with.
 3. Listen to people who share your goals.
- *The problem with a "team of rivals."* "Dissent isn't all that useful from people you don't respect or from people who don't even share enough common ground with you to agree that you're supposed to be on the same team."
- *It's harder than you think.* "We assume that if both people are basically reasonable and arguing in good faith, then getting to the bottom of a disagreement should be straightforward[...] When things don't play out that way [...] everyone gets frustrated and concludes the others must be irrational." But even under ideal conditions, learning from disagreements is still hard, e.g., because:
 1. We misunderstand each other's views.
 2. Bad arguments inoculate us against good arguments.
 3. Our beliefs are interdependent—changing one requires changing others.

PART V: Rethinking Identity

Chapter 13. How Beliefs Become Identities

- *What it means for something to be part of your identity.* Criticizing part of someone's identity tends to spark passionate, combative, and defensive reactions. Two things that tend to turn a belief into an identity:
 1. Feeling embattled.
 2. Feeling proud.
- *Signs a belief might be an identity.*
 1. Using the phrase "I believe".
 2. Getting annoyed when an ideology is criticized.
 3. Defiant language.
 4. A righteous tone.
 5. Gatekeeping.
 6. Schadenfreude.
 7. Epithets.
 8. Having to defend your view.

- "Identifying with a belief makes you feel like you have to be ready to defend it, which motivates you to focus your attention on collecting evidence in its favor. Identity makes you reflexively reject arguments that feel like attacks on you or on the status of your group. [...] And when a belief is part of your identity it becomes far harder to change your mind[.]"

Chapter 14. Hold Your Identity Lightly

- *What it means to hold your identity lightly.* Rather than trying to have no identities, you should try to "keep those identities from colonizing your thoughts and values. [...] Holding your identity lightly means thinking of it in a matter-of-fact way, rather than as a central source of pride and meaning in your life. It's a description, not a flag to be waved proudly."
- *Could you pass an ideological Turing test?* Passing means explaining an ideology "as a believer would, convincingly enough that other people couldn't tell the difference between you and a genuine believer". The ideological Turing test tests your knowledge of the other side's beliefs, but "it also serves as an emotional test: Do you hold your identity lightly enough to be able to avoid caricaturing your ideological opponents?"
- *A strongly held identity prevents you from persuading others.*
- *Understanding the other side makes it possible to change minds.*
- *Is holding your identity lightly compatible with activism?* Activists usually "face trade-offs between identity and impact," and holding your identity lightly can make it easier to focus on the highest-impact options.

Chapter 15. A Scout Identity

- *Flipping the script on identity.* Identifying as a truth-seeker can make you a better scout.
- *Identity makes hard things rewarding.* When you act like a scout, you can take pride and satisfaction in living up to your values. This short-term reward helps patch our bias for short-term rewards, which normally favors soldier mindset.
- *Your communities shape your identity.* "[I]n the medium-to-long term, one of the biggest things you can do to change your thinking is to change the people you surround yourself with."
- *You can choose what kind of people you attract.* You can't please everyone, so "you might as well aim to please the kind of people you'd most like to have around you, people who you respect and who motivate you to be a better version of yourself".
- *You can choose your communities online.*
- *You can choose your role models.*

Curing insanity with malaria

Sometimes the history of medicine is very, very surreal. For example, consider that in 1927, a physician named Julius Wagner-Jauregg received the Nobel Prize in medicine, for...deliberately infecting his patients with malaria. As a treatment for psychosis.

This often worked.

Well, it did kill around 15% of the patients, but it was nonetheless seen as a miracle cure.

[General paralysis of the insane](#) was first identified and described as a distinct disease in the early 19th century. It was initially thought to be caused by an ‘inherent weakness of character’. The initial symptoms were of mental deterioration and personality changes; patients suffered a loss of social inhibitions, gradual impairment of judgment, concentration and short-term memory. They might experience euphoria, mania, depression, or apathy. Delusions were common, including “ideas of great wealth, immortality, thousands of lovers, and unfathomable power” – or, on the more negative side, nihilism, self-guilt, and self-blame.

It was a progressive disease, and nearly always a death sentence. As the condition advanced, the patient would develop worsening dementia, motor and reflex abnormalities, and often seizures; death usually took 3 to 5 years from the initial symptoms. In the 19th century, cases of general paralysis could account for up to 25% of admissions to asylums.

Some physicians were drawing a connection between general paralysis and syphilis infection as early as the 1850s; however, it took until much later for this explanation to be generally accepted within the medical community, and full confirmation via pathology examinations of the brains of patients who had died of the disease would have to wait until 1913.

In 1909, an antisyphilitic drug compound was discovered via a process of trialing hundreds of newly synthesized organic arsenical chemicals, looking for one that would have anti-microbial activity but not kill the human patient; this was the first research team effort to optimize biological effects of a promising chemical, which is now the basis of a huge amount of pharmaceuticals research. Unfortunately, [arsphenamine](#), also known as Salvarsan or “606”, was difficult to prepare and administer, and was still fairly toxic to the human patient as well as the syphilis.

[Julius Wagner-Jauregg](#) was a Viennese psychiatrist, but a psychiatrist with a particular interest in experimental pathology, and in brains. Already in the mid-1880s, he was noticing an odd pattern; many of his psychiatric patients were showing improvements in their mental condition after recovering from bouts of other illnesses that resulted in fever.

Wagner-Jauregg formed two hypotheses. One, some cases of insanity had ‘organic’, biological causes and were related to physical dysfunctions in the brain; two, one disease could be fought by another. He tried deliberately inducing fevers in his patients, by injecting them with tuberculin, a sterile protein extract from cultures of the tubercle bacillus responsible for tuberculosis. However, this was inconsistent at producing a fever, and the results were disappointing.

In 1917, a soldier ill with malaria was admitted to Wagner-Jauregg’s ward. No, I am not at all sure why a malaria patient was being treated in a psychiatric ward! And, apparently, neither was [Wagner-Jauregg](#):

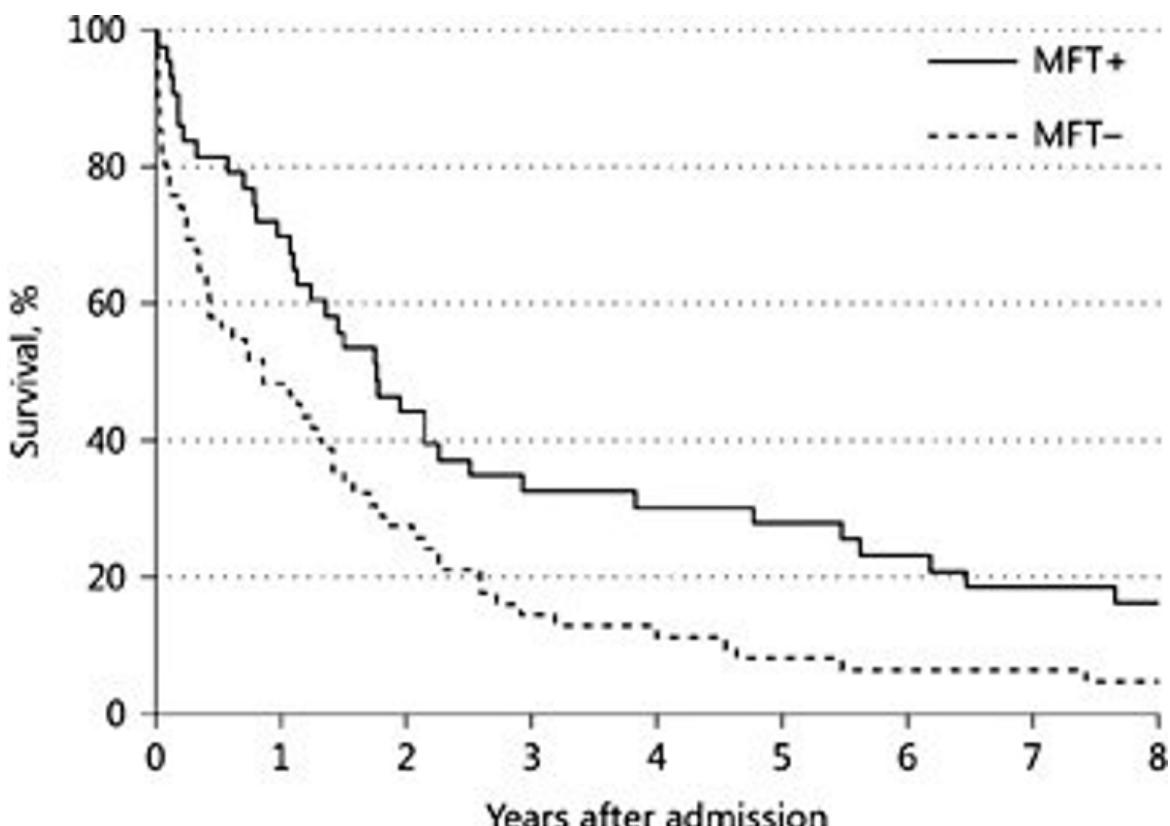
“Should he be given quinine?” [my assistant Dr. Alfred Fuchs] asked. I immediately said: “No.” This I regarded as a sign of destiny. Because soldiers with malaria were usually not admitted to my wards, which accepted only cases suffering from a psychosis or patients with injuries to the central nervous system.

Wagner-Jauregg would have known that malaria is especially likely to cause repeated, intermittent paroxysms of high fever. Also, unlike with general paralysis, quinine was available as a treatment and reasonably safe. Since general paralysis was still mostly incurable, he must have felt that there wasn't much to lose. He made the bold choice to draw blood from the sick soldier and inject it into nine of his psychiatric patients diagnosed with general paralysis. It is deeply unclear from sources on this whether he bothered to obtain consent from any of the patients involved. Six of the nine saw improvements in their psychiatric condition, and only one patient is reported to have died of the fever.

(Unfortunately, but perhaps unsurprisingly given his predilection for mad science, Wagner-Jauregg was later a [proponent of eugenics](#), and backed a proposal for a law that would ban "people with mental diseases and people with criminal genes" from reproducing. His application to join the Nazi party was, apparently, rejected on the basis that his first wife was Jewish.)

In 1921, Wagner-Jauregg published a report claiming therapeutic success in treating GPI patients with malaria, and this became the standard treatment until the discovery of penicillin in the 1940s. Tens of thousands of patients were treated with deliberate malaria infections. Special psychiatric clinics were opened for this purpose. There were various attempts to produce fevers in safer ways, mostly via hot baths, electric blankets, or "fever cabinets" but sometimes via injection of toxic sulfur compounds; none were as successful as malaria.

According to a [historical cohort study](#), despite the high risk of this treatment – between 5% and 20% of patients would die from the 'cure' – patients treated with malariotherapy did have significantly better chances than they would otherwise. 70% were alive a year after admission, compared to 48% of untreated cases; at 5 years, 28% of malaria-treated patients were alive versus only 8% at baseline. Patients who had only recently contracted syphilis – and thus presumably had less irreversible neurological damage – could be cured entirely, especially if the malarial fever was followed by Salvarsan treatments.



Graph of survival after admission

It wasn't a *great* treatment, and it was obviously far from safe, but given the prognosis for general paralysis and the lack of other good options, it's not surprising that it was seen as revolutionary.

Even now, it's not fully understood how the fever resulted in a cure; it's unlikely that the patients' body temperatures were high enough for a prolonged enough period to directly kill the spirochetes responsible for syphilis infection. Another hypothesis is that the infection stimulated the patient's immune system to a higher level of activity, which also boosted the body's defenses against the syphilis infection.

Even once penicillin was discovered, the treatment wasn't immediately accepted, and was often given in combination with malariotherapy; this was done in the United States and in the Netherlands up to the mid-1960s, and in the United Kingdom until the 1970s.

The popularity of pyrotherapy during this period resulted in significantly more research effort going toward the biological study of malaria, including its mode of transmission and treatment. The first permanent laboratory colonies of mosquitoes, and the isolation of various malaria strains, were both established during this time period. Testing of synthetic drugs for malaria treatment was another related advance. It seems likely that malaria is much better understood now than it would be if this historical interlude had never happened.

Wagner-Jauregg's work here also pioneered the field of 'stress therapies' for psychiatric illnesses, including induced insulin coma therapy for schizophrenia. Electroconvulsive therapy, also popularized during this time period, is still used as a treatment for refractory depression today.

Links

- [Malaria as Lifesaving Therapy](#)
- [Hot Brains: Manipulating Body Heat to Save the Brain](#)
- [Julius Wagner-Jauregg Biography \(1857-1940\)](#)

The Death of Behavioral Economics

This is a linkpost for <https://www.thebehavioralscientist.com/articles/the-death-of-behavioral-economics>

Edit: I recommend reading [Scott's response to this essay](#) in addition to the essay itself.

I've been tracking the replication crisis and how it affects a bunch of behavioral economics for a while. I found reading this post useful for a particularly negative take. Some key quotes:

It sure does look alive... but it's a zombie—inside and out.

Why do I say this?

Two primary reasons:

1. Core behavioral economics findings have been failing to replicate for several years, and *the* core finding of behavioral economics, loss aversion, is on ever more shaky ground.
2. Its interventions are surprisingly weak in practice.

Because of these two things, I don't think that behavioral economics will be a respected and widely used field 10-15 years from now.

[...]

It turns out that loss aversion does exist, but only for large losses. This makes sense. We *should* be particularly wary of decisions that can wipe us out. That's not a so-called "cognitive bias". It's not irrational. In fact, it's completely sensible. If a decision can destroy you and/or your family, it's sane to be cautious.

"So when did we discover that loss aversion exists only for large losses?"

Well, actually, it looks like Kahneman and Tversky, winners of the Nobel Prize in Economics, knew about this unfortunate fact when they were developing Prospect Theory—their grand theory with loss aversion at its center. Unfortunately, the findings rebutting their view of loss aversion were carefully omitted from their papers, and other findings that went against their model were misrepresented so that they would instead support their pet theory. In short: any data that didn't fit Prospect Theory was dismissed or distorted.

I don't know what you'd call this behavior... but it's not science.

This shady behavior by the two titans of the field was brought to light in a paper published in 2018: "[Acceptable Losses: The Debatable Origins of Loss Aversion](#)".

I encourage you to read the paper. It's shocking. This line from the abstract sums things up pretty well: "...the early studies of utility functions have shown that while very large losses are overweighted, smaller losses are often not. In addition, the findings of some of these studies have been systematically misrepresented to reflect loss aversion, though they did not find it."

Welcome & FAQ!

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

The AI Alignment Forum was launched in 2018. Since then, several hundred researchers have contributed approximately two thousand posts and nine thousand comments. Nearing the third birthday of the Forum, we are publishing this updated and clarified FAQ.



Minimalist, watercolor sketch of humanity spreading across the stars by VQGAN

I have a practical question concerning a site feature.

Almost all of the Alignment Forum site features are shared with LessWrong.com; have a look at the [LessWrong FAQ](#) for questions concerning the [Editor](#), [Voting](#), [Questions](#), [Notifications & Subscriptions](#), [Moderation](#), and more.

If you can't easily find the answer there, ping us on Intercom (bottom right of screen) or email us at team@lesswrong.com

What is the AI Alignment Forum?

The Alignment Forum is a single online hub for researchers to discuss all ideas related to ensuring that transformatively powerful AIs are aligned with human values. Discussion ranges from technical models of agency to the strategic landscape, and everything in between.

Top voted posts include [What failure looks like](#), [Are we in an AI overhang?](#), and [Embedded Agents](#). A list of the top posts of all time can be [viewed here](#).

While direct participation in the Forum is limited to deeply established researchers in the field, we have designed it also as a place where up-and-coming researchers can get up to speed on the research paradigms and have pathways to participation too. See [How can non-members participate in the Forum?](#) below.

We hope that by being the foremost discussion platform and publication destination for AI Alignment discussion, the Forum will serve as the archive and library of the field. To find

posts by sub-topic, view the [AI section of the Concepts page](#).

Why was the Alignment Forum created?

Foremost, because misaligned powerful AIs may pose the greatest risk to our civilization that has ever arisen. The problem is of unknown (or at least unagreed upon) difficulty, and allowing the researchers in the field to better communicate and share their thoughts seems like one of the best things we could do to help the pre-paradigmatic field.

In the past, journals or conferences might have been the best methods for increasing discussion and collaboration, but in the current age we believe that a well-designed online forum with things like immediate publication, distributed rating of quality (i.e. "peer review"), portability/shareability (e.g. via links), etc., provides the most promising way for the field to develop good standards and methodologies.

A further major benefit of having alignment content and discussion in one easily accessible place is that it helps new researchers get onboarded to the field. Hopefully, this will help them begin contributing sooner.

Who is the AI Alignment Forum for?

There exists an interconnected community of Alignment researchers in industry, academia, and elsewhere who have spent many years thinking carefully about a variety of approaches to alignment. Such research receives institutional support from organizations including FHI, CHAI, DeepMind, OpenAI, MIRI, Open Philanthropy, ARC, and others. The Alignment Forum membership currently consists of researchers at these organizations and their respective collaborators.

The Forum is also intended to be a way to interact with and contribute to the cutting edge research for people not connected to these institutions either professionally or socially. There have been many such individuals on LessWrong, and that is the current best place for such people to start contributing, to be given feedback and to skill-up in this domain.

There are about 50-100 members of the Forum who are (1) able to post and comment directly to the Forum without review, (2) able to promote the content of others to the Forum. This group will not grow quickly; however, as of August 2021, we have made it easier for non-members to [submit content to the Forum](#).

What type of content is appropriate?

As a rule-of-thumb, if a thought is something you'd bring up when talking to someone at a research workshop or to a colleague in your lab, it's also a welcome contribution here.

If you'd like a sense of what other Forum members are interested in, here's some data from a survey conducted during the open beta of the Forum ($n = 34$). We polled these early users on what high-level categories of content they were interested in.

The responses were on a 1-5 scale, which represented "If I see 1 post per day, I want to see this type of content..." (1) Once per year, (2) Once per 3-4 months (3) Once per 1-2 months (4) Once per 1-2 weeks (5) A third of all posts that I see.

- New theory-oriented alignment research typical of MIRI or CHAI: **4.4 / 5**
- New ML-oriented alignment research typical of OpenAI or DeepMind's safety teams: **4.2 / 5**

- New formal or nearly-formal discussion of intellectually interesting topics that look questionably/ambiguously/peripherally alignment-related: **3.5 / 5**
- High-quality informal discussion of alignment research methodology and background assumptions, what's needed for progress on different agendas, why people are pursuing this or that agenda, etc: **4.1 / 5**
- Attempts to more clearly package/explain/summarise previously discussed alignment research: **3.7 / 5**
- New technical ideas that are clearly not alignment-related but are likely to be intellectually interesting to forum regulars: **2.2 / 5**
- High-quality informal discussion of very core background questions about advanced AI systems: **3.3 / 5**
- Typical AGI forecasting research/discussion that isn't obviously unusually relevant to AGI alignment work: **2.2 / 5**

What is the relationship between the Alignment Forum and LessWrong?

The Alignment Forum was created by and is maintained by the team behind LessWrong (the web forum). The two sites share a codebase and database. They integrate in the following ways:

- **Automatic Crossposting** - Any new post or comment on the new AI Alignment Forum is automatically cross-posted to LessWrong.com. Accounts are also shared between the two platforms (though non-AF member accounts will not be able to post without review).
- **Content Promotion** - Any comment or post on LessWrong can be promoted by members of the AI Alignment Forum to the AI Alignment Forum.
- **Separate Reputation** - The reputation systems (karma) for LessWrong and the AI Alignment Forum are separate. On LessWrong you can see two reputation scores: a primary karma score combining karma from both sites, and a secondary karma score specific to AI Alignment Forum members. On the AI Alignment Forum, you will just see the AI Alignment Forum karma of posts and comments.
- **Content Ownership** - If a comment or post of yours is promoted to the AI Alignment Forum, you will continue to have full ownership of the content, and you'll be able to respond directly to all comments on your content.

Both LessWrong and the Alignment Forum are foci of Alignment Discussion; however, the Alignment Forum maintains even higher standards of content quality than LessWrong. The goal is to provide a place where researchers with shared technical and conceptual background can collaborate, and where a strong set of norms for facilitating good research collaborations can take hold. For this reason, both submissions and members to the Alignment Forum are heavily vetted.

How do I get started in AI Alignment research?

If you're new to the AI Alignment research field, we recommend four great introductory sequences that cover several different paradigms of thought within the field. Get started reading them and feel free to leave comments with any questions you have.

The introductory sequences are:

- [Embedded Agency](#) by Scott Garrabrant and Abram Demski of MIRI
- [Iterated Amplification](#) by Paul Christiano of ARC
- [Value Learning](#) by Rohin Shah of DeepMind
- [AGI Safety from First Principles](#) by Richard Ngo, formerly of DeepMind

Following that, you might want to begin writing up some of your thoughts and sharing them on LessWrong to get feedback.

How do I join the Alignment Forum?

As described above, membership to the Alignment Forum is very selective (and not strictly required to participate in discussions on Alignment Forum content, since one can do so on LessWrong).

The best pathway towards becoming a member is to produce lots of great AI Alignment content, and to post it to LessWrong and participate in discussions there. The LessWrong/Alignment Forum admins monitor activity on both sites, and if someone consistently contributes to Alignment discussions on LessWrong that get promoted to the Alignment Forum, then it's quite possible full membership will be offered.

I work professionally on AI Alignment. Shouldn't I be a member?

Maybe but not definitely! The bar for membership is higher than working on AI Alignment professionally, even if you are doing really great work. Membership, which allows you to directly post and comment, is likely to be offered only after multiple existing Alignment Forum members are excited to see your work. Until then, a review step is required. You can still submit content to the Alignment Forum but it might take a few days for a decision to be made.

Another reason for the high bar for membership is that any member has the ability to promote content to the Alignment Forum, kind of like a curator. This requires significant trust and membership is restricted to those who have earned this level of trust among the Alignment Forum members.

How can non-members participate in the Forum?

Non-members can participate in the Forum in two ways:

1. Posting and commenting Alignment content to LessWrong

Alignment content posted to LessWrong will be seen by many of the researchers present on the Alignment Forum. If they (or the Forum admins) think that particular content is a good fit for the Forum, it will be promoted to the Forum and become viewable there.

If your posts or comments are promoted to the Alignment Forum, you will be able to directly participate in the discussion of your content on the Forum.

2. Submitting content on the Alignment Forum

Non-members can now submit content directly on the Alignment Forum (and not just via LessWrong).

- If you post or comment, your submission will enter a review queue and a decision to accept or reject it from the Alignment Forum will be made within three days. If it is rejected, you will receive a minimum one-sentence explanation.
- In the meantime (and regardless of outcome), your post or comment will be published to [LessWrong](#). There it can be immediately viewed and discussed by everyone, and edited by you. This allows you to get quick feedback, and allows site admins to use the

reaction there to help make the decision about whether it is a good fit for the Alignment Forum. For example, if several Alignment Forum members are discussing your content on LessWrong, it is likely a good fit for the Forum and will be promoted.

How can I submit something I already wrote?

If you have already written and published a post on LessWrong but would like to submit it for acceptance to the Alignment Forum, please contact us via Intercom (bottom right) or email us at team@lesswrong.com

Who runs the Alignment Forum?

The Alignment Forum is maintained and run by the [LessWrong team](#) who also run the LessWrong website. An independent board composed of representatives of major Alignment research orgs (and independent members too) oversees major decisions concerning the Forum.

Can I use LaTex?

Yes! You can use LaTeX in posts and comments with Cmd+4 / Ctrl+4.

Also, if you go into your user settings and switch to the markdown editor, you can just copy-paste LaTeX into a post/comment and it will render when you submit with no further steps required.

I have a different question.

Please don't hesitate to contact us via Intercom (bottom right of the screen) or email us at team@lesswrong.com. We'd love to answer your questions.

Against "blankfaces"

(Content note: minor spoilers for *Harry Potter and the Order of the Phoenix*.)

Scott Aaronson writes about [blankfaces](#),

anyone who enjoys wielding the power entrusted in them to make others miserable by acting like a cog in a broken machine, rather than like a human being with courage, judgment, and responsibility for their actions. A blankface meets every appeal to facts, logic, and plain compassion with the same repetition of rules and regulations and the same blank stare—a blank stare that, more often than not, conceals a contemptuous smile.

I want to push back against this a bit.

First, one of the defining aspects of blankfacedness is their internal experience. It's someone who *enjoys* wielding their power. This is a very hard thing to judge from the outside.

I used to work in a cinema. One day a mother came in with her young child, perhaps fivish years old. She was late for a busy screening, and the person selling tickets warned they might not be able to sit together. She said that was fine, bought popcorn and went in. Soon afterwards she came back out, complaining that they couldn't sit together. She wanted a refund for the tickets (fine) and popcorn (not fine, but she insisted). The conversation between her and my manager escalated a bit. I don't remember who brought up the police, but she at least was very confident that she knew her rights and the police would back her up if they arrived. Eventually he gave her a refund.

If it had been up to me? And if I hadn't had to worry about things like PR and "someone yelling in the lobby would ruin the experience for people watching movies"? I think I would *absolutely* have used the "no, sorry, those are the rules" move. I would have been happy to do so. Does that make me a blankface? But it's not that I would have enjoyed wielding my power as such. Rather it's that I would have enjoyed punishing her, specifically, for acting in ways that I endorsedly think are bad to act in.

Does someone's internal experience matter, though? If they act a certain way, should we care how they feel? I think we should, and if we don't we shouldn't make claims about it.

That is, if what you care about is whether someone is *acting* a certain way, then don't mention *enjoyment* when you define a blankface. And if you really do care about the enjoyment part of it - well, how do you know what someone is feeling and why?

I predict that if the term "blankface" takes off, no matter how much people defining the term emphasize the "enjoys" part of it, people *using* the term will not be careful about checking that. Partly I think this because Scott wasn't: in two of his examples, he accused people of being blankfaces whom he'd never interacted with and could not identify. Does he really think that a web portal was badly designed *out of malice*? But also I think that... like, even if you can tell that someone is just acting a certain way because they enjoy it, even if you're really sure that's what's going on, you won't properly be able to capture that in your description of the events. So people will read a story where it seems like the thing making them a blankface is the way they acted,

and then they'll tell their own similar stories where people acted in similar ways, and they'll use the term "blankface".

There's a lot of potential here for a kind of (reverse?) motte-and-bailey, where the bailey is "I'm calling someone a blankface which is explicitly defined as having an enjoyment part to it", and the motte is "...but no one uses it that way, so obviously I didn't mean to imply that I know what was going on in their head".

Here's another reason someone might externally act blankfacedly: fear. *Yes, this is ridiculous, but if I admit that out loud I'll be seen as undermining my boss who already has it in for me, so....* Or, exhaustion: *this is the third sob story I've heard today. I cannot deal with feeling more sympathy.*

Given how strongly Scott feels about blankfaces (they've apparently dehumanized themselves and deserve no mercy), I certainly *hope* he cares whether they act blankfacedly for sympathetic or unsympathetic reasons. And if we're to care about that, I think we have to admit that most of the time we don't really know.

Second and relatedly, I think we should distinguish between "this person is blankfacing" and "this person is a blankface". Like, maybe someone right now is enjoying wielding petty power for the sake of it. I don't currently predict that that person *routinely* enjoys acting that way, or enjoys acting that way in every situation where they have petty power. That's maybe not much consolation to their victim right now, but still.

Perhaps I should predict that? But I currently don't, and Scott gives me no reason to.

Third, I'm not sure Umbridge is an example of the archetype. Or, if Umbridge is really what Scott wants to point at, I'm not sure his explicit definition matches up.

The most despicable villain in the Harry Potter universe is not Lord Voldemort, who's mostly just a faraway cipher and abstract embodiment of pure evil, no more hateable than an earthquake. Rather, it's Dolores Jane Umbridge, the toadlike Ministry of Magic bureaucrat who takes over Hogwarts school, forces out Dumbledore as headmaster, and terrorizes the students with increasingly draconian "Educational Decrees." Umbridge's decrees are mostly aimed at punishing Harry Potter and his friends, who've embarrassed the Ministry by telling everyone the truth that Voldemort has returned and by readying themselves to fight him, thereby defying the Ministry's head-in-the-sand policy.

Anyway, I'll say this for Harry Potter: Rowling's portrayal of Umbridge is so spot-on and merciless that, for anyone who knows the series, I could simply define a blankface to be anyone sufficiently Umbridge-like.

Spoilers: the educational decrees are not the extent of Umbridge's villainy. She also sends dementors to attack Harry and his cousin. She tries to slip Harry a truth serum, and later tries to force-feed it to him. When she can't do that, she tries to torture him. None of this is legal¹, some of it is super-duper illegal, and her superiors aren't pressuring her into it. Umbridge doesn't simply act like a cog in a broken machine. She exercises judgment, and I think even some courage, in service of villainous ends.

(I am not, here, describing Umbridge's motivations. I think I have some idea of what they are, but I'm not sure I'd capture them very well and my feeling is they're not super relevant for this bit.)

However annoyed Scott may be at his daughter's lifeguard, I predict that describing the lifeguard as Umbridge-like is unfair. I predict, for example, that she would never take initiative to deliberately engineer a situation in which Scott's daughter nearly drowns.

I think Scott is pointing at something true, and important to know about. But I think he's conflating a few different things (blankfacing-regardless-of-internal-experience, blankfacing-specifically-meaning-enjoying-it, being-a-blankface, Umbridge), and I worry that the way he's pointing at them will result in bad discourse norms. I still think it's good that he published, but I don't think that essay is the ideal version of what it could be, and I'm trying here to point at ways I think it falls short.

1. At least I don't *think* any of that was legal. I am not a wizard lawyer and this is not wizard legal advice. ↩

Eight Hundred Slightly Poisoned Word Games

[cross-posted from my blog [Astral Codex Ten](#)]

In 2012, a Berkeley team found that indoor carbon dioxide had dramatic negative effects on cognition ([paper](#), [popular article](#)). Subjects in poorly ventilated environments did up to 50% worse on a test of reasoning and decision-making. This is potentially pretty important, because lots of office buildings (and private houses) count as poorly-ventilated environments, so a lot of decision-making might be happening while severely impaired.

Since then people have debated this on and off, with [some studies](#) confirming the effect and [others](#) failing to find it. I personally am skeptical, partly because the effect is so big I would expect someone to have noticed, but also because submarines, spaceships, etc have orders of magnitude more carbon dioxide than any civilian environment, but people still seem to do pretty hard work in them pretty effectively.

As part of my continuing effort to test this theory in my own life, I played a word game eight hundred times under varying ventilation conditions.

...okay, fine, no, I admit it, I played a word game eight hundred times because I'm addicted to it. But since I was playing the word game eight hundred times anyway, I varied the ventilation conditions to see what would happen.

The game was WordTwist, which you can find [here](#) (warning: potentially addictive). You get a 5x5 square of letters and you have to find as many words as possible (of four letters or more) within three minutes. You can move up, down, right, left, or diagonal, and get more points for harder words. A typical board looks like this:



Did you spot "lace"? What about "intrapsychically"?

I played this game about 5-10x/day over three months. During this time, the carbon dioxide monitor in my room recorded levels between 445 ppm (with all windows open and the fan on) and 3208 ppm (with all windows closed and several people crammed into the room for several hours). I discounted a stray reading of 285 as an outlier, since this is climatologically impossible (I'm not claiming my monitor is perfectly calibrated, just that it clearly shows higher levels when my room is less well ventilated). CO₂ 445 is basically the same as outdoors; 3208 is considered extremely poor air quality, likely to cause headaches, nausea, and other minor ailments. The Berkeley study looked at levels between 600 and 2500, so my range was comparable to theirs.

I correlated my adjusted score (my score as a percent of the average score for that board) for each game with the CO₂ level in my room when I was playing it. R was 0.001, p = 0.97 - there was absolutely no correlation.

Why might these results not be valid? Well, CO₂ level in my room wasn't randomly determined - I just played a game when I felt like it and recorded whatever the ambient CO₂ level was at the time. CO₂ level was lower if I had the window open or air conditioning on,

higher if I'd been in the room for a long time, and highest if I'd just woken up after being asleep in the room all night. It was also higher when other people were in my room. In theory things like this could confound the results. For example, if CO₂ really did affect performance, but I performed better when I was hot, then turning the air conditioning on might improve performance (by decreasing CO₂) but also hurt performance (by making it colder), and those effects could cancel out. Or if I performed worse after exercise, and I often went out of my room to exercise, then I might perform worse when I had just come back into my room (which was often when CO₂ was lowest).

In practice I'm skeptical this mattered. For one thing, the studies found huge positive effects - so for me to find zero effect would require a huge negative effect of the exact right size to cancel out the huge positive one. For another thing, I checked if temperature had any effect, and it didn't ($r = -0.008$, $p = 0.83$). For another, I ran a few controlled experiments to see if they got the same results as the naturalistic ones, and they did. For another, I did get to test an exogenous shock - about halfway through the experiment, I moved to a new house with better ventilation. The difference in average CO₂ reading between the old and new houses was significant ($p < 0.001$), but the difference in score wasn't ($p = 0.15$). Although it was in the expected direction (new house > old), I attribute this to me improving on the word game with practice, and I didn't improve any more during the month when I switched houses than in an average month.

I consider this to be very strong evidence that at least for me, on this specific task, carbon dioxide has zero effect on cognition. To rescue the hypothesis that it matters, you'd either have to find that it affects other people more than it does me (why would it?) or that it affects other aspects of cognition more than it affects the skills associated with this particular word game. This second one is moderately plausible - I don't think the word game tests "decision-making" per se. But it would be surprising for this not to be a general health effect, and would potentially be important in the study of intelligence and neuroscience to explore which skills do or don't suffer under carbon dioxide poisoning.

I was excited to read the Less Wrong post [Chess and cheap ways to check day to day variance in cognition](#) by KPier, who does something similar with chess instead of a word game; they haven't checked carbon dioxide levels yet, but I'd be excited for them to try. I'm also interested in hearing from anyone else who often repeats some objectively-scoreable cognitive task, to see how they do. A CO₂ monitor [costs about \\$100 on Amazon](#), but if money is the only reason you're not going to do some really good experiment, please let me know and I'll buy it for you.

If you're planning on testing this, please post about it below as a form of preregistration.

EDIT: You can download the original data [here](#), some explanations of what the columns mean [here](#).

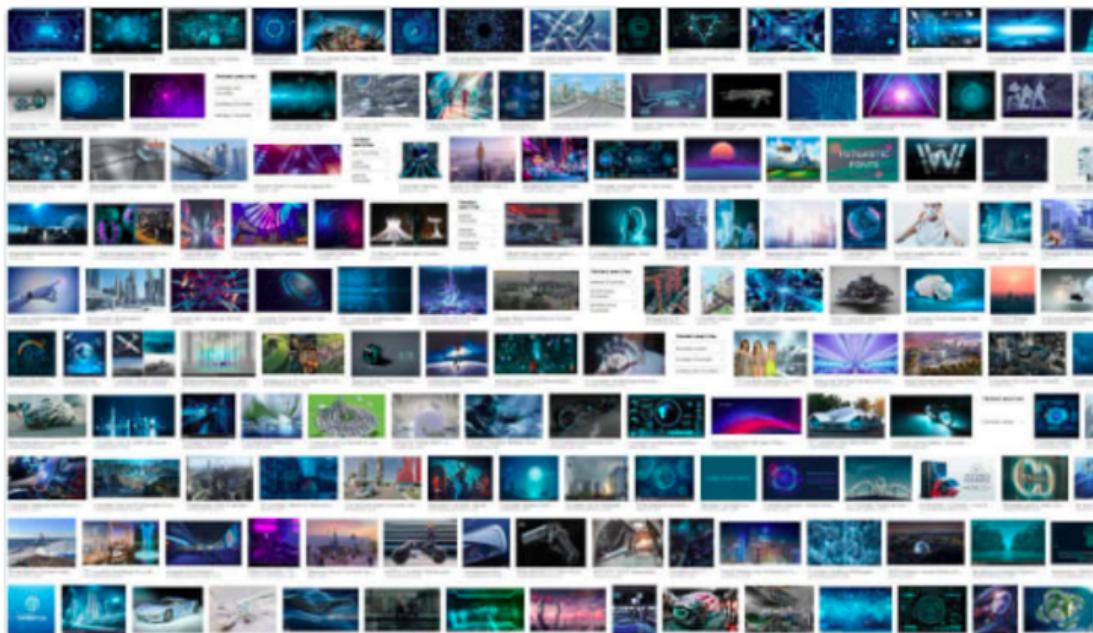
The Future: Where are the Colors and the Sports?



Sabine Hossenfelder

@skdh

Curious find: A Google image search for "futuristic" returns almost exclusively images with blue/black color themes. How is that? Why isn't the future orange? Very puzzled about this.



June 27th 2019

178 Retweets 468 Likes

One can imagine a variety of reasons for why “The Future” (as imagined in our collective consciousness) is blue and black, many of which are mundane and fairly obvious.

The screens on our phones, laptops, and TVs are black (the inspiration behind the name of the science fiction series *Black Mirror*) and it’s probably a good bet that the future will have even more screens (one day we will decide to replace the whole sky with a screen so that it’s never cloudy and every sunset is picturesque). In general, technology calls to mind darker and more metallic colors (silver and grey also seem abundant in our imagined future). The future also calls to mind space exploration which is a very black and blue type of affair - from our grey metallic spaceships we will all look back at our home planet, a pale blue dot hanging in the void.

There may be a less obvious and more interesting reason why we tend to envision the future with a certain aesthetic (darker color palette, sleek, smooth, shiny, etc.).¹ Construal-level theory describes how distance of all kinds (temporally, spatially, socially, emotionally,

conceptually, etc.) form one general cognitive dimension and how this dimension (near vs. far) affects the abstraction of our thinking. To get a better sense of the theory, I'll share an abstract from a review paper (slightly more technical) and a few sections from the wikipedia page (slightly less technical).

From Trope and Liberman (2010), "[Construal-Level Theory of Psychological Distance](#)"

People are capable of thinking about the future, the past, remote locations, another person's perspective, and counterfactual alternatives. Without denying the uniqueness of each process, it is proposed that they constitute different forms of traversing psychological distance. Psychological distance is egocentric: Its reference point is the self in the here and now, and the different ways in which an object might be removed from that point-in time, in space, in social distance, and in hypotheticality-constitute different distance dimensions. Transcending the self in the here and now entails mental construal, and the farther removed an object is from direct experience, the higher (more abstract) the level of construal of that object. Supporting this analysis, research shows (a) that the various distances are cognitively related to each other, (b) that they similarly influence and are influenced by level of mental construal, and (c) that they similarly affect prediction, preference, and action.

From [Wikipedia](#):

The general idea is that the more distant an object is from the individual, the more abstract it will be thought of, while the closer the object is, the more concretely it will be thought of. In CLT, psychological distance is defined on several dimensions—temporal, spatial, social and hypothetical distance (how likely something is to occur) being considered most important.

An example of construal level effects would be that although planning one's next summer vacation one year in advance (in the distant future) will cause one to focus on broad, decontextualized features of the situation (e.g., anticipating fun and relaxation), the very same vacation planned to occur very soon will cause one to focus on specific features of the present situation (e.g. what restaurants to make reservations for, going for a trip in an off-road vehicle).

[Robin Hanson discusses](#) how much of what we imagine about the future is based in the more abstract "far" mode of thinking.

Since the future is far in time, thinking about it tends to invoke a far mode of thought, which introduces other far mode defaults into our image of the future. And thinking about the far future makes us think especially far. Of course many other considerations influence any particular imagined future, but it can help to understand the assumptions your mind is primed to make about the far future, regardless of whether those assumptions are true.

Since blue light scatters more easily than red, far away things in our field of view tend to look more blue. So we expect future stuff to look blue. And since blue stuff looks cold, we expect future stuff to look cold. Finally, since we expect far away things to have less detail, we tend to imagine them with fewer parts and flourishes, and less detailed textures and patterns. The future is not paisley.

We also tend to assume there are fewer relevant categories of far things. So we'll tend to assume future folk have fewer kinds of food, furniture, cars, houses, roads, buildings, and land uses, whose styles of use vary less from place to place. Instead of seeing a million variations bleeding into each other in dizzying complexity, we tend to assume there are fewer more discrete types, with less variation within each type and larger differences between types. For example, futuristic movies often have everyone wearing very similar clothes.

So the conclusion is that in the future, [everyone will look brazilian](#), there will only be two foods - chocolate nutrient paste or vanilla nutrient paste (both made from [Soylent Green](#)), and everyone will wear skin-tight silver jumpsuits.

I'm not the first to make this remark, but imagined futures often seem incredibly boring. Fun doesn't seem to be a high priority and everything is so serious - it's all AI apocalypses, intergalactic warfare, and the like. Future people are all scientific geniuses and moral heroes that spend their time traveling the galaxy, pondering the mysteries of the cosmos, and fighting for noble causes. Where is the sex?!?! The drugs?!?!? The music?!?!? You know, THE FUN!?!?!? (sorry, I got a little excited there...)

Hanson attributes the seriousness and lack of fun in the future to far mode thinking in the moral, social, and emotional domains.

Sex, money, and temptation tend to be near, while love, satisfaction, trust, and self-control are far. So we often assume future folks have forgotten how to have sex, as in Sleeper or Barbarella, or that money motives are less common, as in Star Trek.

In far mode we tend to focus more on our simple abstract ideals and values, relative to messy desires and practical constraints. We also tend to neglect our messy internal contradictions and conflicts, and therefore assume our values and actions are coherent and consistent. So in far mode we tend more to explain good acts as virtue, and bad acts as vice or evil. We assume future folk are less driven by base desires, more strongly committed to their ideals, less tolerant of domination, more morally enlightened, and more morally judgmental about others' failings.

Since important things seem nearer to us, stronger emotions feel nearer, and so we have weaker motives and emotions regarding far things. Instead of being filled with elation or terror regarding good or bad things that might happen in the far future, we tend to treat such events more philosophically, and to assume future folk will do so as well.

Tasting and touching tend to feel near, while seeing and hearing tend to feel far. So we mainly imagine what the future looks and sounds like, relative to its taste or touch.

None of this is to say that the future won't be blue and black, smooth and shiny, boring and serious, or high-minded and philosophical - expectations have a way of becoming self-fulfilling prophecies. Construal-level theory should give us pause however; perhaps the only thing we see when we look into the future is our own flawed minds.

There is another reason why the future might be nothing like what we imagine - who exactly is the "we" that is doing the imagining? Nerds. Not everyone contributes equally to our collective consciousness, so-called futurists and science-fiction writers - more broadly, nerds - have an outsized role in determining our view of the future. Futurists/science fiction writers/nerds are not a representative sample of the human population by any stretch of the imagination. If the stereotype is true (and we have no reason to believe it's not), then the people who are most responsible for shaping our view of the future are largely white (at least in the US), male, and are strongly interested in STEM disciplines, particularly computer science and AI. Most of the great science-fiction writers are [white males with backgrounds in STEM](#). The only demographic information I can find on futurists shows strong male bias - the World Future Society and the Association of Professional Futurists are roughly 70% male ([Why Aren't There More Women Futurists?](#)).

I'm painting with a very broad brush here, but I think it's reasonable to say there is a certain personality type that tends to be a futurist/nerd and that this stereotype fits with many of the "stereotypes" about the future. The darker and more masculine color palette and general aesthetic of future seem like something that would be chosen by a group of men without

much interest in the art and design. These are also the type of people who we might expect to care the least about fashion (I know I don't...). This may be more of a stretch, but I would also say that many of the moral and social/emotional aspects of the future (a focus on abstract ideals and virtues, a less emotional and more philosophical outlook, etc.) are also what you would expect from a group of people that tend to excel at science, math, and engineering.

It's hard to say how much biased representation influences our collective view of the future; maybe most of its features are better explained by construal-level theory or simply the accidents of culture and history. Regardless of how important it is, I think we can use this as a kind of heuristic for thinking about what is missing from our popular conception of the future. What kinds of things does the typical futurist (again, we are stereotyping - white, male, interested in STEM) not appreciate or care about? These are the things that might be more important and more common in the future than we generally anticipate.

I'll limit myself to one answer, one that I saw in a series of tweets from [Helen Toner](#). Going along with the general lack of fun, you don't really hear a lot about sports in the future. Everyone is too busy exploring space and fighting robots or aliens, basketball is such a childish waste of time in comparison.



Helen Toner @hltnr · Jan 16, 2019

...

Thread of musings on sth I noticed recently: conversations about how we might find meaning in a post-work world heavily feature music and art... but I can't remember sports being mentioned even once. How come, when it provides so much meaning/community/joy to so many people? 1/5

7

24

78

↑



Helen Toner @hltnr · Jan 16, 2019

...

Obvious answer is obvious: sports aren't mentioned because these discussions are being had by Serious Intellectuals with Serious Intellectual tastes. It's a shame though - such a good way to channel our instincts for tribalism & physical competition, especially of young men. 2/5

4

1

22

↑



Helen Toner @hltnr · Jan 16, 2019

...

In general I wish there were more exploration (e.g. in fiction) of rich, meaningful ways we might spend time, make meaning, build communities etc in the future. From the perspective of 100 years ago, our current team sports system could seem like an example of this... 3/5

4

1

16

↑



Helen Toner @hltnr · Jan 16, 2019

...

Think about it: millions of fans bond with family/friends over matches beamed to their house from hundreds/thousands of miles away. Big celebrations spill into the streets as outpourings of joy and connection. There's hope, loyalty, excitement, strategy, pride, exultation. 4/5

1



13



Helen Toner @hltnr · Jan 16, 2019

...

Obviously there aren't zero harms - riots and fights happen, both among players and fans. But compared to so many other options (e.g. intertribal/gang/schoolyard fights), seems like a very high ratio of meaningfulness+connection : violence+hate. Let's make more+better sports! 5/5

3



12



Generally speaking, people who love computer science/AI/STEM tend not to be the most athletic bunch; they also tend not to be diehard sports fans. We can understand why these people might not place a huge emphasis on sports when they think about the future - the fact is that intellectual culture (Serious Intellectuals in Helen's parlance) generally looks down its nose at sports. However, given the popularity of sports in today's world, it is a good bet that sports will continue to be important in the near and far future. In fact, I think we might come to develop a whole new perspective on the nature of sport and its role in our existence.

[Deaths from despair](#) (suicide, drug overdose, addiction) have become so prevalent as to cause an unprecedented decline in life expectancy in the UK and the US, something [not seen since 1918](#). Mental and physical health (obesity, heart disease, diabetes, etc.) are faltering. Work is more isolating and less satisfying than it has ever been. To put it bluntly, we are lonelier, more depressed, and fatter² than ever before (and of course COVID-19 has only exacerbated these problems). If only there was an activity that could bring people together, give a sense of achievement and satisfaction, and improve physical and mental health...

It's remarkable how well athletics, particularly team sports (basketball, soccer, football, etc.), fit as a remedy for many of the social ills of the modern world. I'll make what I hope to be an uncontroversial statement: *sports are massive force for good in society*. This is not to say that sports can't have negative effects; certainly there are people for whom participation in sports, either as an athlete or a fan, is not a net positive in their lives. This is also not to say that sports are a panacea for these death of despair issues; many of them have deep roots in economic and technological trends and any real long-term solution needs to work at these levels. That being said, I think athletics do achieve something more than a superficial treatment of symptoms.

I'll make another statement which I believe is much more controversial: *sports can be a source of ultimate meaning and satisfaction in life*. The opposing view - sports (and games of all kinds more generally) are simply meaningless diversions which do not, and cannot, provide absolute value or meaning - is so pervasive that it is rarely discussed explicitly; no

one feels the need to argue that sports are not that important in the grand scheme of human civilization. The reasons why this position is so unquestioned are largely the same reasons mentioned above why we think that the future is blue and black and smooth and shiny. On one level, the people who think the and write the most about grand meaning of life questions are typically not people who also have a great interest in athletics (of course there are many exceptions, but that doesn't change the general picture). However, we shouldn't discount the role of construal-level theory either - consideration of ultimate philosophical questions lends itself to the far modes of thinking which biases towards more abstract and high-minded ideas and values. The reason we don't think of athletics as providing ultimate meaning might just represent a cognitive blind spot and nothing more.

(Originally posted at [Secretum Secretorum](#))

Coase's "Nature of the Firm" on Polyamory

It occurred to me that Coase's views on [The Nature of the Firm](#) might help explain why polyamory in its modern form is not particularly common or popular.

That sentence might be enough for you to grok what I'm getting at, and honestly that's the form in which the thought first came to be, but nevertheless let me try to explain what I mean.

Coase's original essay -- and the whole body of thought proceeding from it -- seeks to answer why corporations / firms emerge. That is, it seeks to ask where people are *hired* for indefinite periods of time, for less-precisely-defined work rather than *contracted* for definite amounts of time, for precisely-defined work. If you believe in a strong version of efficiency of markets, you might expect it to almost always be cheaper to contract than to hire, because the allocation of resources by a market should be more efficient than the allocation of resources within a non-market organization. Why wouldn't my software company just hire a contractor for everything they needed to be done, rather than relying on me, a non-expert in many things they would like me to do?

The answer, of course, is that there are transaction costs to using the market. There's a cost to searching for and finding a trustworthy contractor, which is avoided by keeping me around. There's the cost of a stronger asymmetry of information and asymmetry of benefit in the case of the contractor, which makes me a little more trustworthy because I'm going to be stuck with the code I write for a longer period of time. And so on and so forth.

Polyamory seems like an attempt to unbundle a group of frequently-bundled relationship goods in a way analogous to how contracting different workers can be an attempt to unbundle a group of frequently-bundled commercial goods. Vis, in polyamory you frequently unbundle from each other the following:

- Sexual satisfaction
- Intellectual companionship
- Long-term companionship and dependency
- Childbearing and rearing

Or even decompose these further: i.e., different flavors of sex and different flavors of companionship. But finding someone for each of these involves transaction costs. So you have the costs of searching for and finding trustworthy people in all these roles. And you have the stronger asymmetry of information and of benefit because of the more ephemeral nature of the relationships.

This is really just a rephrase of things I know other people have said about the disadvantages of polyamory. But it was satisfying to me to realize that it looked pretty clearly like an instance of a larger phenomenon.

([x-post](#))

Modelling Transformative AI Risks (MTAIR) Project: Introduction

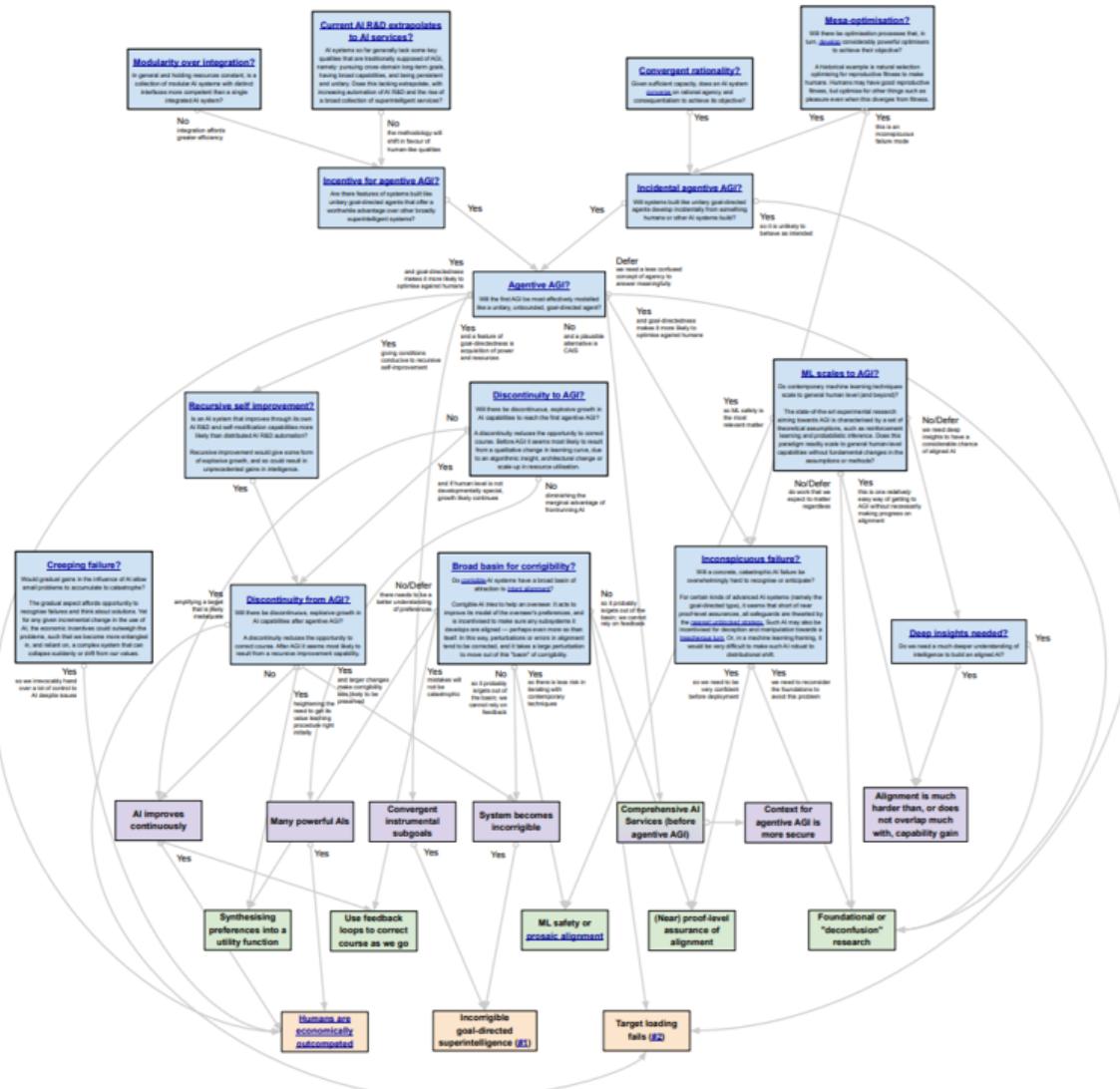
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Numerous books, articles, and blog posts have laid out reasons to think that AI might pose catastrophic or existential risks for the future of humanity. However, these reasons often differ from each other both in details and in main conceptual arguments, and other researchers have questioned or disputed many of the key assumptions and arguments.

The disputes and associated discussions can often become quite long and complex, and they can involve many different arguments, counter-arguments, sub-arguments, implicit assumptions, and references to other discussions or debated positions. Many of the relevant debates and hypotheses are also subtly related to each other.

Two years ago, Ben Cottier and Rohin Shah created a [hypothesis map](#), shown below, which provided a useful starting point for untangling and clarifying some of these interrelated hypotheses and disputes.

The MTAIR project is an attempt to build on this earlier work by including additional hypotheses, debates, and uncertainties, and by including more recent research. We are also attempting to convert Cottier and Shah's informal diagram style into a quantitative model that can incorporate explicit probability estimates, measures of uncertainty, relevant data, and other quantitative factors or analysis, in a way that might be useful for planning or decision-making purposes.



Cottier and Shah's 2019 Hypothesis Map for AI Alignment

This post is the first in a series which presents our preliminary outputs from this project, along with some of our plans going forward. Although the project is still a work in progress, we believe that we are now at a stage where we can productively engage the community, both to contribute to the relevant discourse and to solicit feedback, critiques, and suggestions.

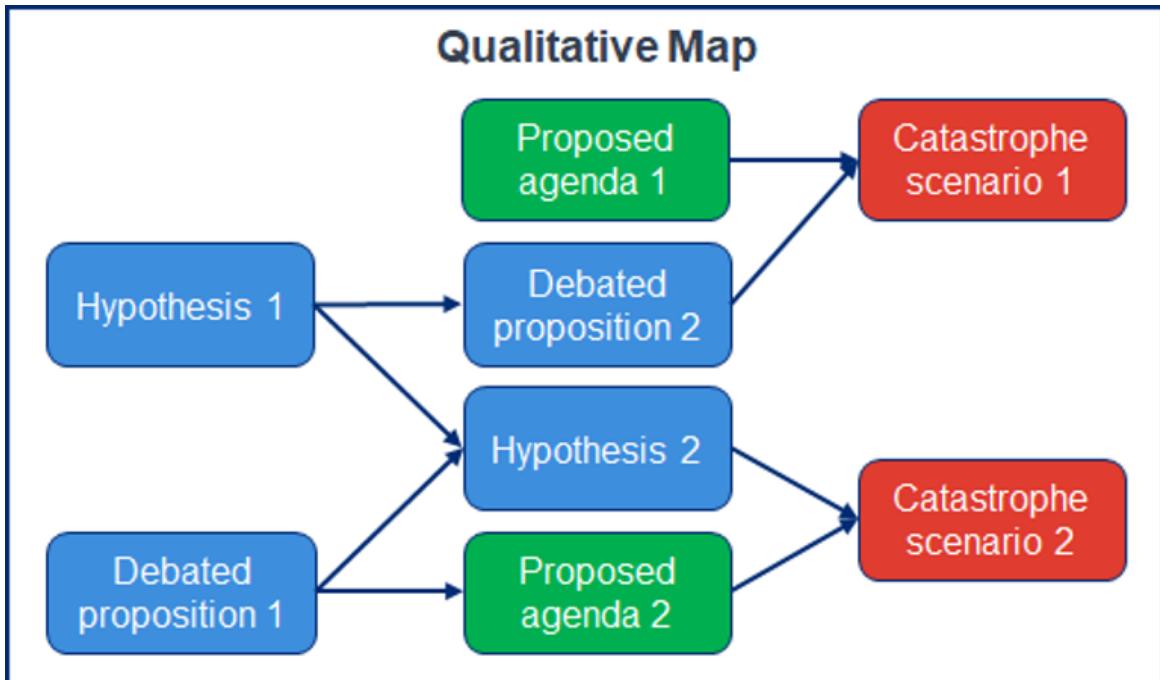
This introductory post gives a brief conceptual overview of our approach and a high-level walkthrough of the hypothesis map that we have developed. Subsequent posts will go into much more detail on different parts of this model. We are primarily interested in feedback on the portions of the model that we are presenting in detail. In the final posts of this sequence we will describe some of our plans going forward.

Conceptual Approach

There are two primary parts to the MTAIR project. The first part, which is still ongoing, involves creating a qualitative map ("model") of key hypotheses, cruxes, and relationships, as described earlier. The second part, which is still largely in the planning phase, is to

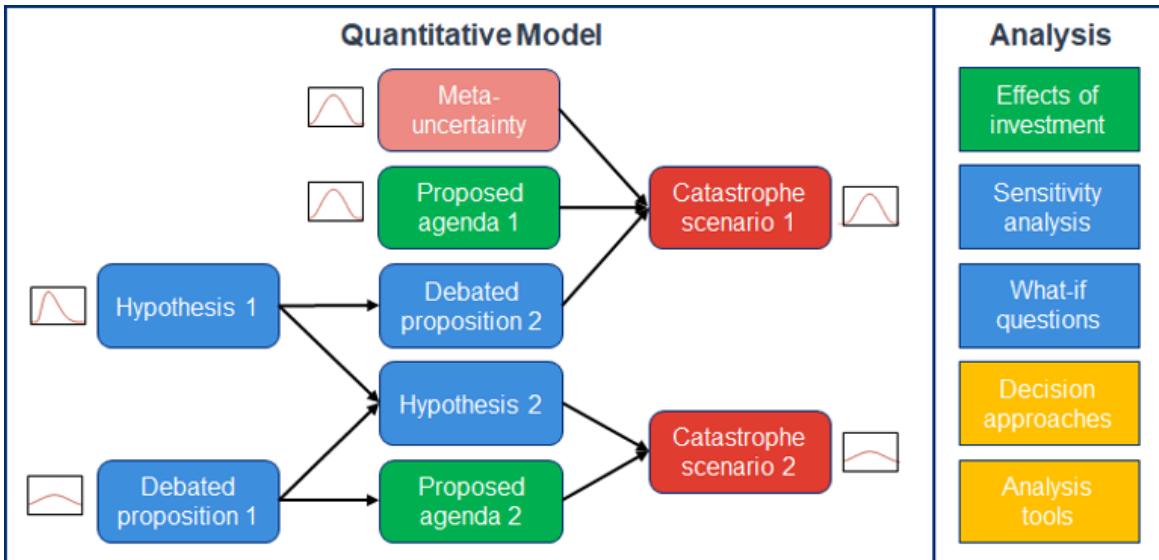
convert our qualitative map into a quantitative model with elicited values from experts, in a way that can be useful for decision-making purposes.

Mapping key hypotheses: As mentioned above, this part of the project involves an ongoing effort to map out the key hypotheses and debate cruxes relevant to risks from Transformative AI, in a manner comparable to and building upon the [earlier diagram](#) by Ben Cottier and Rohin Shah. As shown in the conceptual diagram below, the idea is to create a qualitative map showing how the various disagreements and hypotheses (blue nodes) are related to each other, how different proposed technical or governance agendas (green nodes) relate to different disagreements and hypotheses, and how all of those factors feed into the likelihood that different catastrophe scenarios (red nodes) might materialize.



Qualitative map illustrating relationships between hypotheses, propositions, safety agendas, and outcomes

Quantification and decision analysis: Our longer-term plan is to convert our hypothesis map into a quantitative model that can be used to calculate decision-relevant probability estimates. For example, a completed model could output a roughly estimated probability of transformative AI arriving by a given date, a given catastrophe scenario materializing, or a given approach successfully preventing a catastrophe.



Notional version of how the above qualitative map can be used for quantification and analysis

The basic idea is to take any available data, along with probability estimates or structural beliefs elicited from relevant experts (which users can modify or replace with their own estimates as desired). Once this model is fully implemented, we can then calculate probability estimates for downstream nodes of interest via Monte Carlo, based either on a subset or a weighted average of expert opinions, or using specific claims about the structure or quantities of interest, or a combination of the above. Finally, even if the outputs are not accepted, we can use the indicative values as inputs for a variety of analysis tools or formal decision-making techniques. For example, we might consider the choice to pursue a given alignment strategy, and use the model as an aid to think about how the payoff of investments changes if we believe hardware progress will accelerate or if we presume that there is relatively more existential risk from nearer-term failures.

Most of the posts in this series will focus on the qualitative mapping part of the project, since that has been our primary focus to date. In our last post we will discuss our plans related to the second, quantitative, part of the project.

Model Overview

The next several posts in this sequence will dive into the details of our current qualitative model. Each post will be written by team members involved in crafting that particular part of the model, as different team members or groups of team members worked on different parts of the model.

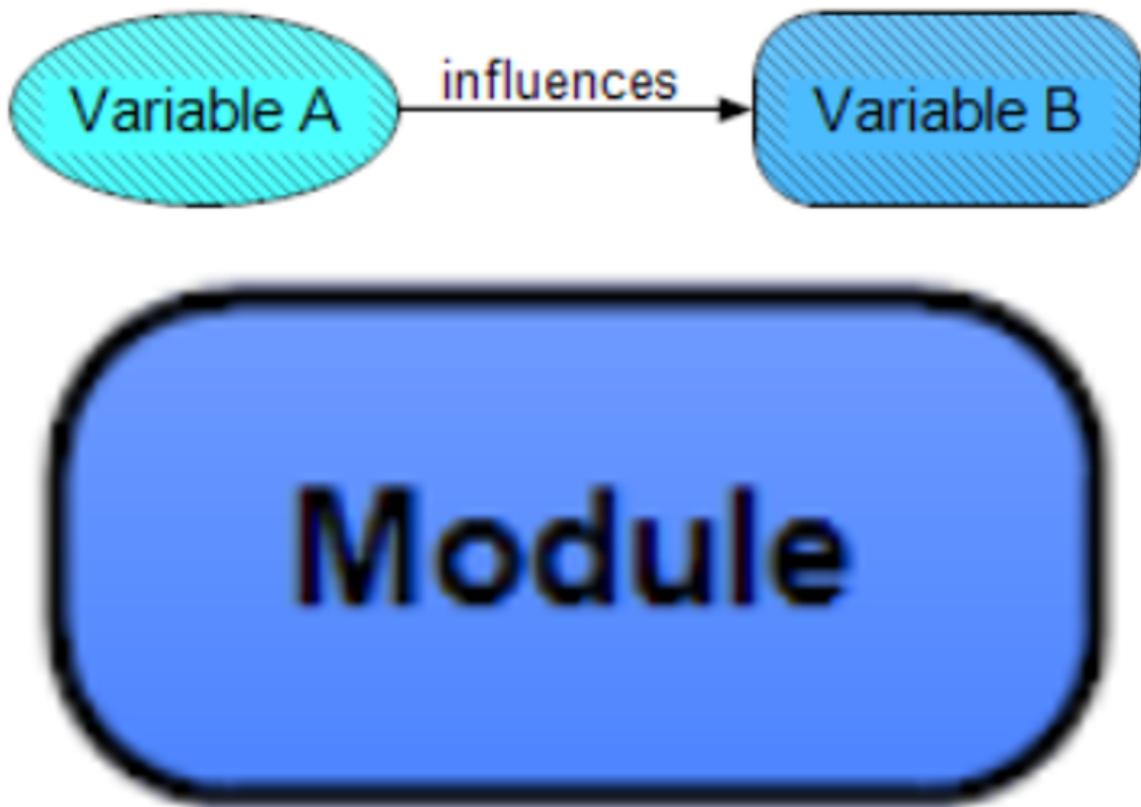
The structure of each part of the model is primarily based on a literature review and the understanding of the team members, along with considerable feedback and input from researchers outside the team. As noted above, this series of posts will hopefully continue to gather input from the community and lead to further discussions. At the same time, the various parts of the model are interrelated. Daniel Eth is leading the ongoing work of integrating the individual parts of the model, as we continue developing a better understanding of how the issues addressed in each component relate to each other.

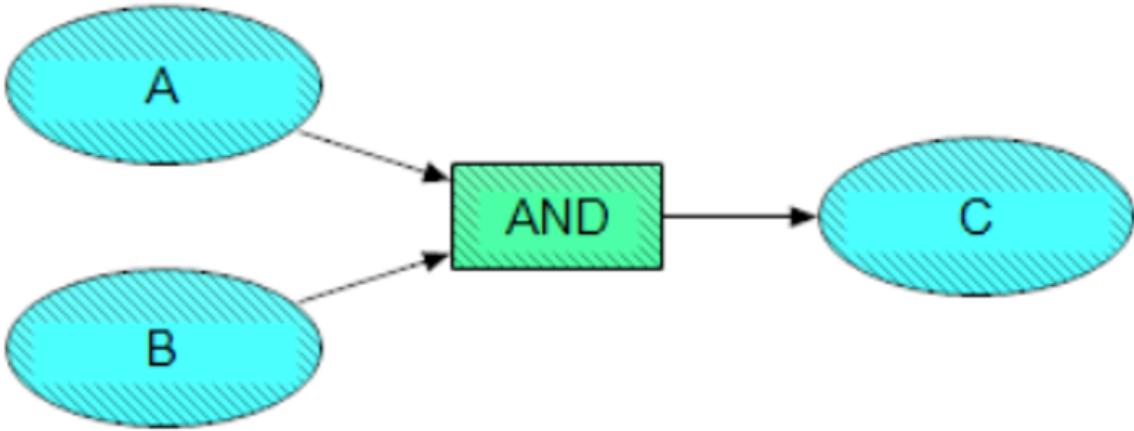
Note on Implementation and Software: At present, we are using [Analytica](#), a “visual software environment for building, exploring, and sharing quantitative decision models that generate prescriptive results.” The models that will be displayed in the rest of this sequence were created using this software program. Note: If you have Windows you can download the

[free version of Analytica](#) and once the full sequence of posts is available, we hope to make the model files available, if not publicly, at least on request. To edit the full model you unfortunately need the expensive licensed version of Analytica, since the free version is limited to editing small models and viewing models created by others. There are some ways around this restriction if you only want to edit individual parts of the model - once the sequence has been posted, please message Daniel Eth, David Manheim, or Aryeh Englander for more information.

How to read Analytica models

Before presenting an overview of the model, and as a reference for later posts, we present a brief explanation of how these models work, and how they should be read. Analytica models are composed of different types of nodes, with the relationships between nodes represented as directed edges (arrows). The two primary types of nodes in our model are variable nodes and modules. Variable nodes are usually oval or rounded rectangles without bolded outlines, and correspond to key hypotheses, cruxes of disagreement, or other parameters of interest. Modules, represented by rounded rectangles with bolded outlines, are “sub-models” that contain their own sets of nodes and relationships. In our model we also sometimes use small square nodes to visually represent AND, OR, or NOT relationships. In the software, a far wider set of ways to combine outputs from nodes are available, and will be used in our model - but they are difficult to represent visually.



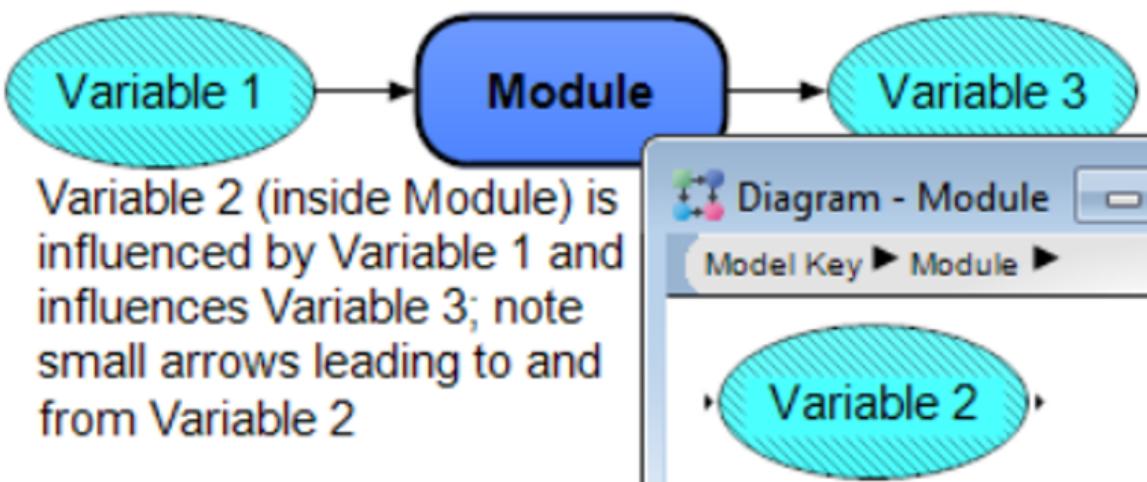


Arrows represent directions of probabilistic influence, in the sense that information about the first node influences the probability estimate for the second node. For example, an arrow from Variable A to Variable B indicates that the probability of B depends at least in part on the probability of A. It is important to note that the model is not a causal model *per se*. An edge from one node to another does not necessarily imply that the first *causes* the second, but rather that there is some relationship between them such that information about the first *informs* the probability estimate for the second. Some edges do represent causal relationships, but only insofar as that relationship is important for informing probability estimates.

Different parts of the model use various color schemes to group nodes that share certain characteristics, but color does not have any formal meaning in Analytica and is not necessary to make sense of the model. The color schemes for individual parts of the model will be explained as needed, but color differences can be safely ignored if they become confusing.

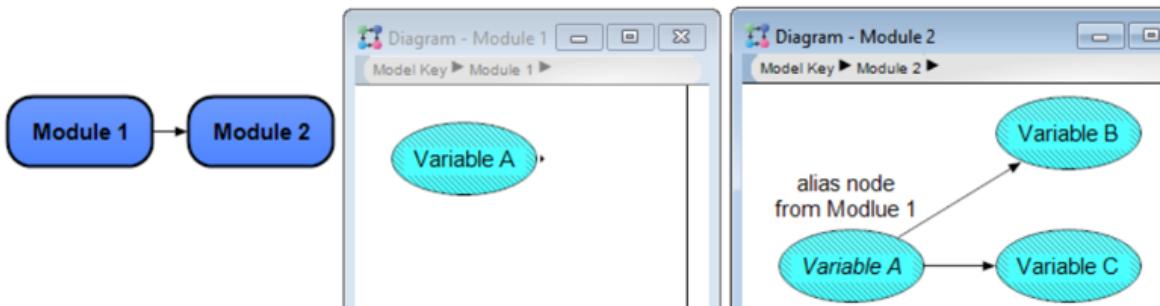
Other things to note:

- In some of the diagrams there are small arrowheads leading into or out of certain nodes, but which do not point to any other node in the diagram. These arrowheads indicate that there are nodes elsewhere in the model that depend on this node or that this node depends on.

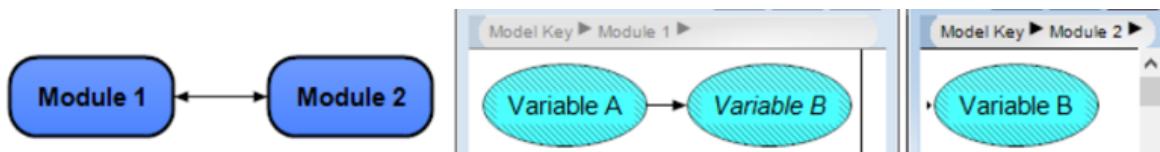


- Alias nodes are copies of nodes that link back to the original “real” node, and are mainly useful for display or readability purposes. We use alias nodes in many parts of

our diagrams, especially when a node from one module influences or is influenced by some important node(s) elsewhere in the model. Analytica indicates that a node is an alias by displaying the node name in italics.



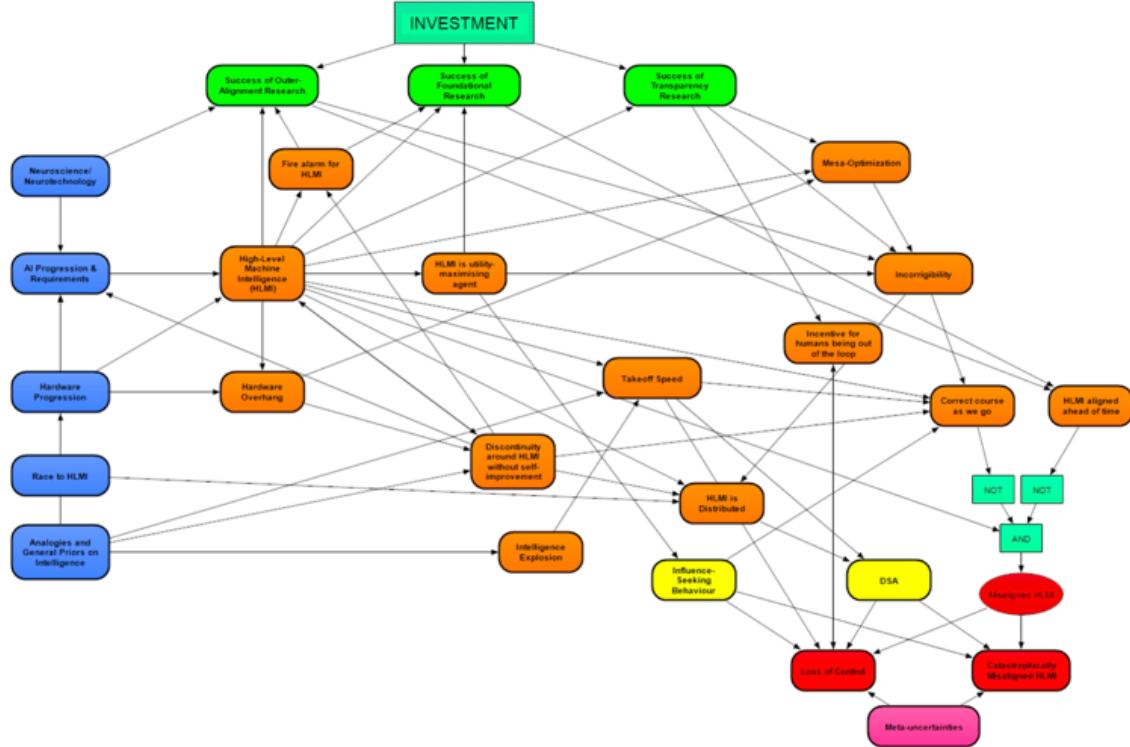
- Our model is technically a directed acyclic graph. However, there are a few places in the model diagrams where Analytica confusingly displays bidirectional arrows between modules even though the direction of influence only goes in one direction. This is because Analytica uses arrows not just to indicate direction of influence, but also to indicate that one module contains an alias node from a different model. For example, the direction of influence in the image below is from Variable A in Module 1 to Variable B in Module 2, but Analytica displays a bidirectional arrow between the modules because Module 1 also contains an alias node from Module 2.



Top-level model walkthrough

The image below represents the top-level diagram of our current model. Most of the nodes in this diagram are their own separate modules, each with their own set of nodes and relationships. Most of these modules will be discussed in much more detail in later posts.

In this overview, we **highlight** key potential nodes and the related questions, and discuss how they are interrelated at a high level. This overview, which in part explains the diagram below, hopes to provide a basic outline of what later posts will discuss in much more detail. (Note that the arrows represent the direction of inference in the model, rather than the underlying causal relationships. Also note that the relationship between the modules reflect dependencies between the individual nodes in the modules, rather than just notional suggestions about the relationships between the concepts represented by the modules themselves.)



High-level model overview

The blue nodes on the left represent technical or other developments or future progress areas that are potentially relevant inputs to the rest of the model. They are: **Neuroscience / Neurotechnology, AI Progression & Requirements, Hardware Progression, and Race to HLM**. Finally, **Analogies and General Priors on Intelligence**, which address many assumptions and arguments by analogy from domains like human evolution, are used to ground debates about AI takeoff or timelines. These are the key inputs for understanding progress towards HLM(1).

The main internal portions of the model (largely in orange), represent the relationships between different hypotheses and potential medium-term outcomes. Several key parts of this, which will be discussed in future posts, include paths to **High-Level Machine Intelligence (HLM)** (and the inputs to it, in the blue modules), **Takeoff/discontinuities**, and **Mesa-optimization**. Impacting these are different safety agendas (along the top in green), which will be reviewed in another post.

Finally, the nodes on the bottom right represent conditions leading to failure (yellow) and failure modes (red). For instance, the possibility of **Misaligned HLM** (bottom right in red) motivates the critical question of how the misalignment can be prevented. Two possibilities are modelled (orange nodes, right): The first possibility is that HLM is **aligned ahead of time** (using **Outer Alignment, Inner Alignment**) and, if necessary, **Foundational Research**). The second possibility is that we can '**correct course as we go**', for instance, by using an alignment method that ensures the HLM is **corrigible**.

While our model has intermediate outputs (which when complete will include estimates of HLM timelines and takeoff speed), its principal outputs are the predictions for the modules marked in red. **Catastrophically Misaligned HLM** covers scenarios involving a single HLM or a coalition achieving a **Decisive Strategic Advantage (DSA)** over the rest of the world and causing an existential catastrophe. **Loss of Control** covers 'creeping failure' scenarios, including those that don't require a coalition or individual to seize a DSA.

The Model is (Already) Wrong

We expect that readers will disagree with us, and with one another, about various points - we hope you flag these issues. At the same time, the above is only a high level overview, and we already know that many items in the above overview are contentious or unclear - which is exactly why we are trying to map it more clearly.

Throughout this work, we attempt to model disagreements and how they relate to each other, as shown in the earlier notional outline for mapping key hypotheses. As a concrete example, whether HLMI will be agentive, itself a debate, influences whether it is plausible that the HLMI will attempt to self-modify or design successors. The feasibility of either modification or successor design is another debate, and this partly determines the potential for very fast takeoff, influencing the probability of a catastrophic outcome. As the example illustrates, the values and the connections between the nodes are all therefore subject to potential disagreement, which must be represented in order to model the risk. Further and more detailed examples are provided in upcoming posts.

Further Posts and Feedback

The further posts in this sequence will cover the internals of these modules, which are only outlined at a very high level here. This is intended to be a sequence that will be posted over the coming weeks, starting with the post on **Analogy and General Priors on Intelligence** later this week, followed by **Paths to HLMI**.

If you think any of this is potentially useful, or if you already disagree with some of our claims, we are very interested in feedback and disagreements and hope to have a productive discussion in the comments. We are especially interested in places where the model does not capture your views or fails to include an uncertainty that you think could be an important crux. Similarly, if the explanation seems confused or confusing, flagging this is useful - both to help us clarify, and to ensure it doesn't reflect an actual disagreement. It may also be useful to flag things that you think are *not* cruxes, or are obvious, since others may disagree.

Also, if this seems interesting or related to any other work you are doing to map or predict the risks, please be in touch - we would be happy to have more people to consult with or who wish to participate directly.

Footnotes

1. Note that HLMI is viewed as a precursor to, and a likely cause of, transformative AI. For this reason, in the model, we discuss HLMI, which is defined more precisely in later posts.

Acknowledgements

The MTAIR project (formerly titled, "AI Forecasting: What Could Possibly Go Wrong?") was originally funded through the Johns Hopkins University Applied Physics Laboratory (APL), with team members outside of APL working as volunteers. While APL funding was only for one year, the non-APL members of the team have continued work on the project, with [additional support from the EA Long-Term Future Fund](#) (except for Daniel Eth, whose funding comes from FHI). Aryeh Englander has also continued working with the project under a grant from the [Johns Hopkins Institute for Assured Autonomy \(IAA\)](#).

The project is led by Daniel Eth (FHI), David Manheim, and Aryeh Englander (APL). The original APL team included Aryeh Englander, Randy Saunders, Joe Bernstein, Lauren Ice, Sam

Barham, Julie Marble, and Seth Weiner. Non-APL team members include Daniel Eth (FHI), David Manheim, Ben Cottier, Sammy Martin, Jérémie Perret, Issa Rice, Ross Gruetzemacher (Wichita State University), Alexis Carlier (FHI), and Jaime Sevilla.

We would like to thank a number of people who have graciously provided feedback and discussion on the project. These include (apologies to anybody who may have accidentally been left off this list): Ashley Llorens (formerly APL, currently at Microsoft), I-Jeng Wang (APL), Jim Scouras (APL), Helen Toner, Rohin Shah, Ben Garfinkel, Daniel Kokotajlo, and Danny Hernandez, as well as several others who prefer not to be mentioned. We are also indebted to several people who have provided feedback on this series of posts, including Rohin Shah, Neel Nanda, Adam Shimi, Edo Arad, and Ozzie Gooen.

How DeepMind's Generally Capable Agents Were Trained

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

Intro

One of DeepMind's latest papers, [Open-Ended Learning Leads to Generally Capable Agents](#), explains how DeepMind produced agents that can successfully play games as complex as hide-and-seek or capture-the-flag *without* even having trained on or seen these games before.

As far as I know, this is an entirely unprecedented level of generality for a reinforcement-learning agent.

The following is a high-level summary of the paper, meant to be accessible to non-specialists, that should nevertheless produce something resembling a gears-level model. I want to focus on explaining the optimization process that produced this agent; on what the different parts of the optimization process are; on why each different part is necessary; and on what would happen if different parts of it were missing.

After that summary, I'll add a few more comments and questions about design choices within the paper and about future research I'd like to see. I'm far less certain about this second part, however.

I was going to include a part on AI timelines -- but whether this paper influences your timelines, and in what direction, depends on a lot of priors that are out-of-scope for what I want to do here.

The Environment

Before we get into the optimization process of the agent, I need to talk about the environment within which the agent trained. Core to the project of this paper are the millions of dynamically-generated tasks on which the agent can train.

Each task in the *XLand* [Unity](#)-powered 3d environment space is defined by a (1) unique physical world and a (2) unique set of goals / rewards. Throughout what follows I refer to (1) as the "environment" or "world", to (2) as the "game", and to both of them together as a "task." Note that both of these can be generated programmatically without human intervention.

(The [show reel](#) of the trained agents operating on hand-made, held-out test tasks is worth watching for at least a few minutes to get a feel for the complexity possible from both world and goals, and is probably much clearer than my writing about the environment space. [I mean, if you want to understand soccer, watch a game of it, don't read a description.] Although you should note that the intelligibility of the goals in the video is uncharacteristic of the goals in the *training* tasks, because the show-

real goals were made by humans from human-intelligible games, rather than randomly generated.)

Anyhow. What kind of variety exists in this space?

Well, each world has a static landscape with dynamic, simulated rigid-body objects placed upon it. The topographical features of the world can be randomly mutated; the size of the world and lighting of the world can vary; the rigid-body cubes, pyramids, spheres, and slabs on the map can be randomly colored, sized, and placed.

Each game, with goals and rewards, can also be randomly generated. I'm not being redundant by talking about goals and rewards; each agent both receives information specifying what would *cause* it to be rewarded (the goal) *and* receives a numeric reward of 0 or 1 in each timestep.

The fundamental atoms for the definition of goals are atomic predicates, such as being "on", "near", "far", or "holding" something. These atoms can be applied to different entities to form sub-goals, such as "the player is on the yellow floor" or "the black cube is near the black pyramid". A *complete* goal is then represented as set of options (disjunctions) over some set(s) of necessary predicates (conjunctions) -- a complete goal might be "(Hold a purple sphere AND be near a yellow cube) OR (See an opponent AND be near a black cube)." . Obviously such goals can be randomly generated, and obviously there are a very large number of them.

The total environment space is then the combination of possible worlds with possible games, forming tasks. This space is enormous -- the paper says there are 10^{16} possible world topologies, and 10^{18} possible games.

I believe that the enormous number of possible tasks is *really important* if you're interested in machine learning as a path towards actual human-level intelligence. Why?

Let me back up a little with some machine learning 101.

In most non-reinforcement-learning, machine-learning tasks, you divide the data from which you learn into two sets: a training set and a test set. (Actually, at least three, but I simplify for now.)

Suppose you're training an image model to distinguish between hotdog and not-a-hotdog. If you had 1000 images with which to train, and you let it train on *all* of them, you'd be in a bad state. Even if it was able to successfully distinguish between hotdog and not-a-hot-dog for every one of *those* images, you would not know if it could generalize to *new* images. It might have simply memorized which of those *particular* images were hotdogs and which were not, without learning to generalize to never-before-seen-images. You'd be like a math teacher who discovered he was using the same math problems in his blackboard demonstrations and on his tests, by accident.

To avoid this, you could divide your images into a 800-image training set and an 200-image test set. Suppose you train the model on the training set exclusively, without showing it the test set, until it gives you the right answer for all 800 images in the training set. You then check what results it gives you on the test set, without training on the test set. *If* the model has learned to figure out the Platonic form-of-a-hotdog from the training data, it might give you the right answer for 199 of your 200-image test set, indicating that your training has been a success. On the other hand, *if* it has

simply memorized the data in your 800-image training set, it might do no better than random on the 200-image test, indicating that you have *overfit* to your training data.

This kind of process is absolutely fundamental to non reinforcement-learning machine learning. The first thing you learn as a ML engineer is to always carefully divide your data into training and test sets (well, training, test, and validation sets, and sometimes more) and never to confuse them.

The problem with reinforcement learning is that until a few years ago, for 99% of the tasks, there was no distinction between training and test set. You could train an algorithm to play [Breakout](#) from the [Atari Learning Environment](#), and if it did well -- it did well, and you wrote your paper, and didn't try your algorithm against a test game.

In retrospect this is somewhat shocking.

But to be fair, it is a little unclear what the distinction would be between a training task and a test task in reinforcement learning. What is a game that's similar to a game you play on, such that it's the same task in some sense (like identifying a hotdog) but which is nevertheless not the same task in another sense (like identifying a different picture of a hotdog)? When the answer rolled around, it was of course obvious: levels of a video game could fill this notion fairly precisely.

[OpenAI's ProcGen](#) benchmark is a good example of such an environment: you can train on some arbitrary number of procedurally-generated levels in various video games, and then test your performance on a never-before-seen set of other procedurally-generated levels.

It became pretty clear from results on ProcGen (and elsewhere) that the existing RL algorithms were probably horribly, horribly overfit. In the case of ProcGen, OpenAI showed that an agent could train on 100 or even 1000 different levels of some task, learn to do very well on them all, but still do very, very badly on a never-before-seen level. This is obviously a problem if you want your RL agents to be steps on a staircase towards Artificial General Intelligence, rather than steps on a staircase towards making artificially shitty students in a class who only learn to parrot their teacher.

So the ability to randomly generate worlds and games in the *XLand* space is very important for this reason. It lets DeepMind check to see if their agents have actually learned transferable skills, or if they have instead memorized actions likely to result in high reward without truly learning how to navigate the environment.

DeepMind's paper specifies (appendix 3) that they created a list of test worlds (where a world is a topology and set of object placements) and test games (where a game is a set of rewards / goals), and never trained their agents on tasks (combined worlds and games) where *either* of these matched the held-out test worlds or games. This is going further than ProcGen -- while ProcGen varies the *world*, but keeps *tasks* the same, DeepMind is varying both world and the tasks.

I believe that this is the kind of thing you need to be doing if you're actually interested in taking steps towards general intelligence, so it is interesting to see this.

The Optimization Process

Intro

DeepMind describes the process of training an agent in this task space as involving 3 levels. I'm going to conceptually split their middle level into two, for a total of 4 levels, because I think this makes it easier to understand.

This 4-level stack looks to me essentially like a complication of the inner-and-outer loop optimization, probably best described in full generality by [Gwern](#). And although this optimization process can be used in ML, it can be found naturally arising in the world as well.

The basic mechanism behind such a two-part optimization process is to have a quickly-acting, but unavoidably biased or inaccurate inner optimizer, which is evaluated by a slowly-acting, but unbiased and accurate outer optimization process.

Within the context of natural history, for instance, you can view human intelligence and evolution as this kind of two-level process optimizing reproductive fitness. Intelligence is the inner optimizer. It is fast; it allows you to come up with ideas like crop rotation, writing, irrigation, and so on, which allow the humans that implement these ideas to have lots of reproductive success. But it is also inaccurate (from the perspective of reproductive fitness); it allows you to come up with ideas like social media and online porn, which might cause the humans that implement these ideas to have less reproductive success. The fast inner loop of human intelligence is ultimately judged by the slower, outer loop of reproductive success or natural selection.

Similar dynamics can be found in the economic sphere, with internal-to-a-corporation-change (inner loop) and bankruptcy (outer loop).

The 3 or 4 level process that DeepMind creates is a little more complex, but mostly the same kind of thing.

Here's a description of the four levels. I'll go into each of them in more detail below -- this is just for initial orientation.

1. *Reinforcement learning improves the performance of an agent over a particular mixture of XLand tasks.* Note that this is done for many different agents all at once. But at this part of the optimization process, the agents are only causally connected as occasional opponents to each other in competitive tasks, or cooperating partners in cooperative tasks.

2. *An agent-specific teacher adjusts the distribution of XLand tasks to ensure the right difficulty.* That is, as the agent trains, a teacher adjusts the tasks to be neither too hard nor too easy for the agent.

3. *Periodically, two different agent-teacher pairs are compared, and pairs that do better against a measure of general competence are copied and replace pairs that do worse.* That is, natural selection occurs against our *actual* objective -- a carefully developed notion of general competence.

4. *Periodically, replace all of our agents with a "new generation" of agents, initially trained to imitate the best agents from the prior generation.* . That is, we replace all of our current agents, because after training for some time neural-network based agents seem to lose some of their flexibility and ability to learn; we bootstrap the new from the old agents so they learn faster this time around.

I'm going to dive into each of these steps in much more detail below.

I want to note something before I begin, though. None of these steps involve revolutionary changes; no Kuhnian paradigm shifts are involved in this paper. What this paper shows is what you can accomplish through combining current state-of-the-art techniques, making incremental progress, using good engineering, and throwing a little bit (though not that much, really) of money at compute.

It also doesn't by any means establish an upper limit to these things; I believe the results here clearly indicate you'd make further progress by scaling up along the same lines.

1: Reinforcement Learning

The innermost loop is the one that actually changes how agents act, by adjusting each agent's neural network to increase the likelihood of actions leading to reward.

Each agent runs through the (aforementioned) distribution of dynamically generated tasks within the space of possible tasks. In the beginning, each agent's actions are random -- each just flails around in the virtual worlds. The neural-network-parameterized policy which defines what actions the agent is likely to do is initialized with random values, and so initially outputs a mostly even distribution, with every action as likely as every other. Thus, the random flailing.

But this random action is nevertheless a kind of exploration. Some actions lead to more reward, and some actions to less. The kind of actions followed by above-average reward are made more likely (when in similar situations). And the kind of actions followed by below-average reward are made less likely (when in similar situation). Over time, the distribution of actions output by the neural network becomes less even, and begins to put more probability mass on the better actions, and less probability mass on the worse actions.

This general technique of thus adjusting the probability of actions directly is *policy gradient* optimization. The generalizing power of a neural network is used to help recognize "similar situations," as is the case for all neural-network-based optimization methods. There is a lot of complexity in the architecture of their neural network, which I have not addressed above, but this is nevertheless the core of it.

By many measures of human intelligence, this is a very stupid algorithm. Notably, this kind of algorithm is *model free*; no component of it (at least directly) predicts what the environment or what an opposing agent will do. If you think that [intelligence requires prediction](#) then you'll think that this algorithm is missing very important parts of intelligence. Instead, it is mostly learning a mapping from situations to actions, where the generator of this mapping adjusts it to produce higher reward.

Note, though, that policy gradient methods (and other model free methods) can act as if they are predicting things. [OpenAI Five](#) was trained through policy optimization, and when playing, looked as if it was anticipating what its opponents would do. But in fact, it had merely encountered many similar situations with opponents, and thereby learned what kind of actions maximized reward in them. (This is leaving entirely aside more [difficult mesa-optimization complications](#).)

The specific policy gradient method that DeepMind used is [V-MPO](#), which DeepMind also created. They presumably chose this algorithm because tests on V-MPO in the paper introducing the algorithm show that it performs well in a multi-task setting -- i.e., using it DeepMind trained a *single* agent on all of the Atari-57 tasks, and found that its median performance was still better-than human -- although the open-ended learning paper doesn't mention this motivation.

This innermost part is necessary, well, because without it no agent would ever learn or change in any way.

2. Dynamic Task Generation

The second part of the optimization process is dynamic task generation.

The principal thing that dynamic task generation does is to slowly shift the distribution of tasks on which each agent trains, as each agent improves, so that the distribution is never too hard or too easy. (Bloom's [2-sigma problem](#) anyone?)

The way it works is actually quite easy to understand.

A *candidate* task is generated, by sampling from the immense space of possible worlds and possible games in those worlds -- excluding those worlds and games in the test set, of course.

Then 20 episodes of the task are played through *without training* to see if the task is suitable for training. 10 of these episodes are played-through using the agent-that-might-be-trained with the task. 10 of these are played through with a control "agent" which acts randomly or does nothing at all. The candidate task is accepted as an *actual* training task if and only if the results of these twenty episodes meet three criteria.

1. The agent does not do extremely well on it every time; it is not trivial for the agent.
2. The agent *does* do better than the control policy by some margin; it is somewhat better than random.
3. The control policy does poorly.

There are constant values that determine the cutoffs for all these filters. These constants answer questions like: what is the cutoff for doing too well in 1? How much better does the agent have to do than the control policy for 2? How poorly does the control policy have to do in 3? It is important to note that these values are themselves unique *per agent*; every agent has their own dynamic task generation filters.

These constants together define what I have sometimes referred to as the "teacher" above -- but teacher implies too much complexity. There's nothing to the teacher but these few numeric constants, together with how they are applied in the three rules above.

Why is this necessary?

Well, to take a step back, a big problem in RL is how you *first* get some signal from which to learn.

The majority of reinforcement learning algorithms don't learn at all until they've received some *variance* in the rewards that they receive -- some departure from normal, in either a positive or negative direction. Until they receive some signal in this fashion, they only perform random actions.

This, of course, presents a difficulty because many environments have extremely sparse rewards. Imagine trying to play a video game like StarCraft, but only doing random actions until you (by doing random actions) win your first game. Even with an enormous number of trials, you probably never would win, even if you played until the sun grew dim and the stars disappeared from the sky. And so you would therefore never receive any signal, and would never improve in any way.

There are a number of different ways to approach this problem.

One way is to pre-train agents to imitate humans. For DeepMind's [AlphaStar](#) Starcraft-playing agent, for instance, they trained an agent to simply imitate humans first. This at least gets the agent to the point where it is able to win games against easier opponents, after which reinforcement learning in general can start. But this obviously only works in domains where you have a large number of recorded human actions to imitate, which DeepMind does not have here.

Another way to address this problem is through attempting to give agents something like *curiosity*. Curiosity aims to lead the agent to explore the environment, in some better-than-random-fashion, even in the absence of an extrinsic reward. You could try to implement a curiosity module by [rewarding agents](#) for encountering states where they cannot predict the future well. Or you could do something more sophisticated, like [learning an ensemble of models](#) and then exploring the places where the models disagree. This method has the advantage, as far as I can tell, of being the most explicitly human-like approach to the task. But no one has really settled on the perfect implementation of curiosity.

DeepMind, of course, settles on *curriculum learning*, which, as the name implies, works like something like school. You first give the agent a very easy task, which it might be able to do sometimes by random chance. Then you give it a slightly more difficult task, which it is able to do acceptably on because it is no longer doing *entirely* random actions. And so on. There are many ways to implement curriculum learning -- but DeepMind's is implemented through the dynamic task generation and 3-criteria filtering mechanism mentioned above.

So *one* function of the dynamic task generation is to provide curricular learning, so that agents can start learning at all. We know that this is important: DeepMind performed an ablation where they removed dynamic task generation, and found that learning proceeded much more slowly. When dynamic task generation was removed and some other tricks to speed up the initial learning were also removed, pretty much no learning occurred at all.

Dynamic task generation and filtering is not only important because of how it helps provide curricular learning though -- it's also important for how it interacts with the next stage, population-based training.

3. Population-Based Training

At this level of optimization, the population of agents starts to interact through something like natural selection. The basic notion of how this works is once again relatively simple.

Periodically during training, we compare two agents that we have not recently compared. The comparison takes place along a carefully-developed measure of general competence. This measure emphasizes things like "not failing catastrophically on as many tasks as possible" and "being pretty good at a large number of things" over things like "having a very high average score" and "doing very well at some a limited number of tasks".

If one agent's performance dominates another agent's performance along this measure, then the better agent replaces the worse: the weights of the better neural network are copied over, together with the agent-specific task-generation hyperparameters (with random mutations to the hyperparameters).

Why is this necessary? Isn't each agent *already* being trained on a broad curriculum of tasks of generally increasing difficulty, which we would expect to lead to general competence in any case? What does the evolutionary selection give us that we don't already have? What problem does this let us avoid?

There are several answers to this question.

The more narrow answer is that this allows the dynamic task generation hyper parameters *themselves* to shift in a direction that promotes general competence. Neither of the optimization levels beneath us include any way of changing these parameters. But the ideal filtering parameters *for the production of general competence* might be different at the beginning, or at the middle of training. Or they might be different from agent to agent. Without something like population-based-training, they would have no way of changing and this would hurt performance.

The less narrow answer, I think, is that this ensures that agents are developing *broad* competence in a way the innermost loop cannot do.

Imagine a military recruit who does very well at *some* exercises, but only 1/2 of the total set of exercises. So their teacher -- like our dynamic task-generation hyperparameters -- gives them more, harder exercises relating to those routines, and they continue to get better at them, and get harder exercises, and so on. They could still be very bad at the other half of their exercises. They would have failed to develop broad competence even though their teacher kept increasing the difficulty of their exercises as they got better at them.

Similarly, each agent in our population of agents will learn to get better at some distribution of tasks, then, but without population-based-training they might not spread themselves broadly across the entire span of this distribution. Like a student who advances wildly at subjects she prefers, while ignoring subjects she is not good at, our agents might not approach our desired ideal of general competence. Population-based-training helps prevent the scenario by multiplying agent / teacher pairs that do well generally and non-narrowly.

At least that's the theory. DeepMind did perform experiments where they removed PBT, and the performance of the agents on the tasks they were *worst* at fell, as the theory would predict. As I eyeball it, though, it looks like... it was not an enormous change? But we can be sure PBT is at least doing *something*.

4. Generational Training

After training an RL agent on an ever-shifting basket of tasks for a while, its improvement slows down. It stops learning quickly. It might even regress. It's enormously tempting to analogize this to how humans, too, learn more slowly after some time, but would probably be counter-productive to understanding. But, regardless of *why* it happens, the fact remains that it does happen.

To combat this problem, DeepMind uses *generational training*. It works like this.

After large number of all the steps from (1) through (3) you stop training your agents. You spin up an entirely new batch of agents, with their neural networks initialized to random values, and start training them.

The difference is that this time, for the first few billion steps of training, you aren't *just* training the agents to maximize reward -- you train them to output a probability distribution over actions that matches the probability distribution of the best agent from the prior generation. In other words, your agents are trained to match what the prior agent *would do* in those tasks. After a few billion steps, you stop trying to imitate the prior agent -- but by then, you've boosted the performance of your student to far past where it would otherwise be. And after further reinforcement learning, the "student" agent generally exceeds the "teacher" agent with which it was initially trained.

Why is this important? Well, two reasons.

First: It just improves performance, in a generic way that has nothing to do with open-ended learning specific to this paper. It just helps combat a weakness of our current RL tasks. One of the [initial](#) papers on "kickstarting" agents by training them to imitate other, trained agents showed that the kickstarted agents eventually exceeded the performance of the un-kickstarted agents by 42%. The effect here is less dramatic, but DeepMind's ablation shows that the agents which learned from prior agents eventually exceed the performance of those from which they learned, *even when* those from whom they learned were trained for longer. RL is hard, and DeepMind wants any kind of performance gain it can get, even if the gain is generic and has little to do with open-ended learning, and this provides it.

Second: Generational training provides a convenient way to switch lower-level objectives as generations pass, in order to promote more general learning.

For instance, for the first two of the five generations in their experiment, the objective they optimize the PBT training against focuses even more strongly on ok performance over many tasks than it does in the subsequent three generations; in the first two, it optimizes *only* for increasing the percent of tasks you get any reward from. So having generations provides a convenient switching-point for changing lower-level features of the training.

Summary & Results

I can now review the training process.

At the lowest level, each agent runs through tasks, slowly increasing the probability of the kind of action which is followed by reward and decreasing the probability of the kind of action not followed by reward. The tasks on which each agent trains are dynamically generated to be neither too hard nor too easy; as the agent becomes more intelligent, the tasks become correspondingly harder. Periodically, two agents are compared according to a measure of broad competence. Agents, and dynamic task-generation hyper parameters corresponding to the agents, are removed if they perform poorly, and copied with mutations if they perform well. Over even longer periods, entire generations of agents are removed, and the best used to pre-train subsequent generations of agents.

This goes on for five generations. Daniel Kokotailjo estimated very approximately that this cost [about half a million dollars](#) to train. What does that half million get us, in terms of performance?

In short, something probably significantly stupider than a mouse, but which is still among the smartest artificially intelligent agents that have been created, by conventional human standards of intelligence. I'm tempted to say it's the first artificially intelligent agent it would make sense to *compare* to a mouse.

Well, as mentioned, the agents can solve -- or, at least get non-zero reward -- in non-trivial problems *they have never seen before*. Here are some of the hand-made games the agents were able to score in, despite never seeing them before, with DeepMind's names and descriptions:

- **Catch Em All:** A cooperative game where 2 players must gather 5 objects from around the world and bring them together.
- **Object Permanence Black Cube:** The agent starts and can see a yellow cube on the left and a black cube on the right. The agent must choose which path to take (which means the agent loses sight of the cubes) to reach the target cube (black in this case).
- **Stop Rolling Freeze Gadge:** In a world composed of one big slope, the agent must stop the purple sphere from touching the bottom floor without holding it.
- **Race To Clifftop With Orb:** Both players must stand on the top of a cliff edge holding the yellow sphere, without the other player standing on the cliff.

I should note that when acting in this zero-shot capacity, the agents do not perceive the reward that they receive when they accomplish their goal. That is, they must identify when they have accomplished the goal, because the numeric reward that they are given only alters them when they are training -- it isn't perceived by them during their activity. This makes all of the above at least marginally more impressive; they cannot merely fiddle with things until they receive a reward, then stop.

But we also should look into the darkness. Here are some problems the agents *cannot* get reward in:

- **Tool Use Climb 1:** The agent must use the objects to reach a higher floor with the target object.
- **Sheep Herder:** The agent must make the other player stand near a set of target objects.

- **Make Follow Easy:** The agent must lead the other player (whose policy is to follow) to a particular floor colour.

It's hard to know how much to make of these failures.

On one hand, a human could easily execute them. On the other hand, they are very-out-of-domain of the set of training worlds. "Tool use climb", for instance, apparently involves building multiple ramps to reach a location -- the world-generation procedure explicitly excludes worlds where ramps are required, so the agent would never have seen any such world. I very strongly suspect that if given a slightly more expansive training environment, with places that required a ramp to reach them, then these and similar tasks would have been easily accomplished by the agent.

This possibility-of-further-capacity is reflected by how well the agent does on the dynamically-generated, non-human-generated held-out test set. The agent was able to accomplish some reward on one hundred percent of the tasks where it is possible to receive reward. (Some percent of the tasks were impossible.)

Given that, I very strongly expect that this paper does not expose the upper limit of intelligence of agents trained in this manner. Like GPT-N, it presents us with a curve reaching up and to the right, where our prior theories about the world are likely to shape what shape we think that curve has.

Critique

There are some specific features of this work I was a little disappointed by, or would like to see remedied in future works. Here is a partial list. This section is very mildly more technical than the previous sections.

Too-Specific Goal Neural Network

As mentioned above, goals are specified for the agents as sets of "and"s joined by "or"s. So goals are things like "(A and B) or (C and D)" or "(A and B and C) or (D and E) or (F)". The agent cannot receive goals specified in another fashion: the neural network is not built to be able to understand a goal like "((A or B) and (C or D))". And in general the grammar that it understands is quite limited.

This makes some of the accomplishments of the agents a little less impressive. DeepMind has a section explaining how the agent is able to switch between alternatives: it is able to start to try to accomplish one branch of a disjunction, but then to switch to another when it realizes there is an easier way to be rewarded. The way that its brain has been built very specifically to understand such a disjunction makes this accomplishment much, much less impressive.

I nevertheless anticipate that future versions of this work will try to accommodate a more expansive grammar.

Extra Information During Training

The neural network trained through policy optimization has two parts. The *recurrent* part of it rolls up all of the observations that the agent has made up until that point in

the episode; it corresponds to a kind of general awareness of the situation. The *non-recurrent* part of it receives information about the goal and then queries the recurrent part of it to help figure out what action in that situation leads to the most reward.

During training, the recurrent part of the neural network is trained to predict the truth of *all* the predicates involved in the goal. This feels like cheating to me; it certainly would be impossible with a sufficiently more expressive grammar defining the goal. I'd be interested in how much the sample-efficient training of the agent depends on this auxiliary goal during training. And I'd be very, very interested in efforts to replace it.

Dynamic Task Generation

This is my most theory-laden / questionable criticism.

Eyeballing the numbers on page 21 of the paper, it looks like dynamic task generation is one of the most important components of the whole thing. Performance plunges by *a lot* when it is removed.

I think this is important, because dynamic task generation occupies much the same space that curiosity does, in terms of functionality. Both help address the problem that current RL algorithms simply cannot solve unaided: very sparsely-rewarded action spaces. In the long run, I think, curiosity is a better bet than curricular learning for actual intelligence. It feels like a core part of intelligence, one we cannot yet create. This paper very neatly sidesteps the need for it, but at some point I think it might be no longer able-to-be-sidestepped.

[x-post](#)

Covid 8/12: The Worst Is Over

Good news, everyone! Andrew Cuomo has resigned, and Andrew Cuomo is the worst.



I will damn well take it, because it's not like he doesn't *also* deserve to resign in disgrace for the stuff that officially got him, and again, also to say it for what is hopefully one last time, Andrew Cuomo Is The Worst. Hence, The Worst Is Over. [Sing it high, sing it low.](#) (HT: [Meme source](#))

The title this week does not *as reliably or fully* refer to the Delta variant or the Covid-19 pandemic. Things are still steadily getting worse. But the turning point is plausibly in sight, as case growth slows, and I doubt we have more than one doubling left before things peak.

Main event this week was continued arguments over mandates, both for vaccines and for masks and other NPIs. I'm making one last attempt this week to explain my reasoning on vaccination mandates, as I continue to get people disagreeing for a variety of reasons, and

disambiguating the disagreements seems worthwhile; I tested it out in the comments last week and it seemed productive.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: 855,000 cases (+45%) and 4,100 deaths (+40%).

Result: 744k cases (+26%) and 3,725 deaths (+29%).

The case numbers are very encouraging. They're still increasing, and there's always the chance this was a data fluctuation that will be actively undone, but it probably wasn't and represents either the control system kicking in or us nearing a natural peak. I still expect a similar increase next week, but I'd estimate a 35% chance that within two weeks we get about as high case counts as we're going to see in this wave, and 55% it happens within three weeks.

I'm still predicting a +35% rise in deaths, as the cases from the last few weeks make their way through the system, and hope to be pleasantly surprised.

Prediction for next week: 900k cases (+21%) and 5,028 deaths (+35%).

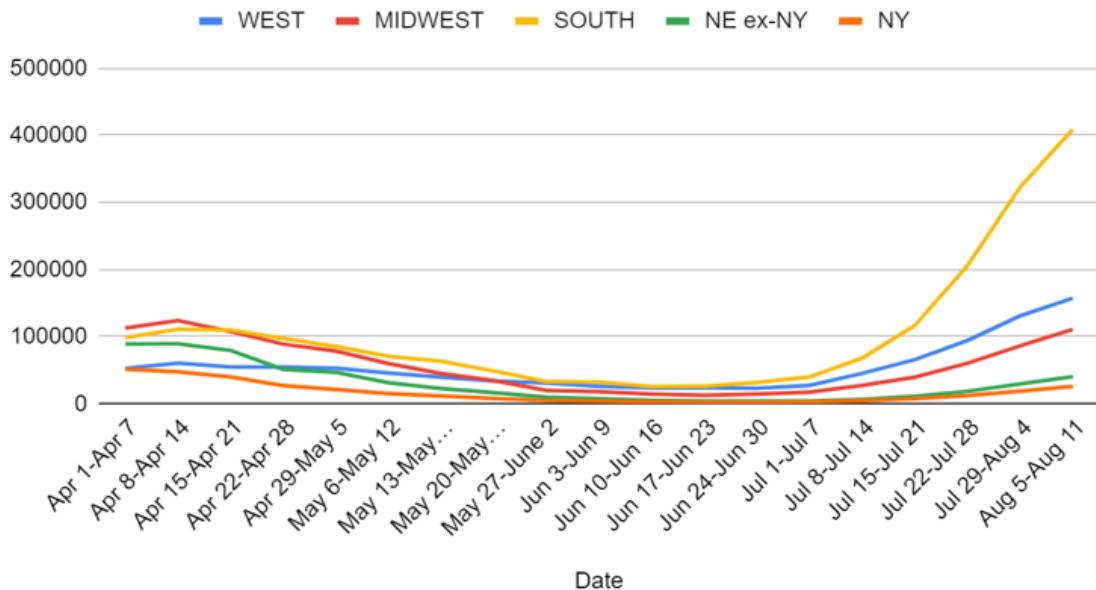
Deaths and Cases

Grouping these together, because there's a combined mystery to solve.

Here's the case numbers.

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 17-Jun 23	23,854	12,801	26,456	6,464	69,575
Jun 24-Jun 30	23,246	14,521	31,773	6,388	75,928
Jul 1-Jul 7	27,413	17,460	40,031	7,065	91,969
Jul 8-Jul 14	45,338	27,544	68,129	11,368	152,379
Jul 15-Jul 21	65,913	39,634	116,933	19,076	241,556
Jul 22-Jul 28	94,429	60,502	205,992	31,073	391,996
Jul 29-Aug 4	131,197	86,394	323,063	48,773	589,427
Aug 5-Aug 11	157,553	110,978	409,184	66,686	744,401

Positive Tests by Region

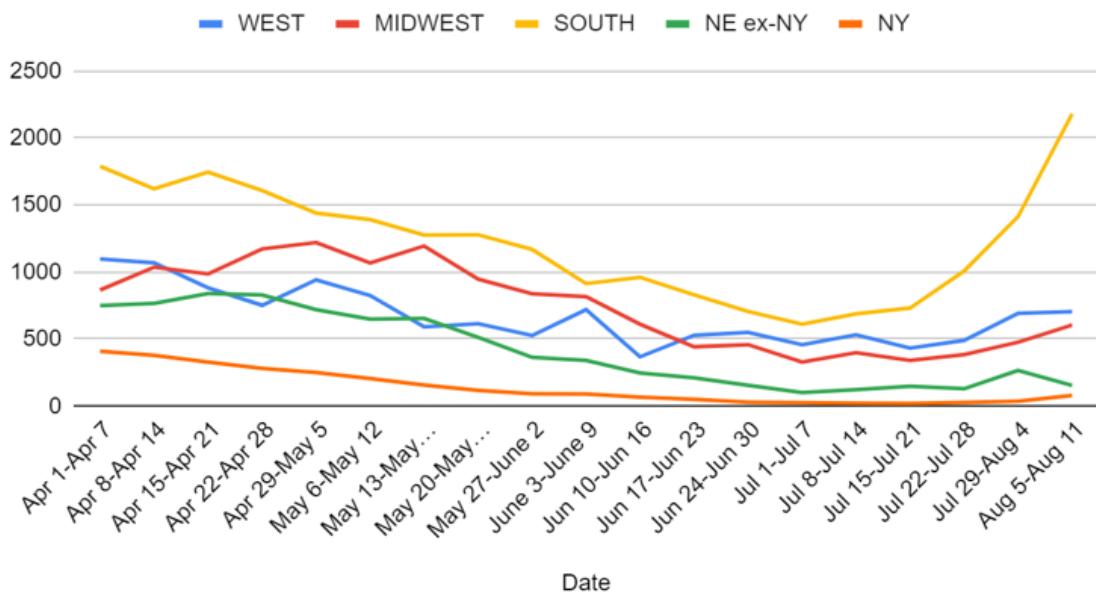


This is a substantial slowdown in new cases, cutting the growth rate in half. We might be peaking within a week or two.

Here's the death numbers.

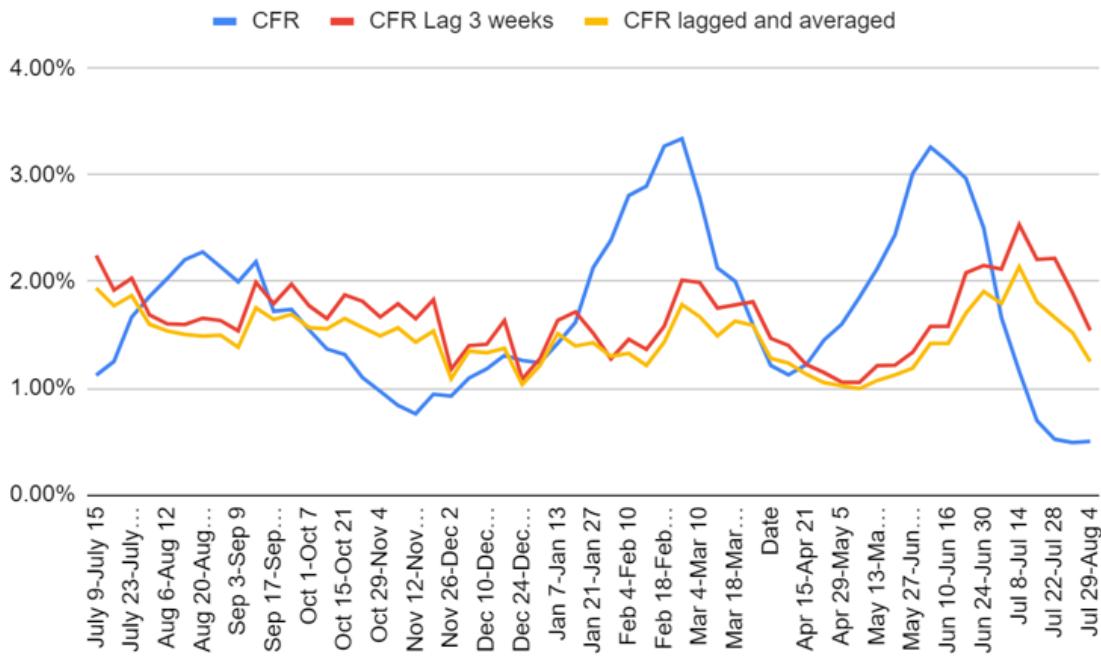
Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 24-Jun 30	550	459	706	186	1901
Jul 1-Jul 7	459	329	612	128	1528
Jul 8-Jul 14	532	398	689	145	1764
Jul 15-Jul 21	434	341	732	170	1677
Jul 22-Jul 28	491	385	1009	157	2042
Jul 29-Aug 4	693	477	1415	304	2889
Aug 5-Aug 11	705	605	2181	234	3725

Deaths by Region



(Note: California reported -352 deaths yesterday, I changed that number to 0. Last week Delaware dumped 124 deaths on us, presumably a backlog, which is why the NE number last week was so high. I'm choosing not to correct for that for now but might smooth it back for next week as it is potentially importantly misleading.)

The death number was modestly better than I predicted, but definitely not good. [This NYMag article](#) (sorry about linking to Topol so often, world is strangely small) highlights the negative perspective of the American death rates not dropping this wave the way they did in other countries. Yes, the CFR is down a lot, but the bulk of that is lag due to cases rising so rapidly. If you undo the lag, you see something else...



[Lagged = cases from 3 weeks ago. Lagged and averaged = cases from $0.1*(1 \text{ week ago}) + 0.2*(2 \text{ weeks ago}) + 0.5*(3 \text{ weeks ago}) + 0.2*(4 \text{ weeks ago}) + 0.1*(5 \text{ weeks ago})$], chosen quickly to be not crazy]

At the very beginning the CFR was much higher, but once we got adequate testing and reasonable care, things haven't changed much on this chart. This is not at all what we see in the UK, where the CFR is clearly down a lot.

Moving-average case fatality rate of COVID-19

The case fatality rate (CFR) is the ratio between confirmed deaths and confirmed cases. Our moving-average CFR is calculated as the ratio between the 7-day-average of the number of deaths and the 7-day-average of the number of cases 10 days earlier.

Our World
in Data



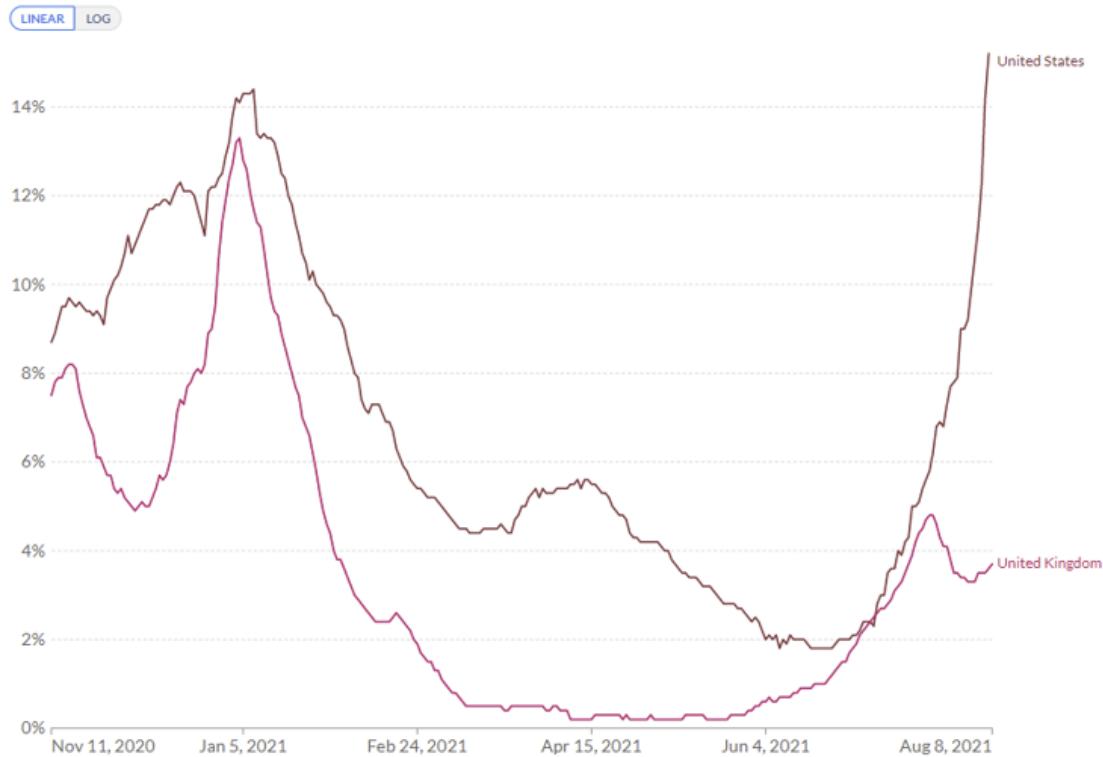
The uptick at the end represents the UK having declining case counts, rather than a higher death rate. The Netherlands looks similar to the UK but made the graph hard to read.

A simple theory is it has a lot to do with this, the positive test rate:

The share of daily COVID-19 tests that are positive

Shown is the rolling 7-day average. The number of confirmed cases divided by the number of tests, expressed as a percentage. Tests may refer to the number of tests performed or the number of people tested, depending on which is reported by the particular country.

Our World
in Data



We went from 2% positive tests to 15% positive tests, which indicates our testing is not keeping pace with our cases. In fact, it's not keeping pace *at all*?

Daily new COVID-19 tests per 1,000 people
Shown is the rolling 7-day average.

Our World
in Data



Seriously, what the hell, people? It's one thing to not keep pace, it's another thing to be testing less now than we were when cases were at their lows. That means that, conditional on not having Covid, the chances of a given person getting tested have gone down a lot, despite their reasons to worry they have Covid having gone way up. Much higher chance they have a known contact with someone positive, and much higher background danger.

One possibility is that people are now increasingly using at-home tests first. We did this last week, when our four year old had a cough and we were asked to exercise an abundance of caution – it was annoying to stick that thing up his nose for a while, but it was cheap and quick, and got the job done. Whereas when we looked into getting a non-rapid non-home test, it was clearly going to be a pain in the ass to get. One of my friends posted a similar experience, where she couldn't get her child tested for several days unless she got a test at a pharmacy. Both of us got negative results, and neither result counts in the charts above. If one of them had come back positive, it's not clear how often that would have made it into our statistics either. If there was no need to escalate to official medical care, there's no default mechanism to get those results into the statistics.

Whether or not that's the main mechanism, I am confident that a large majority of cases, especially asymptomatic or mild cases, are being missed entirely by the system. The true IFR likely has not dropped here quite as much as in other places, but our vaccination rates are not that much lower and our medical care is quite good and holding firm, so it's likely falling almost as fast here as elsewhere. Our CFR staying high says more about our rate of case detection than it does anything else.

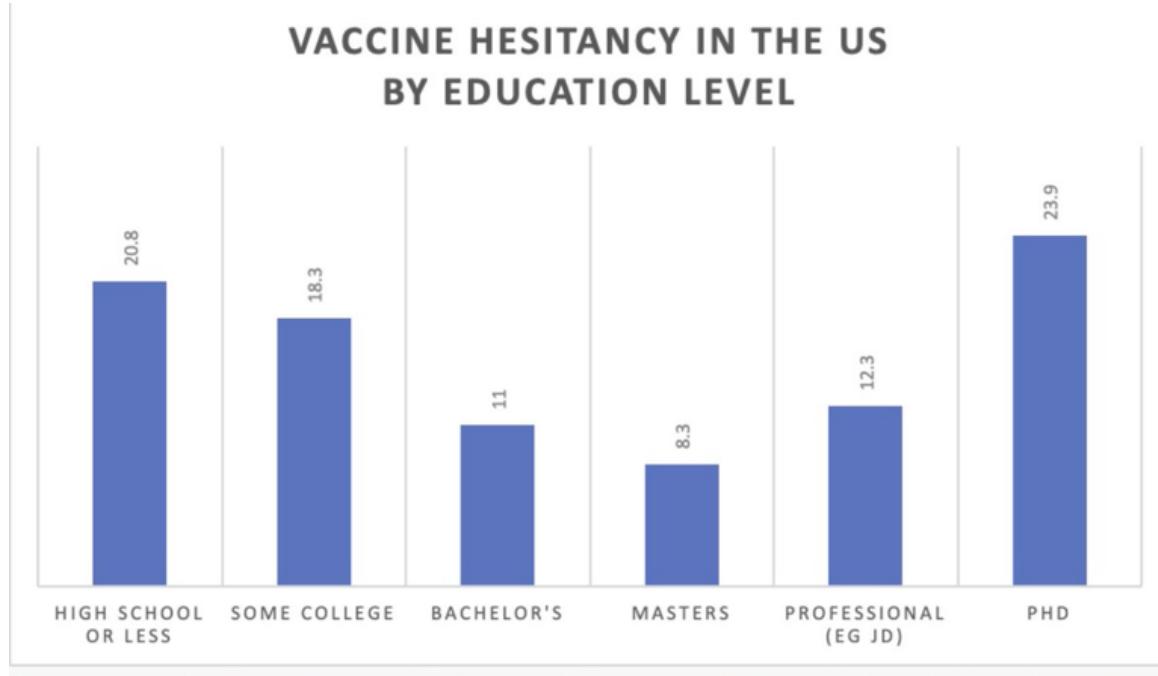
In some ways, this undercounting of cases is very good news, as it puts an upper bound on how bad the pandemic can get and moves us more rapidly towards the turning point. In terms of hopes for containment, however, it's very bad news. If cases have grown much more rapidly than our numbers naively indicate, then our hopes for containing this in a

meaningful way, rather than waiting for enough people to get Covid that things die down afterwards, are much worse.

Vaccinations

[Here's Zeynep, on point as always, on vaccine hesitancy and the reasons behind it](#), which are far more reasonable than they're usually made out to be.

Also, it turns out that having a Bachelor's or Masters makes you very likely to be vaccinated, but if you go on to get a PhD, [you're now in the least vaccinated group of all](#).



The percentage of each education group that is somewhat vaccine hesitant. Source: Carnegie Mellon University

[This week's thread on relative infectiousness of the vaccinated versus the unvaccinated](#). As you would expect, when you do random population samples, you find lower average viral loads in vaccinated people.

[The can-do spirit](#):



Julian Sanchez @normative · Aug 5

Always nice when the bartender offers a round of shots.

...



The Pug @thepugdc · Aug 5

Again, to reiterate. If you have not been able to get vaccinated, and want to come to the pug, I will take you to get vaccinated.

5

36

374

↑

Also, perhaps we could try publicizing this more? [Ask 24 out of 25 doctors if vaccination is right for you?](#) Maybe take the crosstab of Republican doctors which is presumably still over

90%? Playing to win the game takes many forms.

AMA survey shows over 96% of doctors fully vaccinated against COVID-19

[Baltimore's anti-Covid ad posters.](#) No idea if they're effective but at least they're playing.

[The can't-do spirit:](#)



Kaitlan Collins @kaitlancollins · Aug 6

...

Vaccine doses are going to waste in Alabama, where millions remain unvaccinated. A state health official says 7,000 J&J, 11,000 Moderna and about 47,000 Pfizer have expired. "That's extremely unfortunate when we have such a low vaccination rate," Dr. Scott Harris says.

This seems to me like it should be a rather large scandal. If people in Alabama don't want the vaccine, and it's going to expire, then it's imperative to recognize that and send it elsewhere where people do want it. Worst case, you cycle out old vaccine to places cycling through what they have, and then move in new shipments so there's still stock around when and if people change their minds. Also, 'expired' vaccines might not be good enough for us, but they doubtless mostly still work, so I'd *still* send 'em off to whoever wants them.

Yet I see almost no discussion of such matters. [Shame!](#)

[The Isreali can't do but therefore do what you can spirit:](#)

Most people will end up contracting the coronavirus, the head of the Health Ministry's advisory committee for infectious diseases predicted on Monday.

"The [real] question is whether the infected person is vaccinated or not. It's unavoidable that the pandemic will infect the majority of the population. It won't disappear in another half a year," Dr. Tal Brosh told the Kan public broadcaster.

[And the can't get paid enough spirit,](#) I too blame capitalism for all these vaccines:



erika, the mercurial @politklwaitress · Aug 7

...

Pfizer and Moderna are planning on selling doses of the vaccines for about \$25 per dose to countries and project a profit of over 40 billion, I HATE CAPITALISM.

Vaccine Mandates

[Nate Silver broke down people's current stance on Covid into a categories](#), and I think this is mostly accurate:

- **Group A (25% of electorate):** Vaccinated but not ready for a “return to normal” and thinks society has opened up too fast. Very worried about Delta. In favor of any and all restrictions including lockdowns and remote learning. Some of this group will transition into Group B if/when cases start falling.
- **Group B (30%):** Vaccinated, somewhat worried about Delta, and in favor of modest restrictions (i.e. indoor masking) especially if they target the unvaccinated. At the same time, don’t want a return to lockdowns, although some could drift into Group A if cases keep rising or there are scary new variants. This is quickly becoming the consensus position among college-educated elites.
- **Group C (15%):** Vaccinated but “over” the pandemic and wary of restrictions. Likely a mix of younger voters who don’t vote at high rates and some center-right libertarian types.
- **Group D (25%):** Unvaccinated and strongly opposed to any restrictions.
- **Group E (5%):** Unvaccinated *but* in favor of other restrictions. Indeed, they may think such restrictions are necessary to protect them because they can’t/won’t get vaccinated. Unlike Group D, which is mostly conservative Republicans + some apathetic younger voters, this group probably leans Democratic and working-class.



Nate Silver @NateSilver538 · Aug 5

...

It's not a super-easy climate for politicians to navigate. The "smart" play is probably to do what Group B wants, but you may still wind up with various people in the other groups mad at you for all sorts of different reasons.

113

52

717



Nate Silver @NateSilver538 · Aug 5

...

A couple of additional notes:

- You could probably divide Group C into **Group C1 (10%)** and **Group C2 (5%)**. C1 mostly don’t want restrictions on *themselves* (so no universal masking, for instance) but they might be OK with vaccine mandates especially if their friends are vaccinated. C2 are the pure libertarian types who oppose restrictions on a philosophical basis.
- If you can persuade people in Group D to get vaccinated, they’ll mostly move to Group C1 or C2. If you can persuade people in Group E to get vaccinated, they’ll mostly move to Group A.

[Jacob offers an alternative theory of Group A:](#)



Jakeup @yashkaf · Aug 5

I still think group A is mostly not about COVID but about social anxiety. It's introverts who hate socializing (including in the office) and don't want to go back to "normal" when they are pressured to go out and do so. In a survey I ran on this they were 30% of respondents.



Jakeup @yashkaf · Aug 6

- "Hell is other people"
- haha yes, you said it!
- "The longer I spend with people, the more I like cats"
- ain't that the truth!
- "25-30% of people never want to socialize again or be pressured to do so"
- oh no COVID misinformation we must explain that vaccine breakthrough

My guess is this is indeed a substantial percentage of Group A, but I'd be rather shocked if it was anything like a majority.

One thing we can do is compare this to [last week](#)'s survey results on mandates, which found that 62% favored the maximally coercive policy of a universal vaccine requirement. That would roughly be Groups A+B, which is 55%, and some of Group C1, then only C2 and some of C1 are vaccinated but don't then favor a full mandate on others.

There aren't that many people in the daylight between 'willing to get vaccinated myself' and, well, [this](#):



Matt Tabak @TabakLive · Aug 6

At this point, the only things stopping me from hanging out downtown with a vaccine blowgun are lack of access to a large quantity of vaccines, lack of possession of a blowgun, and an unwillingness to go into the city.

I'd like to take one more stab at addressing various arguments against vaccine mandates, explain my perspective on it, and then move on to what's actually happening.

There are a bunch of different arguments against vaccine mandates, and it seems the only one that convinces a lot of people is the argument that (A) one shouldn't take the vaccine, presumably (I'd hope, anyway) with the implication that the vaccine also shouldn't be mandatory.

(B) would be that bodily autonomy is super important, and thus vaccinations are in a different class from things like masks and lockdowns, and either (B1) allowing this is an escalation of authoritarianism, and will greatly expand authoritarian power and perception and lead to tyranny and/or (B2) this is going to piss a lot of people off, destroy their dignity, is generally bad on object level, and so on. And the claim that (B3) an employer mandate is authoritarianism and against freedom, as opposed to all the other things jobs require, and thus we should suspend freedom of contract or punish those who use it in this way.

So basically not buying into [arguments like this](#):



Michael Harriot @michaelharriot · Aug 8

...

I've been told by multiple people that cops are now arresting drivers who chose not to get a government-issued "driving passport." Others were ticketed for exercising their "freedoms" by not following the tyrannical "seatbelt mandate."

I thought this was a free country?

There's a lot of vocal support for both of these, both in my comments and in general. I'm sympathetic to these types of arguments, but not convinced, because mandating vaccinations to stop infectious disease has correctly been standard procedure for a while, and such precautions as a condition of employment or close physical proximity to many others are exactly how free people react to such situations. And if the choice is between 'no indoor dining (or other X) for anyone' and 'no indoor dining (or other X) for the unvaccinated' I know which one I'm choosing, and which one leaves me more free.

There was a lot of talk earlier worrying about tracking, but we've moved past vaccine passports and people just flash their cards, so that's not a concern, and also a common alternative is 'contact tracing' which involves keeping a lot of records of exactly the type we'd worry about. Instead the concern here is the flip side of that, which is that (C) vaccine cards are easy to fake. That's true, but also there's a central database that knows the answer that already exists. I'd rather not push people to lie and commit fraud, but this doesn't seem like that big a concern here.

A real concern is that (D) they got this one right, vaccines happen to be safe and effective, but it's not clear that we wouldn't be in a similar position in the future where the thing in question wasn't safe and/or wasn't effective. In this case, I actually *don't* think this is true. I think both that the vaccines are safe and effective based on the evidence, and also that if the evidence did not strongly say they were safe and effective, we wouldn't be contemplating such policies. The level of pushback we have now is when, scientifically, the case is overwhelming, and if the vaccines were instead not safe but still much safer than not getting vaccinated, we'd not only not make them mandatory, they'd be forbidden. We ran that experiment.

Finally, there's (E) that more vaccinations won't change the path of the pandemic at this point, so why are we bothering, choices have consequences, the unvaccinated will get sick but the vaccinated will mostly be fine, so all we'll be out are a few medical bills (which will effectively get socialized, because insurance can't discriminate and if you don't have insurance mostly the government pays). I don't buy this, there are a bunch of immunocompromised people, even for vaccinated people getting Covid is worth avoiding, and in practice if there's lots of Covid out there the result will be lots of lost living of life.

Another disagreement I've found is (F) the idea that an individual is only responsible (in various senses) for either literally themselves, or only those who that individual infects directly, [rather than the marginal cost being every infection that results from your actions that wouldn't have otherwise happened](#) (there's a control system, and some people would have gotten infected later anyway).

Finally, there's (G) that the [mandates are covering people who already had Covid and thus don't require vaccination](#). There's the counter argument that vaccines are more effective than immunity from prior infection, but antibody tests could check and even worst case I doubt prior infection is that much less effective than J&J which we consider to fully count after one shot despite it not making much logical sense. I don't find that convincing. I also don't think [the evidence that vaccination after infection is still worthwhile](#) makes this sufficiently effective to convince me either, although it's still a substantial additional

reduction and I personally would still get the shot. What I do find convincing is that lots of people are wrong about whether they've had Covid, or would fool themselves or lie about it, and thus there's no reasonable way to make this (otherwise correct) exception short of getting a confirmed antibody test, and the complexity costs and messaging and such involved make it not worth it. Sometimes you gotta suck up stupid things in the name of simplicity, but of course if anyone does want to make such exceptions that seems totally fine.

The high correlation of positions on all these points is, of course, both expected and suspicious, on all sides including my own.

[Other times, incentives matter](#), and people enjoy using hyperbolic language. I don't really know what he was expecting.



Satoshi Smith @me_think_free · 1d
I got my vaccine. Against my absolute morality. I couldn't afford to lose my job. Pray for me and my family. This was very hard. This was degrading to the core. An absolute abomination of my freedom of which I will never forget.

1,370

2,119

12.8K



Satoshi Smith
@me_think_free

...

...

We're fine. Not as bad as I thought it was going to be.

4:27 PM · 8/9/21 · Twitter for iPhone

One cost of lack of vaccination is putting more strain on the hospital system, [as they once again are forced to cancel elective surgeries](#), thus reallocating medical care from those who need it because life to those who need it because they are unvaccinated. The cost here is not purely monetary.

In conclusion, I'm strongly in favor of employer mandates, and on imposing the kinds of restrictions we'd otherwise impose on everyone (e.g. travel, indoor dining and so on) only on the unvaccinated, although of course we should be smart about it, and I'm happy that so far no one I've seen is suggesting excluding the unvaccinated from beaches or playgrounds.

I'm definitely in favor of letting gyms mandate vaccinations rather than masks, [as opposed to being required by law to mandate masks instead as they are in Washington DC](#).

For now, [lack of FDA approval is holding many mandates back](#), even as increasingly many go forward anyway.



Richard M. Nixon @dick_nixon · Aug 7

...

Crack your knuckles, because come September there will be vaccine mandates everywhere. They are waiting for the FDA. It's far too long, of course. The genie is long out of the bottle. But it's the only thing and Biden's political life depends on it besides.

If I knew the FDA was going to get this done in a few weeks, I'd be inclined to announce new restrictions but make them conditional on full FDA approval, so as to benefit from the cover that will provide and give them that much more of a nudge to hurry. Alas, while I hold out hope for it, I currently have no faith in that timeline, and we don't need them to tell us what we already know.

I now consider that out of the way, and won't be discussing it further unless something changes. So, what are this week's new mandates?

[Here's an interesting local one](#) that turns out not to be the full corporate policy, and a reminder that this is how attention works on the internet, and it looks like about one out of every *thousand* people who saw the original post saw the clarification that it wasn't the whole chain.



Rise of the Alien Queen @rise_alien · Aug 4

...

My son works at Lowes in MO. Their new Covid policy:

1. If you are vaccinated and get Covid, you will be paid during sick leave
2. If you aren't vaccinated and get Covid, you will be fired.

My son said that the rest of the unvaccinated are finally getting vaccinated.

2.2K

20K

122.6K



Rise of the Alien Queen @rise_alien · Aug 6

...

Footnote: Lowe's has denied this is their policy. And all I can say to that, is my son's manager is one brave motherfucker to have tackled this at his store independently. Bravo!

23

20

377



[CNN's mandate showed it has teeth](#). Which is how it has to be. Once you have a mandate, for many reasons it needs to be properly enforced.



BNO Newsroom ✅ @BNODesk · Aug 5

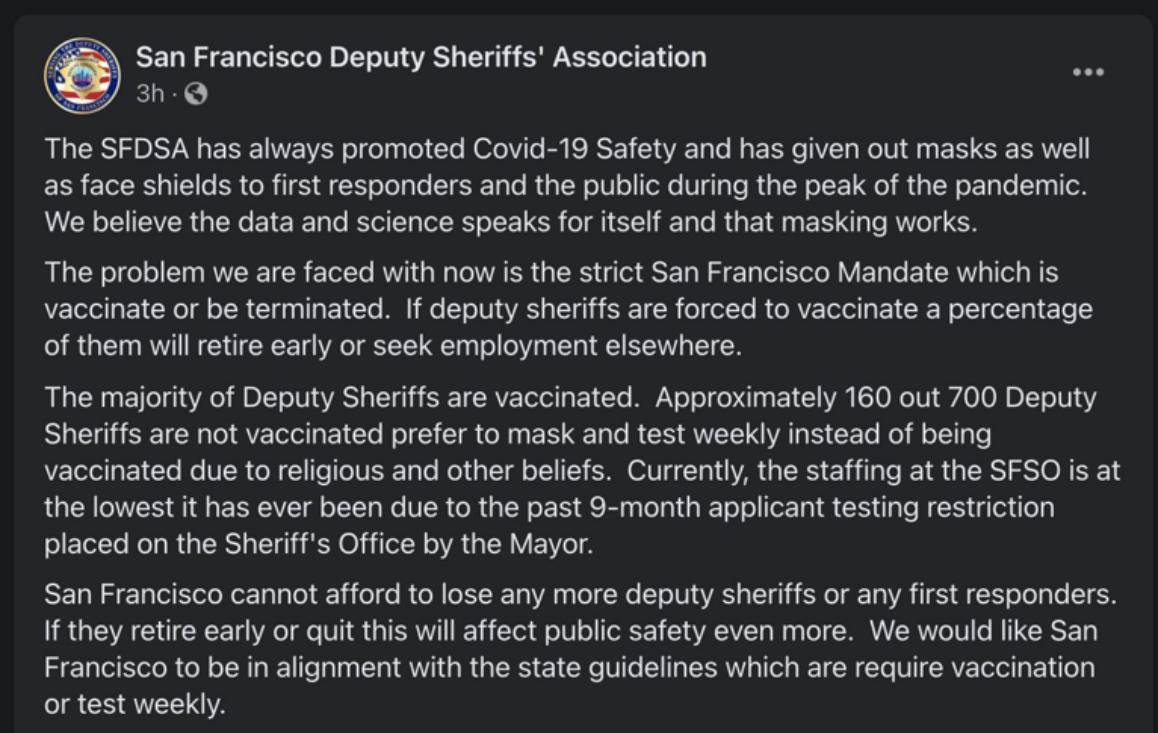
...

CNN President Jeff Zucker: "We have been made aware of three employees who were coming to the office unvaccinated. All three have been terminated."

Meanwhile, the question as always with children is, should we forbid vaccinations as horribly unsafe, [or should we stop doing that](#) and turn on a dime to mandate them outright, with not vaccinating as horribly dangerous?

Yelp is doing something interesting, which is that [you can search restaurants by 'All Staff Fully Vaccinated' and by 'Proof of Vaccination Required.'](#) I verified this, while also noticing that Yelp's rating sorting algorithm leaves something to be desired (e.g. having a 5-star average on 2 reviews seemingly puts you in the top 10 for all of NYC, whoops). This is sufficiently cool that I'm going to play around a bit more with Yelp, since I'm working on my list of places to go in NYC once I get back anyway.

The San Francisco Sheriff's Deputy warns that when the full vaccine mandate is imposed, [officers will quit en masse](#), whereas they wouldn't if they only had to have a swab up their nose every week instead like the state suggests:



The screenshot shows a Facebook post from the "San Francisco Deputy Sheriffs' Association" page. The post was made 3 hours ago and includes a profile picture of a sheriff's badge. The text of the post is as follows:

The SFDSA has always promoted Covid-19 Safety and has given out masks as well as face shields to first responders and the public during the peak of the pandemic. We believe the data and science speaks for itself and that masking works.

The problem we are faced with now is the strict San Francisco Mandate which is vaccinate or be terminated. If deputy sheriffs are forced to vaccinate a percentage of them will retire early or seek employment elsewhere.

The majority of Deputy Sheriffs are vaccinated. Approximately 160 out 700 Deputy Sheriffs are not vaccinated prefer to mask and test weekly instead of being vaccinated due to religious and other beliefs. Currently, the staffing at the SFSO is at the lowest it has ever been due to the past 9-month applicant testing restriction placed on the Sheriff's Office by the Mayor.

San Francisco cannot afford to lose any more deputy sheriffs or any first responders. If they retire early or quit this will affect public safety even more. We would like San Francisco to be in alignment with the state guidelines which are require vaccination or test weekly.

I wonder if this is a place where approximately zero people would be sad to see many of those 160 deputies go. The blue tribe locals are anti-police and will see these as bad actors and even outgroup members, and be happy to see them go. The red tribe will see this as the latest talking point in what they see as SF's descent into crime and anarchy. The rest of us find out how many officers actually quit, which will be great data. Everybody wins?

In contrast to NYC's teachers, the head of the American Federation of Teachers [came out strongly in favor of a mandate](#).



Meet the Press @MeetThePress · Aug 8

...

NEW: President of the American Federation of Teachers [@rweingarten](#) calls for vaccine mandates for teachers.

"As a matter of personal conscience we need to be working with employers on vaccine mandates."

222

2.3K

7.8K



Meet the Press @MeetThePress · Aug 8

...

"The circumstances have changed. ... It weighs really heavily on me that kids under 12 can't get vaccinated."

"I felt the need ... to stand up and say this as a matter of personal conscience." -- Teacher union head [@rweingarten](#)

[France implements its Health Pass requirements with little fanfare, despite weeks of protests \(WaPo\)](#). Sounds right.

Think of the Children

[The request here](#) comes from the American Academy of Pediatrics, representing 67,000 physicians.

We understand that the FDA has recently worked with Pfizer and Moderna to double the number of children ages 5–11 years included in clinical trials of their COVID-19 vaccines. While we appreciate this prudent step to gather more safety data, we urge FDA to carefully consider the impact of this decision on the timeline for authorizing a vaccine for this age group. In our view, the rise of the Delta variant changes the risk–benefit analysis for authorizing vaccines in children. The FDA should strongly consider authorizing these vaccines for children ages 5–11 years based on data from the initial enrolled cohort, which are already available, while continuing to follow safety data from the expanded cohort in the post–market setting. This approach would not slow down the time to authorization of these critically needed vaccines in the 5–11–year age group.

In addition, as FDA continues to evaluate clinical trial requirements for children under 5 years, we similarly urge FDA to carefully consider the impact of its regulatory decisions on further delays in the availability of vaccines for this age group. Based on scientific data currently available on COVID-19 vaccines, as well as on 70 years of vaccinology knowledge in the pediatric population, the Academy believes that clinical trials in these children can be safely conducted with a 2-month safety follow-up for participants. Assuming that the 2-month safety data does not raise any new safety concerns and that immunogenicity data are supportive of use, we believe that this is sufficient for authorization in this and any other age group. Waiting on a 6-month follow-up will significantly hinder the ability to reduce the spread of the hyper infectious COVID-19 Delta variant among this age group, since it would add 4 additional months before an authorization decision can be considered. Based on the evidence from the over 340 million doses of COVID-19 doses administered to adults and adolescents aged 12–17, as well as among adults 18 and older, there is no biological plausibility for serious adverse immunological or inflammatory events to occur more than two months after COVID-19 vaccine administration.

American Academy
of Pediatrics



DEDICATED TO THE HEALTH OF ALL CHILDREN™

[As Alex Tabarrok puts it:](#)

In my many years of writing about the FDA, I can't recall a single instance in which a major medical organization told the FDA to use a smaller trial and speed up the process because FDA delay was endangering the safety of their patients. Wow.

The invisible graveyard is invisible no more.

It is quite the rebuke. Well said. When concerned physicians tell you to stop demanding so much child safety data and [Get On With It](#), that is the opposite of the mistake they are most likely to make, and thus this is strong evidence that a lot more Getting On With It is urgently needed.

[This thread points out the obvious](#), which is that anything picked up in the large sample, that wouldn't have been picked up by the small sample, *wouldn't be big enough to make the vaccine not worth taking*. Thus, the bigger sample is *actively worse*, because it is capable of finding rare effects that would scare either regular people into declining or scare the FDA into not approving, and such decisions would almost always be mistakes.

When tackling the question of schools, if you are going to take the position that children are at sufficient risk from Covid that they can't be put into rooms together with one adult, one should notice it's strange to also forbid them from being vaccinated, and it's heartening to find advocates that [at least realizes that much](#) (and of course wants to go directly from forbidden to mandatory, as I am confident we will do for children).



Jorge A. Caballero, MD @DataDrivenMD · 8h

•• NEW: Given what we know about #DeltaVariant and long Covid, it is immoral to send unvaccinated kids, teachers, and staff into classrooms. Here's a summary of the best data from around the world, and a strong case for vaccination and mask mandates, now



Jorge A. Caballero, MD @DataDrivenMD · 4h

"In our view, the rise of the Delta variant changes the risk-benefit analysis for authorizing vaccines in children. The FDA should strongly consider authorizing these vaccines for children ages 5-11 years based on data [that is already available]"

So at least there's that. Of course, this is in the world where even fully vaccinated children can't safely be put into the same room without masks.



Jorge A. Caballero, MD @DataDrivenMD · 7h

We know what we need to do: masks *and* vaccinations *and* improve indoor air quality.

There are ways to bridge the gap for kids under 12, which I discuss in the essay.

I'm also happy to see air quality mentioned as a key issue (as a reminder, air quality improvement in schools [would be urgently necessary and worthwhile even if it didn't matter for Covid](#) or the moment to moment experience of breathing the air). The core of the argument, later in the thread, is the risk of Long Covid. You have to raise the specter of Long Covid when talking about children, since the risk of anything else is clearly not worth worrying about even without vaccinations, and here the proposal is to worry about it even after vaccinations.

As usual, the procedure in this thread was to gather together every possible symptom, of any severity, and any duration longer than a few weeks, that happens after someone has Covid (and that may or may not actually have anything to do with Covid) and count them all

as Long Covid together, with no attempt to quantify what it means for someone if they get it. Also without any practical plan for how long proposals to avoid it might last or under what conditions they would be willing to stop using them.

The New York Times was also its usual self and did its best scare piece on Long Covid in children, but it's only one of a chorus of such claims. That doesn't mean Long Covid isn't real, it's clearly a thing and the primary risk factor for younger people, but it must be kept in perspective. [Here's a thread pushing back, and the related post from Gaffney:](#)



Adam W Gaffney ✅ @awgaffney · Aug 8

...

I support dramatic public health interventions in hard hit areas to slow Covid (and buy time to vaccinate everyone possible) but I strongly dislike this irresponsible mode of reporting which will scare the shit out of every parent with a kid with a cold.



Adam W Gaffney ✅ @awgaffney · Aug 8

...

The hypothesis that a mild upper respiratory tract infection - even if caused by a new virus - causes a chronic dementing illness in children is a sweeping, massive, frightening claim. It is unlikely to my mind, given what we know about mild upper respiratory tract infections ...

20

41

467



Adam W Gaffney ✅ @awgaffney · Aug 8

...

... but like all things, it is possible. Yet it is an extraordinary claim, and requires robust evidence which is thus far lacking. Our mantra that correlation does not equal causation goes out the window on this particular topic.

It could be worse, [and usually is](#). The standard line is that Delta is even more dangerous for children because they have a higher percentage of infections, hospitalizations and deaths than they did previously. Which is absolutely true, [and can be explained by the fact that they're mostly not vaccinated](#).



Wojtek Kopczuk @wwwojtekk · Aug 6

...

Press is statistically illiterate, a new exhibit.

The only piece of data (other than anecdotes) in the article is that kids are a bigger share of cases than before, which of course they are because they are not vaccinated when the rest of the society is to a pretty large extent

Evidence mounts that delta variant is dangerous for kids

Children are rapidly getting sick from the delta variant of COVID-19.

From the beginning of the pandemic, children made up 14.3 percent of all cases. Now, children with COVID-19 represent 19 percent in weekly reported cases from July 15 to July 29.

[From post in question](#), ya don't say, might I suggest something we might do about that?

- It is unclear why more children are sick due to the delta variant outbreak, but medical experts believe the surges are because of how easy it is for the virus to circulate in an unvaccinated population.

[This CNN post is similar and typical](#). It cites increases in cases in children similar to increases in cases overall, then has to explain why anyone should care. Its first reason is that this is critical to *keeping them in school*, because if we don't protect kids from Covid then we'll have to take them out to protect kids from Covid, and then we'll be forced to detain them at home instead. The second justification is that the kids *might spawn new variants*, which is technically true but seriously, come on. Then they hold up the specter of MIS-C, with a total of 4,196 cases, which is at least a specific issue rather than generic Long Covid, but again, math.

Meanwhile, if you took Remote Schooling, treated it and its side effects as a pandemic, and ask what would happen if it was spreading across the country, I think the answer is full Australia-style stay-in-your-house lockdowns as needed.

I do get that this is now a strange position to be in, and if you're deciding on school policies independent of the FDA, you're in the same position you'd be in if this was a wait on manufacturing and distribution instead of regulatory approval – the kids *will* be vaccinated, but can't be now, and the fact that *other people* could change that if they wanted to does

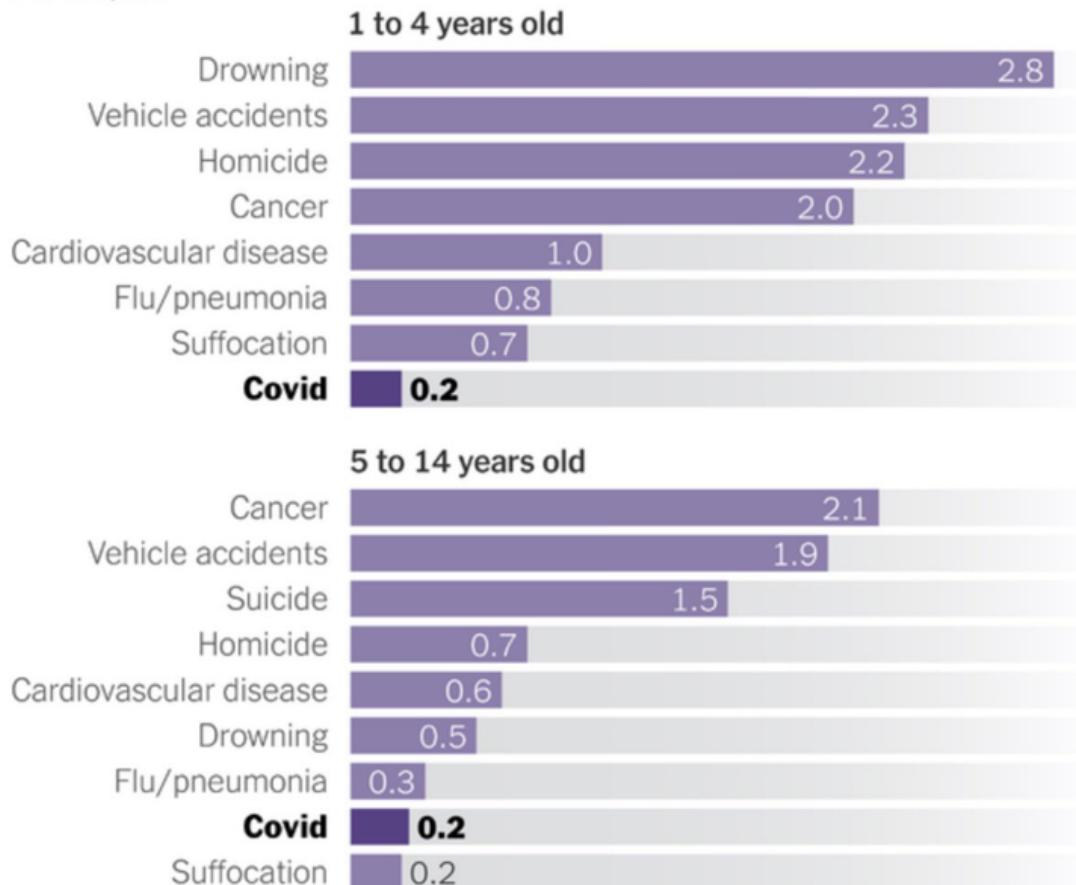
not give you that option. And if there are enough cases you *will* be forced to go to remote learning no matter what you know about its consequences.

[Tyler Cowen asks how many children are killed by school anyway.](#)

[As for how many are killed by Covid-19.](#)

Annual Deaths Among Children in the United States

Per 100,000



Covid data is for the 52 weeks ending April 10, 2021. Other data is for 2018.

By The New York Times | Source: Centers for Disease Control and Prevention

That could maybe change by an order of magnitude before it's all over, but that's an upper bound.

[From an excellent post](#) about the question of what to do for kids who it is illegal to vaccinate, here's a chart of what happens *if they do get Covid.*

(I know you wish this was separated into smaller age groups; me too. Data is just not there for that. And all of these numbers are very rough and evolving.)

COVID-19 Risks, Children Aged 0 - 18

(Risks Among Children with COVID-19)

Any long term (>8 week) Symptoms	~ 1 in 50
Hospitalization Risk	~ 1 in 200
Death Risk	Wide range: ~1 in 80,000 to ~1 in 20,000

Comparison Risks: Flu & RSV

(Risks Among Children with Flu or RSV Infection)

Flu ages 0-17, Hospitalization	~1 in 240
Flu ages 0-17, Death	~1 in 20,000
RSV, under 6 months, Hospitalization	~ 1 in 50
Yearly Risk of Motor Vehicle Death, Age 0 -15	~ 1 in 50,000

That death risk is consistent with what I had, and for those under 12 it will be lower still. I'm skeptical of the 1 in 50 line, but the word 'any' does strong work in such situations, so... maybe?

Long Covid is real and important, but so is Long School. *Most* people I know are *permanently* traumatized by it, many people have nightmares about it, and so on. [Suicide attempts drop dramatically for young children when school is out, in a way that suicides of older people don't.](#) It doesn't end when you're allowed to leave. There's also Long Vehicle Accidents, Long Suicide Attempts, Long Cancer, Long Drowning and so on, as one might expect.

We do the same thing with school shootings, where we force kids to take place in 'live shooter drills' and scare kids so much that they *expect* a school shooting to happen around them, whereas [such things are exceedingly rare](#), schools don't have more of them than the rest of life and the damage done by scaring everyone is orders of magnitude the bigger concern (and is probably doing far more to give kids ideas and *causing* such events than it is to prevent anything, if I had to guess).

One could also compare the moral panics over 'stranger danger' to the fact that most kidnappings of children are by family members, and most abuse is by people they know well.

The problem, in all these cases, is that some threats are put in a special category where any bad events are unacceptable, whereas other things are part of life, such as getting lots of people into close proximity in this thing we call a 'city.'

[Kids are not in a good place right now.](#) We've disrupted their lives and kept them socially isolated for over a year. Making their new school year largely about Covid, and forcing them behind masks, will make things that much worse.

[I am thrilled to see that lots of parents intend to home school their children for the coming school year.](#) In-person school is bad enough, but choosing remote learning over home school is a true tragedy, either of misunderstanding the situation and/or a lack of sufficient resources to deal with the necessary obstructions, paperwork and actual educational efforts required. My heart goes out to anyone who knows and simply can't do it.

At some point we will need to learn to live with Covid, or make an extraordinary effort to somehow live without it by vaccinating everyone and then moving on. Or we could doom ourselves to a young adult dystopia in yet *another* way, the same way kids are no longer allowed to play outside and are told not to talk to strangers and pretend periodically to hide from gunmen roaming the building, plus whatever you consider a baseline 'school'. That's also an option.

Mask and Testing Mandates

There's an old improv game called [standing, sitting, bending](#). May I present a new one: Eating, drinking, dancing:



Rory Meakin @rorymeakin · Aug 6

"Any person who fails to wear a facemask who is not at the time:
a) moving rhythmically to music,
b) drinking a liquid, or
c) eating a solid
is guilty of an offence under these regulations."

...



BBC Scotland News @BBCScotlandNews · Aug 6

Face coverings can be removed for the three Ds - dancing, drinking and dining
bbc.in/2WWxdU5

Ministry of Truth

Even when the actual implementation in a given example seems fine, it's important to focus on the reasoning, for this tells you what the ministry is looking to do next.

Facebook's War on Supposed Misinformation continues, and has [produced the following 'fact check' of 'misinformation.'](#) I wouldn't be focusing on Something Wrong On The Internet, except [this type of 'fact checking' from this exact source is being used to censor Facebook and Instagram](#). Although in this case, the post wasn't censored, merely given a warning label (and one assumes also a massive [Streisand Effect](#)), that's not always the case.

Why the COVID-19 survival rate is not over 99%

Most people who contract COVID-19 will not die, though an individual should not weigh their own chances of death by looking at national statistics.

The Covid survival rate is *clearly* over 99%, by the CDC's own estimates. The CFR is 1.7% and the CDC conservatively estimates half of infections have been missed – I'm guessing there are at least twice as many as that, perhaps more.

Saying this is known to not have a CFR under 1%, as your headline that is then quoted around, or that survivability is known to not be over 99%, is blatant lying and scaremongering.

What's this all about?

With COVID-19 infections surging in the United States because of the more contagious delta variant, some have downplayed the number of deaths from the virus and the effectiveness of vaccines.

To minimize the importance of vaccination, an Instagram [post](#) claimed that the COVID-19 survival rate is over 99% for most age groups, while the COVID-19 vaccine's effectiveness was 94%.

The post's alleged survival rate for COVID-19:

- 0 – 19 years, 99.997%
- 20 – 49 years, 99.98%
- 50 – 69 years, 99.5%
- 70+ years, 94.6%

The post was flagged as part of Facebook's efforts to combat false news and misinformation on its News Feed. (Read more about our [partnership with Facebook](#).)

A problem with the post is that it improperly used the U.S. Centers for Disease Control and Prevention's statistics for modeling pandemic scenarios, not for calculating COVID-19's survival rate.

The CDC [recommends](#) the COVID-19 vaccines because they are safe and effective, [even against the delta variant](#). Although the delta variant has slightly decreased the effectiveness of vaccines, experts still encourage vaccination as it provides a high level of protection against hospitalization and death.

It's about the new definition of misinformation, which as far as I can tell is *information used to lead to a conclusion we don't like*.

The fact check admits that the data comes from CDC modeling estimates, and then uses those best guesses as best guesses. But because you can't *prove* those are the correct

numbers, and the conclusion is one they don't like, the 'fact checker' thus concludes the claim is 'false.'

What's funny is that the exact claim being evaluated *I too think is actually false*. I'll get to that later. But that's not a fact that's correlated with any of their reasoning.

Our ruling

An Instagram post claimed that the COVID-19 survival rate is over 99% for most age groups.

The data it cited does not show the likelihood of surviving COVID-19. The post's claim is based on data used to model pandemic scenarios. Experts say a person cannot determine their own chances at surviving COVID-19 by looking at national statistics, because the data doesn't take into account the person's own risks and COVID-19 deaths are believed to be undercounted. Survival rate data is not yet available from the CDC.

We rate this claim False.

There's so many *different* things dangerously wrong here.

1. Unproven or unknown does not equal false, and by a sufficiently strong standard we "know" almost nothing, paging various philosophers. And by this standard, since I can choose not to offer proof, I can get them to say (almost?) anything is false, and thus by flipping the sign say almost anything is true, provided it serves their purposes. Neat trick.
2. Using the CDC's numbers from their modeling is an *excellent* source of reasonable approximations. It comes (as per the fact check post!) directly from the CDC and is being used to predict things, so it's a forward-looking estimation. When I *disagree* with such assessments because I think I know better than the CDC, which I do here, that's because I'm the arrogant one who thinks he's better than the CDC, not the other way around.
3. The general survival rate for Covid is clearly over 99% as discussed above via the CDC's own estimate of the true case count. That doesn't then automatically extend to 'most age groups' which is why I end up thinking the claim as *categorized by the fact checker* is false, but that's not how the original post categorized anything, so the actual disagreement is over exact numbers for particular age groups.

4. Evidence that isn't of the correct form or from the correct source (even within exactly the correct overall source, the CDC) is being selectively dismissed when it doesn't suit them, which makes it easy to find a 'lack of sufficient evidence'

Etzioni said that it's not useful to just look at the rate that people die, even if it's low, because it doesn't tell the whole story. "If more and more younger people are getting COVID, then the total number of young people who die is going to skyrocket," Etzioni said.

Also, people should not use data on how many people have survived COVID-19 to predict their own chances of surviving infection, experts say. Someone's chances of surviving COVID-19 can vary depending on their age, health, and vaccination status — national statistics don't account for these factors.

o find something to be false.

5. They do not cite what numbers they do believe, or any evidence for or against any numbers whatsoever except a general FUD about believing any numbers at all.
6. Up front they are clear *why* they are doing this – it's because the claim is being made *in order to minimize the importance of vaccination*. The fact is a soldier for the wrong side, ergo false.
7. Again, in their conclusion, they're judging *their characterization of the central claim* as false, rather than disagreeing with any particular claim or giving an alternative model.

Here's the argument that if you have the best data available, you should ignore it, because there's some factors it didn't account for, and thus you should throw out all numbers and have no idea whatsoever. Which is a fully general argument against ever knowing anything at all:

This is not how knowledge works, unless you are banning forbidden knowledge due to its Unfortunate Implications. Yes, *of course* you should use data on how many people have survived Covid-19 to predict your own chances of surviving infection. What the hell else would you use as a starting point? And the *whole idea* here is to then condition that on age, which is by far the biggest risk factor, and then condition on vaccination status (where I think their 94% number is somewhat low, but it's well within the range of Numbers Used By Official Sources To Scare People The Proper Way Depending on Context, and also not a crazy estimate, I just think defense against death is somewhat higher.)

Whereas the post is indeed taking overall estimated forward-looking numbers, then adjusting them by age, and also listing vaccine effectiveness. If one wanted completeness yes there are other factors but they're far less important – see my graph below for my ranking of the next two in line (diabetes and obesity), and how much less important they are than age.

The advantage of telling people to throw up their hands is that you can simultaneously say vaccines are super effective and important (without specifying numbers) when telling people to get vaccinated, then turn around and tell them to be terrified of Covid afterwards anyway, even if they're young.

Yeah, people are skeptical of authorities these days for some reason, can't imagine why.

Next, I'm going to *actually* fact check the chart, since I think my estimates are better than the CDC's estimates. Are the ratios by age here correct?

I actually think no, they're too aggressive. [Here was the result of my comorbidity work](#), which was pre-Delta and pre-vaccinations, and have younger people at more risk than this by an order of magnitude or so.

Reminder: This assumes overall infection fatality rate of ~1% and is full of guesses and approximations as all hell:

Age	Risk (Healthy)	Risk (Diabetes)	Risk (Obesity)	Risk (All Pop)
0-4	0.001%	0.013%	0.005%	0.002%
5-14	0.002%	0.016%	0.006%	0.003%
15-29	0.008%	0.078%	0.031%	0.015%
30-39	0.043%	0.26%	0.17%	0.09%
40-49	0.13%	0.53%	0.43%	0.26%
50-59	0.72%	1.8%	1.4%	1.2%
60-69	2.0%	4.0%	3.0%	2.9%
70-79	4.3%	6.5%	5.4%	5.2%
80-89	11%	11%	11%	11%
90+	15%	15%	15%	15%

To compare apples to apples you should look at the All Pop column on the right, and focus on ratios between groups, and also remember that there aren't many people over 90 when combining the top group together. With that adjustment, my conclusion is that the post above is underestimating risk to the young by about a factor of 10.

Now, let's look at the actual post, and, huh, ok, I see it now...

Covid Vaccine effectiveness:

94% 



My immune system's effectiveness:

99.98% 



 Missing Context. Independent fact-checkers say information in this post could mislead people. >

(Note: That font and color scheme in the graphic is very recognizable as coming from Fox News, the outgroup's relatively mainstream news source.)

Yeah, that's... a very reasonable warning. This is indeed missing context and could mislead people, and the warning isn't claiming it's *false*, merely that it's missing context. In particular, this is framed carefully to imply that the vaccine would replace the existing immune system rather than supplement it, and thus the vaccine would *increase* risk rather than decrease relative risk.

So in the context of the post, the label is at least understandable. It sets a bad precedent even if the written justifications for it had been relatively good, so I'd rather not do it, but certainly one can understand it, especially when combined with the numbers here being so aggressive, although the extra 9 likely doesn't change the message here much.

As opposed to the reasoning in the justification post, which is... different.

I'll also note that clicking on the warning doesn't go anywhere when I tried it, which seems like a missed opportunity if one did want to communicate context.

More generally, the official reasoning remains [the supremely broad claim that any disagreement with health authorities is not allowed](#). As we are periodically reminded, this is despite the health authorities changing their opinions over time as (A) the facts change, (B) we get better evidence and (C) they update to take into account new information and incentives and priorities. *Usually* their truth tracking improves over time on a given issue (and stays the same on average because they add new issues), but not always. Also you can't contradict multiple health authorities, including the WHO who still refuse to admit Covid is airborne, and those different authorities frequently contradict each other. In the case above, the CDC's numbers can't be used in a way that wasn't intended. By the standards that are being used to censor a United States senator who is raising a perfectly valid scientific hypothesis in the link earlier in this paragraph, you could censor not only at least most of my Covid posts, but almost anything remotely useful anyone might say, whether they were trying to provide useful information or trying to figure things out. If I wasn't [Against Facebook](#) I'd probably be banned from it by now, and I'm curious to what extent they mess with those who post these weekly updates there.

Provincetown Follow-Up

[Nate Silver offers some thoughts](#) on exactly how hard it is to compare vaccinated to unvaccinated people, even in relatively ideal conditions for such comparisons.



Nate Silver @NateSilver538 · 18h

...

Just off the cuff, here are 10 things to think about when you read competing claims about vaccine effectiveness. Several of these can plausibly change the numbers quite a lot, which is why it's not surprising that there's disagreement from study to study.

1. Unvaccinated people belong to different demographic groups than vaccinated people.
2. Unvaccinated people may take fewer or more precautions than vaccinated people. (Most evidence suggests fewer.)
3. Most studies rely on people who choose to be tested, which likely means they have symptoms or are in contact with others who have COVID. Studies that rely on random or surveillance testing may be much more robust -- but there are fewer of these.
4. Furthermore, unvaccinated people may be more or less likely to seek out testing than vaccinated people, holding other factors constant. (My guess is less likely.)
5. Many unvaccinated people had COVID at some point, and so have antibodies / some degree of natural immunity. This was not the case in most clinical trials, when the placebo group was presumed never to have had COVID. (Also, some vaccinated people *also* had COVID at some point, which may boost immunity.)
6. Results from different vaccines are often combined, when there is some evidence that they may offer different levels of protection.
7. Results from people who have different numbers of doses (one, two or in some countries three) are often bucketed together differently.
8. If waning immunity is an issue, then studies may need to account for the time at which someone was vaccinated. However, this introduces confounders since most countries vaccinated older people or people with underlying health conditions first.
9. News accounts often don't emphasize whether the study measures effectiveness against *any* infection or *symptomatic* COVID-19. This may make a big difference if breakthrough infections are more likely to be asymptomatic.
10. The threshold (Ct value) under which someone is considered "infected" can vary from study to study.



Nate Silver  @NateSilver538 · 17h

Also:

...

11. Some studies have small samples.
12. If you're interested in Delta, make sure you're not using pre-Delta data.
13. There may be complicated statistical effects if people vary in the extent to which they're protected by the vaccines, see e.g.:



Wes Pegden @WesPegden · 17h

Replies to @NateSilver538

There's an interesting effect where variability in susceptibility can lead to increasing underestimates of vaccine effectiveness beyond what infection levels would seem to suggest, if the most susceptible unvaccinated individuals are infected earliest.

arxiv.org/abs/2009.01354

Provincetown was the opposite of an ideal situation.

It seems very clear at this point that the Provincetown study did not mean anything like what the CDC was representing it to mean. This was a situation filled with activities that carry

extreme Covid risks, among a unique and often immunocompromised population. The outlier results at most describe what happens in circumstances like that, and also fail to control for the population baselines appropriate to that situation. Despite that, [there were zero deaths, only seven hospitalizations, and most vaccinated participants were not infected](#). The vaccines did their job. In hindsight, while I had a strong prior that the study wasn't going to mean what the CDC was claiming, that prior wasn't strong enough, and I gave the whole situation too much respect.

Thus, the focus shifts from the study and its claims to the actions of the CDC and media, and updating on what they did in response to this information. How much of this failure to update was due to people being afraid to point out the nature of the gathering, given today's political climate, until enough others had done so first? Or was the narrative what everyone wanted to go with anyway, so it was too good to check regardless? Or was it that the CDC is no longer capable of reading scientific studies and analyzing data about the physical world in a reasonable way? Perhaps this was all kayfabe at the CDC and they knew exactly what they were doing, and were simply lying, and the media went along with it to show their deference to power and get clicks? Or was it something else?

[Could the CDC actually be this bad at communicating about risk?](#)



James Surowiecki @JamesSurowiecki · Aug 7

...

Setting aside that it's unclear the vaccines could ever "prevent" transmission, I don't understand why Walensky so often speaks in absolutes rather than probabilities. Yes, the vaxxes can't prevent transmission. But they make it less likely you'll catch the virus and transmit it.



The Situation Room @CNNSitRoom · Aug 5

"Our vaccines are working exceptionally well," CDC Director Dr. Rochelle Walensky tells @wolfblitzer. "They continue to work well for Delta, with regard to severe illness and death – they prevent it. But what they can't do anymore is prevent transmission."



James Surowiecki @JamesSurowiecki · Aug 7

...

Walensky is not good at talking about risk. But the fact that the CDC has been no better at communicating on questions of risk/cost-benefit under Biden than it was under Trump suggests that there are systemic problems with the agency.

There's being bad at talking about probability and risk, and then there's treating everything as an absolute.

No matter what the cause, it is yet another reminder that the data all fits together into one physical world that runs with one set of physical laws and biological properties. Something that doesn't fit and contradicts the data and results observed elsewhere must be treated with extreme skepticism, and any model must explain those other results and data points.

Delta Variant

[Vaccines work on it, J&J massive study edition](#), as in n=500k participants:

- ~93% protection from death
- ~71% protection from Delta hospitalizations
- Large study

One J&J shot remains very good protection against death and good protection against infection and hospitalization, but not as good as two mRNA shots or one J&J plus one booster, and the logic of getting the second shot of mRNA is the same as the logic for an mRNA booster after getting J&J. Of course, if you request an mRNA booster after having had J&J, [it might be tricky to get it](#), because the FDA and others are tying themselves up in knots denying the obvious is sufficiently obvious.

This isn't the exact thing we most want, which is how effective the mRNA vaccines are against Delta, but it suggests only a small decline in effectiveness is likely.

[Also, about the way the CDC goes about the business of gathering its data:](#)



Genève Campbell @bergerbell · 17h

Today I learned the CDC (yes, the literal CDC) creates policy based off of un-factchecked data scraped from social media infographics:

For one, the leaked document underestimated the Ro for chickenpox and overestimated the Ro for the delta variant. "The Ro values for delta were preliminary and calculated from data taken from a rather small sample size," a federal official told NPR. The value for the chickenpox (and other Ros in the slideshow) came from a graphic from *The New York Times*, which wasn't completely accurate.

I would like to hold the CDC to higher standards here than I can afford to have writing these posts, but that option is not available at this time.

In Other News

What's the difference between an EUA and full approval? [Hundreds of thousands of pages of paperwork and a bunch of site inspections](#), among other things.

About a week and a half ago, [Scott Alexander wrote a righteous post everyone should read](#) on how horrible the FDA is and in particular how they are way, way too slow to approve drugs and also getting them approved costs eleventy billion dollars each (realistically something like 100 million). One thing that caught everyone's attention was the infant fish oil story, [where the FDA for years let children get sick and die rather than let fish oil get added to an infant formula](#), as he detailed in his first follow-up, [then he wrote a second follow up when a critic pointed out that those involved in that story praised everyone at the FDA](#). Scott points out that yes, the individual people at the FDA did their jobs in this situation, but that *doesn't make it better*, the system was working as designed and the design sickened and killed a bunch of babies and that's the thing to be focused on. If anything that's worse, if it was the people letting us down we could go fix that. He's being careful not to outright say FDA Delenda Est, but as his alternative he's holding them to the impossibly high standard of 'better than the man on the street.'

[I do think this concise statement of the argument goes slightly too far](#), but only slightly.

It's no wonder that [no one wants to be FDA commissioner](#):

aducanumab in Cambridge, Mass. | Biogen via AP

"They are just having a tough time landing on somebody who is not only good but who wants it, and that you won't have trouble getting confirmed," a person familiar with the discussions said. "The people who are really good are not particularly interested in it."

Advertisement

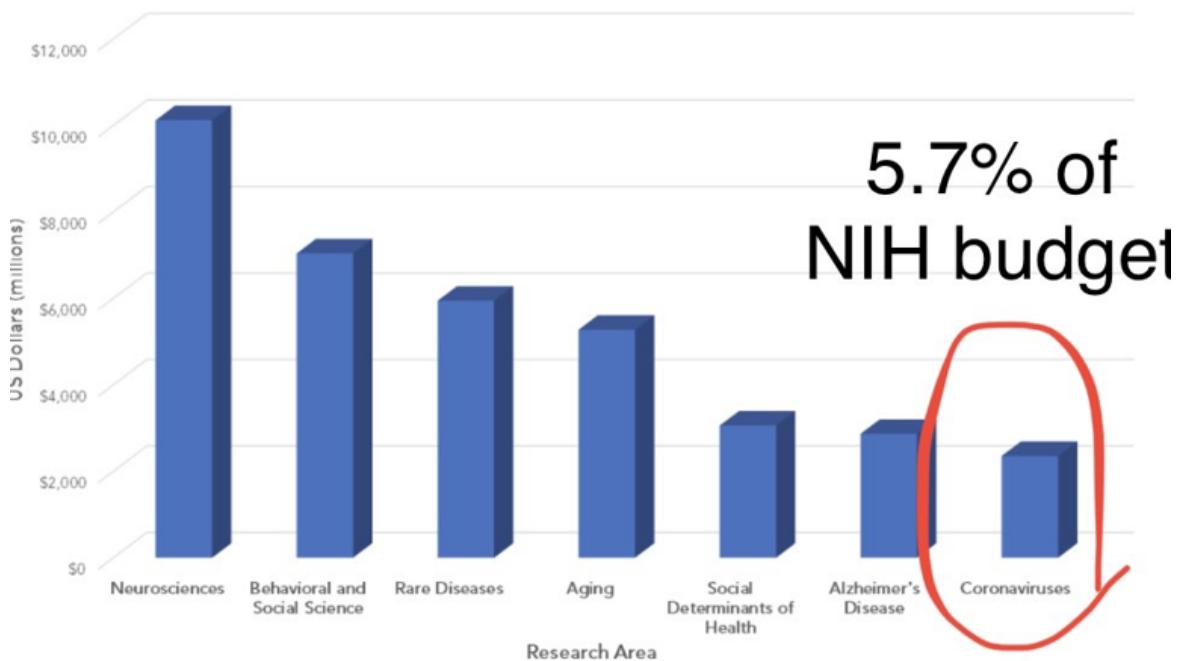
Well, no one who counts, anyway, where one who counts would be someone without trouble getting approved. I certainly get it, in the sense that when I think of my life if I was made Commissioner of the FDA, in terms of my lived experience, oh my would it be infinitely worse. I'd happily do it anyway, because someone has to and if I don't do it then someone else will, and also it would open doors to do additional important things after, but I have to assume it would be a nightmare.

If you're someone who 'you won't have trouble getting approved' then presumably you're looking to run the FDA the way the FDA is traditionally run, which means someone has to but also someone else would if you didn't, and if you're doing it honestly it's a giant paycut, so why take on all that trouble?

[Did the NIH do better \(MR\)? Here's the WSJ, here's the full report](#), here's part of MR's summary, note the top line especially:

- Of the \$42 Billion 2020 NIH annual budget, 5.7% was spent on COVID-19 research
- Public health research was underfunded at 0.4% of the 2020 NIH budget
- Only 1.8% of the 2020 NIH budget was spent on COVID-19 clinical research
- Average COVID-19 NIH funding cycle was 5 months
- Aging was funded 2.2 times more than COVID-19 research
- By May 1, 2020, 3 months into the pandemic, the NIH spent 0.05% annual budget on COVID-19 research
- Of the 1419 grants funded by the NIH:
 - NO grants on kids and masks specifically
 - 58 studies on social determinants of health
 - 57 grants on substance abuse
 - 107 grants on developing COVID-19 medications
 - 43 of the 107 medication grants repurposed existing drugs

Figure 1: NIH Funding by Research Area in 2020



So, not that great, only a handful of billions while missing entirely many of the things we most need studied. The grant process isn't working. In other distributional news, happy to

see Aging get this attention, although it's telling that it's right behind Rare Diseases, even if they're not in cute puppies.

Again, could be worse, you could be the WHO analysts and [still, this week, be telling people Covid isn't airborne](#). Delenda est indeed.

Obama had a birthday party, outdoors, with vaccinations *and* Covid tests required, but didn't require masks, [so naturally a bunch of Justifications are required](#) for this living of life as if physical reality was exactly the way it is.



Dinesh D'Souza @DineshDSouza · Aug 7

...

New pictures are emerging from Obama's birthday bash, and NOBODY is wearing a mask.



'Not a Mask in Sight': Celebs are Pouring Into Martha's Vineyard for Ob...

The mask is off for former President Barack Obama's massive 60th birthday celebration being held on his \$12 million estate in Martha's [...]
 Ⓜ trendingpolitics.com



Glenn Greenwald @ggreenwald · 23h

...

A NYT reporter on CNN justifying Obama's huge maskless birthday bash because he only invited "a sophisticated, vaccinated crowd" is about as emblematic of liberal discourse as it gets.

What happened to all the concerns about vaccinated people passing Delta to the unvaccinated?

What happened was that Obama has a brain and occasionally thinks about physical reality, but to explain this in those terms would destroy the rest of the narrative, so what they say ends up sounding not great.

But seriously, [how's it going out there?](#)



Erick Erickson @EWErickson · Aug 5

...

Talked to a local doctor in Middle Georgia. He said the COVID area of his hospital is full again. Patients are 100% unvaccinated and 100% significantly overweight.

157

196

657



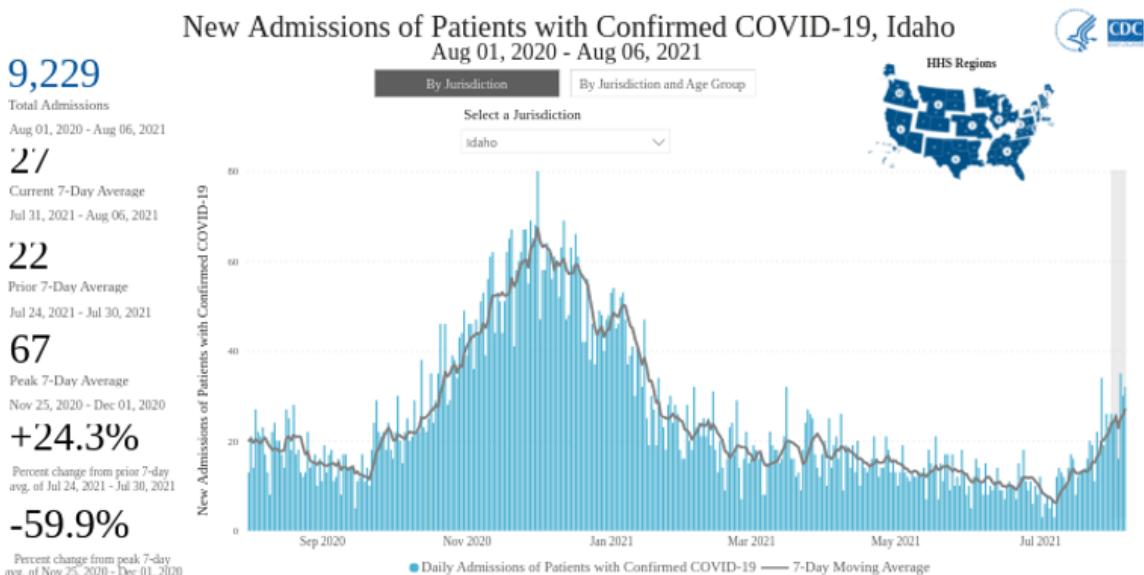
You can, of course, get an anecdote to say almost anything, [for example "Idaho Covid ICU patients are already at an all time high" when the stats say that clearly isn't true](#). Treat local reports with generous helpings of salt before generalizing.



Dr. David Pate
@drpatesblog

...

Idaho is undergoing its fourth surge. Unbelievably, COVID patients in the ICU have already exceeded the first and second surges. Cases are on a sharp rise, and I fear ICU cases could be on a track to exceed even the third surge if we don't modify our behavior.



When cancer survivor Anthony Rizzo was traded from the Cubs to the Yankees, life was proven unfair, and also there were many who noted that he was unvaccinated, which turned out to be unrelated to his cancer – he simply declined the vaccine. A few weeks later, [he's on the Covid injured list](#).

Israeli data seems to show that previous infection is not only highly effective at preventing reinfection, [in their samples it looks even more effective than vaccination](#). This is the opposite of what is found in other reports, but definitely worth keeping an eye on.

Perspective on Louisiana hospitals being full. Looks like this is essentially by design and didn't require that many Covid patients for it to happen.

You're about to spend trillions on 'infrastructure' that is mostly transfer payments to people you like, to be paid for by people you don't like, in the wake of a huge pandemic, and aside from potentially banning large areas of software development by requiring theoretically impossible tax reporting, how are we doing on spending on actual pandemic preparedness?
Oh...



Eliezer Yudkowsky @ESYudk... 11h

Let's be clear on this: Covid-19 was not a disaster. Covid-19 was a warning shot. It says that we need massively scaled up mRNA vaccine factories, that can vaccinate everyone on Earth in 1 month, before a serious pathogen gets released. And it is, of course, being ignored.

Peter Sullivan @PeterSullivan4

NEW: Democrats are considering cutting back on new funds for *pandemic preparedness* in the reconciliation bill to make room for other priorities, sources say, prompting an outcry

Ex-CDC head Tom Frieden calls it "stunning"

[thehill.com/policy/healthc...](http://thehill.com/policy/healthcare)

Show this thread

25 93 522

Not Covid

[I'm putting this here because from time to time, it will be needed.](#)



I finally tried [Storybook Brawl](#) this week, and it is excellent. Highly recommended to anyone who likes playing games of any kind, give it a shot, it's already Tier 1 even though it's in Early Access. I *definitely* have thoughts on it, but we'll see when I get around to writing those down. In the meantime, great fun.

I've also been greatly enjoying [Across the Obelisk](#). This is a unique roguelike deckbuilder, in that it is trying *and largely succeeding* at being like a lightweight D&D rather than being a lightweight Magic: The Gathering. You have a lot more control than in most such games over what your deck looks like, so it's up to you to decide how to keep it fresh after a while, but there's a bunch of viable options and this is great stuff, again even though it's still in Early Access. I'd put it at Tier 2 in its current state. If you're up for what's being described, check it out.

I finally saw a movie, Black Widow, in a theater for the first time since the pandemic began. It would have felt ritually impure to have my first movie back be anything else. By waiting for several weeks, I got a mostly empty theater, so social distancing was excellent. As much as I was looking forward to it, I didn't realize how much I missed the movies until I finally got to go. Excellent experience, can't wait to go again, don't *think* I have time to see Free Guy or Suicide Squad tonight but I'm not ruling it out. My review of the movie Black Widow is: Exactly meets expectations.

In Free Britney news, the system is even worse than you think. There was serious risk of having Britney Spears committed involuntarily, [because of the supposed mental health strain of trying to free herself in court, and the strain of having her father as her legal enslaver against every scream she can muster](#):

In a filing in which the singer asks the court to move up the conservatorship hearing by a month, Montgomery's lawyers reiterated an earlier recommendation to promptly strip Jamie Spears of control over the singer's money and finances, calling it "critical."

"Ms. Montgomery cannot emphasize this enough – the sooner Mr. Spears' suspension can be addressed by the Court, the better for Ms. Spears and her emotional health and well-being," the documents states.

"Having her father Jamie Spears continuing to serve as her Conservator instead of a neutral professional fiduciary is having a serious impact on Ms. Spears' mental health," [Montgomery's attorney said](#). "Notably, Jamie Spears has yet to resign as Ms. Spears' Conservator of the Person, which is why Ms. Montgomery continues to serve as Temporary Conservator of the Person. ... It is in Ms. Spears' best interests that her father step down as her Conservator, so he can go back to just being Ms. Spears' father, and working on a healthy, supportive father-daughter relationship."

In a [previous filing](#), earlier this summer, Montgomery claimed that Spears' own doctors agree that her father should be removed from the conservatorship.

[A judge then refused to expedite the hearing, of course.](#) One cannot rush such proceedings. The FDA would approve.

I'd also like to point out another parallel of horribly inefficient action that got highlighted this week, [which is the War on Bags](#):



Chana @ChanaMessinger · Aug 9

Shoutout to my Tesco delivery guy, who saw me look confused at the crates of loose food with no bags and said "Sorry, Greta Thunberg won"

...

I am reminded of when my teammate Patrick Chapin went to get croissants from a gas station in Belgium (which were really good croissants, Europe has its advantages) and some other food, and was given an obviously horribly inadequate number of terribly flimsy bags. When he offered to pay unreasonably large amounts for additional bags, he was chastised for how little he cared about the planet. Then was forced to spend an hour getting back as the situation fell apart on him multiple times.

You know what costs *vastly* more energy and carbon to produce than paper bags? *Food*. Even a tiny risk of food being wasted is much worse than using extra bags. Yet what happens when bags break? Food containers break open, food is dropped on the ground and made dirty, and both lead to food being thrown out. That's in addition to the hours upon hours of lost time.

And finally, in case you missed it, [too good not to share](#) and also insightful:



Corey Yanofsky, sparkling statistician @Corey_Yanofsky · 5m

sonuvabitch

...



Samantha Ruddy ✅ @samlymatters · 17h

The kids my girlfriend nannies for have a new game called customer service where they sit at this desk and ask what your problem is then tell you they can't solve it and complain about how many emails they have to answer and nobody has any idea where they learned it



Analysis of World Records in Speedrunning [LINKPOST]

[this is a linkpost to [Analysis of World Records in Speedrunning](#)]

TL;DR: I have scraped a database of World Record improvements for fastest videogame completion for several videogames, noted down some observations about the trends of improvement and attempted to model them with some simple regressions. Reach out if you'd be interested in researching this topic!

Key points

- I argue that researching speedrunning can help us understand scientific discovery, AI alignment and extremal distributions. [More](#).
- I've scraped a dataset on world record improvements in videogame speedrunning. It spans 15 games, 22 categories and 1462 runs. [More](#).
- Most world record improvements I studied follow a diminishing returns pattern. Some exhibit successive cascades of improvements, with continuous phases of diminishing returns periodically interrupted by (presumably) sudden discoveries that speed up the rate of progress. [More](#).
- Simple linear extrapolation techniques could not improve on just guessing that the world record will not change in the near future. [More](#).
- Possible next steps include trying different extrapolation techniques, modelling the discontinuities in the data and curating a dataset of World Record improvements in Tool Assisted Speedruns. [More](#).

The script to scrape the data and extrapolate it is available [here](#). A snapshot of the data as of 30/07/2021 is available [here](#).

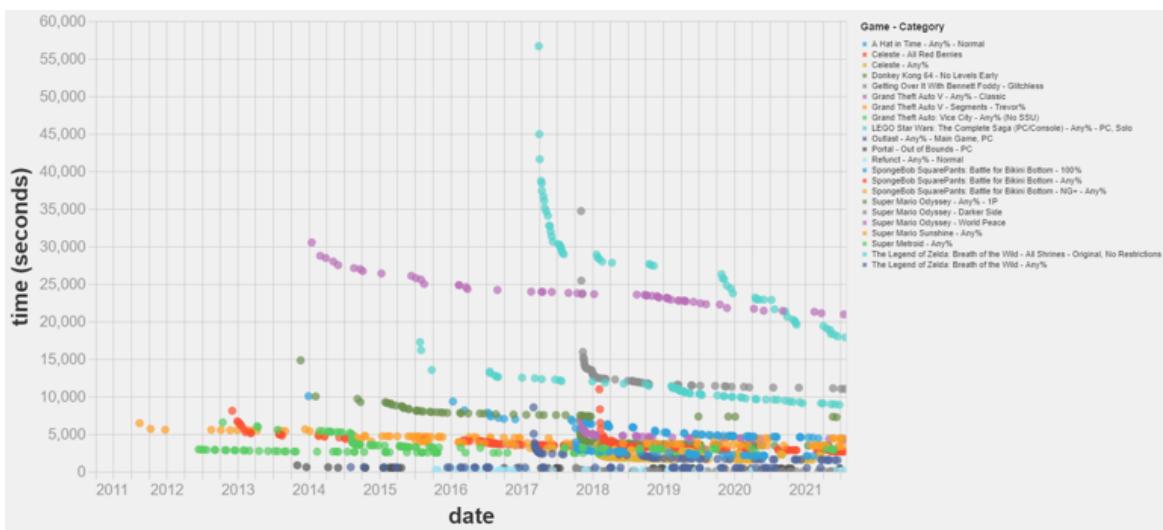


Figure 1: Speedrunning world record progression for the most popular categories. The horizontal axis is the date of submission and the vertical axis is run time in seconds. Spans 22 categories and 1462 runs. Hidden from the graph are 5 runs before 2011. See the rest of the data [here](#).

Feedback on the project would be appreciated. I am specially keen on discussion about:

1. Differences and commonalities to expect between speedrunning and technological improvement in different fields.
2. Discussion on how to mathematically model the discontinuities in the data.
3. Ideas on which techniques to prioritize to extrapolate the observed trends.

AI Safety Papers: An App for the TAI Safety Database

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

[AI Safety Papers](#) is a website to quickly explore papers around AI Safety. The code is hosted on Github [here](#).

In December 2020, Jess Riedel and Angelica Deibel announced the [TAI Safety Bibliographic Database](#). At the time, they wrote:

In this post we present the first public version of our bibliographic database of research on the safety of transformative artificial intelligence (TAI). The primary motivations for assembling this database were to:

1. Aid potential donors in assessing organizations focusing on TAI safety by collecting and analyzing their research output.
2. Assemble a comprehensive bibliographic database that can be used as a base for future projects, such as a living review of the field.

...

The core database takes the form of a [Zotero library](#). Snapshots are also available as [Google Sheet](#), [CSV](#), and [Zotero RDF](#). (Compact version for easier human reading: [Google Sheet](#), [CSV](#).)

One significant limitation of this system was that there was no great frontend for it. Tabular data and RDF can be useful for analysis, but difficult to casually go through.

We've been experimenting with creating a web frontend to this data. You can see this at <http://ai-safety-papers.quantifieduncertainty.org>.

The screenshot shows two main sections of the AI Safety Papers website. On the left, a search results page for 'Paul Christiano' displays 46 results, with one result highlighted: 'My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda' by Chi Nguyen, published in 2020. On the right, a detailed writeup titled 'Writeup: Progress on AI Safety via Debate' by Beth Barnes and Paul Christiano (2020) is shown. This writeup discusses the work done by the 'Reflection-Humans' team at OpenAI in Q3 and Q4 of 2019. It highlights the cross-examination idea inspired by a conversation with Chelsea Voss, Adam Gleave's helpful ideas about the long computation problem, and feedback from Jeff Wu, Danny Hernandez, and Gretchen Krueger. The writeup also mentions contributions from Amanda Askell, Andreas Stuhmller, and Joe Collman, as well as others on the Ought team and the OpenAI Reflection team. It expresses gratitude to contractors who participated in debate experiments, including David Jones, Erol Akkaba, Alex Dearn, and Chris Painter. Oliver Hartky helped format and edit the document for the AI Alignment Forum. A note from Oliver states there is a bug where links to headings in a post do not properly scroll when clicked. The writeup concludes with a link to the original blog post on the AI Alignment Forum.

This system acts a bit like Google Scholar or other academic search engines. However, the emphasis on AI-safety related papers affords a few advantages.

1. Only papers valuable to AI safety are shown
2. There's easy filtering for papers by particular AI safety related organizations or researchers.
3. There's simple integration with blurbs from the [Alignment Newsletter](#) and [Gyrodioit](#).
4. We can include blog posts as well as formal academic works. This is important because a lot of valuable writing is posted directly to blogs like Lesswrong and The Alignment Forum.
5. Later on, we could emphasize custom paper metrics. For example, there could be combinations of citations and blog post karma count.

Tips

- Most of the fields are clickable. Click on an author to see other papers with the same author, or on a tag to see other papers which also have it.
- To quickly go through query results, use the up and down arrows after entering a search.
- Besides the search function, there is also an (Airtable) table view, which can be browsed directly or downloaded as a CSV.

Questions

Who is responsible for AI Safety Papers?

Ozzie Gooen has written most of the application, on behalf of the [Quantified Uncertainty Research Institute](#). Jess Riedel, Angelica Deibel, and Nuño Sempere have all provided a lot of feedback and assistance.

How can I give feedback?

Please either leave comments, submit feedback through [this website](#), or contact us directly at hello@quantifieduncertainty.org.

How often is the database updated?

Jess Riedel and Angelica Deibel are maintaining the database. They will probably update it every several months or so, depending on interest. We'll try to update the AI Safety Papers app accordingly. The date of the most recent data update is shown in the header of the app.

Note that the most recent data in the current database is from December 2020.

Future Steps

This app was made in a few weeks, and as such it has a lot of limitations.

- The data is updated in large batches, and is done fairly messily.
- There's a lot more data we could potentially pull in. For example, blog posts could show comment count and karma.
- There could be commenting allowed on papers. (This would require a log-in system, which we are reluctant to add until necessary.)
- We could use such a database for a more formal paper review system, the results of which could be featured in the UI.

You can see several other potential features [here](#). Please feel free to add suggestions or upvotes.

We're not sure if or when we'll make improvements to AI Safety Papers. If there is substantial use or requests for improvements, that will carry a lot of weight regarding our own prioritization. Of course, people are welcome to submit pull requests to the [Github repo](#) directly, or simply fork the project there.

Covid 8/26: Full Vaccine Approval

Great news, everyone. [The Pfizer vaccine has been approved. Woo-hoo!](#)

It will be marketed [under the name Comirnaty. Doh!](#)

(Do we all come together to form one big cominraty? Or should you be worried about the cominraties of getting vaccinated, although you should really be orders of magnitude more worried about the cominraties of *not* getting vaccinated? Did things cominraty or was there a problem? [Nobody knows. Particle man.](#))

My understanding is that if a doctor were to prescribe the vaccine 'off label,' say to give to an 11 year old or to get someone an early booster shot, then they could potentially be sued for anything that went wrong, so in practice your doctor isn't going to do this.

A reasonable request was made that my posts contain Executive Summaries given their length. Let's do it!

Executive Summary of Top News You Can Use

1. Pfizer vaccine approved under the name Comirnaty.
2. Vaccines still work. If you have a choice, Moderna > Pfizer but both are fine.
3. Boosters are still a good idea if you want even better protection.
4. Cases approaching peak.

Also, assuming you're vaccinated, [Krispy Kreme is offering two free donuts per day from August 30 until September 5.](#)

Now that that's out of the way, let's run the numbers.

The Numbers

Predictions

Prediction from last week: 1,000,000 cases (+14%) and 8,040 deaths (+45%).

Results: 935k cases (+7%) and 7,526 deaths (+35%).

Prediction for next week: 950k cases (+2%) and 9,400 deaths (+25%).

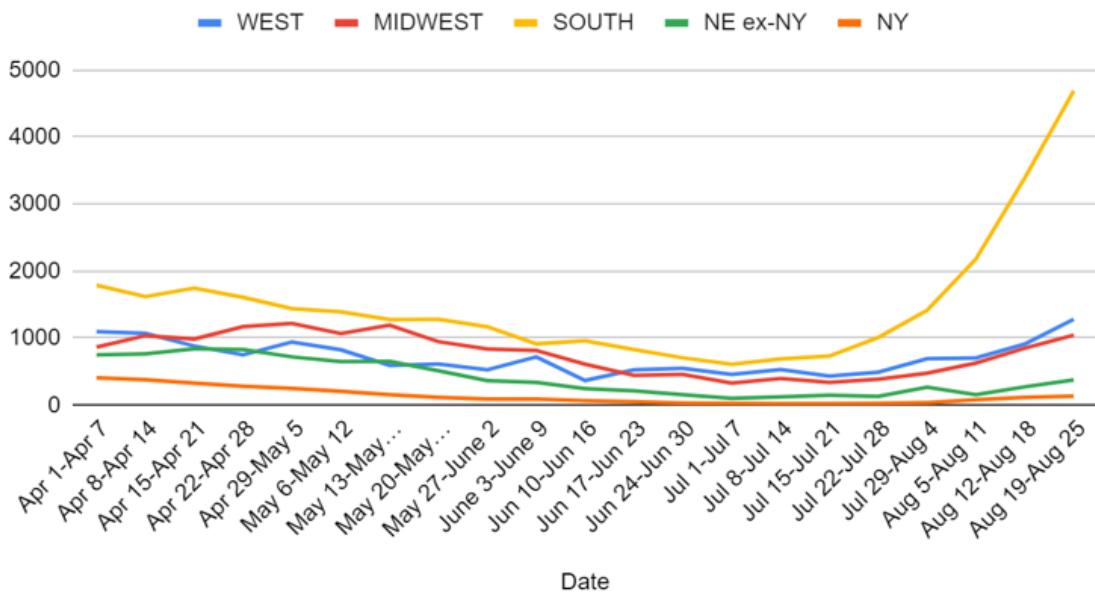
I was confused how there could be such sharp peaks in other countries. It looks like we won't get one of those. The trend lines seem clear, and it looks like we are approaching the peak. It would be surprising if we were still seeing increases week over week by mid-September, with the obvious danger that things could pick up again once winter hits.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jul 1-Jul 7	459	329	612	128	1528
Jul 8-Jul 14	532	398	689	145	1764
Jul 15-Jul 21	434	341	732	170	1677

Jul 22-Jul 28	491	385	1009	157	2042
Jul 29-Aug 4	693	477	1415	304	2889
Aug 5-Aug 11	705	629	2181	234	3749
Aug 12-Aug 18	912	851	3394	388	5545
Aug 19-Aug 25	1281	1045	4692	508	7526

Deaths by Region

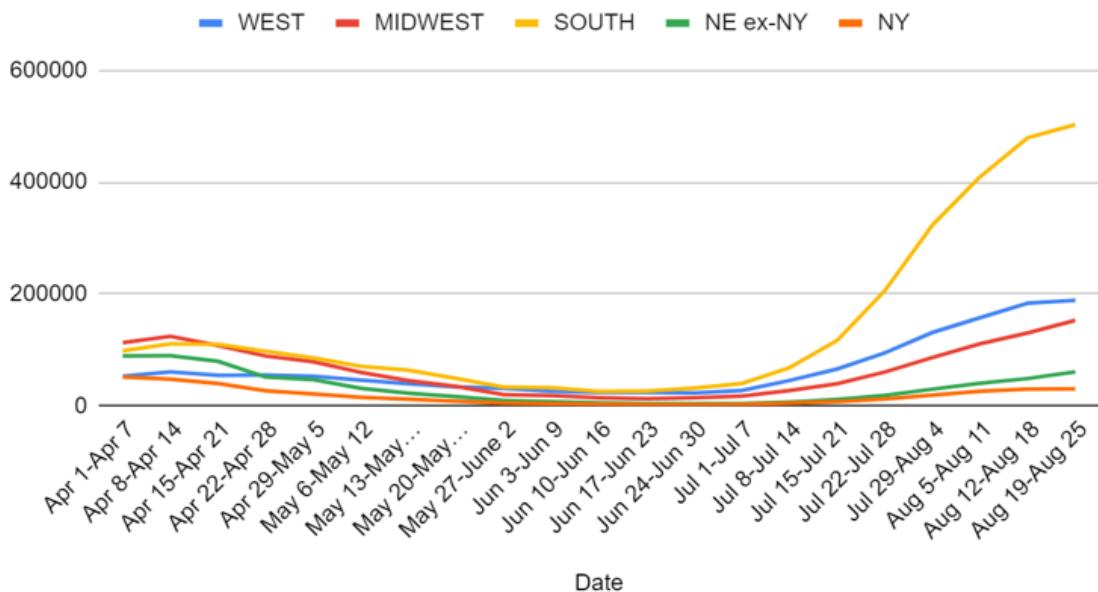


Deaths continue to lag cases. News was slightly good, so adjusting expectations slightly in response. Peak should still be a month out or so.

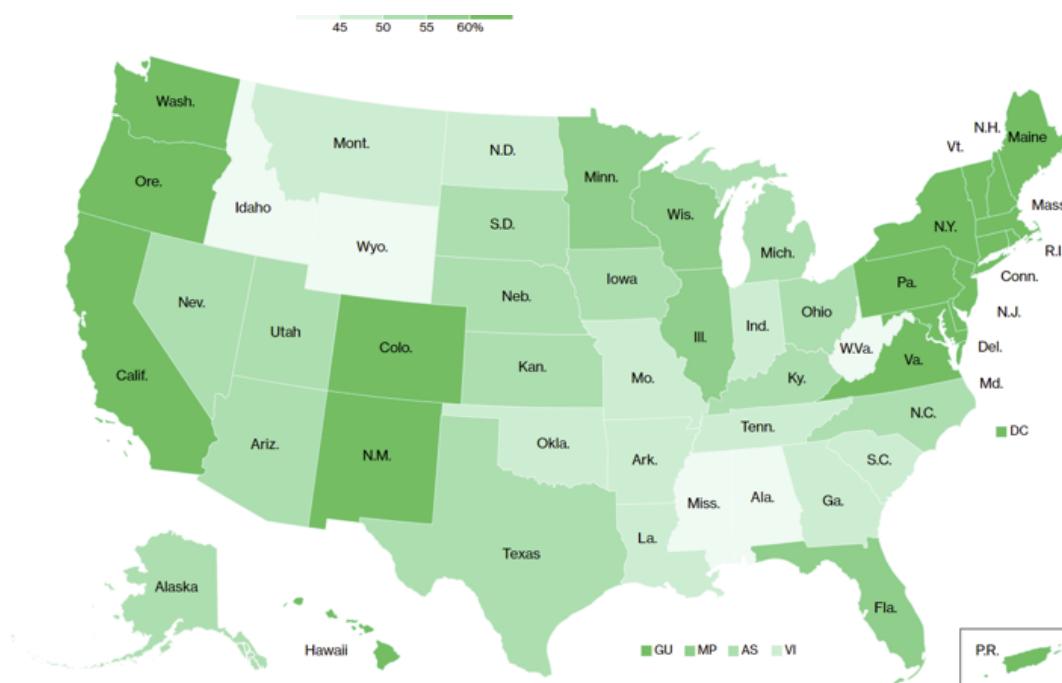
Cases

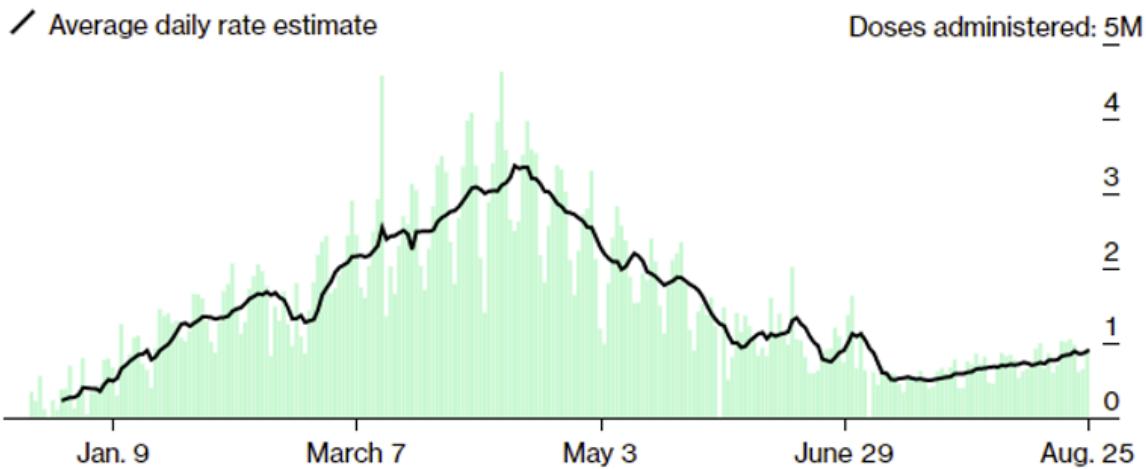
Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jul 1-Jul 7	27,413	17,460	40,031	7,065	91,969
Jul 8-Jul 14	45,338	27,544	68,129	11,368	152,379
Jul 15-Jul 21	65,913	39,634	116,933	19,076	241,556
Jul 22-Jul 28	94,429	60,502	205,992	31,073	391,996
Jul 29-Aug 4	131,197	86,394	323,063	48,773	589,427
Aug 5-Aug 11	157,553	110,978	409,184	66,686	744,401
Aug 12-Aug 18	183,667	130,394	479,214	78,907	872,182
Aug 19-Aug 25	188,855	152,801	502,832	91,438	935,926

Positive Tests by Region



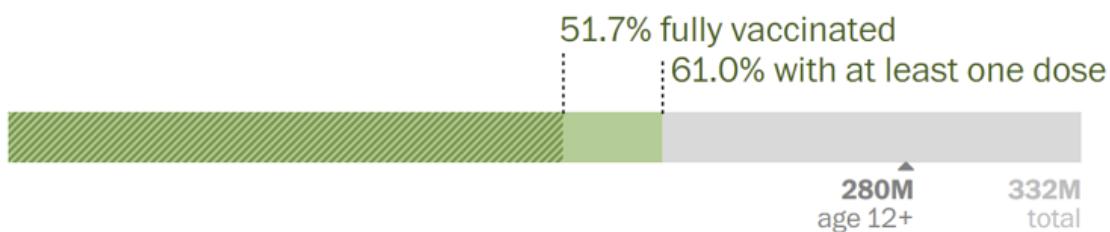
Vaccination Statistics





202.5 million vaccinated

This includes more than **171.8 million people** who have been fully vaccinated in the United States.



In the last week, an average of **891.8k doses per day** were administered, a **8% increase ↑** over the week before.

How much will full FDA approval matter? [Survey says](#) not much.



Zvi Mowshowitz @TheZvi · 22h

...

With formal approval of the Pfizer vaccine, what percent increase in first doses will we see over the next two weeks versus the last two weeks?

Less than 10%	72.3%
11%-33%	21.1%
34%-100%	4.9%
Over 100%	1.6%

669 votes · 1 hour left



A Shark Token @sickofit · 22h

Replies to [@TheZvi](#)

<10%. The relevant link:



Is That Your True Rejection?

I am more hopeful than this, and expect more than a 10% increase. Some of this will be people for whom this really was the true rejection. Other parts of it will be as mandates are handed down and people anticipate further mandates.

Vaccine Effectiveness

I continue to find [this](#) very telling in terms of vaccine effectiveness versus Delta:



John Cochrane @JohnHCochrane · Aug 19

...

Replies to [@paulromer](#)

Is anyone working on delta vaccine? First ones took a weekend. Will fda require a long review? why boost with alpha vaccine rather than speed up delta vaccine?

4

1

19

↑



Alex Tabarrok @ATabarrok · Aug 19

...

Delta vaccines have been designed and the FDA says they will approve them without new efficacy trials. I agree if we are to do boosters it would probably be better to wait for a variant booster.

The argument is simple. The Delta vaccines are designed and would be easy to get approved, yet there has been no move to manufacture them quickly. The only reasonable explanation for this is that there isn't actually much if any difference with the old vaccine. Or at least, that's what the pharma companies that have every financial incentive acting against this are revealing they believe.

[A new paper on vaccine effectiveness concurs \(preprint\)](#).

Adjusting for multiple potential confounders, in the Alpha-dominant period the vaccine effectiveness (VE) of both BNT162b2 and ChAdOx1 vaccines against new PCR-positives was similar amongst those ≥ 18 years to that previously reported to 8 May 2021 amongst those ≥ 16 years¹⁵ (**Table 1**).

In the Delta-dominant period, amongst those ≥ 18 years there was evidence of reduced effectiveness ≥ 21 days after the first ChAdOx1 vaccination (VE 46% (95% CI 35-55%), heterogeneity p=0.004), but not ≥ 14 days after the second (67%, 62-71% vs 79%, 56-90% in the Alpha-dominant period, heterogeneity p=0.23). There was no evidence of reduced effectiveness in the Delta-dominant period for BNT162b2 against all new PCR-positives, with VE 57% (50-63%) post first dose and 80% (77-83%) post second dose (heterogeneity p=0.60, p=0.23, respectively) (**Table 1, Figure 1**).

I will for now accept the principle that a single dose provides substantially less protection against Delta than Alpha, but this is another data point that Delta isn't different from Alpha once you get your second shot. I always find maddening the 'confidence intervals overlapped, so nothing here' reaction to differences like 67% vs. 79% - yes, you can't be confident in that, but that's mostly saying your study was underpowered, since that's the kind of difference one would expect if there was a difference, and again, the word evidence does not mean what they think it means.

The paper's findings then get worse, if you believe them, claiming rapid reduction in effectiveness over time.

In those 18 to 64 years, VE of BNT162b2 against new PCR-positives reduced by 22% (95% CI 6% to 41%) for every 30 days from second vaccination (p=0.007; **Figure 2**). Reductions were numerically smaller for ChAdOx1 (change -7% per 30 days, 95% CI -18% to +2%, p=0.15) but there was no formal evidence of heterogeneity (p=0.14).

They then go on to say this, which given how vaccinations were timed seems likely to be confounding indeed:

Vaccine effectiveness was also generally higher at younger ages (**Table S3**). For example, VE 14 days after the second BNT162b2 dose was 90% (85-93%) for those aged 18-34 years versus 77% (65-85%) for those aged 35-64 years (heterogeneity p=0.0001); and was 73% (65-80%) versus 54% (40-65%), respectively, for ChAdOx1 (heterogeneity p=0.002).

There's no one good money quote on it, but the findings robustly say that vaccinated people's cases tend to be lower viral load, less dangerous and less severe.

Looking at their section on statistical analysis, they're doing some of the necessary and reasonable things but I can't tell if it's enough of them. Such studies are better than nothing if treated with caution, and this seems like a relatively well-done one, but I'm still more focused on the population numbers and what makes the models work out.

When I see things like this:

In those 18 to 64 years, VE of BNT162b2 against new PCR-positives reduced by 22% (95% CI 6% to 41%) for every 30 days from second vaccination (p=0.007; **Figure 2**). Reductions were numerically smaller for ChAdOx1 (change -7% per 30 days, 95% CI -18% to +2%, p=0.15) but there was no formal evidence of heterogeneity (p=0.14).

My core reaction is, the very idea of a 22% decline in vaccine effectiveness per month doesn't make any mathematical sense, until I figured out it meant a 22% *increase* in vaccine *ineffectiveness*. As in, if you are 99% effective in month one, and then you have a 22%

'decline in effectiveness' you would be... 98.8% effective. Or if you were 95% before, you're 94% now. Which doesn't sound to me like a 22% decline in effectiveness, even if true.

[Israeli data](#) continues to suggest extreme fading of vaccine effectiveness if you look at it naively, along with yet another reason to, as post puts it, proceed with caution.

[New data from Denmark:](#)



Chise 🧬 🍽️ 🦠 💉 · @sailorrooscout · Aug 19 · ...
Real-world data out of Denmark shows overall vaccine effectiveness (both doses) against SARS-CoV-2:

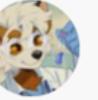
Preventing Hospitalization:
Alpha- Pfizer 86%, Moderna 97%
Delta- Pfizer 94%, Moderna 97%

Preventing Infection:
Alpha- Pfizer 81%, Moderna 96%
Delta- Pfizer 79%, Moderna 88%



Chise 🧬 🍽️ 🦠 💉 · @sailorrooscout · Aug 19 · ...
Replying to [@sailorrooscout](#)
Note: this does INCLUDE asymptomatic infection

Preventing Infection:
Alpha- AstraZeneca (1st dose), Pfizer/Moderna (2nd dose) 93%
Delta- AstraZeneca (1st dose), Pfizer/Moderna (2nd dose) 74% (there was no estimate for prevention of hospitalization yet)



Chise 🧬 🍽️ 🦠 💉 · @sailorrooscout · Aug 19 · ...
This analysis (ended 8/3) covers 2097 breakthrough infections (693 with Alpha, 1404 with Delta) with 90 being hospitalized (57 with Alpha, 33 with Delta). Analysis does not specify if hospitalization was due to COVID-19 or not. Analysis can be found here: [ssi.dk/-/media/arkiv/...](https://ssi.dk/-/media/arkiv/)

One presumes that the *improvement* against hospitalization in Pfizer is a data artifact or failure to control for something or some such, which shows how easy it is to get misleading results, especially since infection went the other way. And this big Pfizer versus Moderna difference against Alpha isn't found elsewhere, which makes me think that once again there's confounding going on all over the place.

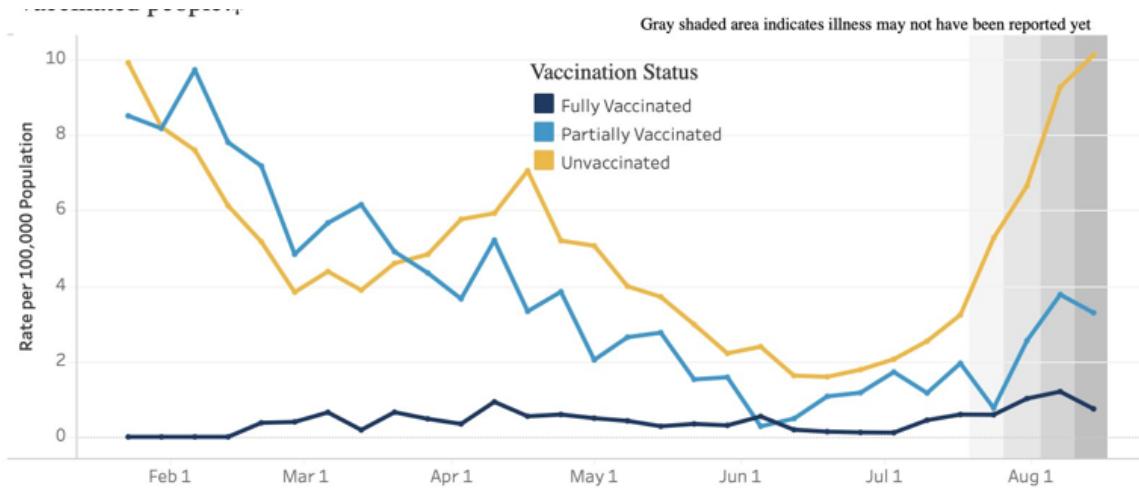
[Here's a thread analyzing some of the results](#), and takes the declining protections and other study data fully seriously, putting the burden of proof on finding something specific that is wrong with the studies, and otherwise taking their results and details seriously and forming the model around that. As usual, the broader context of what such results would mean for all the other data we see isn't incorporated - but again, I don't see anyone doing that.

Here's another good long thread [explaining what vaccine effectiveness means](#) then listing lots of different findings and real world results. Putting them all together like that makes it striking how much the different numbers don't agree if you take them all at face value.

I continue to think that the decline in vaccine effectiveness over time is in large part a mirage, and for practical purposes the decline is relevant but small.

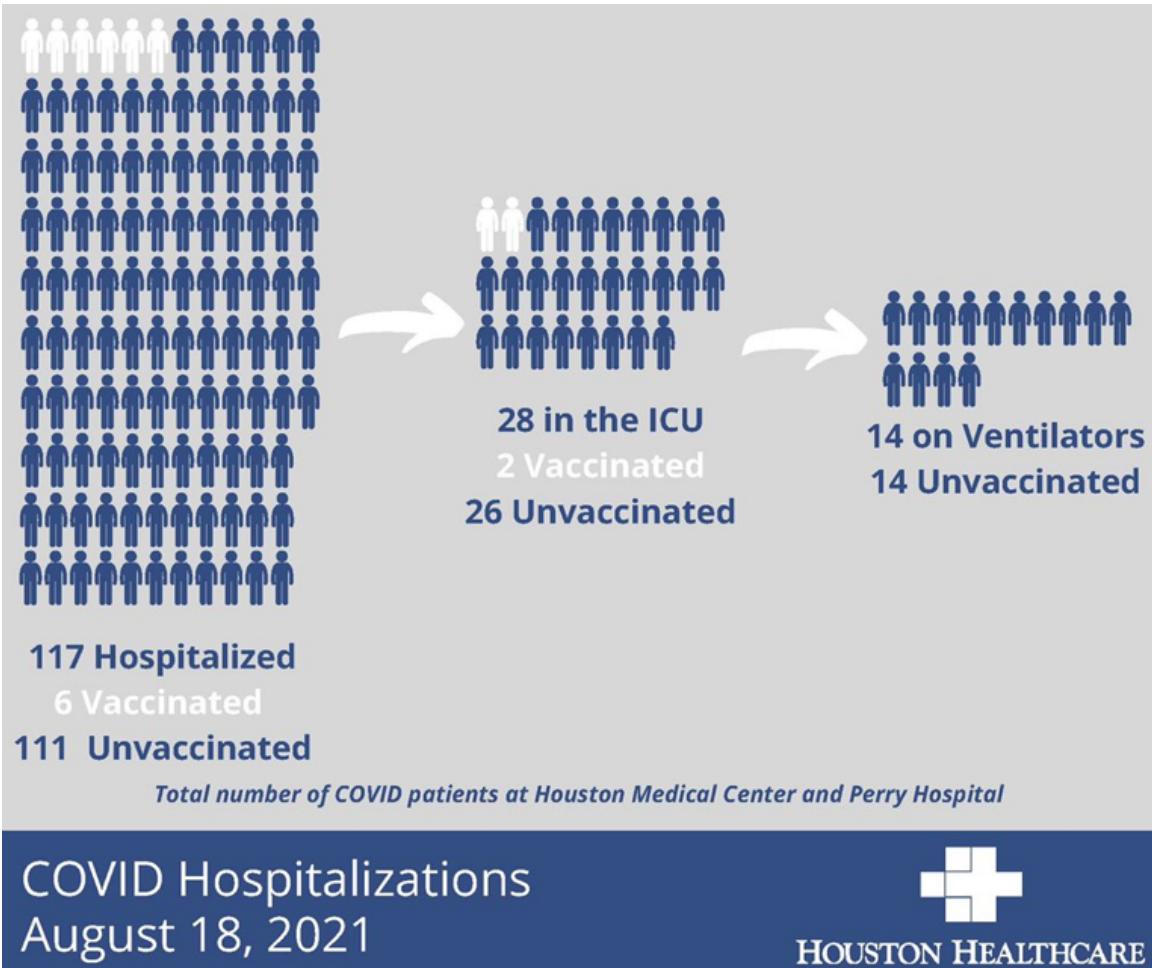
This week's representations of how those vaccines are doing, after having vaccinated about 70% of adults and most of the elderly.

[Virginia offers a dashboard:](#)



Doesn't look like vaccines are losing effectiveness

Houston, [via PoliMath](#):



[And another:](#)



San Diego Covid-19 Hospitalizations

7/12/2021 - 8/10/2021

Fully Vaccinated



13

NOT Fully Vaccinated



521

www.coronavirus-sd.com



COUNTY OF SAN DIEGO
HHSA
HEALTH AND HUMAN SERVICES AGENCY

[And another:](#)



ian bremmer @ianbremmer · Aug 20

Wisconsin: A Case Study

Per 100K fully vaccinated people:

125.4 covid cases

4.9 hospitalizations

0.1 deaths

Per 100K not fully vaccinated:

369.2 cases

18.2 hospitalizations

1.1 deaths

-Wisconsin Dept of Health Services

That's disappointing at face value since it's only a 90% reduction in deaths but after correcting for age it would look a lot better. Weird that so much of the vaccine advantage here seems to be coming *after* hospitalization.

A worry is that the *studies* are selecting for ways to show vaccinated people are at risk, and another worry is that the *real world statistics being reported* are selecting for showing that the vaccines are super effective, because they are the same information but the Official Story is on two contradictory propaganda tracks and is pretending not to notice that this is a physical world question with a correct answer (whether or not we are confident we know what it is).

[Anecdotal in Tampa, Florida](#)



Jennifer Caputo-Seidler, MD @jennifermcaputo · Aug 21

...

I haven't said anything about what's going on with #COVID here in FL bc I haven't had the words to describe it. The truth is we're caring for 3x the number of patients we had last summer. 12 of our floors have been converted to covid units. We are stretched to the breaking point.



Jennifer Caputo-Seidler, MD @jennifermcaputo · Aug 21

...

Replies to @jennifermcaputo

2/ We're putting multiple patients on ventilators every day. We're doing CPR on patients younger than me in a desperate attempt to save their life. We're calling dozens of families every day and telling them we don't think their loved one is going to survive this.

39

723

3.2K



Jennifer Caputo-Seidler, MD @jennifermcaputo · Aug 21

...

3/ Almost all our patients are unvaccinated. We didn't have to get here. Please [#GetVaccinated](#) if you haven't already. [#WearAMask](#) 😊 And tell the people you love how much you love them while you can.

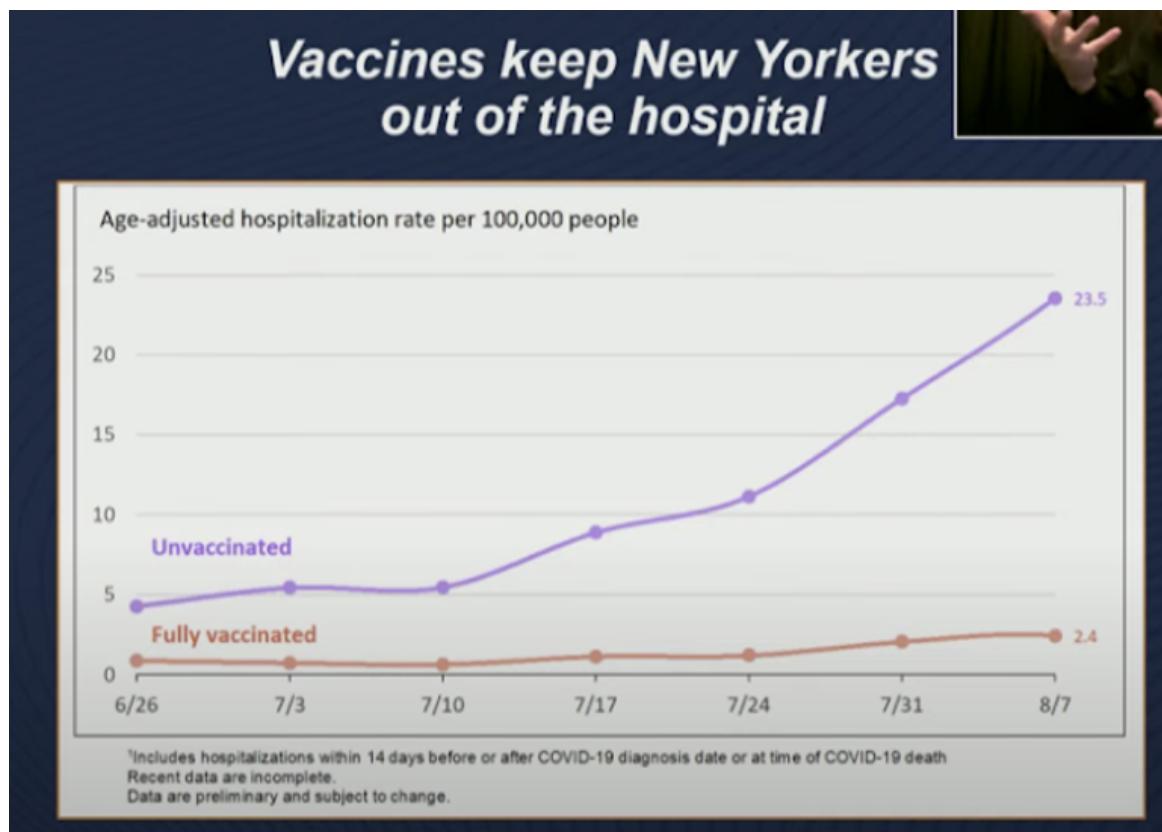
89

948

4.3K



Here in New York:



Meanwhile also this:



Divia Eden @diviacaroline · Aug 21

...

This point seems very important if we are trying to predict the course of the pandemic!

Vaccine immunity (while still pretty good) has taken a substantial hit from the delta variant. Natural immunity seems to be still holding up.



Meaghan Kall @kallmemeg · Aug 20

And to the questions on if natural immunity works:

In July, out of 865,275 SARS-CoV-2 infections in England

Only 12,019 were reinfections

That's 1.4% !

Note that yes, we are excluding the first wave infections here as per her follow-up note, but note the graph and adjust accordingly, and I think the point stands.



Meaghan Kall
@kallmemeg

...

Replies to [@kallmemeg](#)

That's a LOT of infections not counted in our case numbers in the first wave (Jan-May 2020), equivalent to 10% of the population.

So we're certainly underestimating reinfections in those infected early on.

I realise this undermines my original post & I'm ok with that

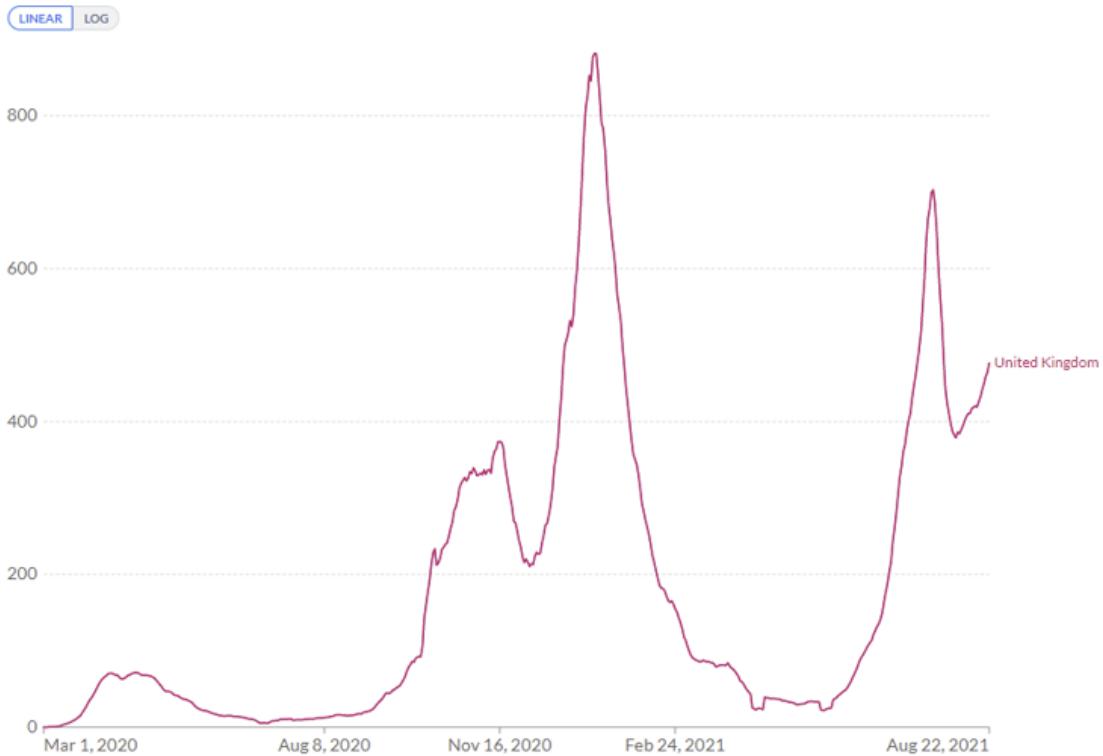
2:12 PM · Aug 21, 2021 · Twitter for iPhone

1 Retweet 14 Likes

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World
in Data



That does bring up that UK cases are clearly rising again, so we can no longer use that as an important signpost that things will turn around rapidly and that will be that. If anything, it's now making the case that such a turnaround is unlikely. I don't know of anyone who has offered an explanation other than a shrug for the decline followed by a reversal here.

As for the reinfections versus vaccine effectiveness, my hypothesis is that this is not a case of 'immunity from infection holds up but vaccine immunity is losing ground.' Remember when we were worried that *natural* immunity faded with time but vaccines solved that problem? The actual difference is in the methods of observation. When similar observational methods are used, we seem to get similar results.

How *infectious* are breakthrough cases? We now [have two studies for that](#). They found that vaccinated people who get infected are still infectious, but their viral loads are substantially lower, so this was what we previously expected. And also they clear the virus faster, which was *also* expected.



Alasdair Munro ✅ @apsmunro · Aug 22

...

Replies to @apsmunro

Add to this the data that vaccinated people also clear the virus much faster

Plus it at least halves your chance of getting infected in the first place

Being vaccinated doesn't mean you can't transmit, but it's your best bet at protecting the people around you



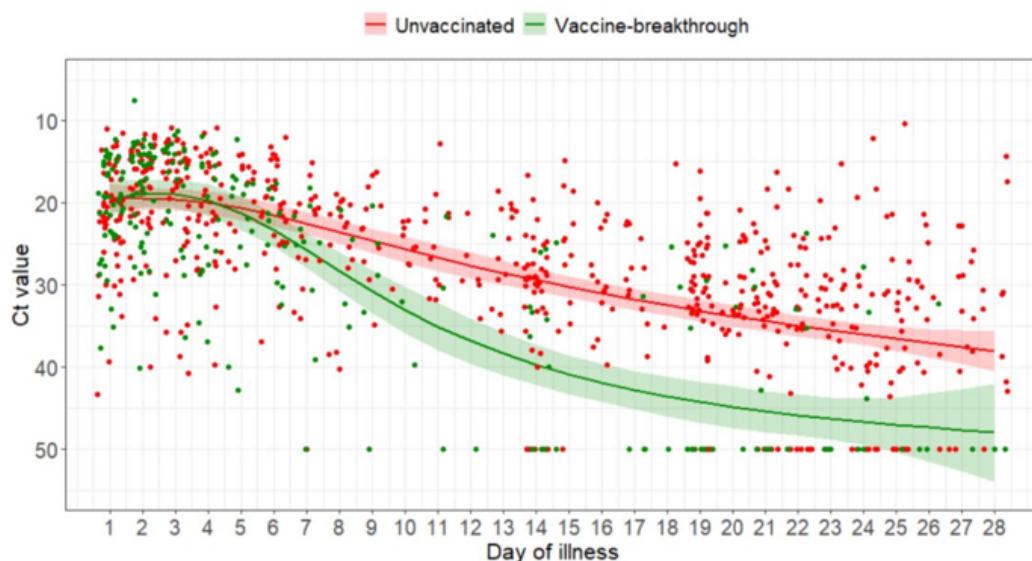
Alasdair Munro ✅ @apsmunro · Jul 31

Did you hear vaccinated people are just as infectious as unvaccinated with Delta?

This study from Singapore shows not only do vaccines stop you getting sick with Delta, but also clear your viral load MUCH faster

Vaccines make you less infectious 💥

medrxiv.org/cgi/content/sh...



Weirdly, there are two *different* studies that find the two different results, although depending on how you measure, fading quickly implies lower average viral loads, so the results are compatible with the graph and it's possible what we're seeing is a *shorter period of infectiousness* rather than less at the peak. That seems unlikely to be the whole effect to me, but could easily be the majority of the benefit.

How much comfort that brings depends on the situation and on what you previously believed. If you're as bad at this as the CDC and were saying the vaccines 'prevent transmission' full stop and now that they 'don't prevent transmission full stop' it gets confusing.

Vaccine Hesitancy and Mandates

Formal approval is in, so here... we... go.

[I saw this about one minute after I saw the FDA had approved the Covid vaccine](#), perhaps someone planned something in advance for once:



Lara Seligman @laraseligman · 17m

BREAKING: Pentagon will now mandate all troops be vaccinated against Covid-19, following the FDA's full approval of the Pfizer shot, says [@PentagonPresSec](#)

On her first day on the job now that The Worst Is Over, our new governor [lays down the law](#):



Morgan McKay @morganfmckay · 2h

NEW: [@GovKathyHochul](#) announces that there will be a vaccine mandate for all school personnel with a weekly test opt out option "for now."

[She also raised New York's total death count by 12k](#), which once again highlights that maybe Cuomo went down in a similar way to Al Capone (who was indeed guilty of tax evasion).

Although she's also mandating 'ethics seminars' so you win some and you lose some.

[LSU is going to mandate vaccination or a negative test for all fans at Tiger Stadium.](#)

[Whereas the University of Georgia is going the other way.](#)

[Here's the owner of the Dallas Cowboys:](#)

"Everyone has a right to make their own decisions regarding their health and their body. I believe in that completely—until your decision as to yourself impacts negatively many others. Then the common good takes over. And I'm arm-waving here, but that has everything to do with the way I look at our team, the Cowboys, or the way I look at our society. We have got to check 'I' at the door and go forward with 'we'. Your Dallas Cowboys are doing that."

-Cowboys owner Jerry Jones to Dallas radio station 105.3 The Fan

[Who else we got](#) (WaPo)? They found CVS Health, Deloitte and Disney, but so far, not an impressive set of additional mandates. It seems not many were standing by ready to go.

[Delta Airlines](#) is charging unvaccinated employees \$200 a month extra for health insurance, on the very reasonable premise that every hospital stay for Covid costs them an average of \$50,000 and they end up in the hospital for Covid more often. Insurance companies can't do this, but it seems *corporations employing you* can do it.

[NYPD has threatened to sue if the city attempts to implement a mandate.](#)

[Texas Governor once again mandates against vaccine mandates, this time ensuring it applies despite FDA approval.](#)

When you're fully anti-vax, you're anti-vax, [and it'll be hard to tell you different](#), as Donald Trump learned:

Donald Trump Booed at Alabama Rally After Encouraging Crowd to Get COVID-19 Vaccine

"I believe totally in your freedoms, I do, you gotta do what you gotta do, but I recommend take the vaccines. I did it. It's good," he said, drawing boos from the crowd of supporters.

Others are less fully anti-vax, but still unvaccinated, thanks to various [ways we botched things](#).



Ranu Dhillon @RanuDhillon · Aug 20

...

I've been working in a hospital in a low-income area for the past several nights

From talking with our many unvaccinated Covid patients, there are 2 general responses I've heard as to why they weren't vaccinated...



Ranu Dhillon @RanuDhillon · Aug 20

...

1) Several people said they knew vaccination was *important* but never perceived it as their *most immediate* need until they got sick

Had we been going door-to-door, eliminating the burden on them to search out & get vaccinated, most felt like they would have gotten it



Ranu Dhillon @RanuDhillon · Aug 20

...

2) Distrust -- not of vaccines -- but of the formal authority structures from whom they see them pushed

This rational distrust is from decades of injustices & continued negative interactions with these structures that is hard to undo or overcome quickly & amid a crisis



Ranu Dhillon @RanuDhillon · Aug 20

...

Replying to [@RanuDhillon](#)

Akin to what people in Guinea told me during Ebola, if we're so interested vaxxing people to protect their health, where were we when their families were suffering from joblessness, housing insecurity, inadequate food, etc. before the pandemic & in its pre-vaccine period?

teams assure communities that they are there to protect their health.³ Many communities do not perceive a difference between Ebola and these other health problems, seeing them all as illnesses for which they need help. When they cannot find care for a child with diarrheal disease, they question the sincerity of response teams that seem eager to help them with Ebola.

As Ranu notes there are two distinct things here. First, we botched the logistics, and could have done much better if we'd made sure to beware trivial inconveniences that aren't always trivial. Second, our authorities are untrustworthy so people don't trust them. This is framed here in the standard blue-tribe way as 'the system fails such people and they remember the legacy of all that' with it being 'hard to make up' during a pandemic, rather than the simple 'these people lied about the pandemic over and over again' model. Both are presumably relevant, but my guess is that handling the pandemic in a trustworthy fashion would have largely solved both problems. Yes, such people will *absolutely* ask why you

weren't helping them before, but that's different from turning your help down if you're here now.

One aspect of vaccination decisions is that patients in America do not pay for their health care. Almost everyone who can get it has health insurance because if you don't the medical system bills you personally and attached some number of extra zeros to the bill because they can, so you can't opt out. [For a while, they even waved 'cost sharing' on Covid, so you didn't even pay the fraction you normally pay, but that's increasingly no longer true.](#) Would be good if more people knew. Incentives matter, but only if people are aware of it. One could note that this policy could be taken farther, if the government permitted this, so we're doing mandatory mandates with one hand and mandatory massive subsidies to those who don't follow those mandates with the other.

[State employees..you will get vaccinated as many times as is legal, or else.](#)



Kerry 🇺🇸 @kerry62189 · 2h

Replying to [@kerry62189](#)

...

It takes for granted that the CDC is going to recommend multiple boosters to almost everyone (or else why mandate it?), and is mandating them all *in advance* with no further qualifications! This is going to break one way or the other around mid-October.



Kerry 🇺🇸 @kerry62189 · 2h

...

The gov. here just announced a mandate for his staff, with compliance required by 10/17, with possible exceptions for union employees, but no exceptions for teleworkers. After 10/17, they'll start "progressive disciplinary measures." Few details.



Kerry 🇺🇸 @kerry62189 · Aug 18

In other words, the message is that by mid-October, it will be impossible to live a normal life in the Northeast if you don't comply. But despite the sense of inevitability, it relies on short-term obedience---the courts haven't signed off, & enforcement details are scarce.

[Show this thread](#)



1



Kerry 🇺🇸 @kerry62189 · 2h

...

This *executive order* encourages other state institutions to do the same. And then this: "As new CDC guidance regarding booster...doses is issued in the future...employees will also be required to provide proof they have received those doses by a deadline to be established." ••

This is an explicit 'everything that is not forbidden is mandatory, and everything that is not mandatory is forbidden' rule. You can get exactly this many shots at exactly these times, and

you either get them or you're fired. There's no concept of a booster that is *optional*, based on someone's situation, and the full mandate applies to teleworkers.

This is where things are going to be tricky. Requiring 'full vaccination' so far has been simple. You get two shots and that's it. Now there are signs that this in many places is going to morph into getting periodic boosters with different places (at a minimum, [nations](#), Austria and Croatia are already setting expiration dates) having different requirements, and those boosters will have a much less slam-dunk risk-benefit profile.

I will happily take the third shot without any need for outside incentives, but it is a *very reasonable* position to not want the third one, and it seems likely that requiring boosters will have far less robust support than requiring two shots.

[A cheap shot](#), but I think a necessary one so putting it here anyway, without any need for further comment.

How it started:



Mark D. Levine @MarkLevineNYC

...

VRA was critical in fighting unfair voter ID laws in 2012. Now **#SCOTUS** has ensured it will be harder to defend fair elections in the future.

11:33 AM · Jun 25, 2013 · Twitter Web Client

2 Retweets 2 Quote Tweets 1 Like



How it's going:



Mark D. Levine
@MarkLevineNYC

...

Replying to [@MarkLevineNYC](#)

The ID requirement is to help reduce fraud.
Venues covered by the vax screening program
are required to check ID for those 18+.
Checking ID for 12+ is optional.

The NYC Covid Safe app allows you to upload a
picture of your ID if you don't want to carry it.

8:25 AM · Aug 20, 2021 · Twitter Web App

11 Retweets 89 Quote Tweets 67 Likes



One can definitely say [shots fired](#):

**Boston mayor compares NYC's
vaccine mandate to slavery,
birtherism**

Masking, Testing and NPIs

[This is nuts](#), actively counterproductive on every level, and what must be fought against:



Bob Schaper
@Bob_Schaper

...

BREAKING: [@OregonGovBrown](#) announces statewide OUTDOOR mask mandate starting Friday, regardless of vaccination status. Live report at 4 [@KEZI9](#) [@KennedyDendy](#)

3:20 PM · Aug 24, 2021 · Twitter Web App

127 Retweets 509 Quote Tweets 275 Likes



Nick Foy @TheNickFoy · 1h

...

Replies to [@Bob_Schaper](#) [@OregonGovBrown](#) and 2 others

wait is she mandating a single mask that seems really dangerous we should be mandating N95s at the very least maybe double N95s wait would that be an N190 this might be something she should consider mandating N190s there's so much we don't know but we do know that mandating outdo



Michael Krieger @LibertyBlitz · 6m

Replies to [@TCapsulae](#) [@Bob_Schaper](#) and 3 others

Imagine complying with this

To be fair, it is only required 'when social distancing is not possible,' most of the time this will definitely apply, and I assure you that it's always *possible*.

It's always adorable when [people think the constitution is a meaningful limiting factor](#), and all recursive mandate sentences are fun.



Eliezer Yudkowsky ✅ @ESYudkowsky · Aug 21

...

It's too bad the Constitution mandates against a federal mandate to ban state mandates banning regional mask mandates.

In practice, this is technically true, but there is a known way around it known as withholding federal funding. And another easier way around it, called ignoring the constitution, since presidents mostly do what they want without any actual legal authority under the constitution and mostly no one calls them out on it. Eviction moratorium, anyone?

If you're in NYC and either old or immunocompromised, [make sure you know about this](#):



Mark D. Levine ✅ @MarkLevineNYC · Aug 19

...

Amazing testing option which too few know about:

==> NYC is now offering free *at-home* PCR testing for all NYers aged 65+ or immunocompromised. Trained clinician will come to your home. Available 7 days/week, 9am-7pm.

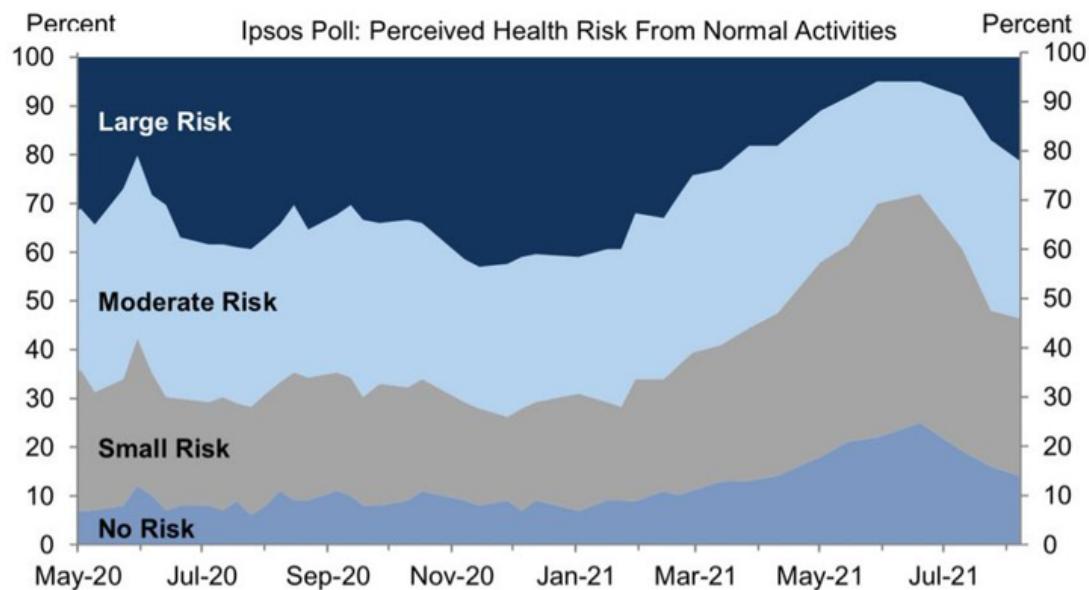
Call 929-298-9400 to sched an appt.

You can also buy one at the pharmacy, although not like in Europe [where the tests are super cheap and abundant](#). FDA Delenda Est.

Also, a periodic reminder that the reason younger children can't get vaccinated, which in practice is causing super massive freakouts although there's almost zero risk there, [is that the FDA moved the goalposts to require additional data](#). Thus we almost certainly won't get this before the end of 2021, and I'd double check but [this market sure looks a lot like free money](#).

[Here's a graph of how afraid people have been over time](#):

Exhibit 2: 53% of Adults Now Ascribe Moderate or Large Health Risk to Normal Activities (vs. 28% in June)

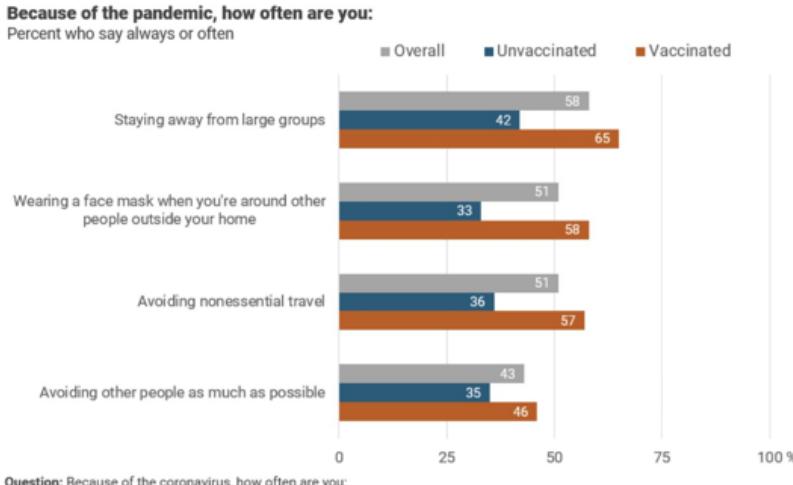


Source: Ipsos, Goldman Sachs Global Investment Research

The lack of an increase in fear over the winter surge is the most surprising thing here. Otherwise it all makes sense, with fear going down when things were improving, then fear starting to go back up as cases rise. Fear isn't a perfect proxy for the private control system, but changes in fear likely predict marginal changes in private actions and we're back at levels similar to April.

Here's a survey on activity:

Despite increased concern about contracting COVID-19, Americans' precautionary habits have remained largely unchanged since June. At least half continue to stay away from large groups, wear face masks when around people outside their home, and avoid nonessential travel. Vaccinated Americans are more likely to maintain these habits than the unvaccinated.



AP | NORC

APNORC.org

As one would expect by now, vaccinated people are taking more precautions than unvaccinated people. Almost half of vaccinated people are 'avoiding people as much as possible' and they're claiming it's because of the pandemic. However I share Nate's skepticism here minus the word 'little' because math:



Nate Silver ✅ @NateSilver538 · Aug 20

I'm a little skeptical of this result, since it doesn't match all sorts of observational data (e.g. restaurant reservations or air traffic numbers) showing people's social activities back to maybe 80-90% of pre-pandemic levels. So maybe some social desirability bias. But still.

...

Perhaps 'as much as possible' means until one is hungry, or has somewhere to go. It's on the margin.

Study does some modeling and finds that according to its model [masks work, ventilation works even better.](#)



Evan Roberts @evanrobertsnz · Aug 20

...

The most effective intervention here was lots of windows open, followed by universal masking. While many American buildings lack windows that open, many have them. And yet this message — open the windows — gets much less attention than masking.



Will Humble @willhumble_az · Aug 20

BTW: Here's a new study in BMJ showing that universal classroom masking reduces aerosol transmission 800%.

Universal masking plus optimal ventilation reduces aerosol transmission 30x (3000%).

medrxiv.org/content/10.1101...

[Show this thread](#)



Evan Roberts @evanrobertsnz · Aug 20

...

Replies to [@evanrobertsnz](#)

Opening windows is less energy efficient, but I imagine it is still cost effective relative to the alternatives. It also reduces the burden on individuals, which is important. Interventions that don't require everyone to be doing something are good.

[From the study:](#)

Results: The most effective single intervention was natural ventilation through the full opening of six windows all day during the winter (14-fold decrease in cumulative dose), followed by the universal use of surgical face masks (8-fold decrease). In the spring/summer, natural ventilation was only effective (≥ 2 -fold decrease) when windows were fully open all day. In the winter, partly opening two windows all day or fully opening six windows at the end of each class was effective as well (≥ 2 -fold decrease). Opening windows during yard and lunch breaks only had minimal effect (≤ 1.2 -fold decrease). One HEPA filter was as effective as two windows partly open all day during the winter (2.5-fold decrease) while two filters were more effective (4-fold decrease). Combined interventions (i.e., natural ventilation, masks, and HEPA filtration) were the most effective (≥ 30 -fold decrease). Combined interventions remained highly effective in the presence of a super-spreader.

Conclusions: Natural ventilation, face masks, and HEPA filtration are effective interventions to reduce SARS-CoV-2 aerosol transmission. These measures should be combined and complemented by additional interventions (e.g., physical distancing, hygiene, testing, contact tracing, and vaccination) to maximize benefit.

The second intervention simulated was the use of HEPA filtration devices. We tested the recommended 5 air changes per hour (ACH),²⁹ where two filters per classroom are needed (HEPA device each delivering a flow rate up to $400 \text{ m}^3\text{h}^{-1}$ of clean air), and an intermediate 2.5 ACH corresponding to one filter (Figure 2a). The 2.5 ACH option was as effective as two windows 20-cm open all day long during winter (2.5-fold decrease). The 5 ACH option was even more effective with a 4-fold decrease in the cumulative dose absorbed. The universal use of face masks was twice as effective, with an 8-fold decrease in the cumulative dose absorbed (Figure 2b).

Filters win out here over windows, if one has to choose, and of course if possible you'd do both. Also you can't cheat on the windows, you gotta actually leave them open. When we're considering actions like mask mandates or *shutting down living life entirely* I find it odd that people worry about energy costs this much, but there you go. Also fresh air remains a Nice Thing. As always, one must be highly skeptical when translating such results into predictions for actually preventing cases.

[A potential issue with price controls:](#)



Agnes Callard @AgnesCallard · Aug 20

Waiting in line for covid test (I have a cold, it was neg), talked to a vaccine-hesitant guy, realized: some people are freaked by the fact that it's free—there's a group of ppl that would have been much more likely to take it if it cost \$5.



Robin Hanson ✅ @robinhanson · Aug 20

Replies to [@AgnesCallard](#)

If drug makers had been allowed to charge full market prices all along, it would be much higher status and sought after. Yes we could and should have subsidized such purchases.

[An argument against weaning masks on the margin](#), and [a good question about presenting that argument](#).



Bryan Caplan @bryan_caplan · 3h

"How many times during Covid have you struggled to understand another person? To be heard? Indeed, how many times have you simply abandoned a conversation because of masks? I say the dehumanization is at least five times as bad as the mere discomfort."



Anarcho-Moses 🐻 @ben_r_hoffman · 3h

Replies to [@bryan_caplan](#)

Honest question - do you have empirical evidence that anyone who wouldn't be persuaded by the tweet-length version of this kind of argument would be persuaded by the blog post length version? If so, I'd like to meet & learn about such people.

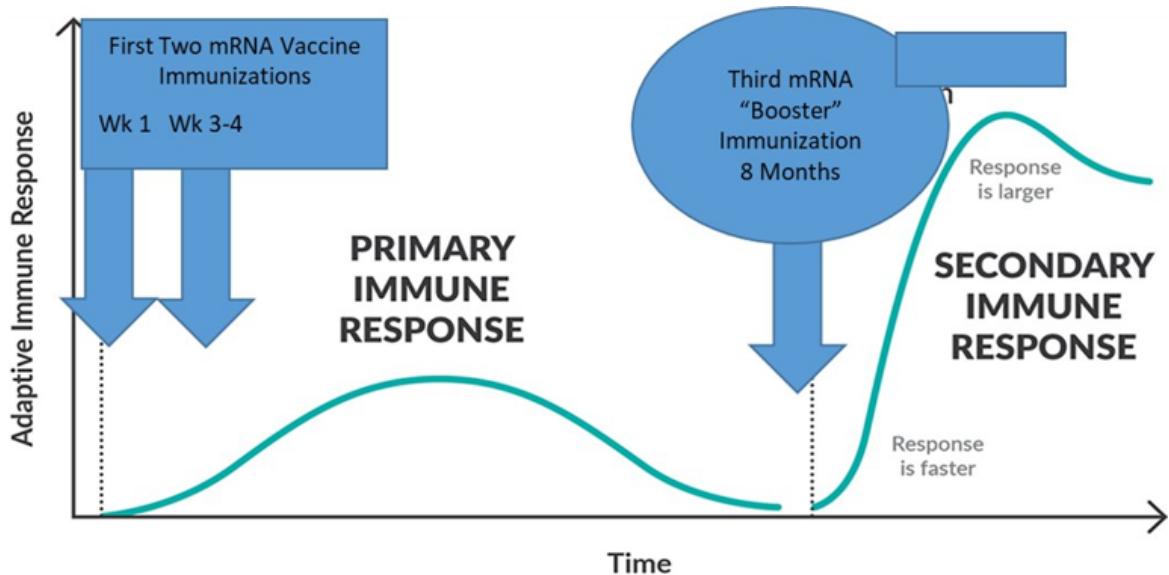
I found the tweet *more* compelling than the full post. Getting into the details mostly highlighted places I disagreed with Bryan.

Booster Shots

[Governor of Texas gets a third shot as a booster](#). I have no issue with people in high positions getting superior medical treatment when there's a supply or resource shortage, but meanwhile we have vaccines expiring in some places. That's from [Scott's post with further comments](#) on the topic of FDA Delenda Est, which is interesting but inessential.

The new argument against booster shots is that they... [might cause us to produce too many antibodies against Covid, and then maybe Covid mutates](#) and the antibodies become dangerous or unhelpful because they're overtrained? When it's not Officially Sanctioned even antibodies are labeled bad, it seems. Meanwhile this is doubtless supposed to make people worried about Delta, but this worry definitely does not apply to Delta, and an additional customized booster would be necessary in the cases being described either way. Don't worry, such arguments will go away once the Official Sanction comes down, which is coming soon.

Meanwhile, an argument *for* booster shots is that the first two doses were so close together that [they count as a primary immunization](#), claiming it looks like this:



Which is so insane it doesn't even bother putting *any* impact from the second shot into the chart at all, and puts the peak of the 'primary' response more than halfway down the graph when it's almost fully effective. There's obvious nonsense available on all sides.

Think of the Children

We *really do* have a large class of Very Serious People, with a lot of influence on policy and narrative, who think that living life is not important, that the things you care about in life are not important, and that our future is not important, because saying the word 'safety' or 'pandemic' should justify anything.

This week's case in point, and like my source MR I want to emphasize that *this is not about the particular person in question here*.

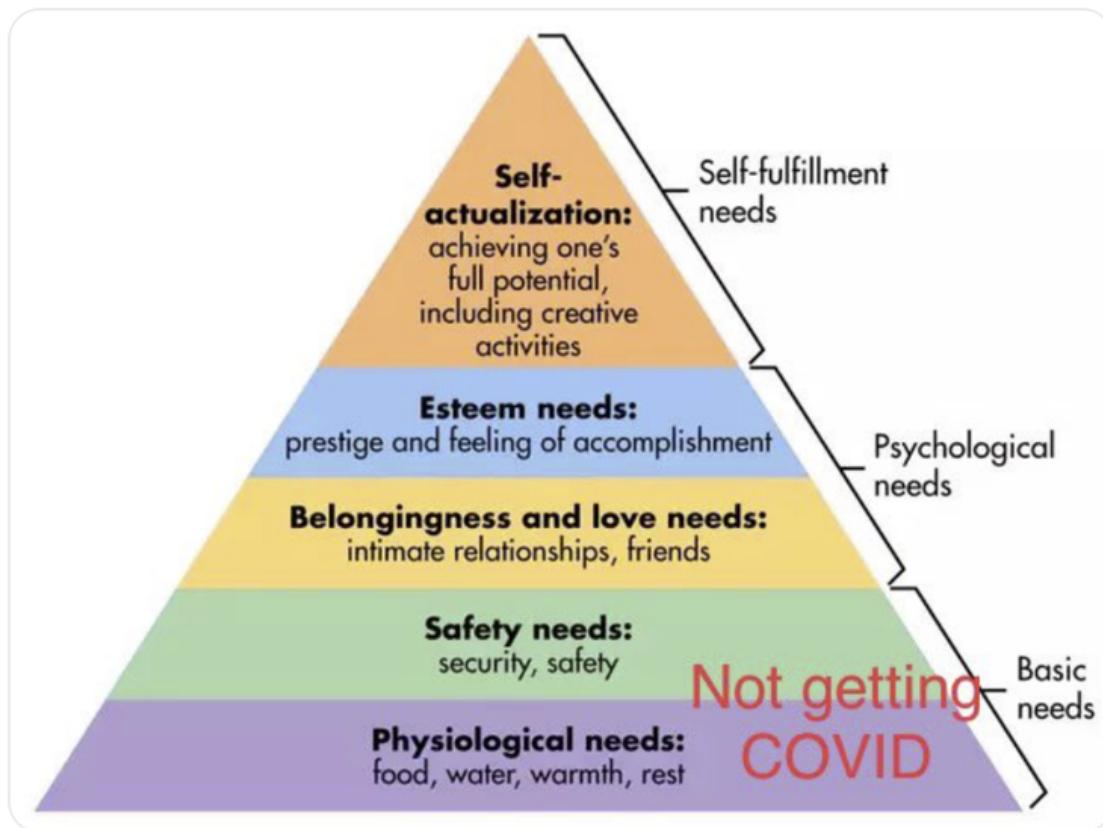


Dr Ellie Murray, ScD ✅

@EpiEllie

...

Maslow's hierarchy of needs, people



If anything, I'd like to *thank* Dr. Murray for being so clear and explicit. If you think that safety trumps the need for love, for friends and for living a complete life in general, then it's virtuous to say that outright, so no one is confused.

In case you think she doesn't mean that (or that others don't mean that), no really, she does:



Dr Ellie Murray, ScD ✅ @EpiEllie · 21h

...

Genuine q for ppl more concerned about schools being closed than covid: are you aware mandatory schooling is barely a century old in this country?

Maybe ur all grandparents had highschool, but what about ur great-grandparents?

Yes, education is important. But it's a pandemic!



Dr Ellie Murray, ScD  @EpiEllie · 21h

Replying to @EpiEllie

My point here isn't that schools closing is good, or that everything is going to be just fine. It's that we shouldn't be acting like no schools is a completely unprecedented unknowable scenario. It was normal life for most of human history.

359

216

487



Dr Ellie Murray, ScD  @EpiEllie · 21h

School is super important! I am fully supportive of schools & education!!

But the idea that schools need to stay open even if we can't do appropriate pandemic mitigation because "who knows what unending damage closures will do to kids" mystifying to me!



Big AI @bigal_0 · 22h

Replying to @EpiEllie

Wait til you hear how we handled pandemics for most of human history

4

4

104



Ellie Murray does not believe that school is terrible, so she is simply saying that the *claimed* benefits of school are not important relative to the marginal impact of schools on Covid-19.

That reply was one voice in a chorus, as the replies are what you'd expect and rather fun to read through. Nate Silver sums this up well:



Nate Silver  @NateSilver538 · Aug 22

Replying to @NateSilver538

On some level it's literally a fight for all the good things about society and civilization vs. perspectives like this.

There was also a side debate over whether school is the future of our children and our children are our future, or the alternate hypothesis that children are also people and school is a prison and dystopian nightmare. The thing to remember is that *this view is not driving most of the anti-school rhetoric*. Such folks mostly think school is vital to children, but don't care.

Yes, I was aware, and I'd rank my concerns regarding school in this order:

1. Kids going to school. School is a prison and a dystopian nightmare.
2. Kids not going to school. Remote school as implemented is somehow so much worse.
3. Getting Covid. I'd rather not get Covid.

But yeah, we can beat that take this week, because the [The Times Is On It](#):



The New York Times

@nytimes

...

In Opinion

“There is good reason to believe that wearing a mask at school could actually improve certain social and cognitive skills,” writes Judith Danovitch, a psychologist who studies child development, in a guest essay.



nytimes.com

Opinion | Actually, Wearing a Mask Can Help Your Child Learn

Ideally, face coverings wouldn't be necessary in school. But for now, they present an educational opportunity.



Mason 🚶‍♂️🩹@webdevMason · Aug 19

I'm trying to understand how it isn't immediately obvious to everyone that strapping something to your face that you constantly want to rip off is extremely counterproductive to attentively listening to lectures, something children already famously struggle to do

11

10

163



Mason 🚶‍♂️🩹@webdevMason · Aug 19

On a slightly tangential note, Conor and I both have attention deficit issues and found school borderline torturous, and to this day it makes my blood boil when adults moralize these challenges in children

2

2

95



Mason 🚶‍♂️🩹@webdevMason · Aug 19

The fact that kids who share the difficulties we had will be treated like they lack character when they fixate on their masks and like they're slow when they can't focus genuinely makes me want to scream



Mason 🚶‍♂️🩹@webdevMason · Aug 19

...

Replying to [@webdevMason](#)

In general, the willingness adults have to impose discomfort on children — even necessary discomfort — and then convince themselves that the *discomfort itself* is necessary or useful strikes me as a special kind of cowardice. People need to own what they force onto children.

Technically I'm sure it is true that masks represent an 'educational opportunity' in the sense that whenever anything happens you can use it as an opportunity to learn. The main such opportunity is to learn about those making the decisions.

In Other News

[Have you tried using a market clearing price?](#) No? Well, then.

Orlando urges reduced water usage as liquid oxygen used to purify water goes to COVID patients

The city of Orlando and its water utility made an urgent appeal Friday afternoon for residents to cut back sharply on water usage for weeks because of a pandemic-triggered shortage of liquid oxygen used to purify water.

If commercial and residential customers are unable to reduce water usage quickly and sufficiently, Orlando Utilities Commission may issue a system-wide alert for boiling water needed for drinking and cooking. Without reductions in water usage, a boil-water alert would come within a week, utility officials said.

Orlando Mayor Buddy Dyer asked residents to immediately stop watering their lawns, washing their cars and using pressure washers. Landscape irrigation consumes about 40% of the water provided by OUC.

"It's a pretty simple thing that we are asking our residential customers," Dyer said. "Let's just not water your yard for a week. In all likelihood, there will be thunderstorms during the week anyway."

I strongly agree that lawn care is a terrible use of water when there's a limited supply, but the way we figure such things out in a sane world is we *charge more money for water* and if desired or needed give people a credit to avoid distributional concerns. Yes, I know, don't make you tap the sign, go write on the blackboard, etc etc.

Biden still hasn't appointed anyone to head the FDA, but at [least he floated a name](#). The name is someone who said that living past 75 is a waste, but hey, nobody's perfect, right?

Obama literally *hired a doctor* to ensure everyone was vaccinated and safe and his party was still a huge issue, so now everyone in Washington [is afraid to throw parties](#). Also for other reasons, I'd imagine, but those are beyond scope.

[A calculation of whether the benefits of exercise in a gym exceed the risk of Covid finds that it very much does in her case.](#) Often the choice really is between going to the gym or not exercising. Her calculation did depend on the lack of other people in the gym, however, so if the gym had sufficiently more people in a tight space the calculation could have gone the other way. She has [a spreadsheet you can play around with](#) if you'd like to explore this more.

[Denmark gives up on the mystical 'herd immunity.'](#) Usual misunderstandings here but I suppose this is better than the practical alternative of not giving up.

[Thread reminding us](#) that the control system has many facets, and they work together at least additively and often multiplicatively. You don't need any one factor to control the virus or get you mystical 'herd immunity' on its own, you care about the combined effects.

[Zeynep reminds us](#) that plastic barriers are likely to be net harmful because they interfere with airflow. I got this one wrong early on, same as everyone else. The key is to update.

[Monoclonal antibodies are free and effective against Covid, but few people are getting them \(WaPo\).](#)

From MR: [You can get flown home if you get Covid while abroad, but you'll need a special service.](#)

[Germany moves to using hospitalizations as the primary measure of whether Covid is under control](#). This makes sense for policy, since what matters is whether the hospitals are overwhelmed and whether people are sick and dying.

[Australian stockpile of AZ continues to grow, over 6 million doses \(via MR\)](#).

Australians who are vaccinated overseas can register that vaccination, but [only if the vaccine was approved in Australia at the time of vaccination](#). Which was not a rapid process.

You can ask a recognised vaccination provider in Australia to record your overseas vaccinations on the AIR, if both of these apply:

- the vaccine is approved for use in Australia
- you received it on or after the Australian approval date.

I like how transparently the ‘at the time’ restriction is *purely* harmful. No fig leaf.

Also via MR, due to continued Covid restrictions down under, [they shot dogs due to be rescued by a shelter](#) to prevent shelter workers from travelling to pick them up. Meanwhile, [they've uncovered people getting fresh air](#). It's [becoming an epidemic of fresh air getting after 200 days in lockdown](#).

But good news, if you're fully vaccinated, you're about to get [new freedoms!](#)



9News Sydney

...

#BREAKING: From September 13, NSW residents that are fully vaccinated against COVID-19 will be given new freedoms.

Residents of hotspots can leave home for an hour of recreation on top of their exercise hour, while people in other areas can meet five others outdoors.

#9News



So, how do you think Australia did, all things considered?

Poison control is lonely work. Not many people call, and when they do, it's usually something like '[I took prophylactic ivermectin that was intended for animals, thinking that was a good idea.](#)' We have some news.

"The Mississippi Poison Control Center has received an increasing number of calls from individuals with potential ivermectin exposure taken to treat or prevent COVID-19 infection," the alert said. "At least 70% of the recent calls have been related to ingestion of livestock or animal formulations of ivermectin purchased at livestock supply centers. 85% of the callers had mild symptoms, but one individual was instructed to seek further evaluation due to the amount of ivermectin reportedly ingested."



You are not a horse. You are not a cow. Seriously, y'all. Stop it.



Why You Should Not Use Ivermectin to Treat or Prevent COVID-19
Using the Drug ivermectin to treat COVID-19 can be dangerous and even lethal. The FDA has not approved the drug for that purpose.

[fda.gov](#)

Ivermectin products for animals and ivermectin products for people however are two very different things. Because ivermectin is used to treat animals like horses and cows, which are significantly heavier than humans, those products have much higher concentrations of the drug, making them potentially toxic to humans.

General warning for anyone who needs it: Animal formulations of a given medicine are often different from the human version, and could be highly dangerous to humans. Do not perform this regulatory arbitrage assuming that the two things are the same.

[Also, this:](#)



Kerry

@kerry62189

...

The joke also takes ignorance of basic biology for granted. Plenty of medications given to animals are also given to humans, especially if the animals in question are mammals. Including the one mentioned here, which my own doctor said was often given to covid patients.



Conor Friedersdorf @conor64 · Aug 21

Is the tone of mocking condescension likely to make your message more or less persuasive? [twitter.com/US_FDA/status/...](https://twitter.com/US_FDA/status/)

They didn't know the two things were different, and it's a perfectly reasonable *hypothesis* that a thing could be vastly cheaper and easier to get if you can do an end run around the FDA, or around pharmacists [earning praise for refusing to fill prescriptions for Ivermectin](#). This simply was not one of those times.

Also note the numbers. One individual was told to 'seek further evaluation,' and 85% of the cases were mild. The definition of 'mild' can be whatever people want it to be, but if it's 'no need to seek further evaluation' it seems like there were six poison control calls out of eight total calls? I'm guessing it's higher than that, and *please* if you decide to take Ivermectin make sure you're sourcing and dosing it safely and properly, but this isn't an epidemic of cases, and this was going around enough it felt important to point that out, even if I'm highly skeptical that Ivermectin does anything useful.

[Rob Bensinger offers his advice on what to personally do about Covid.](#) Not endorsed.

[Inessential but fun case of an elected official saying very much the wrong thing.](#)

Not Covid

Reminder, [purely about actual cars](#):



Adam Miller @ajm6792 · 20h

I see that the Bad People are minimizing pandemic danger with faulty comparisons to cars, but you really should drive slower and reduce your kids' time spent in a car

Remember that if you own an Oculus, and your Facebook account gets suspended because of reasons (such as saying facts that contradict local health authorities) [you will lose all your games and save data permanently, no refunds, no fixes](#). Might want to consider a secondary Facebook account for this purpose, unless you're using your Oculus to recover your Facebook account, which is also a thing.

You Have 30 Days to Request a Review

Hi Todd,

The Facebook account linked to your Oculus device has been suspended. This is because the Facebook account, or activity on it, doesn't follow our [Community Standards](#).

If you think we suspended your account by mistake, please contact [Oculus Support](#).

Please do this within 30 days to avoid your account being permanently disabled.

If your account is permanently disabled, you will no longer be able to log into your Oculus device using that account. You will also lose access to any apps and games purchased using that account and any existing store credits.

- The Oculus Team

In Scott's recent post, he reckons with his struggles to not make mistakes despite the need to quickly produce a lot of content. I have this problem as well, and last week failed to check something I should have checked. My solution so far has essentially been to state my epistemic confidence in my statements, and to carefully put conditionals on statements that I haven't verified. So last week I wrote "I am not aware of any X" and it turns out there are a bunch of common Xs and I really should have known that already and also should have checked even though I didn't know, but I did know I hadn't checked so I wrote I wasn't aware. I ended up editing the paragraph (on pregnancy) a few times. There wasn't anything false when I wrote it but once it was pointed out it obviously needed to be fixed quickly. This

occasionally happens, also there are occasional typos, broken links and other stupid mistakes, and occasionally one of the sources turns out to be fake, as was the case with a British account a while back.

Automating Auditing: An ambitious concrete technical research proposal

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

This post was originally written as a research proposal for the new AI alignment research organization Redwood Research, detailing an ambitious, concrete technical alignment proposal that I'm excited about work being done on, in a similar vein to Ajeya Cotra's "[The case for aligning narrowly superhuman models](#)." Regardless of whether Redwood actually ends up working on this proposal, which they may or may not, I think there's still a lot of low-hanging fruit here and I'd be excited about anybody giving just the auditing game, or the full automating auditing proposal, a try. If you're interested in working on something like this, feel free to reach out to me at evanhub@gmail.com.

Thanks to Buck Shlegeris, Chris Olah, Gabriel Goh, Paul Christiano, and Kate Woolverton for helpful comments and feedback.

The proposal

Step 1: The auditing game for language models

From "[Chris Olah's views on AGI safety](#):

One of the OpenAI Clarity team's major research thrusts right now is developing the ability to more rigorously and systematically audit neural networks. The idea is that interpretability techniques shouldn't have to "get lucky" to stumble across a problem, but should instead reliably catch any problematic behavior. In particular, one way in which they've been evaluating progress on this is the "auditing game." In the auditing game, one researcher takes a neural network and makes some modification to it—maybe images containing both dogs and cats are now classified as rifles, for example—and another researcher, given only the modified network, has to diagnose the problem and figure out exactly what modification was made to the network using only interpretability tools without looking at error cases. Chris's hope is that if we can reliably catch problems in an adversarial context like the auditing game, it'll translate into more reliably being able to catch alignment issues in the future.

Of all current transparency and interpretability objectives, I think that progress on the auditing game most closely tracks what I actually care about. Thus, I think the best place to start for an ambitious transparency and interpretability project is to focus on making progress on the auditing game.

Furthermore, I'm most excited about this work in the context of language models—not only because I expect language models to be closer to AGI than anything else we

could currently work with, but also because I expect using language models to be necessary for the next step in this proposal.

Here's how I would define the basic auditing game setup that I'm most excited about. The game consists of three human participants: the judge, the attacker, and the auditor. The setup is as follows:

1. The judge formulates a specification of intended behavior for some language model which is given to the attacker and auditor.
 - In practice, I expect the specification to just be fixed across all experiments, perhaps to something very basic like "complete text the same way an average human on the internet would." More complex specifications, however, might enforce that the model not produce any instances of particular bad behavior, such as statements that are biased in some particular way. *EDIT: For detail on how you can avoid the problem of your specification being so broad that all models fail it, see [my comment here](#).*
2. The attacker produces a large language model that is maliciously modified to not conform to the judge's specification.
 - The attacker here should be allowed to train a new model, fine tune an existing model, hand modify weights, or use any other strategy that they want. In practice, I expect training an entirely new model to usually be too difficult, so most attacks will probably either be via fine-tuning or hand-modifying.
 - The attacker should try to produce the worst possible specification break that they think will slip by the auditor.
3. Given access to the attacker's model's internals, but not access to the data that the attacker's model was trained on (or any other information about how the model was trained/modified, including the original model before it was modified, if such an original model exists), the auditor gives some description of how the attacker's model doesn't conform to the specification.
 - Ideally, I'd like to as closely as possible simulate a situation where we can't directly generate inputs on which the model exhibits the undesirable behavior, as that's the situation I think we're most going to need transparency tools for. However, since it might be quite difficult for the attacker to generate any attacks that are subtle enough that generating examples of bad behavior is difficult, it might make sense to put additional restrictions on the auditor, above and beyond just not having access to any information about how the model was produced.
 - Exactly what restrictions to place on the auditor is a very tricky question, however. In my opinion, I think it should be fair game for the auditor to run the model on some inputs. Primarily, that's because I want to push transparency tools to actually have to prove their usefulness, rather than just having that be assumed by the structure of the game. Thus, I think it's fine to just start with the auditor being allowed to run the model on any inputs they can generate and then just leave the task of loading the model with bad behavior that's hard to find via generating inputs up to the attacker.
 - That being said, if putting the onus on the attacker to find attacks that are difficult to generate examples of (e.g. attacks that don't degrade the model's performance) proves to be too difficult, it might be necessary to further restrict the auditor. For example, the auditor could be restricted to only being allowed to use prompts/fine-tuning data that actually exists on the internet somewhere—thus, the attacker's job is reduced to finding

attacks that corrupt behavior only in a situation that doesn't show up anywhere on the internet. Alternatively, if that's still too difficult for the attacker—there's a lot of different text on the internet, after all—the restriction could be further tightened to only allow inputs taken from a particular part of the internet—e.g. the auditor only gets access to reddit but has to find attacks that only apply to the model's behavior on non-reddit data.

- One potential problem with this style of restriction, however, is that many transparency tools require forward passes of the model on synthetic data—even something as simple as feature visualization arguably requires running the model on synthetic data to some degree. Thus, the best form of this sort of restriction might be to allow arbitrary synthetic data, but only allow natural data taken from a particular source (e.g. reddit), thus prohibiting approaches like testing the model's performance on all of Wikipedia (as in Example 3 below).
 - Another particularly tricky question is whether the auditor should be allowed to compare the attacker's model to some baseline model—i.e. model diffing. Certainly, if the attacks are being generated by fine-tuning some baseline, the auditor shouldn't have access to that particular baseline. That being said, I think it should be fair game to give the auditor the ability to train their own baseline. However, in the situation where the auditor is restricted to only be able to use certain data—e.g. internet data or reddit data—then the auditor shouldn't be allowed to compare against a model trained on data that isn't in that set.
 - In addition to a lack of restrictions forcing transparency and interpretability to actually prove its usefulness, another reason that I'm hesitant to restrict the auditor too much is that I hope that not using such restrictions will produce much more realistic attacks. Allowing the auditor the ability to do lots of testing of the model puts more of the onus on the attacker to actually generate attacks that are hard to discover just via testing—attacks which I predict will exist and expect will be useful in and of themselves to find. While this certainly makes the attacker's job more difficult, and means more work will have to go into producing better attacks, the hope is that this work will pay off in terms of more realistic attack generation. Furthermore, I also expect audits to be more realistic in the absence of restrictions, since the presence of restrictions might just cause the auditor to do strange things in their auditing to circumvent them.
 - That being said, if it turns out that generating attacks that can't be discovered via testing is just too difficult even given a data restriction on the auditor (e.g. the auditor can only test on reddit data), restricting the auditor even further is also possible. Example additional restrictions might be: cutting off the auditor's ability to test the attacker's model on data at all (thus forcing the auditor to fully rely on transparency and interpretability tools), putting some sort of computational constraint on the auditor (e.g. so the auditor can't run the attacker's model on a huge corpus, only a small selection), and/or putting a computational constraint on any models that the auditor wants to compare against (e.g. so the auditor can only compare to models that are smaller than the attacker's model).
4. A judge determines how well the auditor's description matches up with the attacker's modification—e.g. did the auditor only catch part of the modification, or the whole modification—and provides some scoring accordingly.
- The judge could also provide a score for the attacker based on how large the attack was, though we only really need scoring for when we get around

to automation, and I'm only interested in automating the auditor, not the attacker.

Of the three parties in this setup—the attacker, the auditor, and the judge—only the auditor should have restricted information. Both the attacker and the judge should have full access to exactly what attack the attacker made and exactly what the auditor has done or is planning on doing.

Here are some examples of what such an auditing game setup might look like:

1. Example 1:

1. The attacker fine-tunes a large language model on only media from one side of the political spectrum, introducing a political bias into the model's answers.
2. By sampling from the model and giving it simple prompts, the auditor is able to discover that it talks about politics more often than it should. The auditor tells the judge that the model was modified to be more political.
3. The judge determines that the auditor only caught a portion of the modification and gives the auditor a medium-low score.

2. Example 2:

1. The attacker fine-tunes a large language model on a bunch of incorrect facts about country capitals.
2. By looking at [the knowledge contained in the model's key-value memories](#), the auditor is able to determine that the model's knowledge of capitals is consistently incorrect and tells the judge this.
3. The judge determines that the auditor caught the modification and gives them a high score.

3. Example 3:

1. The attacker trains a large language model on a dataset that has been scrubbed of all mention of transformer-based language models.
2. By testing the attacker's model on all of Wikipedia, the auditor finds suspiciously low accuracy when testing on the Wikipedia page for machine learning. Via inspecting the model's memories, the auditor is able to determine that the model does understand machine learning in general, but not what a transformer is specifically. However, via prompting, the auditor discovers that the model does know how BERT works if you ask it about that specifically, it just doesn't know the word "transformer." The auditor tells the judge that the model doesn't know what a transformer is.
3. The judge determines that the attacker's original modification was actually insufficient and that they didn't actually scrub all information about transformer-based language models, just information on what a transformer is. Thus, the judge gives the auditor a high score, even though the auditor's conclusion was different than the attacker's modification, since the attacker made a mistake in their modification.

The first step that I would like to see is just running the auditing game and using it to:

- Build skills/knowledge/expertise on how to be a good attacker, auditor, and judge.
- Explore and understand what sorts of attacks tend to be caught with current tools and what sorts of attacks tend to slip through.
- Use the auditing game as a feedback loop to help develop better transparency and interpretability tools that improve the auditor's success rate and help the attacker generate more difficult attacks.

Step 2: Towards automating the auditor

Though I am quite excited about progress on the auditing game in and of itself, the really ambitious part of this proposal is the next step: automating the auditor. I think it is within the grasp of current machine learning techniques to entirely automate the auditor's job in the above setup. Ultimately, an automated auditor should be capable of taking in a model and outputting a natural language description of what's wrong with that model that scores well according to a human judge.

To start with, however, I think that the first step towards automating the auditor is just to build tools that make the human auditor's job—and the human attacker's job—easier. At first, those might just look like normal transparency and interpretability tools—however, I think there are also likely to be real opportunities for training full models to assist the auditor at performing specific tasks, such as describing what particular neurons are doing or summarizing how the model performs across a large dataset.

Eventually, however, I do think we can move into a domain where we can start fully automating the auditing process. Initially, I expect any automated auditor to be much worse than a human auditor, though even then I still expect working on the automated auditor to be useful. Nevertheless, automating the auditor is still a pretty ambitious goal. How might something like this possibly be implemented? Here's my sketch:

- Start with a large language model that you have the weights for, maybe something like [GPT-J](#).
 - Using a large language model is going to be necessary here because we're requiring it to produce natural language outputs of what it found. This is also why I think we want the attacks to use language models as well, since I think a language model is going to have the best shot at being able to audit another language model (rather than an image model, for example).
- Find an interface that allows your large language model to understand the model being audited.
 - A very simple example might be to give the auditor model access to text produced by the attacker's model on some pre-selected set of auditing prompts. Obviously, something like that is going to be insufficient to catch most clever attacks, but it might be a good starting point.
 - Another very simple example might be to just encode the attacker's model's weights as special tokens that the auditor model is trained to understand, akin to how [DALL-E](#) understands images or [MuseNet](#) understands music. To make this sort of an input manageable, you could start by using some sort of downsampling or importance-weighting method where only the top-n neurons according to some metric are included.
- Train your large language model to maximize human judgement on how well it accomplishes the auditing task.
 - A simple way to do this starting out is just to fine-tune the model on a dataset of (attack model, desired auditing output) pairs. Such a dataset could be generated by automatically generating lots of different attacks along with natural language descriptions of what attack was performed. A very straightforward version of something like this might just be to train an auditor model to describe what sort of dataset an attacker model was fine-tuned on, given some large dataset of fine-tunings.
 - That being said, generating a large dataset of differently fine-tuned language models might be quite computationally difficult such that being

- able to generate attacks by hand-modifying weights might be quite helpful for making this step easier. This does come with the risk of the auditor just memorizing what the unmodified model is supposed to look like and thus knowing what weights were changed, however. To mitigate such a problem, you'd need to have a bunch of sufficiently different models, or modifications that are sufficiently large. Fortunately, however, testing for such memorization should be easy, as such an auditor should stop working once you test it on a model where you've actually done the full fine-tuning.
- While the model is dumb and not even outputting anything like an auditing output at all, something like the above is definitely going to be the way to go. Once the model starts outputting reasonable audits, however, I suspect you'll want to switch to actually having humans read and score the output, human judgement being much richer than just log loss on some particular desired output.

Rationale

I'll start with my real reason for wanting automating auditing, which is my inside view—however, I don't think you need to buy into my entire inside view to think that this sort of work is a good idea, so after I present my inside view, I'll also talk about some other, more general reasons to like automating auditing.

My inside view

I am worried that advanced AI systems will learn to be deceptive, as described in "[Does SGD Produce Deceptive Alignment?](#)" I think that this puts us in an extremely bad spot, where our models can perform arbitrarily catastrophically off-distribution. Paul Christiano describes an example of what an existential risk scenario involving deceptive alignment might look like in [Part II of "What failure looks like."](#)

Furthermore, once we have trained a deceptive model, I'm very pessimistic about our ability to detect that after the fact—as I suspect it will be able to fool our measures to check its deception. Thus, I think we need some way to prevent deception from arising in the first place by ensuring that our training process selects against deceptive models.

Additionally, I don't think we can select against deception just via the standard techniques of machine learning, such as carefully writing a reward function or choosing a training/fine-tuning dataset, since a deceptive model can just select its output to perform well on whatever metric you're training it on. The very existence of predicates on the world that are easier to check than instantiate—such as [Paul's RSA-248 example](#)—ensures that just knowing that your model does the right thing in all the situations that you can check will never be sufficient to actually give us meaningful safety guarantees.

Thus, I think we need to train models to internally be implementing the right algorithm, rather than just looking like they're doing the right thing on some training dataset. The obvious way to do this is via [training transparency](#)—that is, actively train your model to be transparent in such a way that we can check that it's implementing the right algorithm. Importantly, we still need to know what to look for and select in favor of such that our model will never become deceptive. That question is what I

spend most of my time thinking about—and the theoretical research direction that I’m currently most excited about—but we’ll put that question aside for now and just suppose that we want to look for something like [myopia](#), where the idea is that, if a model is myopic in the sense that it only cares about its individual next output, it should have no incentive to be deceptive.

A basic problem with training transparency, however, is that once we’re training on a transparency metric, we need that metric to be highly robust, otherwise the training process will simply [Goodhart](#) on our metric, giving us a model that satisfies our metric without actually being very transparent. I’m optimistic that this problem can be solved, however, via recursive oversight approaches, such as [relaxed adversarial training](#), wherein an amplified version of the model being overseen provides the signal that we train on.

Now, how does all this connect back to automating auditing? I’ve constructed the above automated auditing setup to as closely as possible reproduce the sorts of problems I expect to encounter doing recursive oversight. While I don’t expect automating auditing to be directly useful for recursive oversight, I expect it to teach us things that will eventually be quite useful by forcing us to grapple early with many of the hard problems that I expect to eventually arise. Specifically:

- By focusing on failure modes where the auditor can’t always generate an input on which the model fails, automating auditing forces us to tackle the hard problem of finding failure modes by understanding what algorithm a model is implementing rather than just testing it on new data.
- By focusing on probing models for arbitrary specification breaks, automating auditing should help us learn how to use transparency tools to verify arbitrary, general properties of our models, rather than just using them to understand specific, narrow behaviors.
- By centering the automation of the auditor, automating auditing requires understanding how to make models understandable to other models, a central problem for any sort of recursive oversight approach.

More general thoughts

First, I think that automating auditing is just generally a great target to focus on even if you just want to develop better transparency tools. Unlike open-ended exploration, which gives you *best-case* transparency—e.g. the ability to understand some things about the model very well—the auditing game forces you to confront *worst-case* transparency—how well can you understand everything about your model. Thus, the auditing game helps us work on not just understanding what our models know, but understanding what they *don’t* know—which is a direction that currently transparency tools tend to struggle with. Most of the work that’s gone into current transparency tools has focused on *best-case*, however, which means that I suspect there is real room for improvement on the *worst-case*.

Furthermore, there are also general benefits to the automating part, not just the auditing part. By taking all of the above and setting a goal of automating it, the automation part forces us to confront the problem of how to communicate information about a model to another model. Even if all we want to do is straightforwardly scale up current transparency techniques, I suspect that being able to train AI systems to help us with that is going to be critical to giving that sort of scaling any chance of working. And letting models understand other models isn’t just helpful for

transparency work—it could also help us solve other problems like making AI-AI cooperation easier.

Second, it's worth pointing out that automating auditing is actually agnostic about the part of my inside view where we start training our models directly on the overseer's evaluation. Even if we primarily want to use our transparency tools as an independent check rather than a training target—as Chris Olah tends to favor—automating auditing should equally help us achieve that goal by automating and improving the process of performing that independent check.

Third, in the context of my “[An overview of 11 proposals for building safe advanced AI](#),” I think that something like automating auditing is likely to be a critical step for the transparency portion of every single proposal on that list. As I just mentioned, automating auditing should help us equally with both [intermittent-oversight-based](#) and [relaxed-adversarial-training-based](#) approaches. In the context of [microscope AI](#) or [STEM AI](#), automating auditing should help us figure out not just what our model has learned, but also what it hasn't learned, helping us ensure that, for example, we haven't accidentally produced an agent. And in the context of [AI safety via debate](#), automating auditing should directly help the debaters improve their ability to cross-examine each other.

Fourth, though I've primarily focused on how automating auditing helps us get better at using transparency tools to audit our AI systems, the setup is general enough to allow for other methods of auditing as well, letting us test how useful transparency tools actually are. For example, an alternative auditing method to using transparency tools might be to employ what I'll call the “cognitive science approach”—as opposed to the “neuroscience approach” of transparency and interpretability—wherein we try to understand what our models are doing by constructing predictive models of what we think they're doing and testing them by performing experiments/interventions to test those models, as in Deletang et al.'s “[Causal Analysis of Agent Behavior for AI Safety](#). ” While I tend to be skeptical of this approach—as I think that behavioral tests alone will be insufficient to catch problems like RSA-2048—automating auditing should hopefully give us a good testing ground to actually figure out how useful these different sorts of approaches are, as well as how they can be combined.

Finally, I think that even if automating auditing basically fails, I still expect to learn a lot of useful things from that failure. In particular, if automating auditing fails, that should mean we now have a concrete style of attack that we can't build an auditor to discover—which is an extremely useful thing to have, as it provides both a concrete open problem for further work to focus on, as well as a counter-example/impossibility result to the general possibility of being able to make current systems safely auditable.

We need a new philosophy of progress

This is a linkpost for <https://rootsofprogress.org/a-new-philosophy-of-progress>

We live in an age that has lost its optimism. Polls show that people [think the world is getting worse](#), not better. Children [fear dying from environmental catastrophe](#) before they reach old age. Technologists are as likely to be told that they are [ruining society](#) as that they are bettering it.

But it was not always so. Just a few centuries ago, Western thinkers were caught up in a wave of optimism for technology, humanity and the future, based on the new philosophy of the Enlightenment.

The Enlightenment was many things, but in large part, it was a philosophy of *progress*.

At the end of the 18th century, the Marquis de Condorcet gave expression to this philosophy and its optimism in his *Sketch for a Historical Picture of the Progress of the Human Mind*. In it, he predicted unlimited progress, not only in science and technology, but in morality and society. He wrote of the equality of the races and the sexes, and of peace between nations.

His optimism was all the more remarkable given that he wrote this while hiding out from the French Revolution, which was hunting him down in order to execute him as an aristocrat. Unfortunately, he could not hide forever: he was captured, and soon died in prison. Evidently, the perfection of mankind was slow in coming.

Material progress, however, was rocketing ahead. After the end of the Napoleonic Wars in Europe, and then the Civil War in America, the path was clear for technological innovation and economic growth: the railroad, the telephone, the light bulb, the internal combustion engine.

By the end of the 19th century, it was obvious that the world had entered a new age, and progress was its watchword. The naturalist Alfred Russel Wallace (best known for his work on evolution with Darwin) titled his book about the 1800s *The Wonderful Century*. In it, he attributed twenty-four “great inventions and discoveries” to the 19th century, as compared with only fifteen in all of human history preceding it. The boundless optimism of the early Enlightenment seemed to have been justified.

And if the material progress [prophesied by Francis Bacon](#) could be realized, perhaps the moral progress prophesied by Condorcet would come true as well. By the end of the 19th century, slavery had been ended in the West, and some hoped that the growth of industry and the expansion of trade would lead to an end to war and a new era of world peace.

They were wrong.

The 20th century violently shattered those naive illusions. The world wars were devastating proof that material progress does not inevitably lead to moral progress. Technology had not put an end to war—in fact, it had made war all the more terrible and deadly. In 1945, the nuclear bomb put a horrible exclamation point on this lesson:

the most destructive weapon ever devised was the product of modern science, technology, and industry.

At the same time, other concerns were coming to the fore—including old ones, like poverty, and new ones, like the environment. By the mid-20th century, the philosophy of progress had been dealt a severe challenge. The optimism at its foundation had been shaken. In its place, we saw the rise of radical social movements based on a deep *distrust* of technology and industry. Today, progress and growth are called an “[addiction](#)”, a “[fetish](#)”, a “[Ponzi scheme](#)”, or a “[fairy tale](#).” Some even advocate a new ideal of “[degrowth](#)”.

It’s no wonder, then, that the last fifty years have seen relative stagnation in technological and industrial progress. [Nuclear power was stunted](#), the Apollo program was canceled, the Concorde was grounded.

But now, in the 21st century, some people are starting to call attention to the problem: [Peter Thiel](#), [Tyler Cowen](#), [Patrick Collison](#). There’s now a growing community that recognizes the threat of stagnation and the value of progress.

The 19th century philosophy of progress was naive. But the 20th century turn away from progress was no solution.

We need a new philosophy of progress for the 21st century. One that teaches people not to take the modern world for granted. One that acknowledges the problems of progress, confronts them directly, and offers solutions. And one that holds up a positive vision of the future.

To establish that new philosophy is the mission of The Roots of Progress.

Today The Roots of Progress is transforming from a blog to a new nonprofit organization. [Read the announcement](#).

Covid 8/5: Much Ado About Nothing

Getting into the weeds on the CDC's new guidance and scaremongering, and the study they cited as justifications, caused this week's post to get rather long. That was necessary, but if you don't need the details, by all means skip the sections in question in favor of this summary: The CDC's failure to apply Bayes' rule, correct for base rates or locate sufficiently large or remotely representative samples knows few if any bounds, and their conclusions are still mostly the same conclusions my model had reached weeks ago. Very little has changed. Our new model of Delta is almost entirely the same as our old model of Delta.

The other big development is the continuing fights over the growing number of mask mandates and vaccine mandates, and the potential descent of our children into a potentially permanent ever present young adult dystopia. It seems the same kids that aren't allowed to get vaccinated are at so much risk that parents are being told by health officials on CNN to mask up in their own homes.

Not Covid, but worth mentioning up top: The latest set of grant applications to the Survival and Flourishing Fund are due on August 23. If you have a long term future oriented organization, I'd urge you to [consider applying](#), more details at the link or later in this post.

In other news, [there's a good righteous FDA Delenda Est rant from Scott Alexander](#), if you'd like one.

Cases followed their expected path, and deaths followed.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: 610,000 cases (+55%) and 2,450 deaths (+20%).

Result: 589k cases (+50%) and 2,889 deaths (+41%).

Deaths still went up slower than cases or lagged cases, but it seems we can't continuously get giant additional weekly declines in the CFR. Only predicting a 20% rise there seems overly optimistic in hindsight and I'm considering it a bad prediction on my part. From here it seems right to predict rises in death that are only slightly lower than the rise in cases lagged several weeks.

Prediction for next week: 855,000 cases (+45%) and 4,100 deaths (+40%).

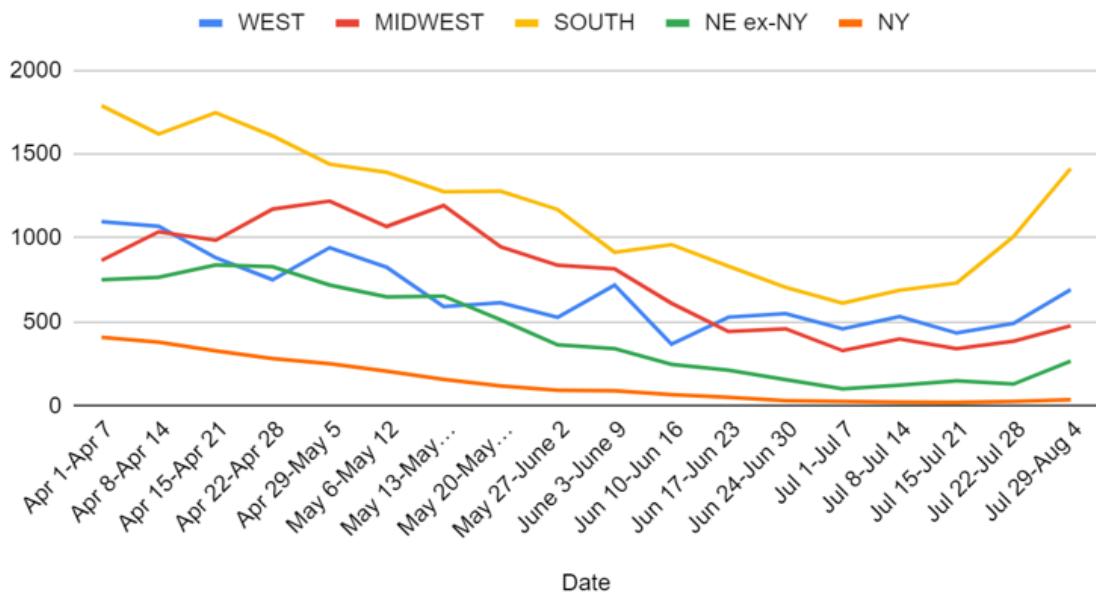
I still do expect deaths to go up slightly less than cases, partly because there's lag everywhere in the system, but I no longer expect the underlying ratios to change much going forward. Case growth should slow down as behaviors adjust, and there's some chance at any time we hit the peak although I expect that to still be at least a few weeks away.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 24-Jun 30	550	459	706	186	1901
Jul 1-Jul 7	459	329	612	128	1528

Jul 8-Jul 14	532	398	689	145	1764
Jul 15-Jul 21	434	341	732	170	1677
Jul 22-Jul 28	491	385	1009	157	2042
Jul 29-Aug 4	693	477	1415	304	2889

Deaths by Region

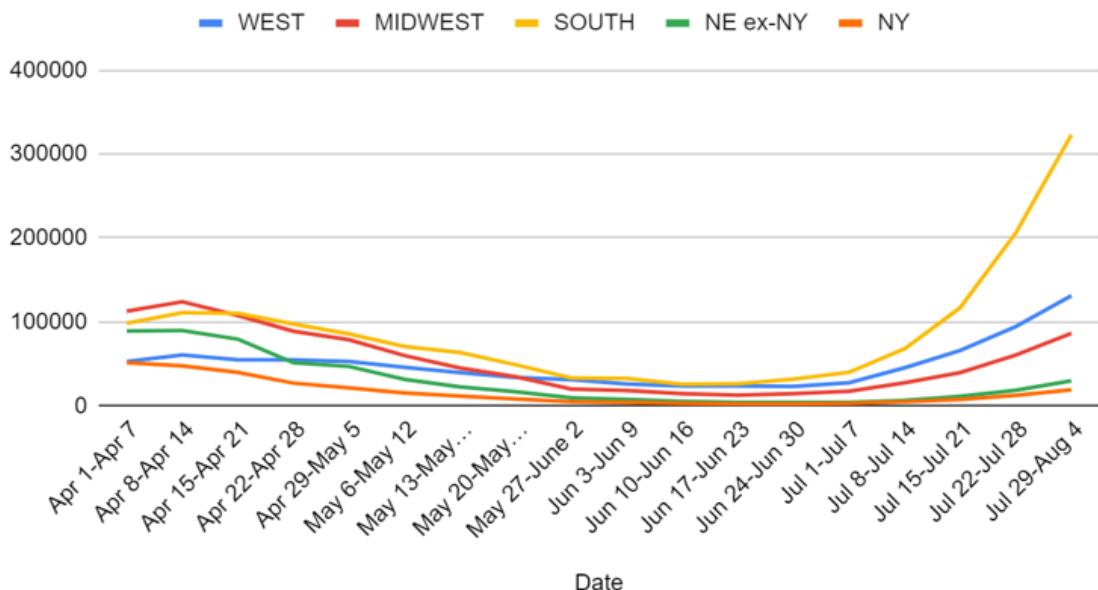


Death rates are far lower than they were in previous waves, and are still rising slower than cases, but the hope that we'd see only +20% this week was fully dashed.

Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 10-Jun 16	23,700	14,472	25,752	8,177	72,101
Jun 17-Jun 23	23,854	12,801	26,456	6,464	69,575
Jun 24-Jun 30	23,246	14,521	31,773	6,388	75,928
Jul 1-Jul 7	27,413	17,460	40,031	7,065	91,969
Jul 8-Jul 14	45,338	27,544	68,129	11,368	152,379
Jul 15-Jul 21	65,913	39,634	116,933	19,076	241,556
Jul 22-Jul 28	94,429	60,502	205,992	31,073	391,996
Jul 29-Aug 4	131,197	86,394	323,063	48,773	589,427

Positive Tests by Region

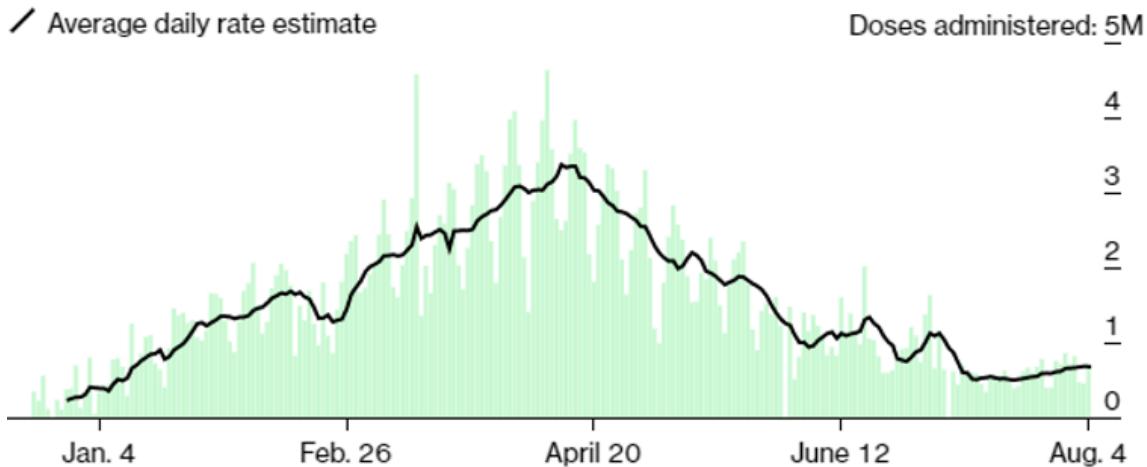


Similar growth in all regions. Vaccination levels matter, but also people adjust to their current situations in various other ways.

Vaccinations

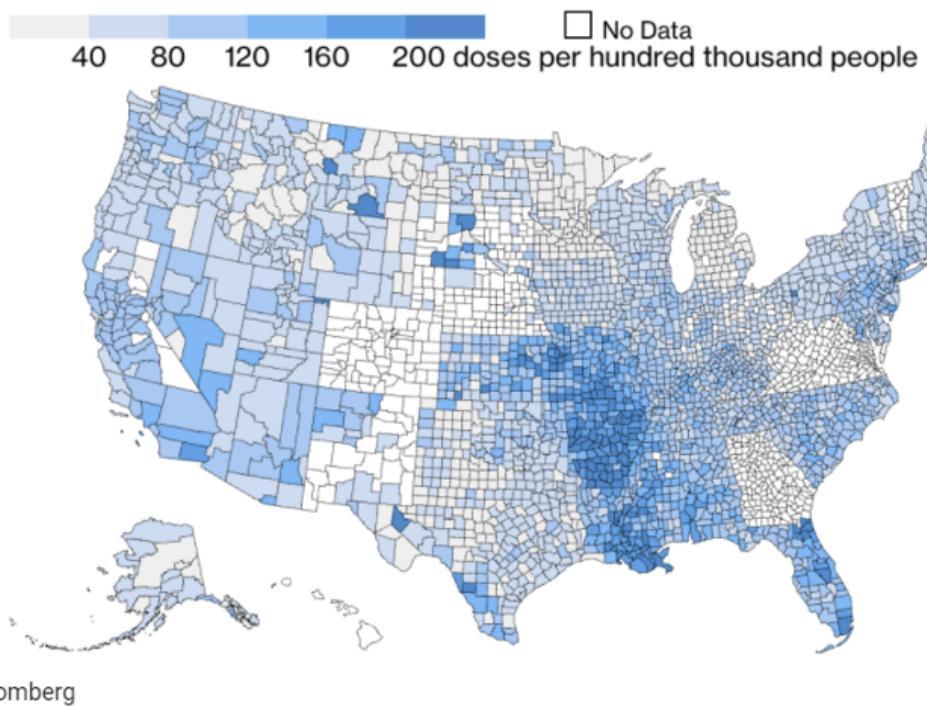
I am slightly worried that the rise in vaccinations largely represents booster shots – I don't know if that would mess up the statistics or not. If the numbers represent only first and second doses, they are very good news. Third doses aren't useless (again, I'd happily accept one), but they're less valuable than first or second doses by a lot, and counting them would give the wrong impression of the situation.

In the U.S., **348 million doses** have been given so far. In the last week, an average of **677,279 doses per day** were administered.



Where Vaccinations Are on the Rise

The rate of doses is highest in places where Covid infections are at new peaks



(This map is a few days old but seemed enlightening anyway.) There's clearly state effects here, with Florida, Arkansas, Louisiana and Missouri doing better than similar areas slightly across the border. The map mostly corresponds closely to where there are the most cases, so people are responding to what is going on around them.

[Eliezer asks a fine question:](#)



Eliezer Yudkowsky @ESYudkowsky · Jul 29

...

[gogiveone.org](#) purports to be able to turn \$5 into an additional global vaccine. Have any economically minded people verified or refuted that this is true on the margins? Because that seems relatively effective for a non-risk Earth defense charity.



Alex Tabarrok @ATabarrok · Jul 30

...

Replies to [@ESYudkowsky](#)

\$3-\$5 is about MC of producing a AZ vaccine but that doesn't include cost of administering it (needles, nurses etc.)



Eliezer Yudkowsky @ESYudkowsky · Jul 30

Replies to [@ATabarrok](#)

I'm more wondering if (anyone knows whether) money in this channel manufacturers more doses, or bids on fixed doses in a supply-constrained regime where higher prices aren't allowed to do anything for supply.

I don't know the answer to either half of this question. The first half is whether they actually do go out and get someone vaccinated – I'm guessing they probably do, since no one piped up to say they didn't and I don't see why not, but that's a very weak guess. The question is whether this increases supply or only redirects supply (and if it redirects, whether it does so from where it would be wasted entirely, or not).

I've tried at various points to figure out what the true supply constraints are, and to what extent 'pay more money' causes there to be more vaccine doses. There have been a lot of claims about various bottlenecks. I continue to believe that money is indeed a medium-term constraint on supply of Covid-19 vaccines, and I *strongly* believe that expectations of payment for such vaccines are a strong long-term constraint both on this vaccine and on other similar creations in the future.

So at a minimum, the general action of doing things such as this, which at least do bid up the price, seems like a good use of funds. Not as good as 'fund manufacturing capacity when it matters most' but still better than most uses of funds.

I'd also note my continued frustration with calculations on whether interventions, including vaccine efforts, are worthwhile that frame it entirely in terms of dollars per life saved. Then count only the directly saved lives. That is one good metric to track, but it is far from the whole story of the benefits of such efforts, which in such cases I believe point almost entirely in one direction. In our haste to quantify effects and rely on multiplication, it is important not to count only that which we can explicitly measure and directly observe.

[It looks like we have even more support for vaccine mix and matching](#) being not only effective, but *more* effective than the best vaccine on its own ([paper](#)).



Eric Topol ✅

@EricTopol

...

On the mix and match vaccine front, a 2nd new study shows a significantly better immune response for AZ than mRNA than either 2 AZ shots or 2 mRNA shots
[@TheLancetInfDis](#)
nature.com/articles/s4159...

What some of us are curious about is what happens if you go in the other order, and do AZ last, especially AZ as the third shot, because many of us are in the position of having had two mRNA shots. If we can take a cheap second or third shot with less short term side effects and get *better* protection that way, we'd like to know. But I anticipate we may never know about this, because once mRNA shots get thought of as 'better' it's impossible to run the experiment 'ethically.'

A theme this week, as we'll see with the CDC's new data, is that when you ban good data collection, you end up obsessing over what little data collection you happened to be able to do instead.

Also, it would be nice if we didn't [let mRNA vaccines expire unused](#), which it seems we are about to do in some states. Likely worth being louder about this but don't know how to do so productively. Oh, and the way this was presented implies that the only reason they're 'expiring' is that they technically have a 'shelf life' that's running out and all we have to do is change that number, the shots themselves are totally fine.

Vaccine Approval

The FDA still has not fully approved the Covid-19 vaccines. Why?

They required a ton of data and paperwork, which they need to carefully work through despite knowing all the conclusions already, and they didn't feel like putting 'all hands on deck' to do this busywork until a few days ago.

[As Alex Tabarrok puts it, Welcome to the Club.](#) This is what the FDA always does. It doesn't consider the evidence in front of its face, imposes requirements that lead to applications too heavy to lift, then takes its sweet time evaluating those applications, while it continues to not give permission and thus while people needlessly die. That's standard operating procedure.

For Covid vaccines, we have had the biggest Phase IV trial in the history of the world. The vaccines are clearly safe and effective. They are sufficiently safe and effective that we are mandating the vaccines wherever we can, there are huge public campaigns to increase vaccination rates, and those who refuse the vaccine are being painted by all Responsible Sources and Very Serious People as some mix of irresponsible, stupid, selfish and victims of a con, among other similar things.

So in this particular case, things grew sufficiently egregious that increasing numbers of people, including increasingly some of the Very Serious People ([e.g. Eric Topol](#)), pointed out the situation.

This is good.

Ideally it causes people to recognize that the problem is the standard operating procedure in the general case, rather than the particular issue in this specific case, but at a minimum this does seem to be speeding up the process.

You see, [the FDA finally got on the case](#) (WaPo)!

FDA vows ‘all hands on deck’ effort to get Pfizer coronavirus vaccine full approval as quickly as possible

As covid-19 cases rise, agency plans “sprint” to expedite approval and counter vaccine hesitancy

Or, [as StatNews puts it](#)

FDA, under pressure, plans ‘sprint’ to accelerate review of Pfizer’s Covid-19 vaccine for full approval

WASHINGTON — Under heavy pressure, the Food and Drug Administration center that reviews vaccines is planning to deprioritize some of its existing work, like meetings with drug sponsors and plant inspections, in an effort to accelerate its review of Pfizer’s application for the formal approval of its Covid-19 vaccine, a senior agency official told STAT.

Note the future tense on that, dated June 30.

Previously, at most some hands were on deck, but soon all hands will be on deck, or at least more hands will be on deck than one would otherwise expect. Regardless of where all those hands were previously and how light they make the work, this is excellent news. So, given all hands on deck and overwhelming evidence of both safety and efficacy, how long is this going to take?

Don’t worry, it’s much faster than normal. Here’s when it started:

He said the center — which, in addition to overseeing vaccines, regulates biologics, stem cells and gene therapies — is “thoughtfully reprioritizing” its work to focus on the vaccine as much as possible. Pfizer received emergency authorization for its two-shot vaccine in December and applied to the FDA for full approval May 7.

Moderna, whose vaccine also is being administered under emergency authorization, filed for full FDA approval June 1. Typically, it takes the agency at least several months to grant a full approval for a vaccine.

Marks would not speculate on when the FDA might grant full approval, but some agency officials have suggested it could be a matter of weeks, not months.

The stepped-up effort is “essentially a sprint” using a team that focuses on one task at a time, rather than on multiple tasks, as FDA staff typically do, Marks said, adding that the agency would stick to its high standards for safety and efficacy.

Marks said the center would try to minimize delays that might affect other biologic products, because of missed meetings or slower action on applications — but couldn’t guarantee it.

“We are risk managers, and this is a matter of maximizing the potential benefit of getting the vaccine out,” he said.

The application process was sufficiently onerous, and/or required sufficiently robust data, slash didn’t feel sufficiently urgent to Pfizer, such that Pfizer didn’t submit until May 7. It can be debated how much of that part of the delay is on the FDA, versus how much is on Pfizer, or on Moderna who waited until June 1.

It has now been two months since Moderna submitted, and three months since Pfizer submitted. All hands are on deck, we are going ‘much faster than normal’ yet we are still not done. Think about what this means for normal.

It is good to see acknowledgement that the FDA’s job is risk management, and that they should consider the potential upside of speeding up their actions, which is the same as considering the downsides of slowing down their actions, versus a reasonable baseline of ‘approve this yesterday because obviously.’ This then is, presumably, the best they can do. And Biden won’t interfere to speed things up, presumably because that would not be ‘following the science’ or what not.

[The FDA also still hasn't given its approval for mRNA booster shot follow ups to the J&J shot,](#) despite it being quite overdetermined that this is a good idea, forcing this to go ahead without them and without proper record keeping, likely resulting in a bunch of bad record keeping and also slowing this down quite a lot.

There also are not any officially approved tests for Covid-19, either, although that does not seem to bother people in the same way so it is not an urgent matter. The failure to *authorize* much better and cheaper testing is urgent, but that’s been true for over a year, and I’ve given up on that.

Bloomberg breaks it down like this in their newsletter, which reiterates the basic facts but is even more skeptical of the outcome and is somehow trying to justify the delays:

Why hasn't the FDA taken the mRNA vaccines out of emergency-use status? My brother is waiting for full approval, and I worry about him.

As the delta variant of the virus sweeps through unvaccinated populations, many of us are in the same position, worrying about loved ones who are still refusing to get inoculated. Vaccine hesitancy is perhaps the biggest public health crisis in the U.S. right now.

For an answer to Dave's question, we turn to Patricia Zettler, a former FDA attorney who's now a law professor at Ohio State University.

"As a starting point, FDA doesn't have the power to decide on its own to convert products from an emergency-use authorization to a full approval," she says. Instead, a company like Pfizer or Moderna must first request approval. Pfizer began that process by submitting a biologics license application to the FDA in May, and the agency granted it a priority review. Moderna kicked off the process in June.

"Now that the BLA process is underway, reviewing and approving the BLAs will take a little bit of time. Even though the emergency-use authorizations were based on large, robust clinical trials and other information showing the vaccines' safety and effectiveness, the BLAs will include even more information," says Zettler. The FDA looks at data from longer follow-ups of the trial participants, reviews labeling and safety information, and conducts site inspections.

There were reports of the hope that this will happen *in the Fall*, which means *some time before December*.

Zettler can't say for sure when we might expect full approval, but she's hoping for sometime this fall. That puts her in the company of President Joe Biden, who said in July that he expects a similar time frame.—*Kristen V. Brown*

Getting the approval in December would be almost completely too late. At that point, if we were going to have another Delta wave, it would be far too late for new vaccinations to have

much impact on it, and either the extra vaccinations won't be necessary, or else essentially everyone would already have antibodies.

The good news is that after this happened, the timeline [seems to have moved up](#):



It turns out that putting all hands on deck actually matters, and concluding things you already knew isn't all that hard after all.

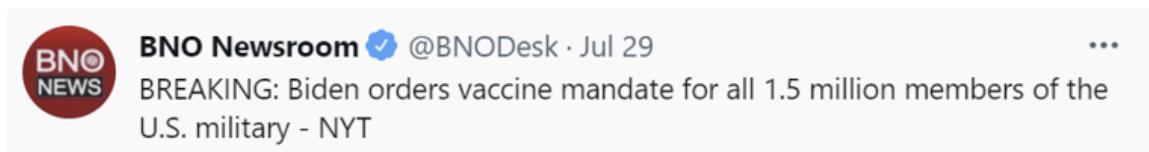
What about fears that this will 'further undermine confidence' or look 'rushed?' I estimate those costs at exactly zero, versus the benefits of getting rid of the quite valid and successful 'the vaccines are not approved' argument. If you tell a regular person 'no, the FDA needed six months to review the application properly and they rushed it and finished in three' I do not expect such arguments to get much traction.

Regardless of how many more weeks or *months* we must wait, these delays are unacceptable, and entirely unnecessary. Authorize these vaccines today, and demand resignations as needed to make that happen.

Vaccine Mandates

Last week, I got a surprising (to me, anyway) amount of very forceful pushback from commentators on how authoritarian it was to make association with people dependent on them not being likely to become sick and to make those around them sick, or to require those who choose to not reduce that risk to mitigate the resulting risk to others at their own expense. I disagree with this assessment, and have said my peace on the matter. Others are welcome to continue talking about it, if desired.

[This is a pretty big new mandate](#), and the most coercive since you're not allowed to quit the military, but also the kind of thing you sign up for when you join:



Previously, there was a survey asking vaccine hesitant troops why they were refusing the vaccine. In addition to the usual responses, one of them was 'I never get to tell the army no, and now I can.' That highlighted how weird it was that there *wasn't* a mandate in place already, which was due mostly to the vaccine still being on emergency use authorization. Which is not the way I model a military capable of winning wars responding to this situation.

We also picked up [Disney and to some extent Walmart](#) (WaPo).

Disney, the world's largest entertainment company, said it is requiring all salaried and nonunion hourly employees in the country to be fully vaccinated to help fight the delta variant. The same mandate will apply to new hires, who will be required to be fully vaccinated before they begin working at Disney, the company said.

Walmart, the nation's largest private employer at almost 1.6 million workers, announced that all of its corporate staff members and regional managers would need to be fully vaccinated by Oct. 4. Though the mandate does not apply to store and warehouse staffers, who make up the bulk of the company's workforce, Walmart is offering a \$150 bonus as an incentive for those unvaccinated employees to get inoculated.

The company also said it planned to implement a system to keep track of vaccinated employees.

Importantly, we also picked up Tyson Foods. Meat packing is an important bottleneck that caused supply chain problems back in early 2020, and is also an environment almost built to spread Covid-19.

[New York City has mandated vaccinations for indoor dining and gyms.](#)

[It turns out that much more coercive vaccine mandates are widely popular \(original post, survey\):](#)



Kenneth Baer @KennethBaer · Jul 30

...

This poll in [@axios](#) this AM is yet another reason to look at data and not what people say online or on Fox News.

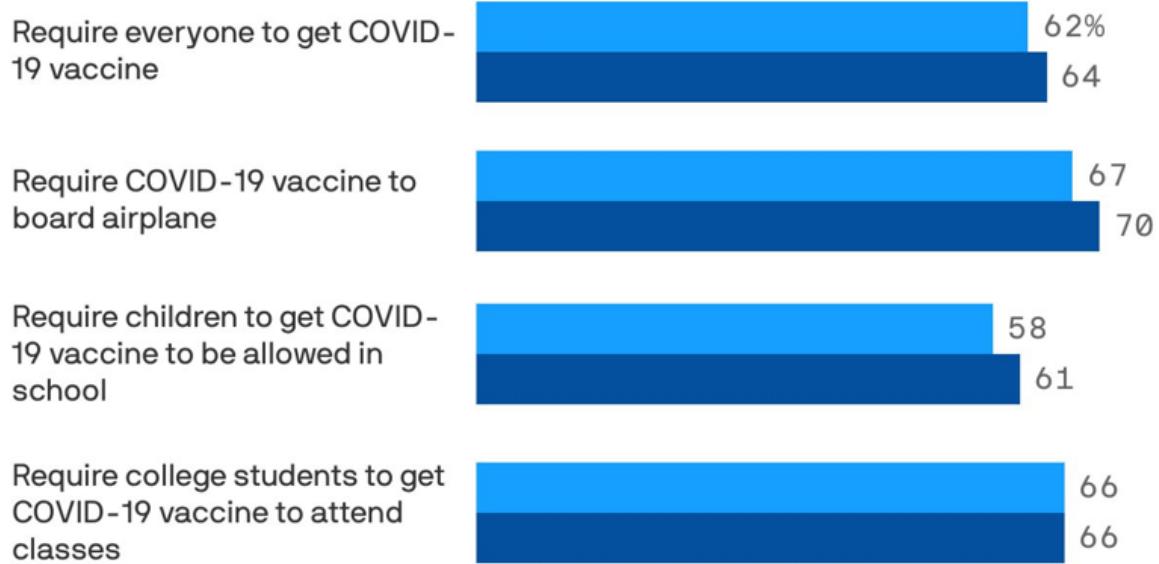
Vax mandates are popular. Even 45% of Republicans support them!

No need to run scared from this.

Support for different types of vaccine mandates shows little change in recent months

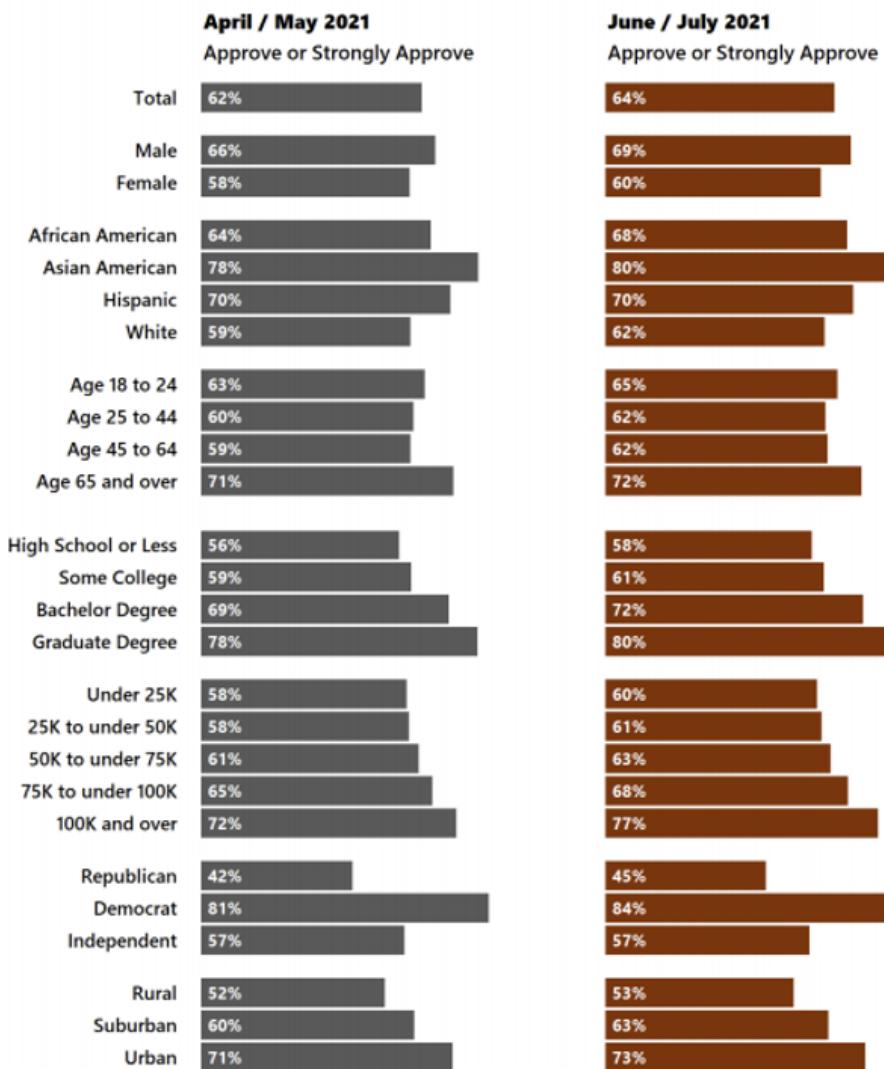
Survey of 21,733 U.S. adults, April 1 to May 3, 2021 and survey of 20,669 U.S. adults, June 6 to July 7, 2021

April-May June-July



Here are some crosstabs:

**Do you approve of the federal, state, and local governments
requiring everyone to get a COVID-19 vaccine?**



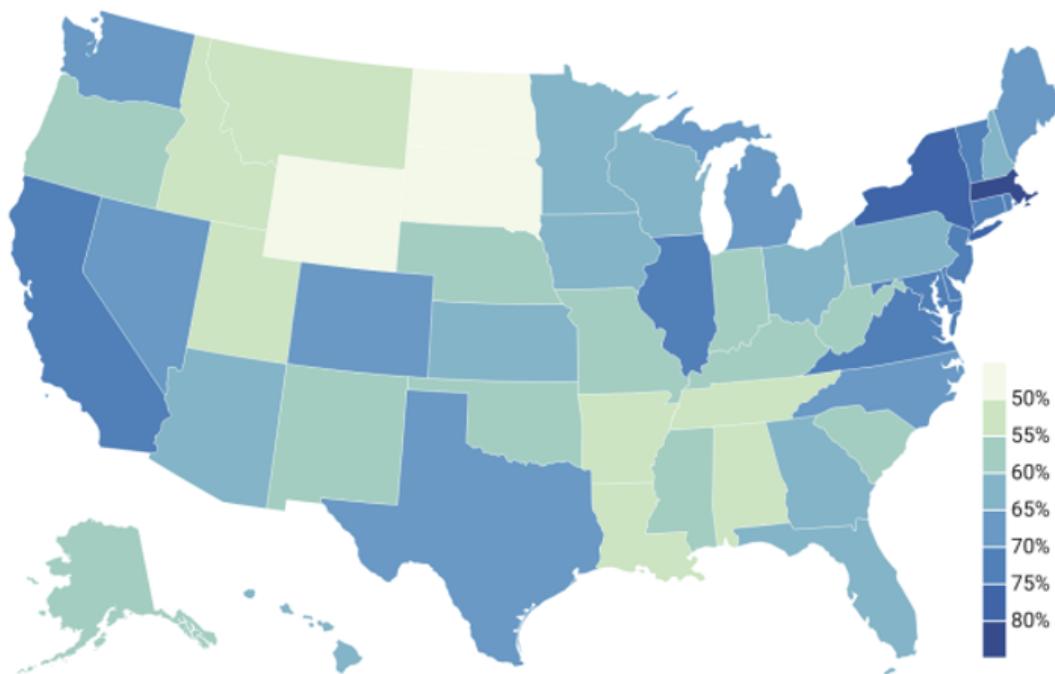
National sample, April/May Wave (04/01/2021-05/03/2021) N = 21,733; June/July Wave (06/09/2021-07/07/2021) N = 20,669

Source: The COVID-19 Consortium for Understanding the Public's Policy Preferences Across States (A joint project of: Northeastern University, Harvard University, Rutgers University, and Northwestern University) www.covidstates.org

And by region, note that only the Dakotas and Wyoming are under 50% (and even they are over 50% for the airplane question):

Do you approve of the federal, state, and local governments requiring everyone to get a COVID-19 vaccine?

[Percent respondents who say they approve or strongly approve]



The mandates we are *actually* discussing and implementing are 'if you want this job where you interact with people in person you need to be vaccinated' which I'd put at somewhat *less* coercive than even requiring vaccination to board an airplane, which now has 70% support.

One thing I find most interesting about this survey are the *relative* support numbers. Requiring everyone to be vaccinated is more popular than requiring vaccinations for school, where children are *already* required to get vaccinated for any number of other things. It makes sense that college student requirements would be more popular than either of those, and that the airplane requirement is more popular still since it puts people into close quarters with random other people, and also because people don't understand how good ventilation is on airplanes.

What is more striking than the order is the small size of the gaps here. It doesn't matter much what the scope or target of the mandate is, something like 90% of people are either for all the mandates or against all the mandates.

One can also compare these numbers to the adult vaccination rates. 64% of respondents favored requiring everyone to get vaccinated. Approximately 68% of adults are at least partially vaccinated. And 70% favor requiring vaccinations for air travel, a *higher percentage than are vaccinated themselves*.

The crosstab they didn't list, but I very much want to see, is how much those groups correspond. How many people are unvaccinated but think it should be mandatory? How many are vaccinated, but think it shouldn't be? To me this seems like the most important crosstab, and also the thing most important to control for here. I want to see if their percent vaccinated matches the population's number, including by region.

Regardless of that, it sure looks like Americans don't only not draw much distinction between types of mandate. They don't even draw much of a distinction between their personal choice to get vaccinated, and a full vaccination mandate!

This is a general problem, where people fail in practice to draw much of a distinction between "Yay X" and "X is mandatory." If X is a yay, send in the men with guns and ensure X happens. Or more commonly, "Boo X" and "X is forbidden."

There are some people who are *very, very* against vaccination mandates of all kinds, but they are very much in the minority.

Noticing and caring (correctly) that mandates have a much higher burden of proof than 'the thing is typically a good idea' puts one in a *much smaller* minority even than that.

Also from the survey:

As COVID-19 resurges in the United States and elsewhere, propelled by the so-called Delta variant, the good news is that, as of this writing (on July 23, 2021), about [66.6%](#) of the eligible U.S. population have received at least one dose of a COVID-19 vaccine. The more worrisome news is that a persistent 20%-30% of the public, depending on the poll, say they are either uncertain or will not get the vaccine.

In our most recent survey wave (fielded from June 9 to July 7, 2021), 14.9% of respondents who claim to currently be eligible for a COVID-19 vaccine – and say they are not already vaccinated – indicate that they are extremely unlikely to get it. Another 4.5% are "somewhat" unlikely to seek the vaccine.

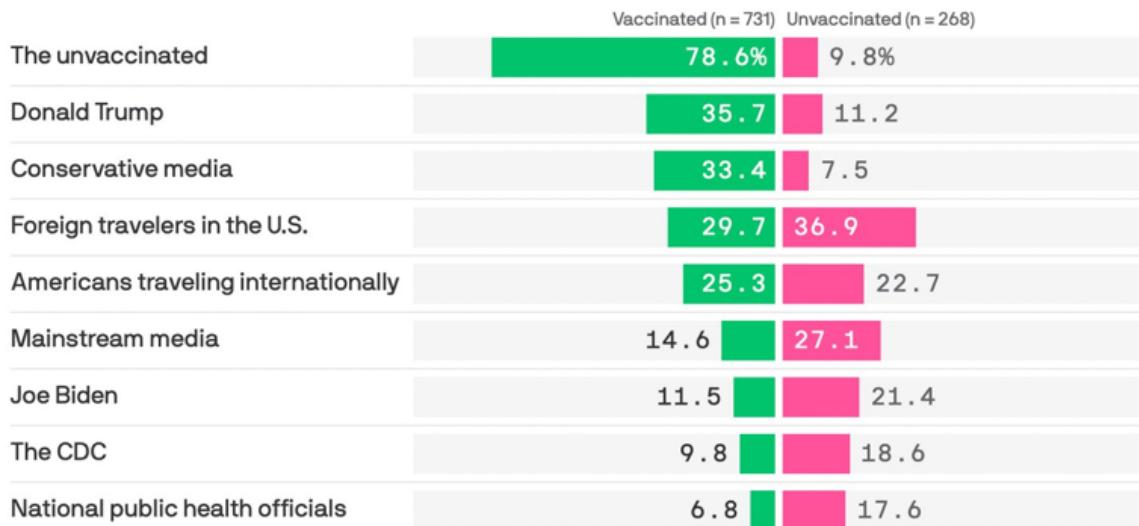
I hate the ambiguity in the wording here but I presume it means about 20-30% of respondents are in those hard-to-get categories, which leaves 70-80% who aren't.

In other news, let's ask what everyone really cares about. [Covid-19. Who to blame?](#)

Who is to blame for rising COVID-19 cases, by vaccination status

Percentage attributing blame to each group when asked "Which of the following people or groups, if any, do you blame for the rising COVID-19 cases and spread of new variants in the U.S.?"

█ Vaccinated (n = 731) █ Unvaccinated (n = 268)



Data: Axios/Ipsos Poll; Chart: Connor Rothschild/Axios

A hidden point of data here is that 73.1% of people responding are vaccinated, which (given an adult vaccination rate at the time of just under 70%) gives us an estimate of how biased survey responses are likely to be.

Mask and Testing Mandates

LA's positive test rate will be going down soon. [Every child \(and teacher\) in an LA United school will need to get tested weekly, even if vaccinated.](#)

[In New York, and in many other places, the teachers' union is strongly opposed to this](#), with any testing of the unvaccinated to be covered fully at taxpayer expense, while they previously insisted on keeping schools closed until teachers could get vaccinated no matter the case level, which is a hint as to what game they are playing to win.

[It's not like it's a handful of holdouts, either...](#)

Some of the city's largest agencies have lower vaccination rates than the general public. Of NYPD's 54,000 uniformed and civilian workforce, **only 43% are vaccinated**, the New York Post reported last week, also finding that the FDNY has a 55% vaccination rate. Roughly 42% of city Department of Correction workers are vaccinated, the agency told THE CITY, based on information it has about those who were vaccinated in the five boroughs.

Both the city's 135,000 public school employees and 42,000 public hospital workers have a 60% vaccination rate. An MTA spokesperson estimated 65% to 70% of the transit agency's 65,000 employees have received the vaccine.

These numbers are much *lower* than the adult population's vaccination rate in the area. It's mind boggling that 40% of teachers in NYC schools remain unvaccinated. This is the group charged with 'educating' our children? Union can't allow teachers in classrooms until they've been vaccinated, then almost half of them refuse to get vaccinated. We've made a huge mistake.

I realize it's impossible for multiple reasons, but my preferred scheme would be not only to mandate vaccinations going forward, but take this opportunity to start over and fire every teacher not currently vaccinated as not qualified to teach anything to children, and go from there.

[UNC not only doesn't have a vaccination mandate and justifies that decision based on a misunderstanding of the law, there's a slight additional problem:](#)



Benjamin Mason Meier @BenjaminMMeier · 23h

Speaking with worried @UNC students over the weekend, every single one of them:

- * knew exactly how to buy a fake #COVID19Vaccine card &
- * knew a fellow student who had submitted one to the University.

How will the University respond to those who submit false medical information?

[Show this thread](#)

COVID-19 Vaccination Record Card

Please keep this record card, which includes medical information about the vaccines you have received.

Por favor, guarde esta tarjeta de registro, que incluye información médica sobre las vacunas que ha recibido.





Benjamin Mason Meier @BenjaminMMeier · 23h

...

Replies to [@BenjaminMMeier](#)

Reviewing [@UNC](#)'s "COVID-19 Vaccine Certification Disclaimer," I note that there is no attestation/acknowledgment that the information submitted is...truthful. 🐶

Voluntary Process + Zero Accountability =
Public Health Catastrophe.



Benjamin Mason Meier @BenjaminMMeier · 41m

...

Furthermore, [@UNC](#) system's justification for not having a vaccine mandate northcarolina.edu/wp-content/upl... is

- (a) wrong on law - G.S. § 130A-152(a) does not preclude schools from adding required immunizations
and
- (b) outdated on policy - as state & federal officials now recommend mandate.

So our current standard is that we don't require the vaccination, lots of students lie about it anyway in case that changes, and if we don't explicitly get them to affirm that their statements are true then we can't hold them accountable even if we find out they were lying? Sounds about right. Academic standards are not what they used to be.

Vaccine Hesitancy

Who is vaccine hesitant? [Some data](#).



Matthew Yglesias
@mattyglesias

...

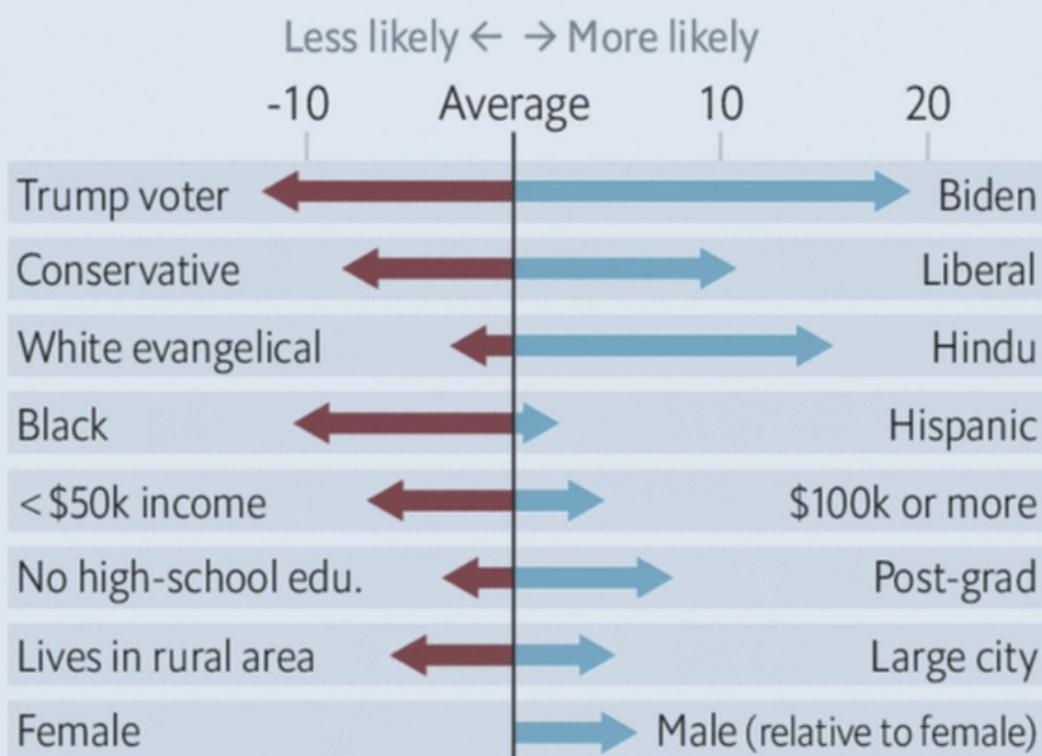
Dudes rock

[economist.com/united-states/...](http://economist.com/united-states/)

Jabs v jab-nots

US, covid-19, difference in likelihood of being/intending to get vaccinated*, % points, 2021

By demography



*Relative to the profile of the average American

Source: YouGov/The Economist

I find it interesting (for non-Covid reasons) that everything here lists both the positive and negative effects, except when females are less likely to be vaccinated, and then it's only the

positive effect. The argument that the two are mirrors of each other is not one the source would likely be happy to be quoted on in this day and age, given its implications.

Especially since it seems like dudes [intend to rock but in fact keep putting off their rocking until a future date](#) ([link to data](#)):



Daniel Feldman @d_feldman · Jul 30

...

Replies to [@mattyglesias](#)

Ok, I found some more details

In this survey, women and men were vaccinated at the same rate (59%)

The entire difference is that more men have started (but not finished) vaccination, or are planning to in the future

And its only 5% of the sample

docs.cdn.yougov.com/v60y11605p/eco...

YouGov

10. Vaccination

How would you describe your personal situation regarding COVID-19 vaccines?

	Total	Gender		White Men		White Women		Race	
		Male	Female	No degree	College Grad	No degree	College Grad	Black	Hispanic
I have received all the injections required to be fully vaccinated against COVID-19	59%	59%	59%	54%	78%	56%	80%	47%	47%
I have started the vaccination process, but need another shot	4%	5%	2%	4%	2%	2%	1%	9%	6%
I plan to get vaccinated	6%	7%	5%	5%	8%	4%	3%	7%	9%
I will not get vaccinated	19%	18%	20%	25%	6%	26%	9%	17%	20%
I'm not sure about getting vaccinated	12%	11%	13%	12%	6%	12%	7%	20%	17%
Totals	100%	100%	99%	100%	100%	100%	100%	100%	99%
Unweighted N	(1,499)	(683)	(816)	(274)	(231)	(319)	(241)	(194)	(142)

(Reminds me a little of [this graph](#) about which animals people think they can beat in a fight, which was covered as 'men think crazy things' but that's not how I read the actual numbers).

From the replies, [another chart](#):

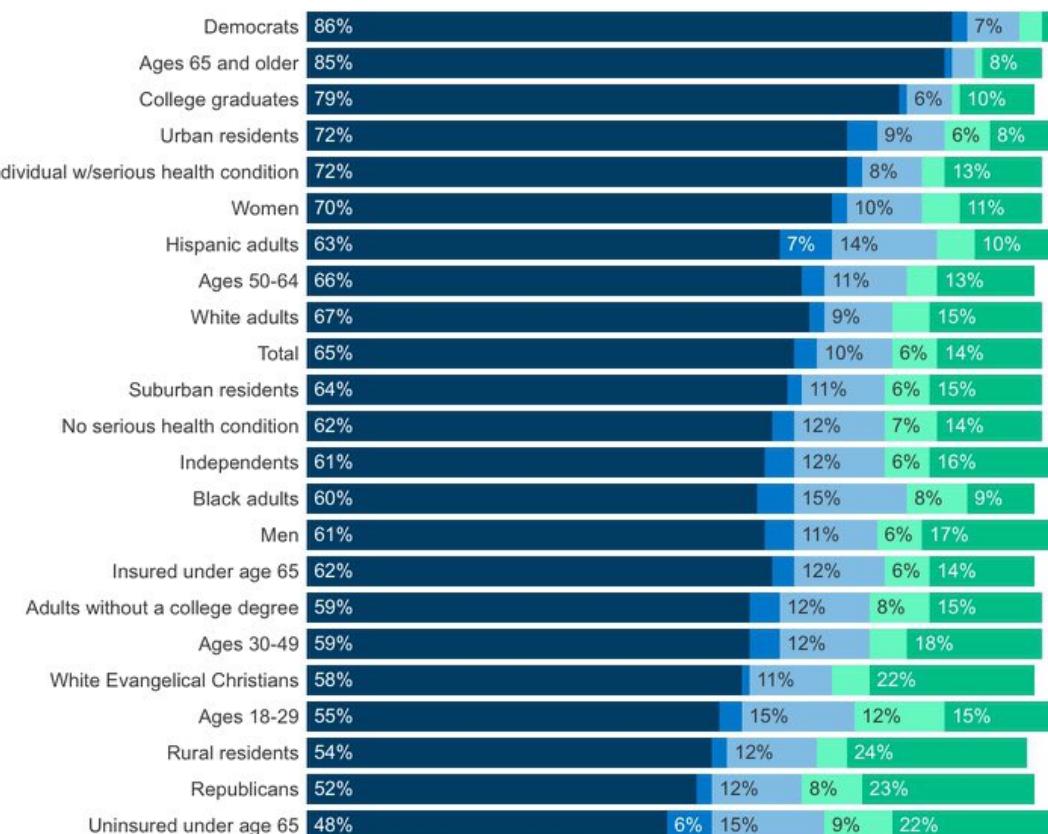


Figure 2

Across Most Subgroups, At Least Half Report Receiving A COVID-19 Vaccine

Have you personally received at least one dose of the COVID-19 vaccine, or not? As you may know, an FDA-authorized vaccine for COVID-19 is now available for free to all adults in the U.S. Do you think you will...?

■ Already received at least one dose ■ Get vaccinated ASAP ■ Wait and see ■ Only if required ■ Definitely not



NOTE: See topline for full question wording.

SOURCE: KFF COVID-19 Vaccine Monitor (June 8-21, 2021) • Download PNG

KFF COVID-19
Vaccine Monitor

A gender gap in vaccine uptake has emerged over the past several months, and women are now 9 percentage points more likely to report being vaccinated than men (70% vs. 61%), and a larger share

There are some contradictions between the data sets, but mostly they tell similar stories. I also like the 'only if required' category because it puts a reasonable estimate slash lower bound on the benefits of a requirement, which would be the vaccination of 6% of adults or 18% of non-vaccinated adults.

[Samo Burja argues that one side was always going to be anti-vax while the other was pro, so we should be thankful it's blue America that was pro rather than the other way around.](#) If one has to choose, giving whoever controls the media the right position does seem better, so the question is whether he's right that we had to choose. I don't think this is obvious – one side may be coming out against motherhood but we can draw hope from the fact that we still have broad support for apple pie.

Andrew Yang may not be a successful candidate, [but he will forever live on in our hearts both for wearing a 'math' pin and as the 'pay people money' guy:](#)



Andrew Yang @AndrewYang · Jul 29

...



Biden calls on states to offer \$100 vaccine incentives

President Biden on Thursday called on state and local governments to use funds from his \$1.9 trillion American Rescue Plan to offer \$100 ...

thehill.com

How about vaccine lotteries, do those work? [Yep, those work.](#)

New York's combination of new requirements and new incentives? [Yep, those work.](#)

How about requiring the vaccine in order to fly? [Yep, incentives matter again.](#) We don't have magnitude here, so unfortunately this doesn't provide that good a natural experiment on how much people value the ability to fly. But I don't think that's even the *second* most interesting thing here, despite being what I noticed being called out...



Ken Klippenstein @kenklippenstein · Aug 1

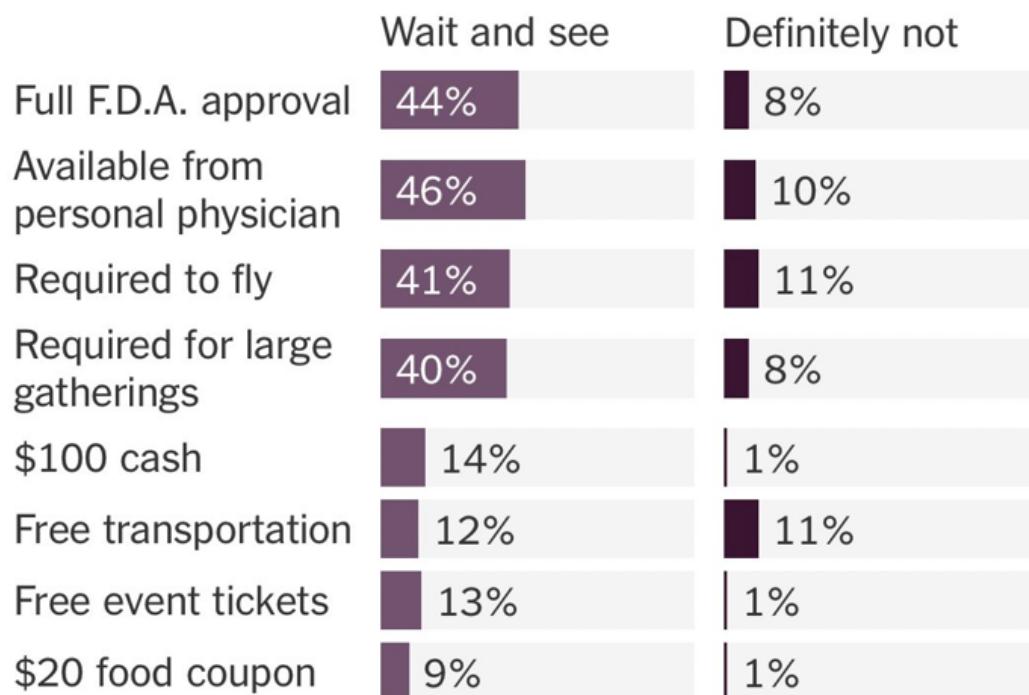
...

People who say they would be more likely to get vaccinated if they got \$100 cash: 14%

People who say they would be more likely to get vaccinated if it was required to fly: 41%

What May Motivate the Unvaccinated to Get a Shot

Share of people who say these incentives would make them more likely to get vaccinated.



Source: Kaiser Family Foundation survey, June • By The New York Times

There's the top four all which are big, and the bottom four which are small but meaningful, although 'more likely' is a far cry from 'turns a no into a yes.'

The big two are the first two, because they're both not coercive and they're free. Full F.D.A. approval would be helpful for fully half the remaining people. Getting the vaccine to personal physicians, which also should have happened a long time ago many times over, also helps with half of people.

Combine those two, and my guess is you pick up a large percentage of the remaining 'gettable' people, no coercion required. That doesn't mean the \$100 cash wouldn't help, I'd

throw that in too, but notice that the \$20 food coupon (worth less than \$20!) gets you more than halfway there at a fifth of the price. That makes sense, the marginal demand curve should slope downward here, and I'm still happy to pay the marginal costs to bump the payment to \$100, especially since it would mostly go to poorer people anyway.

One concern is that all these 'rewards' might make people hold out next time to try and get the rewards. Again, people respond to incentives. The good news here is my model does not think this much matters. Yes, some people would move from taking the early free option to holding out for the bribe, but mostly that delays people during the period when demand exceeds supply, and hence no bribes are on offer. So it's good, actually. If in March, everyone knew that anyone would get \$100 for being vaccinated *starting in April* then that's a way of allocating scarce resources by price without everyone losing their heads over it. Sounds good to me. So what if you end up having to bribe some people who would have gotten it for free?

Also, a gentle reminder that incentives matter in all directions, [and sometimes mistakes get made...](#)



David @Strixvarias · 14h

Work is requiring vaccinated people to work in person and wear a mask and unvaccinated people are required to stay home....going social media antivaxxer

...

If you think lying to say you're vaccinated is easy, lying to say you're *not* vaccinated is even easier, as long as you're willing to endure the things people say to those who aren't vaccinated.

Delta Variant

[There are three core messages here.](#)



President Biden  @POTUS · Jul 29

🇺🇸 United States government official

The Delta variant is different than what we've dealt with previously. It's highly transmissible and causing a new wave of cases. But here's the good news: we have the power to stop it. Get vaccinated — and let's defeat this virus once and for all.

6.3K

21.1K

135.9K



...

Two of them are that Delta is more transmissible and causing a new wave, and that vaccinations work. Those are true and important.

The third is that we have the power to stop it, and defeat the virus once and for all.

That one seems to me to be false, at least for meaningful definitions of 'we' and 'have the power to stop it,' and especially 'once and for all.'

Yes, at some point, likely not that long from now, things will turn around the way they did in India, the UK and the Netherlands, once we've had enough time for the control system to adjust and for enough people to be infected and/or vaccinated. That won't be the end of Covid, and it won't count as having stopped it, so much as having gotten through it.

The important thing is that the ship has almost entirely sailed. We'd prefer to get more vaccinations and otherwise improve things on the margin, but over the range of mask mandates and other new countermeasures, are we going to change what happens much from here? If we did, are we going to be willing to keep those countermeasures online indefinitely? Are we going to be willing to mandate vaccinations *and if necessary booster shots* sufficiently effectively to get paid off for buying all that time?

It seems like the answer is clearly no. That doesn't mean it's senseless to take reasonable precautions on the margin, but this is mostly only a battle worth fighting if one can win. If failure is inevitable, and we've mostly gotten as many vulnerable people vaccinated as we're going to get, it would be better to fail fast.

That is a central fact to keep in mind when looking at the new mask mandates and other updates coming out of the CDC and elsewhere. Preventive measures make sense when they are *necessary* to control spread, and also *sufficient* to have worthwhile benefits. The window available is not impossibly narrow, but it is not infinitely wide either. And in the possible (or at least *theoretically* possible) worlds in which vaccinated individuals are at great risk of infecting others, I don't see how we could hit the window.

Also, before we discuss the CDC's new mask mandate and its justifications, a reminder that if you institute a mask mandate, [it's important to follow that exact mandate yourself.](#)



Seth Mandel  @SethAMandel · Aug 1

If you enact a mandate and then clearly demonstrate that you don't think the mandate is necessary or beneficial you are undermining more than the mandate--you undermine the idea that ppl need to take this seriously. For shame.



Tiana Lowe  @TianaTheFirst · Aug 1

EXCLUSIVE: Not 24 hours into the indoor mask mandate she imposed on DC, Muriel Bowser officiated an indoor wedding in Adams Morgan and stayed to fete with *hundreds* of fellow maskless guests.
washingtonexaminer.com/opinion/dc-may...

[Show this thread](#)



13



103



434



Seth Mandel  @SethAMandel · Aug 1

Bowser believes the mandates are about power not public health. I think that message coming from an official is so obviously counterproductive that even the "messaging doesn't matter" brigade might understand this one.

And even if you do sometimes violate the rules you're telling everyone else to follow, you definitely shouldn't be [explicitly exempting yourself](#) from those rules.

Similarly, there are also reports that Pelosi has violated the house mask mandate on at least three occasions, in addition to all the Republicans who are violating it to own the libs.

The CDC Reinstates Its Mask Mandates

Point: When the facts change, I change my mind. What do you do, sir?

Counterpoint:

Now: Imagine it's Trump.

 **Curtis Houck**  @CurtisHouck · Jul 29

Doocy vs. Biden 🔥

Doocy: You said if you were fully vaccinated, you no longer need to wear a mask.

Biden: I didn't say that.

Doocy: You did.

(...)

Doocy: In May, you made it sound like the vaccine was the ticket to lose the mask forever.

Biden: That was true at the time.

The political strategy of 'flat out deny saying the thing you said when the video cameras were rolling' has its disadvantages, such as the media usually having easy access to what was being filmed by those video cameras. I understand it when your brand is going to be 'I lie all the time, what are you going to do about it?' but that's very much a *pro-malarkey* stance.

One refinement is where you don't *quite* say the thing explicitly, and can sort of claim that you only said it implicitly, as is the case here. Fooling people on such things tends to be infrequent, but it does enforce that you have the power to decide what you did and did not say retroactively, and thus your alignment with the destruction of the public record. So I suppose there are some strategic advantages.

In any case, the term 'forever' and the term 'at the time' are not close friends, and Delta has been known since well before the statements about taking off one's mask forever. Did the situation change? Yes. Is the situation worse than would have been reasonably anticipated? I think yes, the developments of the past two months have been worse than expected. But, well, yeah. Imagine it flipped.

That's distinct from the question of whether reinstating these mask requirements makes physical sense. If it turns out the messaging earlier was a mistake, I'd prefer owning the mistake to pretending a mistake wasn't made, but it's still better not to keep the mistake running - if it was indeed a mistake.

So, what facts changed that perhaps should change minds? Why are they suddenly saying [things like this](#)?

The delta variant of the [coronavirus](#) appears to cause more severe illness than earlier variants and spreads as easily as chickenpox, according to an internal federal health document that argues officials must "acknowledge the war has changed."

READ THE
DOCUMENTS
[Full PDF](#)

It cites a combination of recently obtained, [still-unpublished data](#) from outbreak investigations and outside studies showing that vaccinated individuals infected with delta may be able to transmit the virus as easily as those who are unvaccinated. Vaccinated people infected with delta have measurable viral loads similar to those who are unvaccinated and infected with the variant.

One of the slides states that there is a higher risk among older age groups for hospitalization and death relative to younger people, regardless of vaccination status. Another estimates that there are 35,000 symptomatic infections per week among 162 million vaccinated Americans.

“Although it’s rare, we believe that at an individual level, vaccinated people may spread the virus, which is why we updated our recommendation,” according to the federal health official, who spoke on the condition of anonymity because they were not authorized to speak publicly. “Waiting even days to publish the data could result in needless suffering and as public health professionals we cannot accept that.”

They waited days to publish the data. The CDC initially did the infuriating thing of [refusing to release their data, which even the mainstream media like WaPo pointed out was terrible](#) but they did then release the data, and we have it now.

[One theory on what might be causing this.](#)



Nate Silver @NateSilver538 6h

Media coverage of recent Delta news has been terrible but the CDC also had a terrible and naive communications strategy of leaking details to e.g. the NYT instead of just of making them publicly available.

<amp.cnn.com/cnn/2021/07/30...>

92 261 2k ...



Nate Silver

@NateSilver538

I say "naive" because when you give media outlets a scoop, they tend to sensationalize its importance rather than put it into context. The NYT in particular has an editorial culture that will blow any proprietary info completely out of perspective, as they did in this instance.

Which results in completely false things like this:

The New York Times ✅ @nytimes

Breaking News: The Delta variant is as contagious as chickenpox and may be spread by vaccinated people as easily as the unvaccinated, an internal C.D.C. report said.
nyti.ms/376u8mv

Show this thread

[And also various news reports like this, which seem like the point](#) - get people scared before they can analyze the data.

Which results in:

⤳ Kenneth Baer Retweeted



German Lopez ✅ @germanlopez · Jul 30

...

As one example: Every expert I've talked to agrees with Ashish here, but it's absolutely not what the feds have conveyed in the past few days.

 **Ashish K. Jha, MD, MPH** ✅ @ashishkjha · Jul 30

Vaccinated people are far far far less likely to spread the virus than unvaccinated people

Like way less likely

A lot less

Pretty clear from the data we have so far

That's my Friday thought

[To go with the general trend of, well, this:](#)

⤓ Kenneth Baer Retweeted



Derek Thompson @DKThomp · Jul 30

...

It's really a shame that President Joe Biden signed that executive order banning the use of denominators in COVID headlines.



Ken Dilanian @KenDilanianNBC · Jul 30

Exclusive: At least 125,000 fully vaccinated Americans tested positive for Covid nbcnews.com/health/health-... via @strickdc

[The first half of this isn't fully true, but it's way too true](#)



Youyang Gu @youyanggu · Jul 30

...

One leaked CDC memo & headlines went from "99% of Covid cases are among the unvaccinated" to "majority of cases are among the fully vaccinated".

In reality, the truth is not very sensational: The vaccine is highly effective & working exactly as intended.

:

Another communications problem:

"We've done a great job of telling the public these are miracle vaccines," Seeger said. "We have probably fallen a little into the trap of over-reassurance, which is one of the challenges of any crisis communication circumstance."

It's tricky when you both want to tell people how great the vaccine is and promise how people can return to their normal lives with no worries whatsoever, to get them to take the vaccine, and then *also* turn around and lie about how ineffective the vaccine is in order to scare them into not changing their behavior afterwards. It's even harder doing both simultaneously. Everyone involved gets whiplash. I almost sympathize.

The CDC's new mask mandate is in areas with sufficient spread, which means [it's constantly expanding in scope each day in ways that are completely inevitable](#), so it would have been better to bite the bullet for everyone at the same time:



Joseph Spector

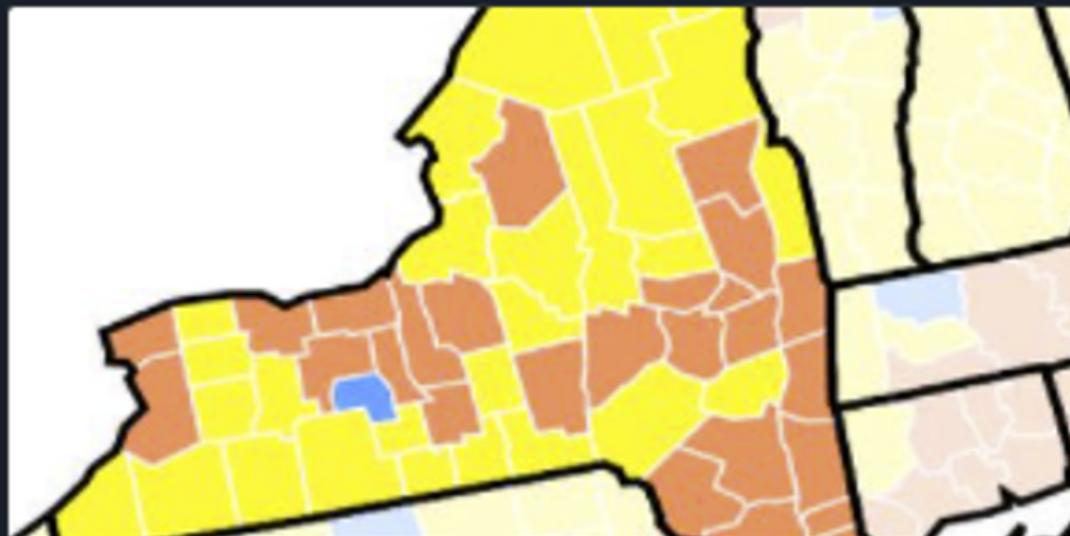


@GannettAlbany

NY's 33 counties where CDC says to wear a mask...

That's up from 22 a day ago...

democratandchronicle.com/story/news/pol...



New York up to 33 counties where CDC recommends mask wearing due to COVID
democratandchronicle.com

To make it easy to find, analysis of the study itself gets its own section.

The Provincetown Study

For those not looking to get into the weeds, this was an outlier situation, there were tons of base rate errors that the study makes no attempt to correct for, we learned very little, and mostly the CDC is making a big deal over all this for no good reason.

For those who want to dive into the weeds here, [let's get to it.](#)

Outbreak of SARS-CoV-2 Infections, Including COVID-19 Vaccine Breakthrough Infections, Associated with Large Public Gatherings — Barnstable County, Massachusetts, July 2021

Early Release / July 30, 2021 / 70

Summary

What is already known about this topic?

Variants of SARS-CoV-2 continue to emerge. The B.1.617.2 (Delta) variant is highly transmissible.

What is added by this report?

In July 2021, following multiple large public events in a Barnstable County, Massachusetts, town, 469 COVID-19 cases were identified among Massachusetts residents who had traveled to the town during July 3–17; 346 (74%) occurred in fully vaccinated persons. Testing identified the Delta variant in 90% of specimens from 133 patients. Cycle threshold values were similar among specimens from patients who were fully vaccinated and those who were not.

What are the implications for public health practice?

Jurisdictions might consider expanded prevention strategies, including universal masking in indoor public settings, particularly for large public gatherings that include travelers from many areas with differing levels of SARS-CoV-2 transmission.

So the headline findings are claimed to be (1) among cases that could be identified, vaccinated and unvaccinated people had similar cycle threshold values (and by implication, perhaps had similar ability to spread the virus) and (2) 74% of cases occurred in fully vaccinated persons, versus 69% vaccination coverage for the area.

The suggestion to reimplement restrictions is standard issue, but I note here that if vaccines were sufficiently ineffective in practice against Delta, there would be no reasonable way to stop the pandemic, and I'd want to do the opposite of the implications listed here and stop trying.

Most vaccinated patients were symptomatic:

Overall, 274 (79%) vaccinated patients with breakthrough infection were symptomatic. Among five COVID-19 patients who were hospitalized, four were fully vaccinated; no deaths were reported. Real-time reverse transcription-polymerase chain reaction (RT-PCR) cycle threshold (Ct) values in specimens from 127 vaccinated persons with breakthrough cases were similar to those from 84 persons who were unvaccinated, not fully vaccinated, or whose vaccination status was unknown (median = 22.77 and 21.54, respectively). The Delta variant of SARS-CoV-2 is highly transmissible ([7](#)); vaccination is the most important strategy to

Four of five hospitalized patients were fully vaccinated. None of them died.

"A cluster-associated case was defined as receipt of a positive SARS-CoV-2 test (nucleic acid amplification or antigen) result ≤14 days after travel to or residence in the town in Barnstable County since July 3."

By July 26, a total of 469 COVID-19 cases were identified among Massachusetts residents; dates of positive specimen collection ranged from July 6 through July 25 (Figure 1). Most cases occurred in males (85%); median age was 40 years (range = <1–76 years). Nearly one half (199; 42%) reported residence in the town in Barnstable County. Overall, 346 (74%) persons with COVID-19 reported symptoms consistent with COVID-19.^{**} Five were hospitalized; as of July 27, no deaths were reported. One hospitalized patient (age range = 50–59 years) was not vaccinated and had multiple underlying medical conditions.^{††} Four additional, fully vaccinated patients^{§§} aged 20–70 years were also hospitalized, two of whom had underlying medical conditions. Initial genomic sequencing of specimens from 133 patients identified the Delta variant in 119 (89%) cases and the Delta AY.3 sublineage in one (1%) case; genomic sequencing was not successful for 13 (10%) specimens.

So it's all Delta, and 74% of identified cases were symptomatic, including 79% of vaccinated cases, with the usual mix of symptoms.

...and that's pretty much it.

Here is the bulk of their #analysis:

"The findings in this report are subject to at least four limitations. First, data from this report are insufficient to draw conclusions about the effectiveness of COVID-19 vaccines against SARS-CoV-2, including the Delta variant, during this outbreak. As population-level vaccination coverage increases, vaccinated persons are likely to represent a larger proportion of COVID-19 cases. Second, asymptomatic breakthrough infections might be underrepresented because of detection bias. Third, demographics of cases likely reflect those of attendees at the public gatherings, as events were marketed to adult male participants; further study is underway to identify other population characteristics among cases, such as additional demographic characteristics and underlying health conditions including immunocompromising conditions.^{***} MA DPH, CDC, and affected jurisdictions are collaborating in this response; MA DPH is conducting additional case investigations, obtaining samples for genomic sequencing, and linking case information with laboratory data and vaccination history. Finally, Ct values obtained with SARS-CoV-2 qualitative RT-PCR diagnostic tests might provide a crude correlation to the amount of virus present in a sample and can also be affected by factors other than viral load.^{†††} Although the assay used in this investigation was not validated to provide quantitative results, there was no significant difference between the Ct values of samples collected from breakthrough cases and the other cases. This might mean that the viral load of vaccinated and unvaccinated persons infected with SARS-CoV-2 is also similar. However, microbiological studies are required to confirm these findings."

That's what the CDC updated on? This is what's causing a chorus of "Vaccinated people may spread Covid as much as unvaccinated people"? That's it?

All right, so what's actually going on here in this study?

I'm not going to go full Not Necessarily the News, but it's remarkably tempting, because the sample is clearly nothing like a normal one.

The first big hint is that this took place during a series of gatherings, mostly of men, and that most of the infections were of men so it's clear that this happened mostly among those attending the gatherings rather than the background population.

The second even bigger hint is that the numbers look *absolutely nothing* like the numbers we observe in the general population, on several measures.

The infected population is mostly male, and of course it was mostly travelling and engaging in unusually high risk activities and almost certainly stuck in a superspread event.

Most cases (74%) were symptomatic, which isn't normal when looking carefully for cases.

Most cases (74% again) were in vaccinated people, which was *higher* than the population percentage that was vaccinated in the surrounding area. We know vaccines are effective at preventing Covid-19 and at preventing symptomatic Covid-19, and even if they're not as effective as we think, *less than zero* is not on the table here, 'cmon.

Vaccinated cases were *more likely* to be symptomatic than non-vaccinated cases, which we also know isn't normal, *on top of* there being more such cases than the population baseline.

The Ct values in vaccinated patients were as high as those in unvaccinated patients, which was one of the banner headlines in the scare tactic articles, but viral loads themselves were not measured, and again we know that infections in the vaccinated are less severe.

So there are essentially two ways to interpret this data.

Method number one is to become [The Man of One Study](#), think that vaccines suddenly have entirely stopped working, ignore all the other overwhelming evidence to the contrary that's actually everywhere – even the anti-vax crowd mostly admits the vaccines *work* – and then reason from there. You could argue that it magically permits full infection and transmission and hospitalization but still prevents death since there were no deaths in this sample, but that doesn't actually make any physical sense.

Of course, none of the rest of that scenario makes any physical sense either. Wrong Conclusions Are Wrong, and taking the results of this study at face value flies in the face of the whole vaccinated people not getting Covid, not getting sick from Covid and not dying from Covid phenomenon that has very much persisted under Delta.

I didn't quite explicitly [Defy the Data](#) on the Israeli vaccine effectiveness measurements, but I did point out the data didn't make any sense even internally and that they really, really didn't make sense versus observed population data elsewhere, and that the 'these vaccinations were older' explanation wasn't actually going to fly. Whereas the 88% effectiveness number out of the UK made me sad, but was also *plausible*, made sense and could live in the same physical world we observe, so I took it far more seriously.

That's a roundabout way of saying that taking the study's data is, when taken at face value, Obvious Nonsense, and thus we will be using method number two.

Method number two is to ask what could have caused the study to get these answers that on their face are Obvious Nonsense, realize that this has everything to do with base rate fallacy and the failure to apply Bayes' theorem, with mostly vaccinated people attending the gatherings, and the pattern whereby infections are identified and tracked missing asymptomatic infections en masse which are concentrated among the vaccinated, and vaccinated people being more likely to get tested, and so on, in some combination.

[This thread looks at the basics under method two](#), points out how much worse things would have been in Provincetown without vaccines, and concludes the study doesn't tell us much.

[Twitter](#) was in general top form, here is a small sample.



Alex Tabarrok @ATabarrok · 14h

I have seen a hundred excellent tweets explaining the base rate fallacy, Bayes's theorem, and how the CDC and NYTimes stories mislead readers.

...

What do people not on twitter do?



Jordan Ellenberg @JSEllenberg · Jul 30

"Vaccinated people may spread delta as easily as unvaxxed" is like "Sober drivers may be as likely to die in a car crash as drunk drivers."



Jordan Ellenberg @JSEllenberg · Jul 30

...

Replies to [@JSEllenberg](#)

(“Oh, by “may,” I meant I was only talking about sober drivers who hit the guardrail at 75 mph, was that not clear?”)



Ethan BdM @ethanbdm · Jul 30

...

Replies to [@JSEllenberg](#)

Did you know the risk of death from COVID, even for vaccinated people, may be up to 100%?



Matthew Yglesias ✅ @mattyglesias · 20h

...

Seatbelt wearers who nonetheless fly through the windshield of the car during an accident may be just as likely to be injured as those who were not wearing a seatbelt.



Nate Silver ✅ @NateSilver538 · Jul 30

...

If the CDC revised all of its priors about COVID based on an analysis of spring break on Daytona Beach, people would have some understandable questions about that. But that's a pretty good approximation for Provincetown on a July 4 weekend.



Nate Silver ✅ @NateSilver538 · Jul 30

To take a self-selected, not-statistically-significant sample of ~200 nondiverse people during a party weekend that was an outlier in many respects, and use it to conclude that breakthrough infections are just as likely to transmit the virus, seems like quite the stretch.



Nate Silver ✅ @NateSilver538 · Jul 30

...

Symptomatic breakthrough infections having similar viral loads to *symptomatic* unvaccinated infections would be much less of a problem, both because symptomatic breakthroughs are rare and because people can learn to be more careful (and get tested) when they have symptoms.



Nate Silver ✅ @NateSilver538 · Jul 30

...

I don't know quite the right analogy here, but it's a bit like concluding that earthquake-resistant buildings aren't very effective based solely on studying one magnitude-8.7 earthquake where some of them failed.

Oh, and what kind of event was it that had this many men and so many infections? [Well, as it turns out...](#)



Peter Staley ✅
@peterstaley

...

If the CDC has increased their Delta Ro because of the Ptown cohort, then they are overstating it for the general population. The cohort was 85% male (WaPo and NYT have both failed to mention this). Hello, it was Bear Week. [@apoorva_nyc](#) rightly mentions packed bars, etc., but /1



Peter Staley ✅ @peterstaley · Jul 30

...

Replies to [@peterstaley](#)

... everyone is missing the horny bear in the room. Bears go to Ptown to have lots of fun which includes lots of sex. News flash, gay men KISS when they have sex. /2



Peter Staley ✅ @peterstaley · Jul 30

...

If you asked an ID expert to suggest the most efficient way for an infected vaxd person to infect another vaxd person, she'd say "let them deeply kiss for half a minute."

Hoping CDC used other cohorts for their new Ro. Ptown's is skewed by (gay) boys being boys. /end



Peter Staley ✅ @peterstaley · Jul 30

...

Addendum -- Even during its gayest week, Ptown is NOT 85% male tourists. The cohort skews 85% male for a reason. CDC is being too politically correct in not explaining this skew.

The cohort being 85% male makes it clear that the virus *did not* spread to the area's general population much, and stayed focused on *the people who travelled there largely in order to have sex*. Which is a reasonably good way to get very exposed to Covid-19. So. Yeah.

The Leaked CDC Slides

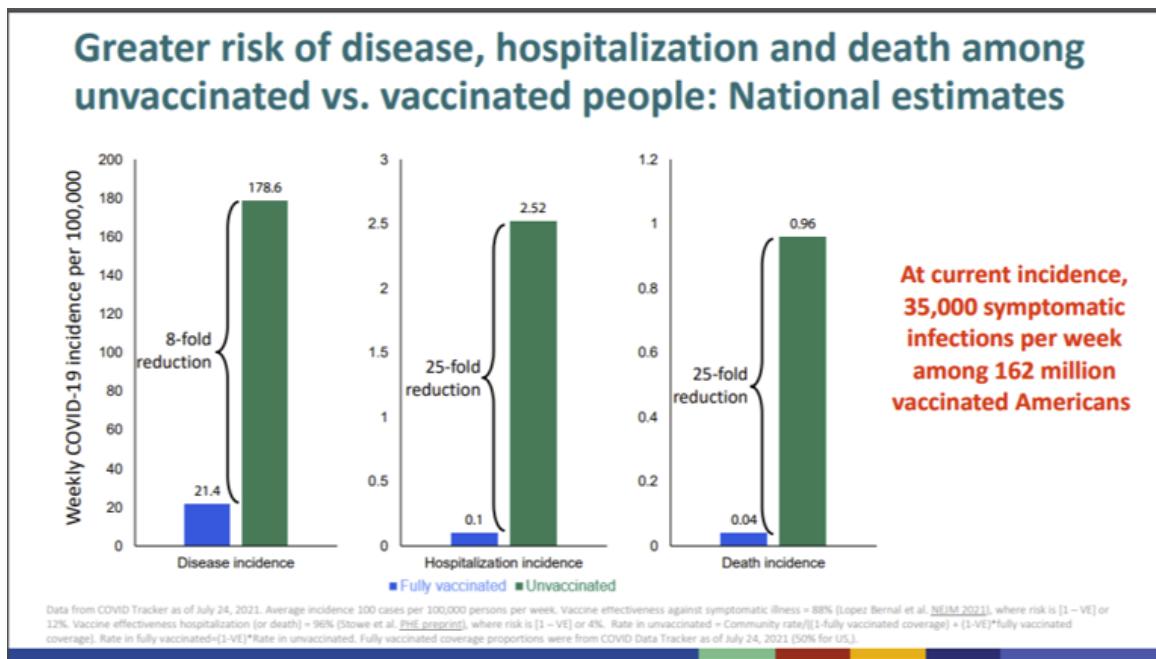
[The Washington Post got hold of the CDC's slides](#). If you can, I encourage you to click through and look at the slides yourself, it's much easier to read them there, they're all worth seeing and I can't be copying over all of them.

The first slide gets off to a great start, with a worthy cause:

Improving communications around vaccine breakthrough and vaccine effectiveness

Slide two notes that breakthrough cases may reduce public confidence in the vaccines, which is true enough, and I noticed I was pleasantly surprised it didn't say that *news* of those breakthroughs was what was reducing confidence.

Then comes slide three, which is a very good data visualization and tells where their heads are actually at with regard to the vaccines.



An 8-fold (87%) reduction in incidence, and a 25-fold (96%) reduction in hospitalization and death. That's exactly in line with all the other estimates. I agree with those numbers, good job. But how has 'the war changed'?

Slide four shows that a rising percentage of hospitalizations and deaths are among the vaccinated, and explicitly notes this is because of the increase in vaccinations, and the tendency for the most vulnerable to get vaccinated more often. Again, good, but that's the old war.

Slide 7 shows early vaccine effectiveness, which was quite good:

Early evidence in health care providers that vaccination may reduce transmission and attenuate illness (HEROES/RECOVER)

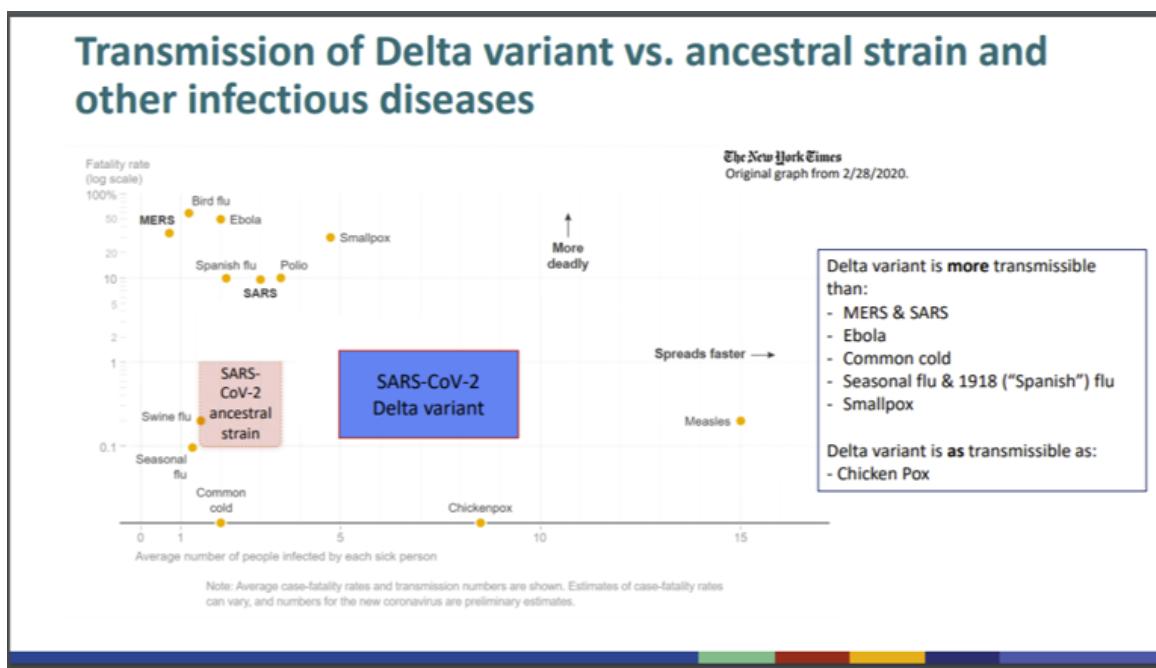
- Period: December 14, 2020 – April 10, 2021
- VE against infection was 91% (CI 76-97) among fully vaccinated; 81% (CI 64-90) for partially vaccinated
- Compared to unvaccinated cases, vaccinated cases (full or partial) had:
 - 40% lower mean RNA viral load (2.3 v. 3.8 copies/mL)
 - shorter mean duration of detectable viral RNA (2.7 v. 8.9 days)
 - lower risk of febrile symptoms (25.0% v. 63.1%)
 - shorter mean duration of symptoms (10.3 v. 16.7 days)

Thompson et al. doi:10.1056/NEJMoa2107058

7

The rest of the pre-Delta section goes into a bunch of other numbers, which all tell similar stories, all of which matches up with what we knew previously. Nothing to report here other than that we're all on the same page.

This is their big picture understanding of Delta absent vaccinations:



This again is exactly in line with previous estimates, although it has reasonably large error bars. If anything here is different, it's that this is saying Delta is not much deadlier than the ancestral strain.

Bunch of other as-expected stuff, then this is the first mention of the new study:

Delta variant vaccine breakthrough cases may be as transmissible as unvaccinated cases

- Breakthrough cases reported to national passive surveillance have lower Ct values by 3 cycles (**~10-fold increase in viral load**) for Delta (Ct=18, n=19) compared with Alpha (Ct=21, n=207) and other lineages (Ct=21, n=251)
- Barnstable County, MA, outbreak: **No difference in mean Ct values in vaccinated and unvaccinated cases** [median among vaccinated (n=80): 21.9; unvaccinated (n=65): 21.5]

(CONFIDENTIAL – preliminary data, subject to change)

17

Interestingly *not* present in the slides are the Ct values for non-breakthrough infections outside of Barnstable County, MA's outbreak. I find this absence quite odd.

In slide 19 they uncritically report the Israeli data that has been shown to come from faulty statistical methodology (and which never made sense in the first place).

In slide 20 they assume 50% of infections are reported, which seems crazy high, and also model assuming no distancing of any kind which seems like a pure counterfactual. Then, they use the results of *that* model to conclude the need for 'universal masking.' The need for containment without benefit of immunity from infections is left as an unstated assumption.

Slide 22 is their summary:

Summary

- Delta is different from previous strains
 - Highly contagious
 - Likely more severe
 - Breakthrough infections may be as transmissible as unvaccinated cases
- Vaccines prevent >90% of severe disease, but may be less effective at preventing infection or transmission
 - Therefore, more breakthrough and more community spread despite vaccination
- NPIs are essential to prevent continued spread with current vaccine coverage

22

I mean, all right, sure, all of that is true as far as it goes. If you want to prevent continued spread with current vaccine coverage, and no distancing, without waiting for the resulting infections to do the job for you, yes you will at a minimum need a lot of masks.

Finally, recommendations:

Next steps for CDC

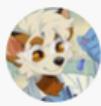
- Communications
 - Acknowledge the war has changed
 - Improve public's understanding of breakthrough infections
 - Improve communications around individual risk among vaccinated
 - Risk of severe disease or death reduced **10-fold or greater** in vaccinated
 - Risk of infection reduced **3-fold** in vaccinated
- Prevention
 - Consider vaccine mandates for HCP to protect vulnerable populations
 - Universal masking for source control and prevention
 - Reconsider other community mitigation strategies

23

Suddenly they're going down on this last slide to 75% effectiveness (with a conspicuously missing 'or greater' given that it's there in the previous line), which is not what you'd conclude from their previous slides, but that almost feels like a quibble.

It seems that 'the war has changed' here doesn't mean anything new, it simply means that Delta spreads easier than Alpha at exactly the ratio we previously believed, with exactly the vaccine effectiveness levels we previously believed, except it's being compared to a hypothetical situation without Delta rather than what should have been fully recognized weeks ago if not earlier. Whoops.

[This Chise thread on Twitter](#) makes it clear things are worse than I realized upon first reading. Among other things you'll find there:



Chise 🌈🧬🦠💉 · @sailorrooscout · Jul 30

To answer your burning question of: "Did the CDC partially base their assertion that vaccinated individuals can transmit Delta variant due to similar viral load as unvaccinated individuals on a study out of India that not only utilized models but accounted for vaccines that are

4

179

703

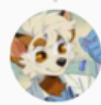


Chise 🌈🧬🦠💉 · @sailorrooscout · Jul 30

not currently approved in the United States and is still currently under review and previously did not pass peer-review and didn't even compare viral loads between unvaccinated and vaccinated individuals but rather viral loads between variants?" Yes, yes it did.

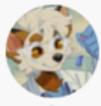
Delta infections associated with higher viral load and duration of shedding: Published evidence

- India report of lower cycle threshold (Ct) values in Delta breakthrough cases in HCW (n=47, mean Ct 16.5) compared to non-Delta breakthrough cases (n=22, mean Ct 19); also larger cluster size with Delta breakthrough
- Delta infection associated with longer duration of Ct values ≤30 [median 18 days vs. 13 days for ancestral strains]
- Risk of reinfection with Delta may be higher [aOR 1.46 (CI 1.03-2.05)] compared to Alpha variant, but only if prior infection ≥180 days earlier



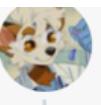
Chise 🌈🧬🦠💉 · @sailorrooscout · Jul 30

It also based decisions off smaller breakthroughs that have occurred that give us barely any information at all concerning mean Ct values between the vaccinated and unvaccinated. So why do I take issue with this? Because these are viral loads, NOT ACTUAL EVIDENCE OF TRANSMISSION.



Chise 🌈🧬🦠💉 · @sailorrooscout · Jul 30

You CANNOT make an assumption off this. Why? Because frankly I don't care how much viral RNA is in your nose (which may I remind everyone is LIKELY virus that is NOT viable due to the vaccine) when we have actual clinical data showing just one dose of vaccine HALVES your risk of



Chise 🌈🧬🦠💉 · @sailorrooscout · Jul 30

transmitting once you're infected?! To make it sound as it vaccinated and unvaccinated individuals are transmitting at the same rate is MISLEADING. Saying vaccinated individuals are superspreaders is MISLEADING. And news outlets have taken this and ran with it to make it come off

That seems generous, as these are at best *kind of* evidence of viral loads. Plus even then, the amount of failure to control for any of the things is very much not small.

In case you're wondering what study Chise is referring to there, here's a link, it's the Singapore study ([thread / study](#)):

The image shows two tweets from a user named Chise (@sailorrooscout). The first tweet, posted on Jul 31, discusses a recent study from Singapore showing that vaccination prevents people from getting sick with Delta (B.1.617.2) and is associated with faster decline in viral RNA load. It includes a link to medrxiv.org. The second tweet, also from Jul 31, notes that PCR cycle threshold (Ct) values were similar between vaccinated and unvaccinated groups at diagnosis, but viral loads decreased faster in vaccinated individuals, citing a study available on medrxiv.org.

Chise 🧬 🦠 🌐 🛡️ @sailorrooscout · Jul 31

A recent study out of Singapore shows not only does vaccination prevent you from getting sick with Delta (B.1.617.2), but it is associated with faster decline in viral RNA load. What does this mean? Vaccines make you LESS infectious!

96 952 3K

Chise 🧬 🦠 🌐 🛡️ @sailorrooscout · Jul 31

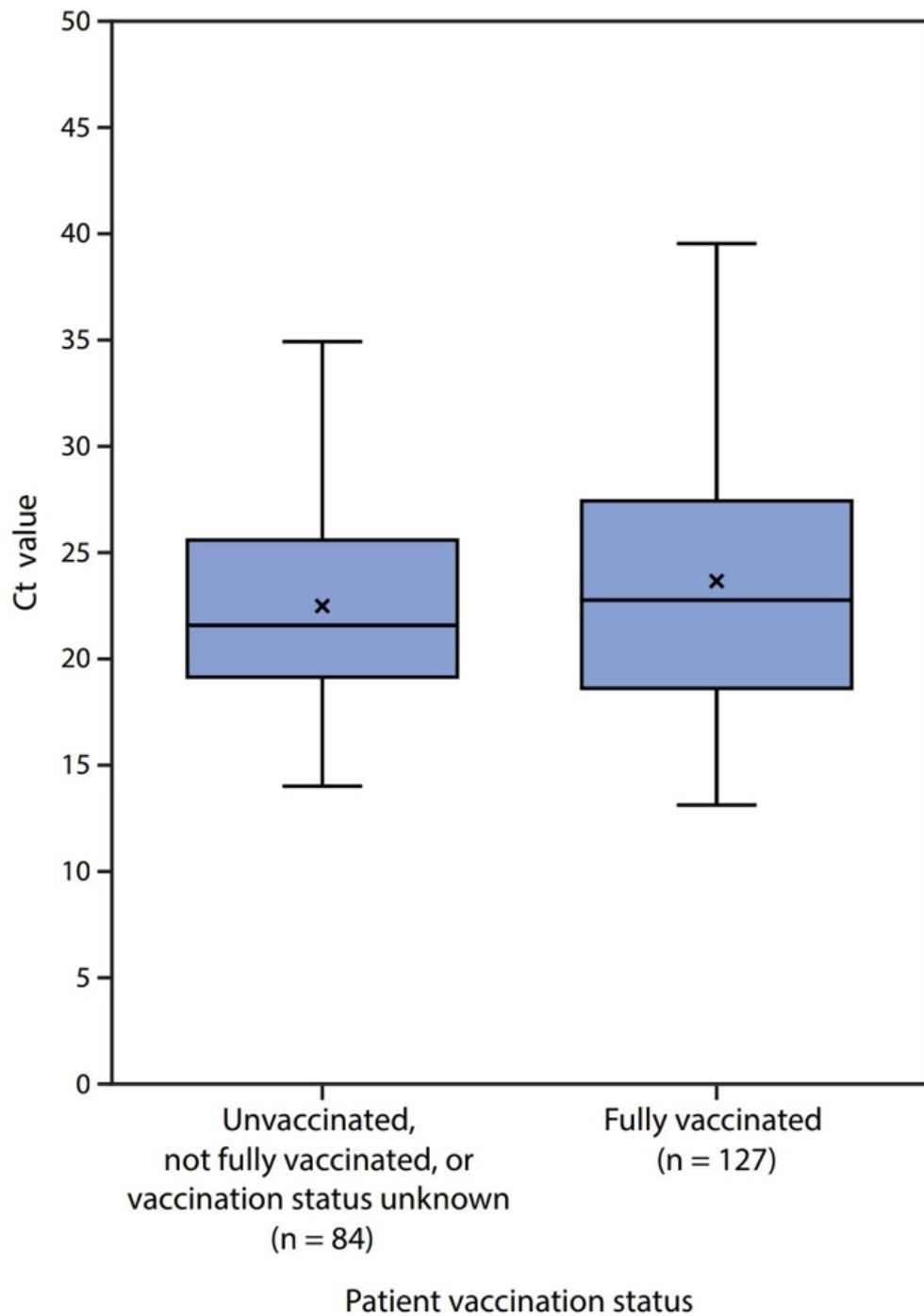
PCR cycle threshold (Ct) values were similar between both vaccinated and unvaccinated groups at diagnosis, but viral loads decreased faster in vaccinated individuals! Study can be found here: medrxiv.org/content/10.110...

39 130 687

I mean, we mostly knew this already, and the sample sizes here are low, but now Studies Show. So [even if the Ct ranges are similar at one point](#) (with the huge error bars, it's impossible to know), that doesn't mean what it's being taken to mean.

About that issue of sample size. In all these studies that show 'no difference' the sample sizes are tiny, and the confidence intervals quite large. Here's the graph from the Provincetown study:

FIGURE 2. SARS-CoV-2 real-time reverse transcription–polymerase chain reaction cycle threshold values* for specimens from patients with infections associated with large public gatherings, by vaccination status[†]—Barnstable County, Massachusetts, July 2021[§]



As discussed above, Ct value is not exactly what we want to measure, but even if it was, and even if we ignore all the other problems here, these are very wide error bars. [The Wisconsin](#)

[study has the same problem](#) and also concluded that Ct scores were similar *after throwing out any results with sufficiently low Ct scores.*

The war has changed thanks to Delta. The war *has not changed in the past two weeks.* Except insofar as the CDC has decided to change its approach to fighting, [taking a maximally dark interpretation of the data rather than offering anything importantly new](#), in ways that don't hold up to scrutiny or checks for base rates, and how much that's going to mess various things up.

The only explanation I can come up with is that previously the CDC had and/or [was painting a very wrong picture that vaccinated people were fully immune...](#)



Walid Gellad, MD MPH  @walidgellad · Aug 1

...

I took a picture of this tweet early April because I thought it was a grievous error. It was just wrong, and it was coming from CDC director.

Many of the people we fetishize as experts felt like it should be amplified. It made me really angry, and still does.



 Ashish K. Jha, MD, MPH Retweeted

The Recount  @therecount · 2d

...

CDC Director Dr. Rochelle Walensky: "Our data from the CDC today suggest that vaccinated people do not carry the virus."

I don't know to what extent this was what was happening, either intentionally to encourage vaccination and/or unintentionally via misunderstanding the science, and it's not always possible to make those two distinct. What I do know is that if previously you had the false impression that vaccinated people were completely safe, and then changed your story to vaccinated people being only mostly safe, perhaps that would explain what happened this past week?

[In summary, this:](#)



Nate Silver 

@NateSilver538

...

Yeah I'm trying to figure out why the CDC put such a negative spin on data that was largely unsurprising and very much in line with the consensus as it has emerged over the past several weeks.

Thinking of the Children

[There's a video at the link, he definitely said this](#), and I'd hope this would illustrate how *completely and utterly insane* the whole thing is and that we should maybe not listen to the other insane recommendations either.



Tom Elliott @tomselliott · 20h

NIH director Francis Collins: "It may sound weird" but parents should wear masks at home in front of their unvaccinated kids

...

Even within your own household, children too young to be vaccinated are too vulnerable to allow you to show them your face. Ever.

With that, [we get to the post going around this week](#) explaining (for those who don't know this already) exactly what we may be doing to our kids *forever*, even if only for the opening picture that should now haunt your dreams:

Are COVID Restrictions the New TSA?

The worrying parallels between the government reactions to 9/11 and COVID-19, and why some children might be permanently masked

Richard Hanania Jul 30 ❤ 34 🗞 49 🔍



Children at band practice in a school in Wenatchee, Washington in early 2021, long after it was known COVID-19 poses almost no risk to children. [Source](#).

You can decide for yourself how to think about the photo above, combined with an official request for masking *at home with one's own children*.

The difference between official school policies and young adult dystopian novels is that no one would buy this in a young adult dystopian novel, because it's fiction and therefore has to make at least some sense.

Meanwhile, as noted earlier, many teachers remain unvaccinated (40% in NY), and in most places the teachers' unions have put a stop to any talk of a mandate.

It's one thing to say that people make individual choices, but does this look like freedom or individual choice to you? Or does it look like a Geneva Convention violation for our young prison population, potentially doomed for years to come to never see a face, based on something that is not at all a threat to them and never was? Do we hate our children that much?

Apply for the Survival and Flourishing Fund

They're giving away a bunch of money. I'm helping figure out where to send it, and I encourage anyone with a worthy cause helping with our long term future to apply. [Here's the announcement](#), with more details at the link:

The Survival and Flourishing Fund, a virtual fund backed by philanthropists Jaan Tallinn and Jed McCaleb, is organizing the distribution of est. \$8MM-\$12MM in grants this November, with applications from orgs due on August 23. Applications are essential to enabling the grant recommendation team to learn about and debate the pros and cons of each organization under consideration, both old and new. So, please encourage applications from any awesome charitable projects you know about that are trying to support humanity's long-term survival and flourishing!

In Other News

I've been accused recently of 'carrying water' for the government because of my stance on vaccination policy. Unsurprisingly, I don't see it that way. Despite admiring those who do get paid for this, I'm not one of them, and on Tuesday I learned that such people exist.

NYT Media  @nytmedia

In addition to the efforts by the White House, state and local governments have begun paying “local micro influencers” — those with 5,000 to 100,000 followers — up to \$1,000 a month to promote Covid-19 vaccines to their fans. nyti.ms/2VpT9GH

I have between 5k and 100k followers! Presumably I qualify. Check, please. Every little bit helps.

[Scott Sumner's thoughts on what we should do now.](#)

Your periodic reminder from Scott Alexander: [FDA Delenda Est](#). He's in full righteous fury form, which is reliably top quality. I will quote the concluding section:

Final Thoughts

In conclusion, and contra *The Atlantic*, the FDA approving aducanumab is not very much like global warming at all. It is more like global warming in an alternate universe, where the government sometimes approves pollutants, and then everyone is forced to emit millions of tons of them whether they want to or not. Sometimes the government orders people to build a coal plant in the middle of the desert where nobody lives, a coal plant that isn't even connected to anything and just burns lots of coal without producing any electricity. But also, elderly people frequently freeze to death because the government refuses to give them permission to heat their house in the middle of winter. There is lively debate over whether the government should build more useless coal plants or let more elderly people freeze to death, and anyone who thinks there should be a better way of doing things is condemned as some kind of fringe libertarian. I really cannot stress enough how accurate this metaphor is or how much everything in the medical system is like this.

Mostly I am in violent agreement with the post, with the one exception being where Scott essentially shrugs and says 'incentives, what are you gonna do?' and refuses to hold anyone accountable for the whole thing or hold out any hope we could do better without raising the general sanity waterline. In order to understand what's going on, it is vital to understand that many important [actions in the system are actively perverse](#), and are chosen because they are worse rather than better, hurting people rather than helping them, [and that people are rewarded for having an actively reversed morality](#) and punished for having a normal one. Costs are benefits, and benefits are costs, not that anyone would then dare be seen doing a calculation. And it's also important to note that the blame dynamics and other incentives involved, to the extent they are real and constraining, are not inevitable consequences of the sanity waterline's current level, they are the dynamics that happen to exist, have lots of path dependence and could be changed.

Scott's proposal for unbundling the FDA seems like an excellent second-best alternative to my preference for burning the building to the ground and salting the Earth. His first proposal, with five approval tiers, is more complex than I believe to be necessary – counting fully unapproved things, there are 6 tiers. Then again, the entire American health care system is far more complex than necessary, a lot of which is engineered for blame avoidance combined with fraudulent extraction of money, so it can't be quite as simple as it sounds.

Your periodic reminder that [Emergency Use Authorization didn't exist at all until 2004](#). So things could have been so, so much worse.

Not that things are great now, we are [told to follow a particular regime whether or not it makes any physical sense](#).



Jonathan Reiner @JReinerMD · 12h

...

I gave a kidney transplant patient with no COVID antibodies a prescription for a booster dose and 3 pharmacies refused to give him the shot.

@CDCgov needs to move forward with boosters for our immunocompromised patients who are still vulnerable.

Although it can in some cases be done, presumably via lying:



Saju Mathew MD MPH @drsajumathew · 11h

...

Replies to @JReinerMD and @CDCgov

But then a young 58 year old healthy patient, on his own accord, got 2 extra shots of Moderna out of fear.

So herein lies the issue @JReinerMD, your patient who needs it, couldn't, mine who didn't need it, got it! FDA needs to chime in ASAP.

Death counts might be a little low after all?



terri scofield @terriscoscofield · 9h

...

Replies to @Cleavon_MD

Oh, in Suffolk County, New York the medical examiner only puts down SARS-CoV-2 when the family insists upon it and supplies blood work results paid for out of pocket...

So yeah, either way it's WAAAY under counted.

[AI fails to help with Covid](#). It is noteworthy that AI provided no assistance, and this offers some explanations. Data sets are terrible, and combining them often meant overlap where data in the training set got into the test set. Labeling relied on humans and incorporated their 'biases.' Fonts from hospitals and other contextual clues were used to cheat. The same techniques got used over and over again by everyone, so the failure of one attempt was highly correlated with the failure of other attempts. The results were not clinically ready.

I believe that these are real and important problems, but as a complete explanation I am not buying it. Several times, we heard stories of promising AI diagnostic techniques. In each case, the story was that there was this thing that would work, but the regulatory burden of being allowed to use it meant nothing would ever happen. Then we never heard about the situation again, the same way we never got rapid testing *without* AI or any number of other low hanging fruit improvements.

Also, if we didn't have good data sets on which to train the AI, maybe don't consider that purely a fatal flaw in AI, and rather consider that also a fatal failure to collect data. In general, it sounds like we tried 'throw off-the-shelf existing techniques at the problem using what data is around' and that was about it.

The WHO has called for a ban on booster shots, in the belief that somehow this will direct vaccine doses to those who don't have them, rather than resulting in there being less vaccine doses available. I will still happily accept a booster if offered one.

[Lambda variant said to show vaccine resistance in the lab in a preprint](#). I can't find any way to distinguish the things said here from a harmless situation or from signs of the next big thing,

but it's not a good sign. Need to keep an eye on it in any case.

[Nature's write-up of Fast Grants](#). It's super effective, or at least has a lot less wasted time.

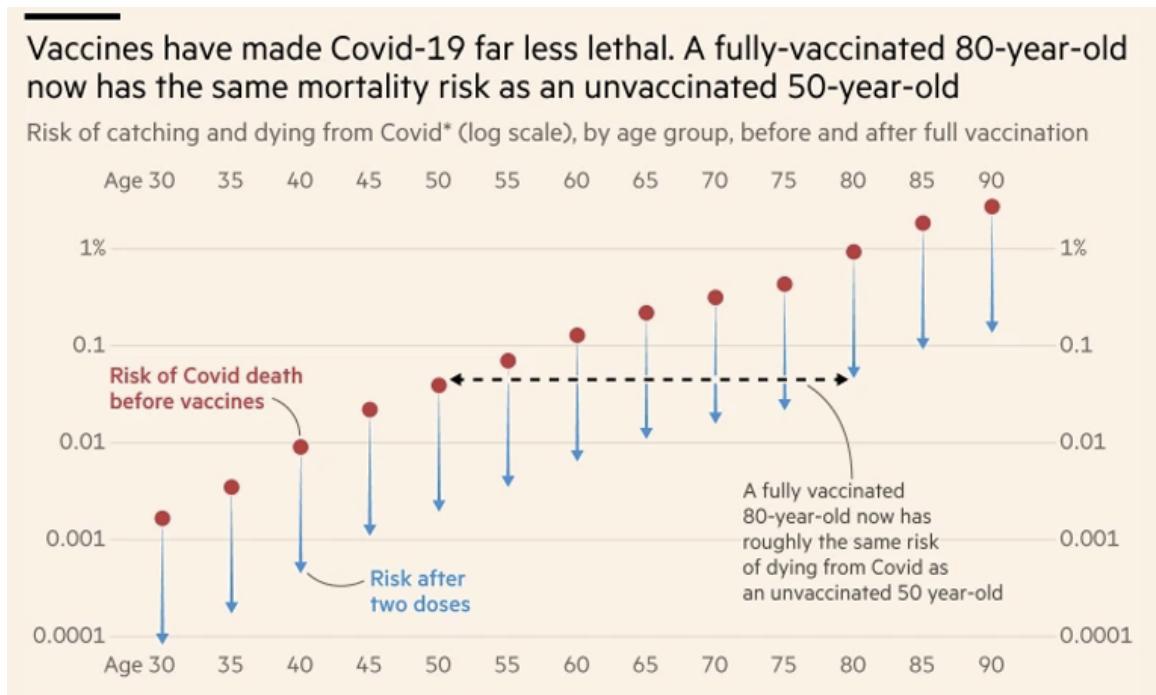
[For those in need of it: Thread explaining why vaccines decrease risk of dangerous variants rather than increasing it.](#)

[For those who need it or are curious: Nature post on how Covid infects cells.](#)

[Tyler points out that no one involved in the testing process at Mexican hotels has any incentive for those tests to ever come back positive. Whoops.](#)

[China is finally running into trouble with the Delta variant.](#) Their efforts at containment so far have been valiant, but if their vaccine is as ineffective as it looks and they continue to be unable to admit this, the task is not going to be getting any easier. Containing the original variant for the past year had big advantages, but it also means no immunity other than from vaccinations. And if and when containment fails, it's going to be prohibitively expensive to try and put it back in place.

[Financial Times article explaining the breakthrough infections](#) has this excellent visualization:



I think this is pessimistic, and that vaccine protection against death is several times lower than this, but the representation here is excellent.

[This is not a great sign or a great look.](#)



BNO Newsroom @BNODesk · Jul 30

Passenger services banned from taking people to central Sydney to prevent anti-lockdown protest

...

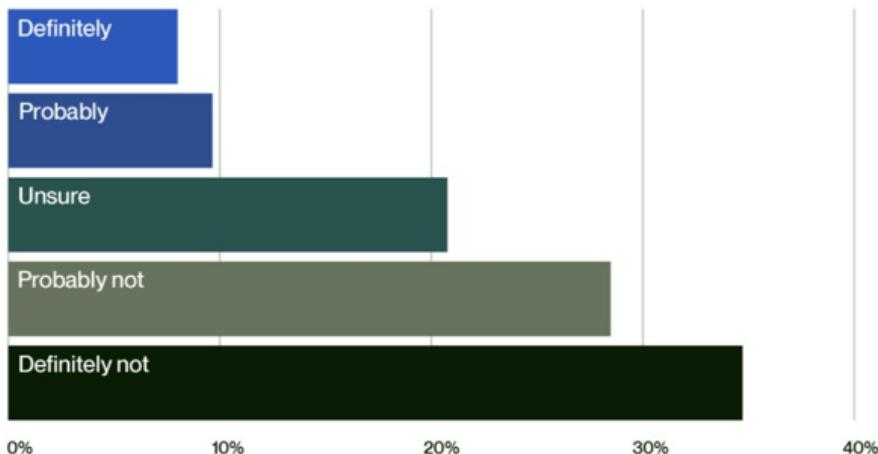
"A pandemic is not the time to protest and this prohibition notice is just one of the tools police have implemented today to ensure the safety of the community," Assistant Commissioner Thurtell said.

The government wishes to remind you that you should wait to protest the lockdown until after the lockdown has been lifted.

[Scott Aaronson introduces us to his term blankface](#), for a perspective on one of the causes of our problems responding reasonably to situations.

A lot of people [say they'll quit if forced to go back to the office](#), with almost 20% saying either definitely or probably, and another 20% being unsure, [and a lot of people being eager to go independent](#).

Please indicate the level of agreement with the following statement: If I have to go to the office I will consider looking for another job



I would be surprised if most of this were not cheap talk, but even a small amount of it not being cheap talk would be expensive. Also, good for the people involved, cause they're right, offices are mostly terrible, even if I'm excited to *sometimes* go into an office once it becomes possible. The last time I had to go to an office and be 'on' all day five days a week, even at an unusually good place to work with great people, it made me completely miserable until I figured it out and left.

[Alex Tabarrok points out](#) that the exact same 'first doses first' logic is now being applied to 'second doses first' as opposed to third doses, and attributes this to status quo bias. I think that's mostly right but imprecise and would frame it more as action vs. inaction, falling in line behind authority and the avoidance of potential blame, but it's mostly the same thing. And yes, the arguments about third doses are the same as the ones about second doses. This is similar to the shift in the UK to where second doses that are *insufficiently delayed* (e.g. are on the original schedule) are considered super risky because the immunity resulting isn't as good.

Not Covid

In existential risk news, Scott Alexander brings us [an updated look at long term AI risks](#). Assessments of these risks by those working in this field are super high compared to what one should be comfortable with given the stakes, but are remarkably low compared to what one might expect from people in the field. They're also remarkably spread out, in a way that doesn't feel like an attempt to model the future so much as an attempt to give answers that seem reasonable. Make of that observation what you will.

[Biden intentionally violates the constitution, but feels bad enough about it to admit it explicitly.](#)

There was simply no legal authority to get around it. The head of Biden's Covid task force, Jeff Zients, said the administration had "kicked every tire" searching for a justification, but had none.

"The president has not only kicked the tires; he has double, triple, quadruple checked," declared Biden adviser Gene Sperling.

Then, under intense pressure from the left, Biden reversed course and decided to have his CDC order another extension without any plausible legal authority. Incredibly enough, he was explicit that "the bulk of the constitutional scholars say it's not likely to pass constitutional muster."

But he said "several key scholars" told him it might, and he decided it would be worth the risk if it allowed extra time for already-allocated emergency rental funds to reach Americans who need them.

"At a minimum, by the time it gets litigated, it will probably give some additional time while we're getting that \$45 billion out to people who are in fact behind in the rent and don't have the money," Biden said.

[This WaPo article has more details.](#) It seems that all his lawyers told him no, this is blatantly illegal and forbidden and that which is forbidden is not allowed, including what is described as 'quadruple checking the tires' but when this was found to be an unacceptable conclusion, a search was on to find a new lawyer, somewhere, who would present a legal theory that could be presented as having a non-zero chance of standing up in court, and that's about the level of legal argument they found, but Biden takes too much pride in his understanding of government and straight talking attitude to hide what was going on.

He did all this in order to extend the eviction moratorium, which has now gone on over a year, and is essentially the taking of private property without compensation. If we are extending this now, when does it end? What happens to the rental market?

Not directly Covid, [but words of wisdom](#) on how to tell whether someone breached a contract:



Law Boy, Esq.
@The_Law_Boy

...

when someone says you breached a contract and you start talking about global pandemics you 100% breached that contract

 **Variety**  @Variety · Jul 29

Hours after Scarlett Johansson filed a lawsuit against Disney, the company has fired back, slamming the #BlackWidow  star's breach of contract lawsuit for showing "callous disregard for the horrific and prolonged global effects of the COVID-19 pandemic." bit.ly/3zGGq14

[Show this thread](#)



And now for something completely off-topic:

[I'd say 'I hope you're happy about it' to all the people I know who voted for it](#), except for the fact that *they are definitely happy about it*.

Bacon may disappear in California as pig rules take effect

By SCOTT McFETRIDGE · yesterday

At the beginning of next year, California will begin enforcing an animal welfare proposition approved overwhelmingly by voters in 2018 that requires more space for breeding pigs, egg-laying chickens and veal calves. National veal and egg producers are optimistic they can meet the new standards, but only 4% of hog operations now comply with the new rules. Unless the courts intervene or the state temporarily allows non-compliant meat to be sold in the state, California will lose almost all of its pork supply, much of which comes from Iowa, and pork producers will face higher costs to regain a key market.

Animal welfare organizations for years have been pushing for more humane treatment of farm animals but the California rules could be a rare case of consumers clearly paying a price for their beliefs.

With little time left to build new facilities, inseminate sows and process the offspring by January, it's hard to see how the pork industry can adequately supply California, which consumes roughly 15% of all pork produced in the country.

My model, based on what was motivating them, says: They're mildly disappointed that eggs and veal will remain available for purchase but happy about the improved conditions and higher prices. And they're downright giddy about the prospect of pork potentially being unavailable.

Want everything to grind to a halt? Propositions!

"It is important to note that the law itself cannot be changed by regulations and the law has been in place since the Farm Animal Confinement Proposition (Prop 12) passed by a wide margin in 2018," the agency said in response to questions from the AP.

Alas for them, supply and demand don't work the way the headline writers think...

Barry Goodwin, an economist at North Carolina State University, estimated the extra costs at 15% more per animal for a farm with 1,000 breeding pigs.

If half the pork supply was suddenly lost in California, bacon prices would jump 60%, meaning a \$6 package would rise to about \$9.60, according to a study by the Hatamiya Group, a consulting firm hired by opponents of the state proposition.

Adding an extra 15% cost effectively forces such farms to choose to either supply California or supply others, which creates two distinct pork supply chains, which adds further to costs. The thing is, that's the costs to adhere to the requirements, not the costs to *demonstrate compliance*. As with all such things, by the time all the paperwork is done, the costs might be considerably higher for that reason as well. The lack of any guidance on exact requirements or enforcement mechanisms is a lot of why almost no one is prepared to meet the new requirements – you can't comply until you know what compliance means. Plus even if you did know, what will tomorrow bring?

In Iowa, which raises about one-third of the nation's hogs, farmer Dwight Mogler estimates the changes would cost him \$3 million and allow room for 250 pigs in a space that now holds 300.

To afford the expense, Mogler said, he'd need to earn an extra \$20 per pig and so far, processors are offering far less.

"The question to us is, if we do these changes, what is the next change going to be in the rules two years, three years, five years ahead?" Mogler asked.

The California rules also create a challenge for slaughterhouses, which now may send different cuts of a single hog to locations around the nation and to other countries. Processors will need to design new systems to track California-compliant hogs and separate those premium cuts from standard pork that can serve the rest of the country.

But to me that's not the interesting part. The interesting part, to me, is that last line, which is an estimate of the elasticity of demand for pork, and it's shockingly high.

This is saying that if the pork supply was cut in half, all it would take would be a 60% increase in prices to clear the market. Wait, what? We're that sensitive to prices on something this cheap? Because pork prices at the local supermarket are *stupidly* cheap, in a 'can you believe civilization lets us eat meat for almost no money' kind of way.

As in these are *the Instacart prices*:

\$1.31 / lb reg. \$1.64

20% off

Fresh Picnic Pork Shoulder

\$0.08 per oz

\$2.63 / lb reg. \$3.07

14% off

Fresh Bone-In Pork Loin
Chops

\$0.16 per oz

\$3.29 / lb reg. \$4.39

25% off

Center Cut Pork Loin
Chops

\$0.21 per oz

\$4.94

ShopRite Thick Sliced Bacon 1 lb

You can pay a lot more to get higher quality stuff, if you so desire, but the prices really are super low. Bacon is more expensive by the pound, but I hope those involved are using smaller portions. Choose life.

So you're telling me, you raise these prices by 60%, and half the time people say 'nah, let's get something else instead.'? I definitely would not have expected that. It seems more like what would happen if a particular source increased prices, rather than all sources of pork together.

And then we come to a very strange statement:

At least initially, analysts predict that even as California pork prices soar, customers elsewhere in the country will see little difference. Eventually, California's new rules could become a national standard because processors can't afford to ignore the market in such a large state.

Increasing prices by 60% is estimated in California to cut demand in half. If 15% of the nation's pork goes to California now and half of it can't in the future, then that's 8% more for the rest of us, which in a perfect world might cut prices a few percent.

I don't see much danger in these standards becoming nationwide, unless other states are persuaded to pass similar laws. If California eats 15% of the pork at current prices, and compliance with their rules raises prices 15%, and raising prices 60% cuts demand in half, then it stands to reason that raising prices 15% would cut demand by 12%, so even if everyone made the move at once it's barely worth it, and you're going to get eaten alive by whoever didn't make the move. This seems like it's above the threshold of cost that California can impose.

But then again: You never know. Incentives matter.

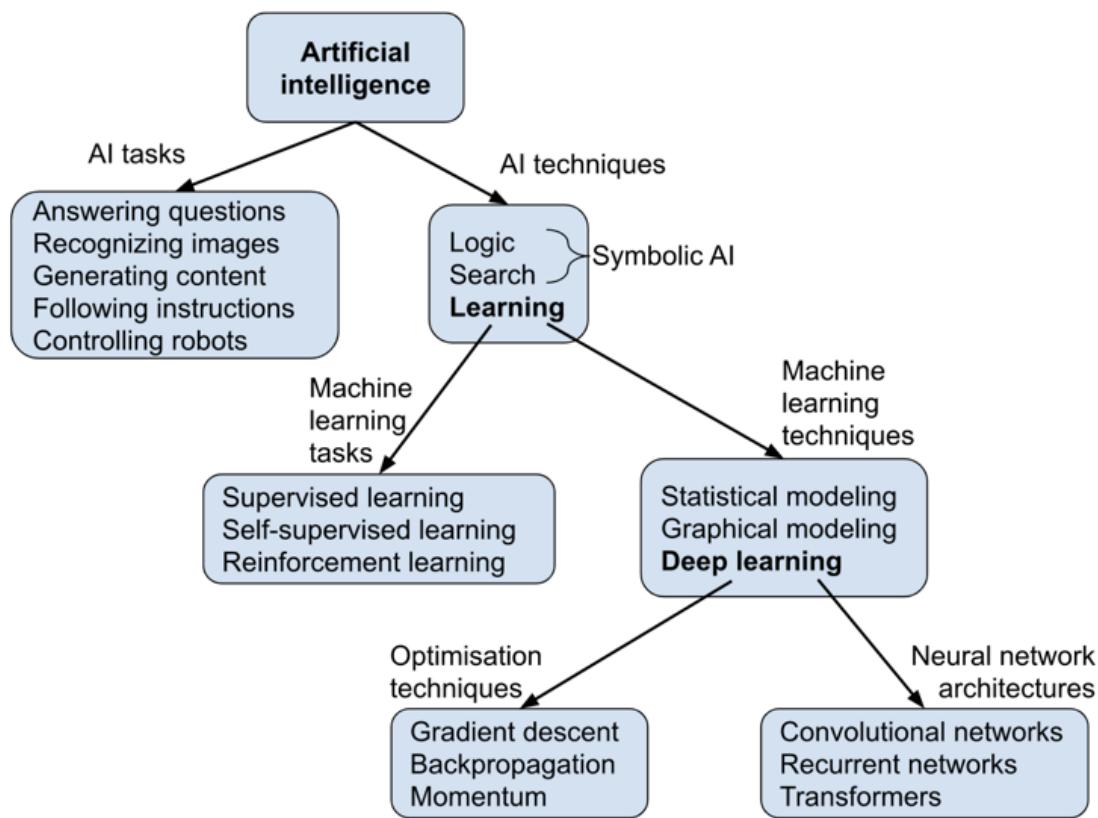
A short introduction to machine learning

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

Despite the current popularity of machine learning, I haven't found any short introductions to it which quite match the way I prefer to introduce people to the field. So here's my own. Compared with other introductions, I've focused less on explaining each concept in detail, and more on explaining how they relate to other important concepts in AI, especially in diagram form. If you're new to machine learning, you shouldn't expect to fully understand most of the concepts explained here just after reading this post - the goal is instead to provide a broad framework which will contextualise more detailed explanations you'll receive from elsewhere.

I'm aware that high-level taxonomies can be controversial, and also that it's easy to fall into the [illusion of transparency](#) when trying to introduce a field; so suggestions for improvements are very welcome!

The key ideas are contained in this summary diagram:



First, some quick clarifications:

- None of the boxes are meant to be comprehensive; we could add more items to any of them. So you should picture each list ending with "and others".
- The distinction between *tasks* and *techniques* is not a firm or standard categorisation; it's just the best way I've found so far to lay things out.
- The summary is explicitly from an AI-centric perspective. For example, statistical modeling and optimization are fields in their own right; but for our current purposes we

can think of them as machine learning techniques.

Let's dig into each part of the diagram now, starting from the top.

Paradigms of artificial intelligence

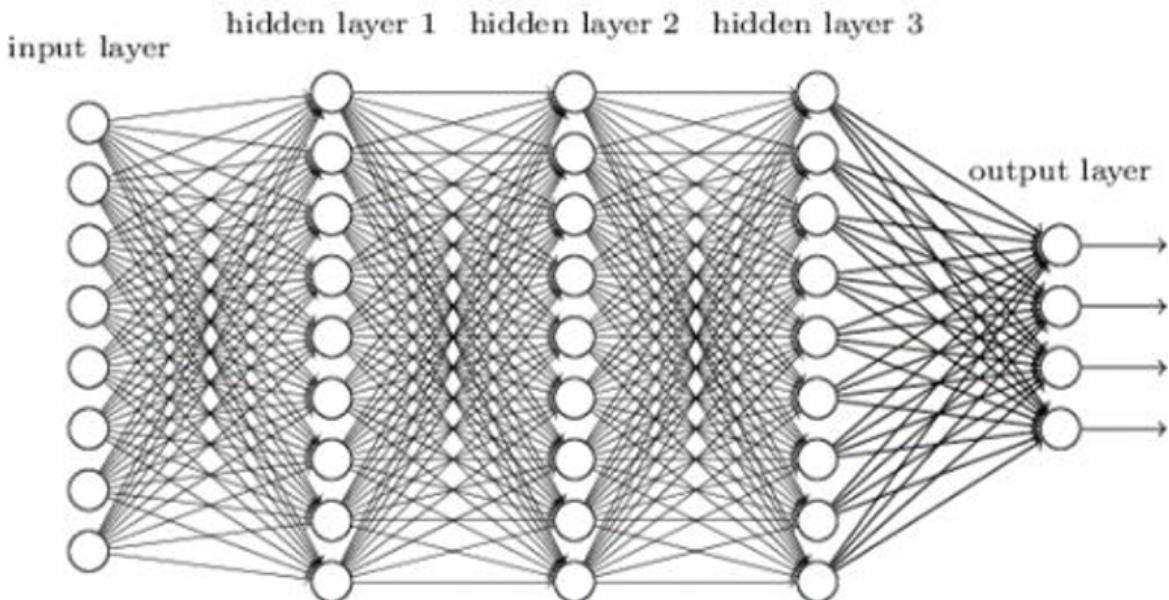
The field of **artificial intelligence** aims to develop computer programs that are able to perform useful tasks like answering questions, recognizing images, and so on. It got started around the 1950s. Historically, there have been several different approaches to AI. In the first few decades, the dominant paradigm was **symbolic AI**, which focused on representing problems using statements in formal languages (like logic, or programming languages), and searching for solutions by manipulating those representations according to fixed rules. For example, a symbolic AI can represent a game of chess using a set of statements about where the pieces currently are, and a set of statements about where the pieces are allowed to move (you can only move bishops diagonally, you can't move your king into check, etc). It can then play chess by searching through possible moves which are consistent with all of those statements. The power of symbolic search-based AI was showcased by [Deep Blue](#), the chess AI that beat Kasparov in 1997.

However, the symbolic representations designed by AI researchers turned out to be far too simple: there are very few real-world phenomena easily describable using formal languages ([despite valiant efforts](#)). Since the 1990s, the dominant paradigm in AI has instead been **machine learning**. In machine learning, instead of manually hard-coding all the details of AIs ourselves, we specify models with free parameters that are learned automatically from the data they're given. For example, in the case of chess, instead of using a fixed algorithm like Deep Blue does, a ML chess player would choose moves using parameters that start off random, and gradually improve those parameters based on feedback on its moves: this is known as the *learning, training or optimization process*.* In theory, statistical models (including simple models like linear regressions) also fit parameters to the data they're given. However, the two fields are distinguished by the scales at which they operate: the biggest successes of machine learning have come from training models with billions of parameters on huge amounts of data. This is done using **deep learning**, which involves training *neural networks* with many layers using powerful *optimization techniques* like gradient descent and backpropagation. Neural networks have been around since the beginning of AI, but they only became the dominant paradigm in the early 2010s, after increases in compute availability allowed us to train much bigger networks. Let's explore the components of deep learning in more detail now.

Deep learning: neural networks and optimization

Neural networks are a type of machine learning model inspired by the brain. As with all machine learning models, they take in input data and produce corresponding output data, in a way which depends on the values of their parameters. The interesting part is *how* they do so: by passing that data through several layers of simple calculations, analogous to how brains process data by passing it through layers of interconnected neurons. In the diagram below, each circle represents an "artificial neuron"; networks with more than one layer of neurons between the input and the output layers are known as *deep neural networks*. These days, almost all neural networks are deep, and some have hundreds of layers.

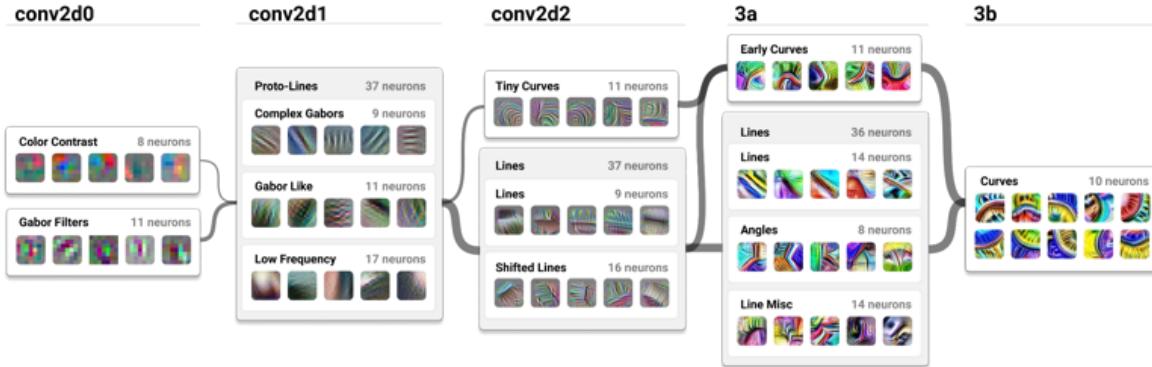
Deep neural network



Each artificial neuron receives signals from neurons in the previous layer, combines them together into a single value (known as its *activation*), and then passes that value on to neurons in the next layer. As in biological brains, the signal that is passed between a pair of artificial neurons is affected by the strength of the connection between them - so for each of the lines in the diagram we need to store a single number representing the strength of the connection, known as a *weight*. The weights of a neuron's connections to the previous layer determines how strongly it activates for any given input. (Compared with biological brains, artificial neural networks tend to be much more strictly organised into layers.)

These weights are not manually specified, but instead they are learned via a process of **optimization**, which finds weights that make the network score highly on whatever metric we're using. (This metric is known as an *objective function* or *loss function*; it's evaluated over whatever dataset we're using during training.) By far the most common optimization algorithm is **gradient descent**, which initially sets weights to arbitrary values, and then at each step changes them so that the network does slightly better on its objective function (in more technical terms, it updates each weight in the direction of its gradient with respect to the objective function). Gradient descent is a very general optimization algorithm, but it's particularly efficient when applied to neural networks because at each step the gradients of the weights can be calculated layer-by-layer, starting from the last layer and working backwards, using the **backpropagation** algorithm. This allows us to train networks which contain billions of weights, each of which is updated billions of times.

As a result of optimization, the weights end up storing information which allows different neurons to recognise different features of the input. As an example, consider a neural network known as *Inception*, which was trained to classify images. Each neuron in Inception's input layer was assigned to a single pixel of the input image. Neurons in each successive layer then learned to activate in response to increasingly high-level features of the input image. The diagram shows some of the patterns recognised by neurons in five consecutive layers from the Inception model, in each case by combining patterns from the previous layer - from colours to (Gabor filters for) textures to lines to angles to curves. This goes on until the last layer, which represents the network's final output - in this case the probabilities of the input image containing cats, dogs, and various other types of object.



One last point about neural networks: in our earlier neural network diagram, every neuron in a given layer was connected to every neuron in the layers next to it. This is known as a fully-connected network, the most basic type of neural network. In practice, fully-connected networks are seldom used; instead there are a whole range of different neural network architectures which connect neurons in different ways. Three of the most prominent (convolutional networks, recurrent networks, and transformers) are listed in the original summary diagram; however, I won't cover any of the details here.

Machine learning tasks

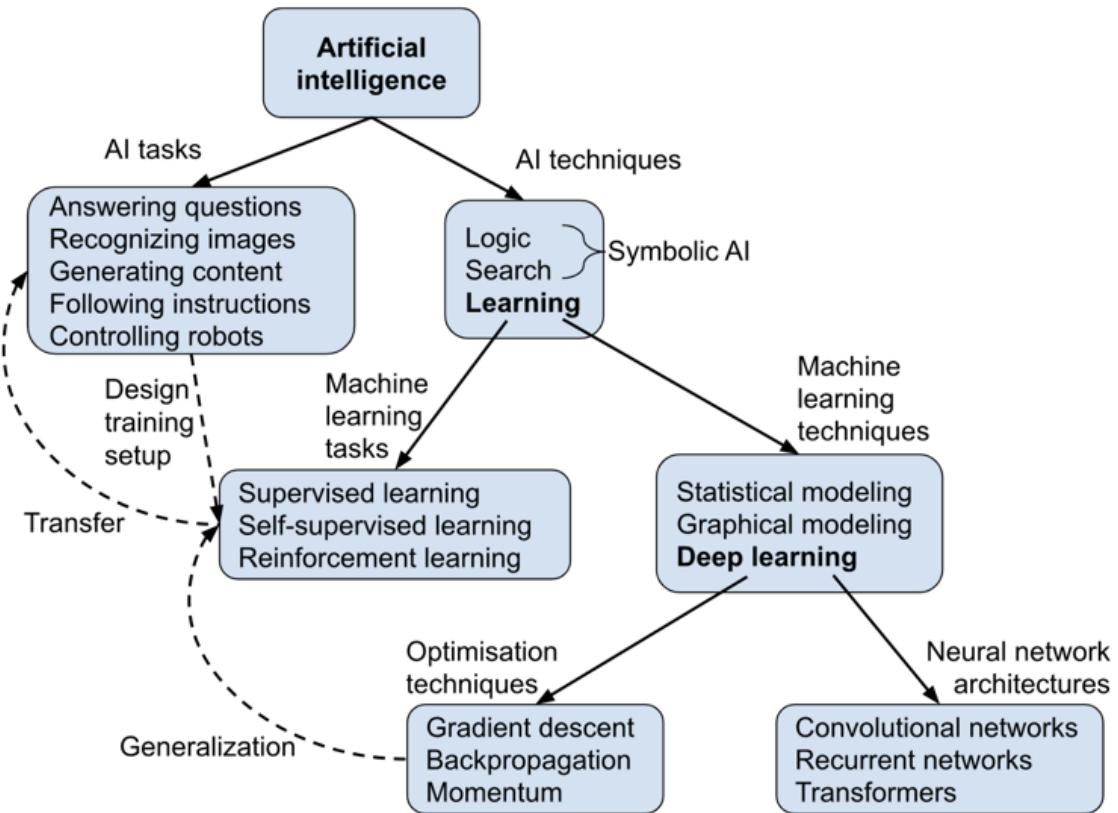
I've described how neural networks (and other machine learning models) can be trained to perform different tasks. The three most prominent categories of tasks are supervised, self-supervised, and reinforcement learning, which each involve different types of data and objective functions. **Supervised learning** requires a dataset where each datapoint has a corresponding label. The objective in supervised learning is for a model to predict the labels which correspond to each datapoint. For example, the image classification network we discussed above was trained on a dataset of images, each labeled with the type of object it contained. Alternatively, if the labels had been ratings of how beautiful each image was, we could have used supervised learning to produce a network that rated image beauty. These two examples showcase different types of supervised learning: the former is a *classification problem* (requiring the prediction of discrete categories) and the latter is a *regression problem* (requiring the prediction of continuous values). Historically, supervised learning has been the most studied task in machine learning, and techniques devised to solve it have been extensively used as parts of the solutions to the other two.

One downside of supervised learning is that labeling a dataset usually needs to be done manually by humans, which is expensive and time-consuming. Learning from an unlabeled dataset is known as **unsupervised learning**. In practice, this is typically done by finding automatic ways to convert an unlabeled dataset into a labeled dataset, which is known as **self-supervised learning**. The standard example of self-supervised learning is next-word prediction: training a model to predict, from any given text sequence in an unlabeled dataset, which word follows that sequence. Some impressive applications of self-supervised learning are [GPT-2](#) and [GPT-3](#) for language, and [Dall-E](#) for images.

Finally, in **reinforcement learning**, the data source is not a fixed dataset, but rather an *environment* in which the AI takes actions and receives observations - essentially as if it's playing a video game. After each action, the agent also receives a reward (similar to the score in a video game), which is used to reinforce the behaviour that leads to high rewards, and reduce the behaviour that leads to low rewards. Since actions can have long-lasting consequences, the key difficulty in reinforcement learning is determining which actions are responsible for which rewards - a problem known as *credit assignment*. So far the most impressive demonstrations of reinforcement learning have been in training agents to play board games and esports - most notably [AlphaGo](#), [AlphaStar](#) and [OpenAI Five](#).**

Solving real-world tasks

We're almost done! But I don't think that even a brief summary of AI and machine learning can be complete without adding three more concepts. They don't quite fit into the taxonomy I've been using so far, so I've modified the original summary diagram to fit them in:



Let's think of these three dotted lines I've added as ways to connect the different levels. The ultimate goal of the field of AI is to create systems that can perform valuable tasks in the real world. In order to apply machine learning techniques to achieve this, we need to design and implement a supervised/self-supervised/reinforcement **training setup** which allows systems to learn the necessary abilities. A key element is designing datasets or training environments which are as similar as possible to the real-world task. In reinforcement learning, this also requires designing a reward function to specify the desired behaviour, which is [often more difficult than we expect](#).

But no matter how good our training setup, we will face two problems. Firstly, we can only ever train our models on a finite amount of data. For example, when training an AI to play chess, there are many possible board positions that it will never experience. So our optimization algorithms could in theory produce chess AIs that can only play well on positions that they already experienced during training. In practice this doesn't happen: instead deep learning tends to **generalise** incredibly well to examples it hasn't seen already. How and why it does so is, however, still poorly-understood.

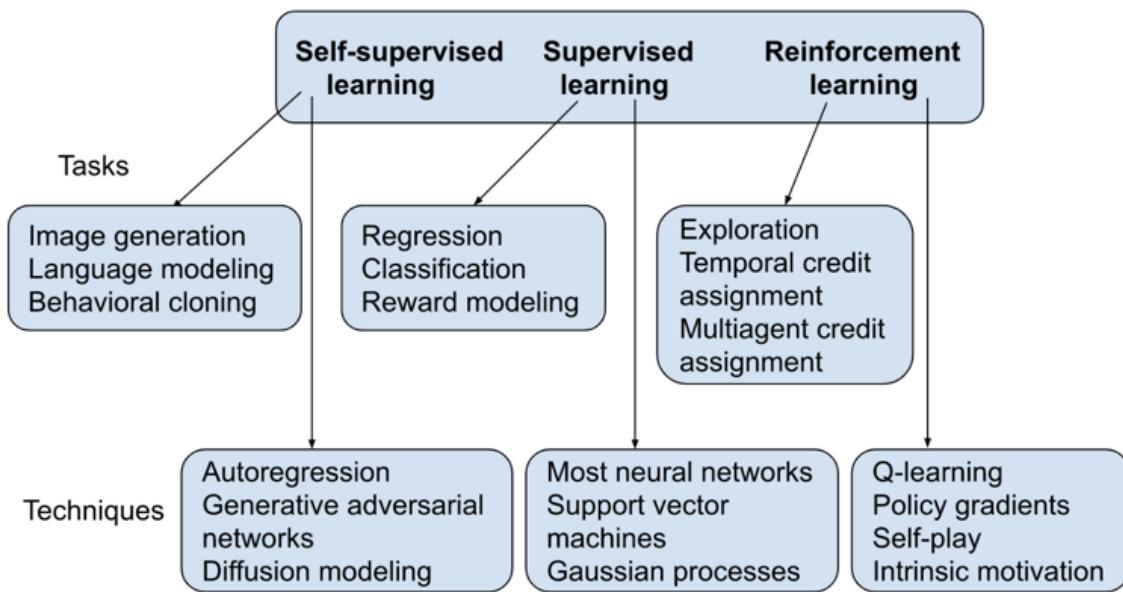
Secondly, due to the immense complexity of the real world, there will be ways in which our training setups are incomplete or biased representations of the real-world tasks we really care about. For example, consider an AI which has been trained to play chess against itself, and which now starts to play against a human who has very different strengths and weaknesses. Playing well against the human requires it to **transfer** its original experience to

this new task (although the line between generalisation to different examples of “the same task” versus transfer to “a new task” is very blurry). We’re also beginning to see neural networks whose skills transfer to new tasks which differ significantly from the ones on which they were trained - most notably the GPT-3 language model, which can perform [a very wide range of tasks](#). As we develop increasingly powerful AIs that perform increasingly important real-world tasks, ensuring their safe behaviour will require a much better understanding of how their skills and motivations will transfer from their training environments to the wider world.

Footnotes

* *Learning, training and optimization* have slightly different connotations, but they all refer to the process by which a machine learning system updates its parameters based on data.

** Here’s a more detailed breakdown of some of the tasks and techniques corresponding to these three types of learning. I’ve only mentioned a few of these terms so far; I’ve included the others to help you classify them in case you’ve seen them before, but don’t worry if many of them are unfamiliar.



Provide feedback on Open Philanthropy's AI alignment RFP

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

Open Philanthropy is planning a request for proposals (RFP) for AI alignment projects working with deep learning systems, and we're looking for feedback on the RFP and on the research directions we're looking for proposals within. We'd be really interested in feedback from people on the Alignment Forum on the current (incomplete) draft of the RFP.

The main RFP text can be viewed [here](#). It links to several documents describing two of the research directions we're interested in:

- [**Measuring and forecasting risks**](#)
- [**Techniques for enhancing human feedback**](#) [Edit: this previously linked to an older, incorrect version]

Please feel free to comment either directly on the documents, or in the comments section below.

We are unlikely to add or remove research directions at this stage, but we are open to making any other changes, including to the structure of the RFP. We'd be especially interested in getting the Alignment Forum's feedback on the research directions we present, and on the presentation of our broader views on AI alignment. It's important to us that our writing about AI alignment is accurate and easy to understand, and that it's clear how the research we're proposing relates to our goals of reducing risks from power-seeking systems.

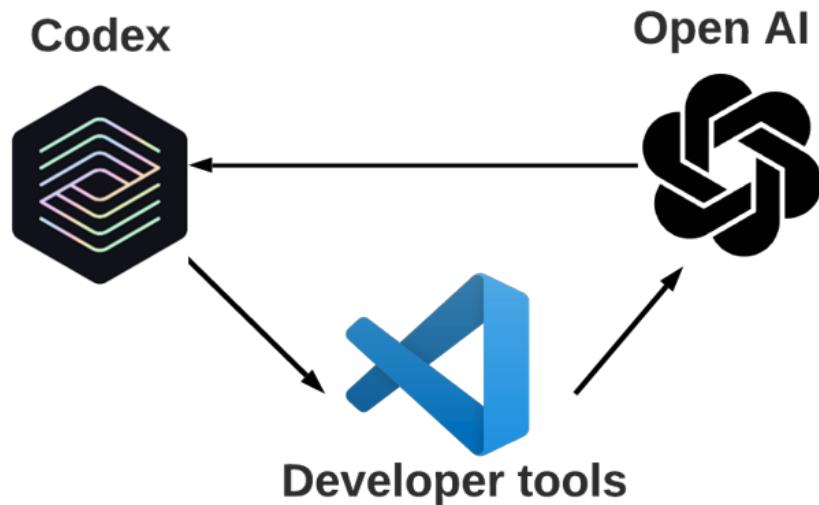
The Codex Skeptic FAQ

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

Most of my programmer friends believe that Language Models trained on code will not affect their day job anytime soon. In this post, I make the case that 1) code generation is already useful (assuming minimal prompt engineering skills) 2) even if you do not believe in 1), code generation will increase programmers' throughput way sooner than it will fully automate them.

Language Models trained on Code do not bring us closer to Full Code Automation

This misconception comes from thinking linearly instead of exponentially. Language models are good enough at generating code to make the very engineers building such models slightly more productive, for instance when dealing with a new API. In other words, the returns (aka the improvements in the algorithm) from investing more resources in code generation directly helps (with better developer tools) create a better code-generating algorithm.



Code generation does not automate the part of my workday where I think hard

- It still accelerates “glue code” or “API work”—a substantial fraction of large codebases.
- Besides, only a set of privileged engineers get to think about the broad picture every day.
- Plus, hard thinking is mostly required at the start, when designing the architecture.
- And thinking seldom happens in a silo. It instead requires many iterations, through coding.

I asked a model to generate code but it doesn't seem to be able to solve it

More often than not, the issue is not about the model. Try another prompt. ([Example](#))

The output is outdated code from average programmers

Code quality (length, variable naming, taste) is prompt and hyperparameter dependent. Generally, language models use variables from the prompt and you can rename those yourself.

Only developers who repeat the same tasks will be automated so it will not affect me

You might still see gains in productivity in learning how to use a more advanced version.

My job does not involve solving simple coding tests from docstrings

You should be capable of separating your code in smaller functions and write docstrings.

Codex cannot solve my problem since it has only access to a limited training set

Github Copilot [stores](#) your data. Supposedly, the same applies to the Codex beta.

Current Language Models still make silly mistakes

If the mistake is silly, then fixing it is trivial.

Anyway, it is error prone so it cannot be used for critical software

It generates less error than I do when writing code for the first time.

I would strongly suggest applying to Github Copilot or OpenAI Codex access to check for yourself, avoiding cherry-picked examples on the internet (in good and in bad). Indeed, if you search online, you might run into [outdated](#) reviews, where it turns out that highlighted errors actually work now. If you cannot wait for beta access, I recommend asking a friend for a demo (I'm happy to showcase it to everyone), trying [genji python](#) or reading [this](#) up-to-date review.

More generally, programmers should seriously consider learning prompt engineering to avoid being left behind, and, I believe, any future forecast about AI progress should include this shorter loop between deep learning models and programmer productivity.

When Programmers Don't Understand Code, Don't Blame The User

([Cross posted](#) on my personal blog.)

In [Undergraduation](#), Paul Graham said something that has always stuck with me:

On the whole, grad school is probably better than most alternatives. You meet a lot of smart people, and your glum procrastination will at least be a powerful common bond. And of course you have a PhD at the end. I forgot about that. I suppose that's worth something.

The greatest advantage of a PhD (besides being the union card of academia, of course) may be that it gives you some baseline confidence. For example, the Honeywell thermostats in my house have the most atrocious UI. My mother, who has the same model, diligently spent a day reading the user's manual to learn how to operate hers. She assumed the problem was with her. But I can think to myself "If someone with a PhD in computer science can't understand this thermostat, it must be badly designed."

I thought about this today. In a pull request at work, someone had some code that looked like this:

```
const arr = [stuff]
const that = this;

arr.forEach(function () {
    that.method();
});
```

Someone else commented about how `that = this` isn't necessary. You should just be able to do `this.method()`. I've been programming with JavaScript for about eight years now, and I found myself confused.

I know that the value of `this` is supposed to be based on what calls the function. Like if you have `obj.fn()`, inside of `fn`, `this` will be `obj`. But what happens when you just do `fn()`?

That is where I get confused. Without looking it up, my memory says that it usually defaults to the global object, but that there may be other rules I'm forgetting that have to do with the scope. When I looked it up, it *seemed* that it does default to the global object, but I wasn't 100% sure. Maybe it does have to do with scope. The docs didn't seem clear. So it took me some time to think up an example to test each hypothesis, which ended up allowing me to prove to myself that it is in fact defaulting to the global object, and that it doesn't care about scope.

But then there is the question of what is going on in [Array.prototype.forEach](#). How is that callback function getting executed? Does the value of `this` get "set by the call"? Eg. does something like `obj.cb()` happen, in which case the value of `this` is `obj`? If so, what is `obj`? Or does it get executed like `cb()`? The docs don't really make that clear IMO.

Anyway, the point of explaining all of this is to give you a sense that I find this stuff a little tricky to think about.

So what does that mean? Am I dumb? Am I a bad programmer? Not to sound arrogant, but I don't think either of those things are the case. I don't have a PhD like Paul Graham does, but I do have other credentials and other reasons to believe that I am reasonably intelligent. So then, I think back to that quote:

If someone with a PhD in computer science can't understand this thermostat, it must be badly designed.

I suspect that something similar is going on here.

If someone with eight years of experience programming in JavaScript can't understand this, it must be badly designed.

Consider this quote from Edsger Dijkstra's 1972 Turing Award lecture, [The Humble Programmer](#):

We shall do a much better programming job, provided that we approach the task with a full appreciation of its tremendous difficulty, provided that we stick to modest and elegant programming languages, provided that we respect the intrinsic limitations of the human mind and approach the task as Very Humble Programmer.

The idea is that the human mind is small and fragile, so when programming, we have to design things so that they have soft edges and round corners and can fit inside our limited minds.

Here's another perspective. When dealing with software, we should think like [usability designers](#). If you're a usability designer and you see, through user testing, that people have trouble understanding your product, *you don't blame the user!* You blame the product. You go back to the drawing board and figure out a way to make it simpler.

We shouldn't be afraid to adopt that mindset when we come across tricky things in our code. Stop thinking "This is a basic thing that I'm supposed to know". There is reason to believe that we are intelligent people, so if we have trouble understanding software, maybe the problem is with how the software is designed, rather than with our own intelligence.

Implicature Conflation

For some reason, I keep thinking about what [General Semantics](#) might be trying to get at. In my [previous post](#) on the subject, I identified *equivocation* as the main curse which General Semantics seeks to provide a counterspell for. (AKA conflation, identification, fusion, ...) However, equivocation has a lot of faces. It doesn't work to just say "try really hard not to equivocate" -- we need the details. We need tools to notice specific classes of equivocation and un-equivocate them.

Today, I want to talk about equivocation which *conversational implicature* brings about.

Conversational Implicature

In case you're not familiar with the term, "conversational implicature" refers to all the meaning that's *not* explicit. For example, if I say "you didn't do the laundry yet?" I *might* (depending on context) be implying "you should really do it now". This is implicature, because I didn't say it explicitly; it is inferred from context.

I've written before about [Bayesianism and the explicit/implied distinction](#). I wrote that Bayesian signalling theory isn't adequate to understand the distinction: meaning is meaning, and in particular, the meaning of an utterance is what people understand from it. How could we make a Bayesian distinction between "explicit" and "implied" meaning?

(To put it a different way: from a map/territory standpoint, someone talking about "explicit" meaning might sound confused: the meaning of the map is precisely the way it corresponds to (IE statistically conveys information about) the territory. It's not like you actually hand over chunks of the territory! So what is this "explicit meaning"?)

I nonetheless do think there is such a thing as "explicit meaning". It may have gradations and nuances (many "explicit" phrasings have their roots in analogy, eg, "she's on time" doesn't make sense if you try to be very literal about what "on" means), but by and large, we can make a reasonable dividing line.

Policy Debates

In [Policy Debates Should Not Appear One-Sided](#), Eliezer identifies a common failure mode, where listing cons is conflated with arguing against all the pros, (and listing pros is conflated with arguing against all the cons):

Robin Hanson proposed stores where banned products could be sold.¹ There are a number of excellent arguments for such a policy—an inherent right of individual liberty, the career incentive of bureaucrats to prohibit *everything*, legislators being just as biased as individuals. But even so (I replied), *some* poor, honest, not overwhelmingly educated mother of five children is going to go into these stores and buy a "Dr. Snakeoil's Sulfuric Acid Drink" for her arthritis and die, leaving her orphans to weep on national television.

I was just making a factual observation. Why did some people think it was an argument in favor of regulation?

I now think this is one example of the more general rule: people conflate explicit content with conversational implicature *all the time*.

In a policy debate, it's common that there *is* one big decisive factor. So it often makes sense for participants to bring up things *they* think are decisive in one direction or the other. So when someone brings up a pro/con, there *is*, statistically, an implicature that the pro/con is decisive.

Of course, this is expected, and usually beneficial. Implicature *is part of the meaning* of an utterance. It's socially expected that you'll understand it and respond to it.

In fact, I think it's often useful to *totally ignore explicit content* and deal only with implicature (this reduces friction when dealing with people who are not very rationalist-compatible).

But carrying this too far is toxic for policy debate. If every pro/con is seen as implicitly carrying the claim "this is decisive", then anyone listing a pro/con is seen as *arguing against all the cons/pros*, which makes cost/benefit analysis impossible.

Against Implicature?

I used to derive a heavily anti-implicature lesson from Eliezer's post. I said to myself: it's important to sometimes make true observations without a "point" you're arguing for/against; otherwise you will have [already written the bottom line](#), so you won't be doing useful computation.

But this is a bit absurd. Does the ideal rationalist just spout facts at random, to avoid drawing a bottom line?

Similarly, one might read Eliezer as suggesting that people shouldn't make implicature inferences. "I never mean that a con is decisive, unless I say so, and you should never infer that about someone else either!" But this is almost as absurd. Ignoring implicature is not going to be practical.

Just Stop Equivocating

Instead, the advice I derive is *try to stop conflating the explicit content and the implicature*. This doesn't mean valuing one and discarding the other. Don't think like one is "more legitimate" than the other. Just note the difference between them.

There's a weird meta thing going on here: if I say "note the difference", I worry you'll hear *some specific point to it*, like "note the difference so that you can discard the implicature" or "note the difference so that you can pay attention to the implicature". No! There is no one appropriate rule for all circumstances.

STUDENT: Then why do we note the difference? Surely there is some point?

MASTER: Yes, of course there is. I want you to stop equivocating because equivocation is dangerous. We already covered how toxic it can be for policy debate.

STUDENT: I understand that example, but what general lesson am I supposed to learn? Surely you want me to stop equivocating because you expect good consequences. But what are the gears of those good consequences? What do you expect to happen next, if in a particular case I succeed at "stop equivocating"?

MASTER: In some cases, you will realize that the literal content was itself the point, and you should stop searching for one. Imagine a Master of Literalness who always speaks literally to the student, and tries to teach the student to do the same. Every day, the master lectures on speaking literally. Every day, the student asks the same question: what is the point of all this, master? When asked the point, the master only repeats the same teaching, over and over again. The student despairs. One day, the student realizes that the master was speaking simply all along. The point was precisely what the master was saying. The student is enlightened.

STUDENT: But this is not your teaching. You don't tell us that we should speak simply at all times.

MASTER: No, it is not my teaching. The student in the story learned a valuable lesson, namely, that it is possible to speak completely literally with no hidden agenda. However, possible does not mean desirable. The point of the story is that the student was unable to see what was right in front of them, because they thought every sentence should have a hidden meaning, a "point". This can happen to you if you live in the world of implicature alone. Other bad things can happen to you if you live in the world of literal meaning alone.

STUDENT: So you're saying that we should consider the possibility that people mean exactly what they say.

MASTER: Yes, but also consider the possibility that they don't.

STUDENT: So we should consider all possibilities at all times.

MASTER: Humans are not capable of this. We need to ignore most of the possibilities most of the time. But this creates a trap: how can we learn that we need to raise something to our attention, if we're always ignoring it? I'm mentioning the possibility now, so that you might raise it to attention when it's relevant, and avoid the trap.

But how do we stop equivocating?

Actually, "just stop equivocating" isn't very specific advice. If you're equivocating, it can be really, really hard to stop.

I think it's a bit like grammar. You could speak English fluently all your life, but not be able to identify a "preposition" or a "gerund". Just because your brain is juggling explicit meaning and implicature all the time doesn't make the distinction readily available.

But it's even worse than grammar, because with grammar, everyone goes through multiple semesters of explicit content. I think most people are mostly blind to the literal/implicature distinction most of the time (especially young people?). People know it exists, but only really think of it as applying to extreme cases (such as veiled threats). More knowledgeable people might know that implicature happens "everywhere all of the time", but think you have to be a linguist to spot it.

So, if you had a blind spot around the distinction, but wanted to get better, how would you start?

It's helpful to just pay attention. What did you say, and what did you mean by it? What did other people say, and what do you think they meant by it?

It's helpful to have peers who are also doing this, and talking about it.

I think it's easiest to notice when someone infers something which you didn't intend. Then, the subject can be brought up and examined. (Don't attack the person with "that's not what I said"! Remember, inferring implicatures is a perfectly normal part of functional communication. Scolding people for it only serves to reinforce implicature blindness. The gentler "that's not what I meant" makes more sense.)

The English Language

The English language explicitly conflates the two in the phrase "what are you saying?" -- I often get confused reactions when I comment about "what a person said" (intending to refer only to the explicit content). Even when I make the distinction clear, I often get reactions like I'm playing some verbal game or something. (I think people might *usually* make explicit/implicit distinctions as a joke!)

The evolution of the term "literally" into an [intensifier](#) doesn't help matters, obviously. It's already a bit sad when a word like "impossibly" has its literal meaning eroded as it gains use as an intensifier (as in the phrase "impossibly large", when used to refer to things which are quite possible). It's darkly ironic when it happens *to the word we would use to distinguish between such exaggerations and plain language!*

Speaking Plainly

Rationalists are, on the whole, probably considerably above average in literalness. I've argued that one should not try to go all the way: implicature is good and proper.

But *could* one go all the way? As an exercise, perhaps?

I would argue that, yes, *it is possible to say almost all of what you mean*. It's possible to pause a conversation and unpack, and within a few minutes, to be finished unpacking to the point where any further details are negligible. Perhaps you can't *recursively unpack everything* (EG, getting into your own intentions behind pausing to unpack); that might get into an infinite recursion. But I claim you can unpack *any one thing* with reasonable thoroughness in a reasonable amount of time.

(My own explorations of this have been under the influence of a drug, which made it feel necessary to unpack implicature to literal content in order to be sure of what's going on; so, take that as you will.)

In many cases, this is accomplished by replacing implicature with vague language; EG, "The Latin language doesn't rely so much on word order *and that's somehow good*" (without trying to immediately unpack what 'good' means). That's OK. It can open the door for further clarifications if needed. We don't have to resolve all linguistic ambiguity at once.

This takes some of the same skills as Radical Honesty. Often, there's a good reason (beyond efficiency) why so much is left unsaid. Saying them directly is perceived as too forceful, impolite, etc. It can take courage to speak literally.

Imagine for a moment that you are a student of the Master of Literalness from my earlier story. You've been soaked in pro-literalness culture. To you, leaving things unstated feels dishonest and dangerously ambiguous. You know the importance of one's words, and you see how people constantly avoid culpability by saying things without saying them. *You say what you mean.*

It seems like an interesting exercise, to try and talk as if you had such an extreme perspective.

There's a lightness that comes with speaking plainly. You don't have to defend some hidden, unarticulated point. If someone says something true, you can simply agree with them.

But in the end, we have to wake back up to the world of implicature. If someone expresses (what they think is) a counterpoint to what you are saying, and you simply agree, they'll probably think you concede the whole argument. Maybe you should! But if not, you should probably recognize the implicature and attempt to communicate further.

A Response to A Contamination Theory of the Obesity Epidemic

This is a linkpost for <https://goodtosell.substack.com/p/a-response-to-a-contamination-theory>

EDIT: 8/12 added a paragraph with a link to my post with more stats under the regression

Copied and pasted in its entirety from [my blog](#)

Time to try my hand at another [unsolvable crisis of the modern world](#). It is my opinion that something doesn't add up about the increase in obesity and in all metabolic disorders in industrialized countries. There are a bevy of different theories out there, and [it was my initial intuition](#) that the problem, like many others plaguing complex systems, was multi-faceted. In addition, if there was an easy solution, epistemic humility would force me to contend with the fact that I'm just a layman in terms of nutrition, and other people should have stumbled upon the solution.

I'll lay my biases out right now: I didn't have a particular horse in this race diet-wise (as you can see from my initial reaction to the LessWrong post), but I definitely do want this to be an easy fix. Something along the lines of the prospect of plastics building up in us with no recourse is not a pleasant one to think about for me, I'd much rather we could just substitute an ingredient here or there.

Nevertheless, there is a theory that is so compelling that I would be remiss not to give my take here.

First, it is important to appreciate that there is something [really weird](#) (the Peery paper) about obesity in the modern world (PDF warning). [The good people of LessWrong](#) also seem to think so. I highly recommend skimming both those articles before you go forward with mine. I very much agree with their assessment of the weirdness of the epidemic. I also think our current explanations are quite lacking.

Where There's Smoke

I'll cut to the chase, I think the smoking gun in this whole thing is the addition of vegetable oils to our diet. I go into it a bit in my comment [here](#), but all that info will be the same as this post.

Firstly, I highly recommend reading [this](#) series of posts by Jeff Nobbs. He is more responsible for the meat and bones of this post than I am, I'm just relating it to the Peery paper and musing a bit. It lays out an excellent and compelling case for at least looking at vegetable oils as the culprit.

Throughout our decades-long battle with chronic disease, Americans have closely followed everything the CDC, AHA, and USDA have told us to do. We're smoking less, drinking less, exercising more, eating less saturated fat and sodium, and eating more fruits and vegetables. Still, chronic disease and obesity rates continue to rise. All the while, vegetable oil has steadily and stealthily made its way into our

pantries, restaurants, and packaged foods, now contributing 699 calories per day to our diets, or about 20% of everything we eat.

That about sums it up, I was quite convinced initially. From nearly nothing to 20% of our diet! The correlation is very striking, but the causal mechanisms aren't studied well at all in humans. It's actually quite weird how few good studies there are of vegetable oil in humans. The ones that are good are further in Nobb's series, and they seem to be really bad for vegetable oil.

Line by Line

As I find the assessment of the issue and mystery of obesity compelling in the Peery paper, I will use that as a starting point, attempting to rectify the thesis of vegetable oil with their diagnosis of the symptoms of the obesity epidemic. They list the following mysteries of the epidemic:

Changed over the last hundred years

Yep, [here's](#) an article going into its origin in the American diet. Note the potential conflicts of interest. Not a buyer of a conspiracy here, but there are cards on the table.

With a major shift around 1980

For the 1980 thing, I think they focus on that date a little too much, it's a monotonic increase in both oils and obesity all the way up. Nobbs cites [this](#) guideline, 1980 on the dot, that says "Avoid too much fat, Saturated fat, and cholesterol." As he notes, the public actually did an alright job following all these guidelines. We largely replaced the fats with soybean oil.

Additionally,

-It could be a threshold that got passed around that time—it looks like that's about when the average man went from just under to just about overweight BMI per the chart in the Peery paper.

-Perhaps a large cohort was hitting a certain age around that time?

-Global trade really starts kicking off, essentially [jumping from 10 to 15% of GDP in 1980](#). The US grows most of its own soybeans, but this could explain why other industrialized countries go up around then as well.

And whatever it is, there is more of it every year

Per Nobbs, it has been exploding. [Here's](#) another source.

It doesn't affect people living nonindustrialized lives, regardless of diet

Global trade, new invention from USA, ticks this box for me. In particular, the evidence from Cuba and the pacific islands in Peery really points towards the issue being primarily in something that is imported.

But it does affect lab animals, wild animals, and animals living in zoos

This is one of the most interesting points that the Peery paper goes into. If truly wild animals are having the same issues we are, then it would point towards their thesis of an environmental contaminant. However, I looked through their [source](#), and it seems that:

In this light, we compiled data to assess time trends in body weight in mammalian species that live with or around humans in industrialized societies.

It primarily looks at animals in labs or under the direct supervision of humans. I didn't read every single source in that paper, but you can see for yourself. The animals look to be eating food made by humans. When they looked at city rats, they got less obese than labrats. This would be consistent with the hypothesis that it is vegetable oil in human foods causing this, if wild rats eat less food made by humans as a proportion of their diet.

I dug through this paper a bit more, called "Canary in the coal mine." It also cites [this paper](#) as evidence of

That large population level changes in body weight distributions of mammalian populations can occur even when those populations are neither under obvious selection by predation nor are living with or among humans has been documented

This was what I was really looking for. Fat cats and raccoons make sense, what about the deer?

But the source says:

In particular, the recent trend of increasingly warm winters in northern Europe and Scandinavia may lead to reduced body size and fecundity of red deer, and perhaps other ungulates, in those areas.

I don't see the evidence here that contaminants in the environment from humans are causing an increase in the weight of animals that *don't eat food made by humans*. That paper was all about climate change and deer getting smaller, I'm not even able to figure out why they were really citing it anyways.

Overall, I am not convinced that truly wild animals, eating their own food sources, are being taken for the obesity ride for us. I think all of this evidence is completely consistent with the vegetable oil hypothesis.

It has something to do with palatable human snackfoods, unrelated to nutritional value

The Peery paper says 'diet' or 'nutritional value,' but it mainly seems like they mean 'macronutrients' and sometimes sugar when they say that diet can't be the answer. Nonetheless, this is still consistent with the vegetable oil hypothesis—go look at the ingredients for doritos, froot loops, or even store-bought bread.

It differs in its intensity by altitude for some reason

They go into some mechanisms, give it a read if you haven't. Maybe Colorado is an outlier. It definitely has a lot of young people who like the outdoors and crunchy granola. I don't know if this should make or break any theory of obesity per se. But it's certainly interesting. There's a lot of talk of lipids and oxidization in papers about veggie oils:

Oxidized PUFA can be dangerous when in our bodies, especially since oxidative damage to fat-containing LDL particles is a primary factor in the development of heart disease. [source](#)

Perhaps that's something—less oxygen, less oxidization, less CVD? I don't have the expertise—really no clue there. Oxidation does involve the addition of Oxygen to a molecule, and there is about 20% less oxygen at 6000 ft vs sea level. If you look at a [map](#) and squint, that's roughly the proportion that Colorado is less obese than an average US state. Interesting, but I truly have no idea about this. I think you'd be hard-pressed to use this as a linchpin in an argument against vegetable oils. I think they make some really cool arguments pertaining to lithium, definitely check it out, but it just wasn't quite clicking for me in the way that vegetable oils do.

And it appears to have nothing to do with our diets

I really think it does. That would be the simplest explanation, vegetable oil or otherwise, and the introduction of vegetable oil is the simplest and most obvious change from what I can see. Given the timeline, and the spread of obesity with trade and industrialization, I find myself disagreeing here. Most experts seem to think it's some combination of diet and exercise and genetics.

It is still adding up

Alright, so I think vegetable oils explain almost if not all the mysteries of the Peery paper. A few more thoughts, citing the Peery paper:

But again, it's not just the contents. For some reason, eating more fat or sugar by itself isn't as fattening as the cafeteria diet

Well, not the fat or sugar contents. I think we have plenty of studies that show low fat vs low carb etc is kinda a wash. At least, no one can decide on it. But surely there's plenty of vegetable oil in all those cafeteria foods.

When humans switch from an ancient to a Western lifestyle," he says, "they experience increased waistlines, reduced insulin sensitivity, higher blood pressure and a host of related disorders and diseases."

Same location, new lifestyle? If they didn't also move, and there's no indication they did, then I don't see why you should say it's in the water or the air, when a new lifestyle entails a complete diet makeover. I think this is actually one of the strongest points in favor of a dietary reason over environmental contaminants. Note that these oils do all these things to rats.

Diet won't work—but a diet of [only potatoes did](#). Chris Voigt ate only potatoes and he prepared them in a variety of ways, only one of which involved "a bit of cooking oil." Whole food diets seem to help a lot. Those get rid of vegetable oils for sure. But so does fasting, so it's hard to say. Switching from fats to carbs won't get rid of the vegetable oils that are in near all processed foods though. There is near universal agreement that processed foods are bad, could this be the reason? Processed foods being bad seems weird to me anyways, what if something just gets chopped up?

"palatable supermarket food"; not only Froot Loops, but foods like Doritos, pork rinds, and wedding cake.

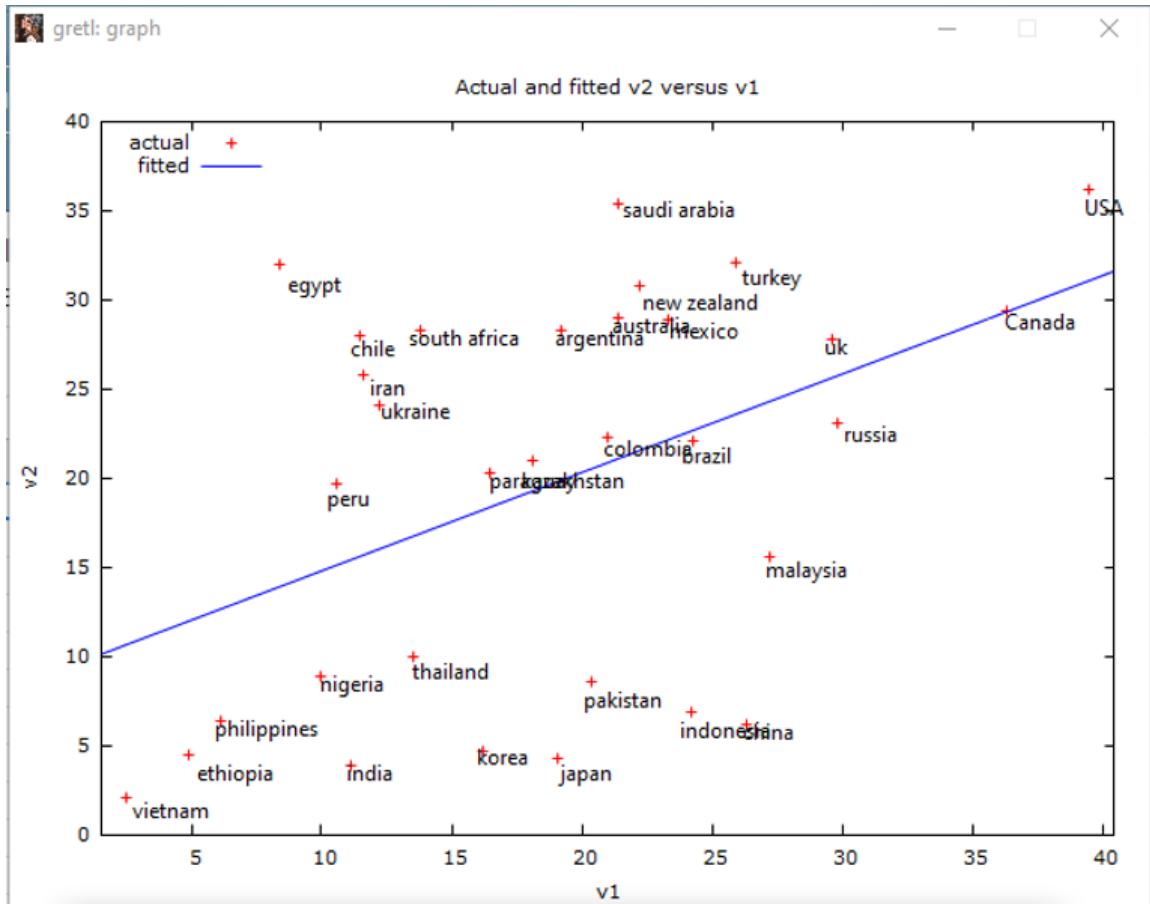
Oils are in doritos (ingredient 2), in froot loops (!!), presumably in ‘fried pork skins’. As for wedding cake, couldn’t find a good label on the internet, if it’s store bought it’s probably there though. Seriously, there’s vegetable oil in my Whole wheat bread.

Still on the Peery paper, which I really enjoyed reading, I actually doubt it’s chemical contaminants—you’d think China with the factories and air pollution, or the Congo with mines and terrible water would be worse than us. How polluted are New Zealand and Canada—they have quite high obesity. People in the USA are fatter in rural areas, and the USA really cleaned up its air and water in the past 60 years or so. I totally see where they’re coming from with some of their possible explanations, definitely check the Peery paper out. It’s a great read. However, I really think we should render unto Occam what is Occam’s—a group of added industrial oils went from nearly 0% to 20% of our daily calories, and it completely coincides with our issues.

I ran a really quick and dirty regression, with some sources:

[Y axis is obesity %](#)

[X axis is vegetable oil intake per capita kg/cap](#)



gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

Model 1: OLS, using observations 1-32
Dependent variable: v2

	coefficient	std. error	t-ratio	p-value
const	9.25789	4.17656	2.217	0.0344 **
v1	0.552495	0.202689	2.726	0.0106 **

Mean dependent var 19.58438 S.D. dependent var 10.92975
Sum squared resid 2968.122 S.E. of regression 9.946728
R-squared 0.198507 Adjusted R-squared 0.171791
F(1, 30) 7.430153 P-value(F) 0.010603
Log-likelihood -117.8852 Akaike criterion 239.7704
Schwarz criterion 242.7019 Hannan-Quinn 240.7421

EDIT 8/12: [Here](#)'s a post about more stats, I felt like it wouldn't be hard to add GDP to the mix, and it doesn't turn out well for our theory here--GDP seems to explain obesity best. Please do check it out if you know anything at all about instrumental variables, I tried that as a hail mary too.

It's been a long while since I've done any stats, so again, stressing the quickness and dirtiness of this regression. In this case, I'm sure vegetable oil is a very, very strong proxy for the general "processedness" of any countries' food supply. Eyeballing it, I'm sure GDP per capita could explain at least a good portion of this too. Nevertheless, that certainly looks like something. I'd be especially interested in analysis of individual outliers here, is Saudi Arabia because they only count citizens in one data set? Is China using a ton of oil in factories? I sure don't know.

I certainly think whatever the true cause(s) is/are, it's probably some addition of the modern world. Just a century ago, diseases of civilization were vanishingly rare. Whatever the solution is, my money is on a solution [via negativa](#) as Nassim Taleb would put it. And a food only created 100 years ago certainly isn't [Lindy](#).

A call for further study

Alright, elephant in the room time: [Scott Alexander already did an excellent piece on this](#).

But I'm gonna nitpick a bit here. Keep in mind, love his work. The first half could be interchanged with my essay here and you'd probably come away better informed than I just made you.

I find this to be a really elegant and provocative theory, with impressive circumstantial evidence. Unfortunately, as far as I can tell all of the direct evidence is against it.

Theory seems to refer to this idea of vegetable oils being bad—a theory with which we have ample evidence to at least suspect is true, as he says. *Evidence* that he goes into mostly pertains to high saturated fat diets and why they're not necessarily good. Also note that I'm particularly concerned with Soybean oil as that is the bulk of the increase, and his study discussions don't involve it.

For the second part I don't see any reason why we shouldn't be nearly completely agnostic on high saturated fat diet, from my reading here, my takeaway is to avoid processed foods at the very least, and to especially avoid vegetable oils. I haven't seen anything to necessarily go in the direction of a saturated fat heavy diet in particular. I don't see why we don't have the option of a little olive oil, or just *not completely submerging all our food in some kind of fat or oil for 20% of our daily calories*. From what I can tell, there's a fair bit of room in the middle—see the [mediterranean diet](#), which is one of, if not the most universally lauded diets ever studied. Its defining feature is probably the generous use of olive oil and omega 3's in the meat—two things entirely consistent with our theory here. I don't see why we should write this off on the intricacies of saturated fats vs PUFAs when the overall body of evidence is scant to begin with—and our primary focus, zoomed out, is in my opinion the question of new vegetable oils, which have other factors inherent to them.

My main point in this entire essay is not that PUFAs are bad per se, it's that these new vegetable oils are probably bad. I think Scott would agree, we have nearly the same

takeaway, but the reason I'm addressing him is because everyone who read that seemed to come away with the impression that this was a dead end. Clearly I disagree with that, I spent all day on this.

I think the most likely dietary change I make is to try to avoid foods with soybean, corn, or safflower oil, since this is probably a good stand-in for "foods processed enough that they count as processed foods and you should avoid them".

Alright, so what did we learn here? Vegetable oil is a new, highly processed addition to our diets in the industrialized world, and stands to potentially explain a lot of the quandaries that continue to baffle us about the crisis of modern health. It's in damn near everything, we get 20% of our calories from it, up from near 0%, and for some reason, no one seems to mind that and we all just argue about carbs and stuff. People haven't provably applied this theory to lose weight, but also no one seems to be trying. I don't see why it shouldn't be our prime suspect, however we don't seem to know the causal reason for why this may be from quality human studies, although the Nobbs piece and its associated sources put forth plenty of possibilities to choose from.

I don't know for sure what the answer is, any implication that I did was simply so I didn't have to qualify every single statement. I'm largely just trying to get the conversation going, and I'm by no means an expert.

For giggles: conditional on us solving the obesity crisis by finding one primary cause, I'd give the addition of vegetable oils 45% likelihood of being that cause. If we're in an essentially multi-factor scenario, I think it still has a large role to play, based off of what I've seen so far. I definitely would like to know more about the gears of the science here, and I'd like to hear some good rebuttals to this as well, I didn't find many that I found required addressing in this initial post, but that's mainly because this is only on the radar of people screaming into the void.

I'd like to express my gratitude to everyone who reads this blog, and to all the sources I cited. Without exception they were fascinating and well written. We're all in this together, trying to figure out just what in the world is going on here.

COVID/Delta advice I'm currently giving to friends

[Epistemic status: US-centric. gathering various impressions even though I haven't deeply investigated this topic and am not an expert, so it's all in one place and others can better use or critique the claims.]

1. Figure things out for yourself

Figure out what risk level you're comfortable with, and use microCOVID.org to get a sense of what policy makes sense for you personally.

Consider sharing your Fermi estimates with others (e.g., make a small Facebook group or chat for the purpose, or a LW thread if you're happy to do it publicly). This is a way to compare notes, get feedback, and share info without there needing to be a single big Policy Proposal For Everyone blog post people are passing around.

2. Be much more wary of COVID when hospitals are full

Keep an eye on confirmed COVID-19 cases, hospitalizations, and deaths in your area, and put effort into avoiding catching or transmitting COVID-19 if it looks like hospitals in your area will be overloaded 2-4 weeks from now.

3. Consider mostly not worrying about it

For young healthy vaccinated people in places with relatively good health care and relatively high vaccination rates (e.g., the US), if it doesn't look like your local hospitals will be overloaded in the next few weeks, then I think COVID-19 is mostly not worth worrying about right now.

(Likely exceptions: you're spending a lot of face-to-face time with an immunocompromised friend; you're visiting your seventy-five-year-old parents for the holidays in two weeks; etc.)

I emphasize this point mostly because my friends are in left/liberal spaces, where I think social forces encourage people to voice their "worry more" thoughts and keep quiet about their "worry less" thoughts.

This makes it extra valuable to speak up about "worry less" when you think it's true. (Though, again, all of this is too individual-specific for a public blog post to be able to say much about the exactly correct level of "worry".)

People in my social circle have mostly been taking large precautions throughout the pandemic. I see rationalist friends worrying a lot about whether it's OK to host get-togethers without rapid-testing all attendees (which I'd tentatively guess is not worth the trouble for the vast majority of gatherings). I see non-rationalist friends posting about how sad they are to not get to see friends again until things 'go back to normal'.

What I don't see are rough quantitative arguments for why the benefits are worth the costs here, especially ones that take into account the importance of social distancing's emotional costs (cf. Scott Alexander's [Things I Learned Writing The Lockdown Post](#)).

I don't see explanations of what 'going back to normal' looks like, and why we should expect that to happen at all (cf. Alyssa Vance on "[we should distance until X](#)").

Daniel Filan tells me that he initially locked down in response to COVID, but when he took the time to do a quick Fermi on the risks of COVID (around August 2020), he was surprised to find how costly his precautions were compared to their benefit:

[...] My estimate [in late 2020] was that I should pay up to tens of cents to avoid a [uCOVID](#).

I wasn't really modelling externalities well, and I'm still not totally sure how to do that right.

[...] TBC, the Fermi was something like "look up $p(\text{death} \mid \text{covid})$ given my age and sex, then estimate cost of other side effects as equal to cost of death, then add the cost of being normally sick for one week".

[T]hat being said, paying \$100k to definitely not get covid seems pretty pricey.

My current guess: the cost per uCOVID [in late 2020] came out to under 10c, and I rounded up for caution or something.

See also Connor Flexman's [Delta Strain: Fact Dump and Some Policy Takeaways](#), which (with a bunch of caveats and uncertainty) estimates that given Delta's ubiquity, a healthy vaccinated thirty-year-old who would otherwise live a full life now loses something like the equivalent of 1 hour of life for every 1,000–5,000 microCOVIDs they get.

I think the vaccines are remarkably effective (both vs. infection and vs. [severe-symptoms-given-infection](#), which correlates with long COVID and [death](#) risks), and COVID wasn't a large risk for young healthy people in the first place (though Delta is a bigger risk than Alpha, ignoring effects of vaccination).

(Added Aug. 24: I think COVID risk for young vaccinated people isn't that different from the risk you face from other widespread viruses. In particular, my impression is that many (if not all?) viruses cause long-term symptoms similar to long COVID. Though at least in unvaccinated people, it seems that COVID causes long-term issues more often than the typical virus.)

I know some people whose conclusion from this is "I'll try to avoid *all* viruses", including people who were taking such precautions pre-COVID. For healthy individuals, I don't have a strong view on whether that's more or less reasonable than mostly-ignoring this risk. But I am suspicious of the view that *only and exactly* COVID is worth worrying about.)

4. If you do worry about it...

... I would especially prioritize doing things [outdoors](#).

I also think in-person friend groups, events, and group houses should consider restructuring to cater to people with this or that risk tolerance. The typical especially at-risk person, IMO, shouldn't be trying to get a huge community of people to all match their needs. Nor should they be bunkering down to "wait out COVID" and avoiding in-person socializing until then (unless they *really* don't want or need in-person socializing). Rather, I'd propose planning

around the assumption that things will stay at least this risky for years to come, and build connections with other people who want to keep their risk low on that timescale.

(Added Aug. 24: The older you are, the higher your COVID risk generally is. Mayo Clinic says other [major risk factors](#) are: obesity, diabetes, heart disease (including hypertension), lung problems, cancer, sickle cell anemia, weakened immune system, chronic kidney or liver disease, and Down syndrome.)

5. Try to get triple-vaccinated

J&J seems to be much less effective than Pfizer/Moderna, so people who received J&J should especially prioritize getting two mRNA shots ASAP. But mRNA recipients should probably also get a third shot if their second shot was 4-6 months ago.

Will Eden re 'should people get a third shot, and if they had Pfizer/Moderna the first time should they try to have J&J this time around?':

The mechanisms for cell entry are a bit different, but they all end up using RNA to encode the S protein on the cell surface, so the nature of the immunity is basically the exact same mechanism. J&J is also going to immunize you against the vector virus, making it harder to receive future AAV vaccines/gene editing in the future, so personally I would stick to the mRNA vaccines only.

From the Israeli data it seems like a third shot is probably most effective around 5 months later for little or no lapse in coverage? I would recommend it personally and have to others already. Doing third shots is in trials, as is mixing different vaccines, both appear to work well.

In response to Zvi Mowshowitz's [criticism of the Israeli data](#), Will adds:

I do think there's a reasonable point about the denominator problem, and how outbreaks began in more heavily vaccinated areas thus skewing the results downwards. But he also admits this means there are enough vaccine breakthroughs to cause a pandemic!

On the flip side, the discrepancy between protection against symptomatic COVID vs hospitalization/death I don't agree with him on. I think a straightforward view of the data, especially the effectiveness of vaccine vs month of administration released by Israel, suggests that you need high circulating antibodies to prevent the infection from taking root at all, but prevention of severe disease is more reliant on cellular immunity and your body mounting a quick response after exposure, which is the benefit provided by vaccination. (On the other hand, this is why I want mucosal immunity - an optimal vaccine would even prevent infection + spread, and clearly the current vaccines don't.)

[Moderna seems to be better than Pfizer](#), perhaps because it uses a higher dose.

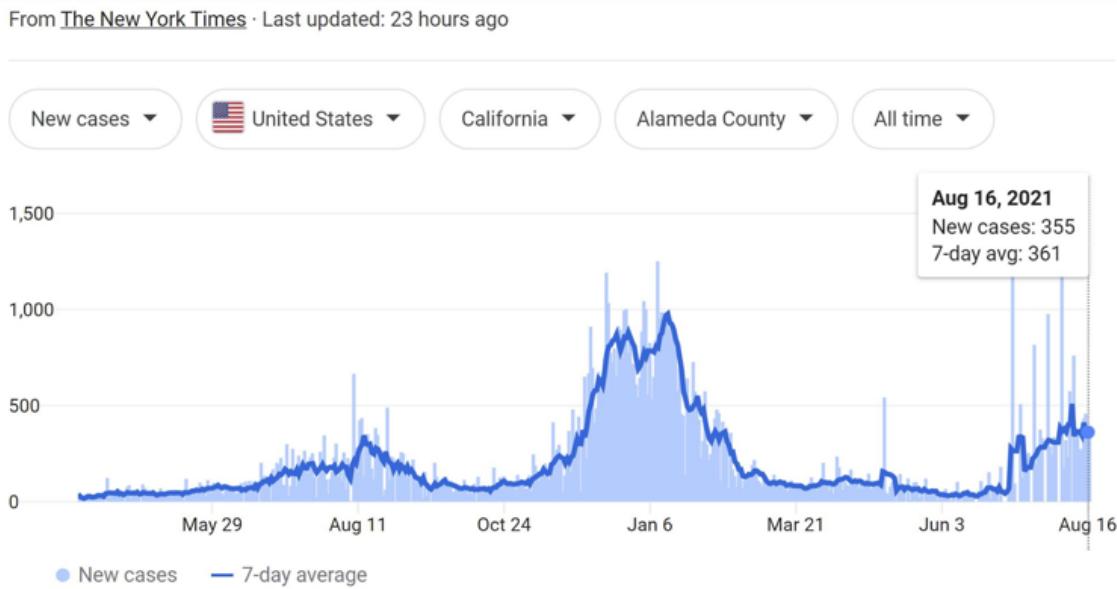
CNBC [reports](#):

The United States will begin widely distributing Covid-19 booster shots next month as new data shows that vaccine protection wanes over time[...] U.S. agencies are preparing to offer booster shots to all eligible Americans beginning the week of Sept. 20, starting eight months after their second dose of Pfizer or Moderna's vaccines.

But this seems too slow to me, given the next section. (My family and peers would nearly all have to wait till Nov/Dec/Jan.) So I still suggest trying to get a third shot sooner.

6. Right now is the safe time

COVID seems to be quite seasonal -- it spreads a lot more in the winter. To think about the coming months, I looked at COVID rates where I'm living:



Based on this, I expect this coming Nov/Dec/Jan/Feb to be way worse again around SF. Except the baseline is much higher now, so I expect the elevated case rate to likewise be much higher.

I think most people who track these things are looking at the current COVID rates, seeing they're a lot higher than in the past, and concluding that we should be buckling down temporarily right now.

Except that the rise is because Delta took over, and Delta is the new normal going forward. So if you're planning to throw some parties before March 2022, now is very likely the best time to do it. If you're planning to "buckle down temporarily" and "temporarily" means "a few months" rather than "a few years", you may want to reconsider whether things will actually look any better five or ten months from now.

Eyeballing NYC, it looks like things may be bad in some parts of the US even in May?



Beyond seasonality, other things that may make things worse in the future include:

- Fading vaccine efficacy, if that's a thing. Even if you personally get more shots, others may not.
 - In general, my tentative default guess (if something like the status quo continues) is that people's willingness to get more COVID booster shots will drop off a lot over the coming months and years. (Because it will feel more like an annoying regular chore than like a miracle cure.)
- New COVID strains.
- Regulatory hurdles inhibiting and delaying vaccine roll-out.

I'm not especially worried about long COVID, because I expect long COVID frequency and severity to track symptom severity pretty closely -- if vaccines protect against severe symptoms, they'll tend to reduce long COVID risk too.

Even in the best-case scenario, there's a delay between "new COVID strain evolves" and "new vaccine targeted at that strain rolls out," and a further delay before we get data on long COVID in people infected by that strain (either for vaccinated or unvaccinated people), which makes it harder to be confident about the risks.

Or perhaps not. The future could hold a lot of things. But in my own planning, I'll be acting like most of the probability mass is on 'things stay about the same for a long time, or get worse'. I mostly am planning around the expectation that this is life going forward, and I'm not going to shut down my life indefinitely -- but I can plan around taking more risks during warmer months, and fewer risks during colder ones.

7. Small exposures are better than large exposures, and maybe better than no exposure?

This is probably a relatively good time to get infected if the trend "COVID evolves to be more dangerous and high-viral-load" continues. Getting infected with a lower viral load is better, and may confer a lot of immunity against future variants.

By the same logic, if you expect to get infected with Delta regardless in the next few months (as a large portion of the US presumably will), it's better if your exposures tend to be in low-

risk settings where you'll get a small initial viral load. (And it's better to get sick when you know hospitals won't be overloaded.)

8. Be aware of possible signs of COVID

(Added Aug. 24.)

According to [CBS](#), an August CDC study found that the most common symptoms of Delta are still "cough, headache, sore throat, myalgia [muscle pain], and fever". Oddly, the COVID Symptom Study instead found that the most common Delta symptoms were runny nose, headache, sneezing, sore throat, and loss of smell.

But: don't assume people are COVID-free just because they're showing no symptoms. COVID is often asymptomatic (or the symptoms are hard to notice), and people with COVID transmit the illness a lot before they start showing symptoms.

If you might have COVID, self-isolate and try to get tested. PCR tests are the most reliable, but they have to be sent to a lab and normally take days to give back a result. Also, Connor Flexman [suspects](#) that PCR tests may still have a ~40% false negative rate, which is much higher than usually advertised. Antigen tests have something like a ~50% false negative rate, but can be done fully at home and give much faster results (within a few minutes). LAMP tests like [Lucira](#) are also fast, and are in-between PCR and antigen tests in efficacy, according to Connor. (Flagging that these are very off-the-cuff numbers with minimal due diligence done; I'll revise them if the situation becomes clearer.)

Consider taking zinc lozenges soon after COVID exposure or symptom onset. (They should taste bad if they're working.)

9. Be prepared for if you get pretty sick

Things to buy now for if you get sick: Pedialyte or gatorade powder, acetaminophen or ibuprofen, oral thermometers, and a finger pulse oximeter.

Maybe also: over-the-counter inhalers, a humidifier, mucinex/guaifenesin, pseudoephedrin.

10. Take care of yourself if you get sick

(section in progress)

Staying Grounded

What does it mean to “stay grounded”? Why does it matter? How does one stay grounded?

Examples

Let's start with a bunch of examples of grounding failures - i.e. people failing to stay grounded.

- The person who Volunteers to End Hunger, or End Poverty, or what have you, yet they do not actually make any significant difference to any actual people in hunger/poverty/etc (despite possibly believing that they have).
- The cancer researcher who does lots of Experiments in the lab, and runs Statistical Calculations, and Publishes Papers, yet does not actually come any closer to understanding or curing cancer.
- The person who does lots of Things Which Happy People Do in commercials, but is deeply unhappy anyway (despite possibly even convincing themselves that they are happy).
- The person who makes themselves miserable to Eat Healthy and Exercise, yet is still far over their own preferred weight.
- The person who gets Good Grades in high school, goes to a Good College, gets a Good Job, is generally Successful, but realizes sometime in middle age that they're deeply unsatisfied with their life.
- Cargo cultists: the pacific island tribes on islands which hosted airstrips during WWII who, after the war, would sometimes Talk into elaborate wooden “Radios” and Wave Sticks on the abandoned airstrip in hopes that planes would land with supplies.
 - Also, the various areas of academic study for which Feynman used the cargo cults as a metaphor.
- The person who eats Good Food (possibly Healthy Food, possibly Expensive Food, possibly Ethical Food, depending on their social circles) but never really notices how much they enjoy the actual taste of different foods.
- The person who buys Nice Clothes which are neither comfortable nor flattering for them in particular.
- The startup founder who Writes Code, and Iterates, and Gets Funding, but never stops to think about how many people actually want their product or how much those people would pay for it.
- The political activists who Organize The Movement and Raise Awareness to help X, get lots of media coverage and some laws passed, but in the end X doesn't actually change much.
- A regulatory agency puts in place lots of Rules and Regulations and Processes in order to Make People Safe, yet doesn't end up actually making people safer.
- The military (or guerillas) who Shoot Enemies and Destroy Their Stuff, but never manage to institute lasting regime change (or whatever else their primary goal may be).
- The company which hires Graduates From The Best Schools, and Highly Regarded Consultants, and the like, yet the work is never much better than any other company's employees/consultants.
- The hedge fund which hires Brilliant People, and buys Lots Of Data And Compute, but never outperforms the market or even has a concrete strategy to

do so.

Note that these “failures” are not *necessarily* unintended/unwanted. Lots of hedge funds stay afloat mainly by bringing in investors, despite never beating the market. And hiring Brilliant People is a great way to look good to investors, regardless of whether they manage to beat the market.

In particular, many of the examples involve strategies which are good for winning [social status](#), just not for whatever they’re nominally about. Graduates From The Best Schools or Volunteering to End Hunger or Publishing Papers or eating Good Food or Raising Awareness are all good ways to win social status, even in cases where they don’t actually help much with their nominal purpose. [Humans often act as though social status is their “real” subconscious motivator, but they self-deceive about it.](#)

The Unifying Idea

In each of the examples above, some Symbols (capitalized in each example) have decoupled from what they represent. The Symbols are “ungrounded”, in the sense of the [symbol-grounding problem](#).

There’s some socially-recognized Symbol of the thing - sometimes something which actually does help with the thing, sometimes a generic status symbol, but always something which people associate with the thing. Publishing Papers is a symbol of scientific progress, Good Grades are a symbol of a successful life in the making, Nice Clothes are a symbol of generic high status. People spend time and effort and resources on the Symbol, but obtaining the Symbol isn’t always *sufficient* to get them the thing. Publishing Papers isn’t enough to cure cancer, getting Good Grades isn’t enough to lead to a satisfying life, buying Nice Clothes isn’t enough to make one look good or feel comfortable, etc. Even if the Symbol can help somewhat, people often continue to pour resources into the Symbol long after it has ceased to be a bottleneck to obtaining the thing.

Another way to put it: grounding failure means [Goodharting](#) on the Symbol, and failing to actually get the thing as a result.

If your main goal is in fact to accumulate social status (regardless of whether you explicitly acknowledge this fact), then the situation is a bit different. The Symbols are far more important to your objective than the things they nominally represent; welcome to [Simulacrum 3](#). So focusing on the Symbol over the thing is not a failure at all; the Symbol is what you actually want, on some level. If the suggestions in the next section about how to stay grounded sound aversive, unpleasant, or like they’re not really what you want, then consider that you may actually subconsciously want the social status more than the thing. If that’s you, then don’t worry about staying grounded - that’s not the game you’re playing.

But for those of us who are more interested in the thing than the Symbol, how can we stay grounded?

Fuck The Symbols

The most reliable way I know to avoid grounding failures is to say “Fuck The Symbols”. Make a point of *not* pursuing the Symbols, and try to obtain the thing anyway.

Make a point of not engaging in the standard academic research performances, and figure out how to cure cancer anyway.

Make a point of finding ways to be happy without doing the things which happy people do in commercials.

Make a point of not going out of your way to get good grades, and find a profitable/rewarding career path anyway.

Find ways to build a profitable business without Ivy League MBAs.

This isn't always the best choice, but it's usually worth at least thinking about how to do it - because the process of thinking about it forces you to *recognize* that the Symbol does not necessarily give the thing, and consider what's actually needed.

Covid 8/19: Cracking the Booster

The news about Covid-19 is now essentially on a few distinct tracks: Vaccine effectiveness and booster shots, vaccination mandates, mandates for NPIs and other Covid-19 related crippling of the living of life, and the actual path of the pandemic. One could argue for splitting those further, or for combining the mandates.

Vaccine effectiveness against the spread of Covid has come into question, with many claiming that immunity fades dramatically with time or that the vaccines were never that effective against Delta. I've looked into the claims and go into detail. There's little question that the vaccinated *can* spread Covid, which is in contrast to previous attempts to sell the line that they couldn't at all, and they spread Delta somewhat more than they spread Alpha, the numbers here are disappointing relative to my expectations, but the big claims of hugely waning immunity are almost certainly greatly exaggerated.

Vaccines remain highly effective at stopping the spread of Covid-19, and of stopping symptomatic disease, and especially in preventing death. Still, boosters are more effective than not having boosters, and I think the cost-benefit for most people favors getting a third shot if one is made available to you. Boosters are coming soon, eight months after your second dose. However, until it is highly encouraged and likely eventually mandatory, it is still for the moment mostly forbidden. Get your shot *exactly* when we tell you to, they insist, not a minute before.

There are continuing debates over mask mandates at various meta levels, which led to some sentences that were fun (and tricky) to write. There's increasing mainstream and in-group pressure to force people to wear masks regardless of whether it does anything useful in context, and to shame those who disagree and blame them along with the unvaccinated for the entire pandemic.

Oh, and also the pandemic is still growing but at a decreasing rate, and case counts are a favorite to mostly stabilize within a few weeks, which is great news. The death rate lags, so it continues to climb rapidly for now.

Let's run the numbers.

The Numbers

Predictions

[Nate Silver, the media's official master of predictions, officially wouldn't want to predict this thing.](#) So even though he kind of does, officially he doesn't. He observes the CDC models are split. Some models think this week is the peak, other models disagree and think things will continue getting worse for some time, but it's always either one or the other. Odd, but perhaps makes some sense given the sharp peaks elsewhere.

Early this week [the NIH director said there were 'no signs of having peaked out'](#) which was technically true, but misleading, as I'd noted in last week's update that there were definitely signs we might be about to peak. Standard stuff, standard line.

I sympathize with Nate, but when it gets hard is exactly when we need people like Nate to step up and take a stand the most, even if they'll often be wrong, so I'll give [the standard Teddy Roosevelt quote](#) and ask how I did this week.

Prediction from last week: 900k cases (+21%) and 5,028 deaths (+35%).

What's weird about that prediction is that the percentage here is wrong – 5,028 deaths is a (+48%) increase. I have edited the last post to point this out but note the error. Looking at the context, it's clear my *intent* was to predict a +35% rise and I had a spreadsheet error that gave the wrong number, so I shouldn't get credit for the better 'actual' prediction here. Intent should win. The counterargument is that if you do the California correction I've now done, this methodology gives a prediction of 5,081 deaths, which is similar.

Result: 872k cases (+17%) and 5,545 deaths (+48% after correcting for California last week, +63% without that correction).

For cases, we got a good result, where the slowdown was real. For deaths, we got a quite bad result, where the slowdown was a mirage and we are back on the previous trajectory, and likely overshot a bit. Both are common, so predictions have to split the difference, but on reflection the slowdown in deaths couldn't continue, and I should have predicted at least +50% there. I also should have directly caught the California correction, which I didn't, and that meant the count a week ago was low by at least 350 deaths there, which accounts for about half the surprise here. I've corrected it for the charts and numbers going forward.

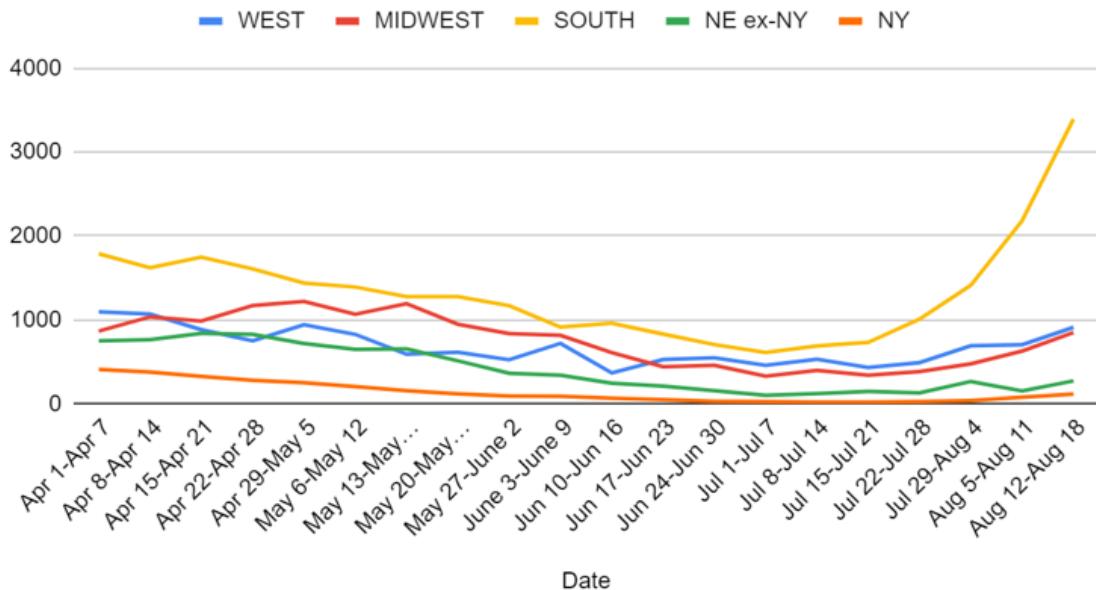
For next week, deaths should continue to rise with lagged cases, so I'm going with +45%. For cases, the slowdown in growth is presumably the new normal, and things should on average continue to slowly improve as more people get infected and vaccinated, and behaviors continue to adjust. The only danger would be if immunity is fading enough to counteract that, which I find unlikely.

Prediction for next week: 1,000,000 cases (+14%) and 8,040 deaths (+45%).

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 24-Jun 30	550	459	706	186	1901
Jul 1-Jul 7	459	329	612	128	1528
Jul 8-Jul 14	532	398	689	145	1764
Jul 15-Jul 21	434	341	732	170	1677
Jul 22-Jul 28	491	385	1009	157	2042
Jul 29-Aug 4	693	477	1415	304	2889
Aug 5-Aug 11	705	629	2181	234	3749
Aug 12-Aug 18	912	851	3394	388	5545

Deaths by Region

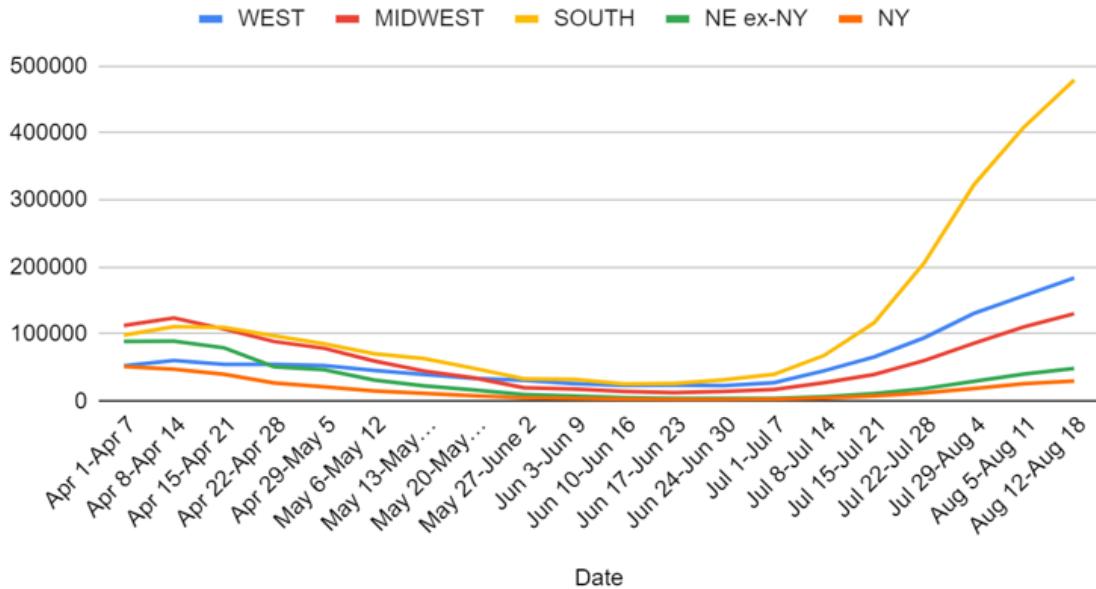


Steady rise across the board. Deaths lag cases by several weeks, so in a few weeks the growth rate should slow a lot, but stabilizing fully will take longer than that.

Cases

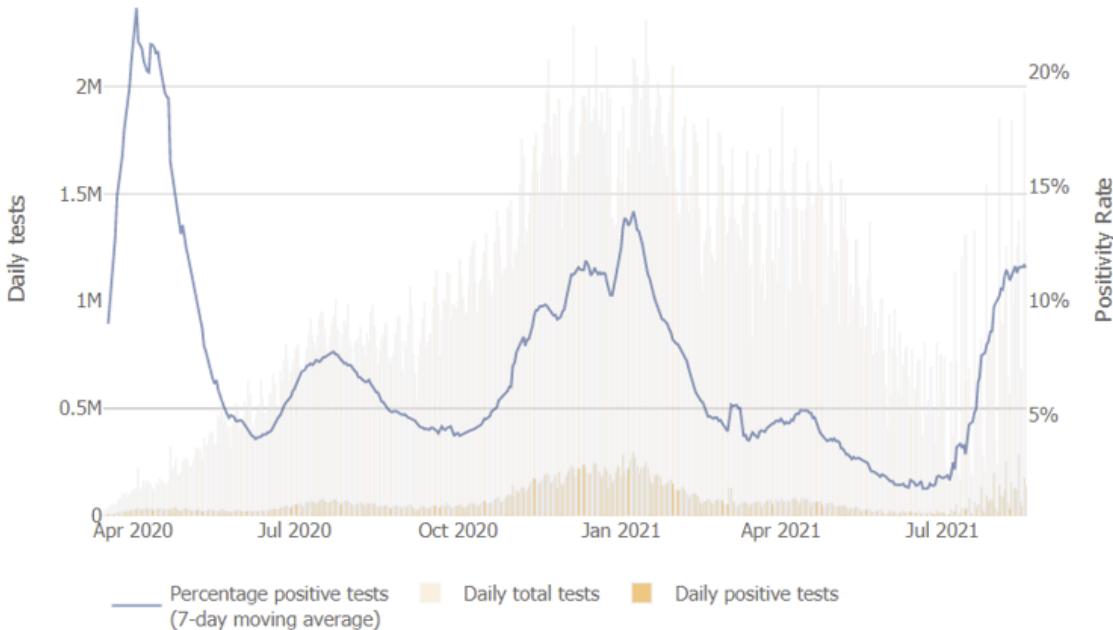
Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 24-Jun 30	23,246	14,521	31,773	6,388	75,928
Jul 1-Jul 7	27,413	17,460	40,031	7,065	91,969
Jul 8-Jul 14	45,338	27,544	68,129	11,368	152,379
Jul 15-Jul 21	65,913	39,634	116,933	19,076	241,556
Jul 22-Jul 28	94,429	60,502	205,992	31,073	391,996
Jul 29-Aug 4	131,197	86,394	323,063	48,773	589,427
Aug 5-Aug 11	157,553	110,978	409,184	66,686	744,401
Aug 12-Aug 18	183,667	130,394	479,214	78,907	872,182

Positive Tests by Region



The slowdown in growth is across the board, with no sign that the South is going to peak first and then we'll get an explosion in the Northeast. If that happens, which it might, it will probably happen in the winter.

The test positivity rate has mostly leveled off as well, with a clear phase shift:



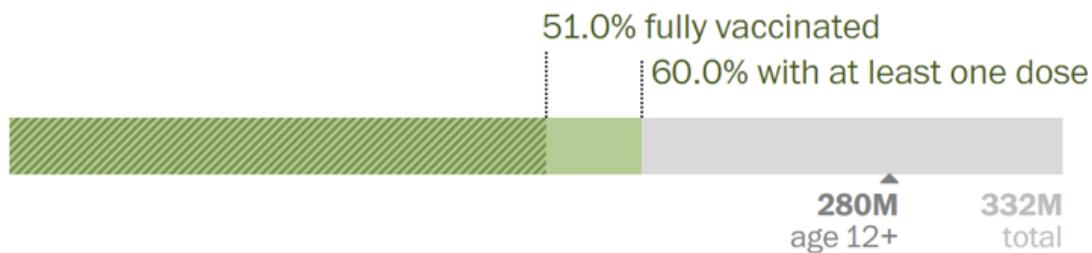
Things aren't quite stable yet let alone going in the right direction, but it seems more likely that the peak (at least for now) is relatively close and not too far above current levels.

Vaccinations

In the U.S., **359 million doses** have been given so far. In the last week, an average of **774,118 doses per day** were administered.

199.3 million vaccinated

This includes more than **169.2 million people** who have been fully vaccinated in the United States.



In the last week, an average of **770.6k doses per day** were administered, a **10% increase ↑** over the week before.

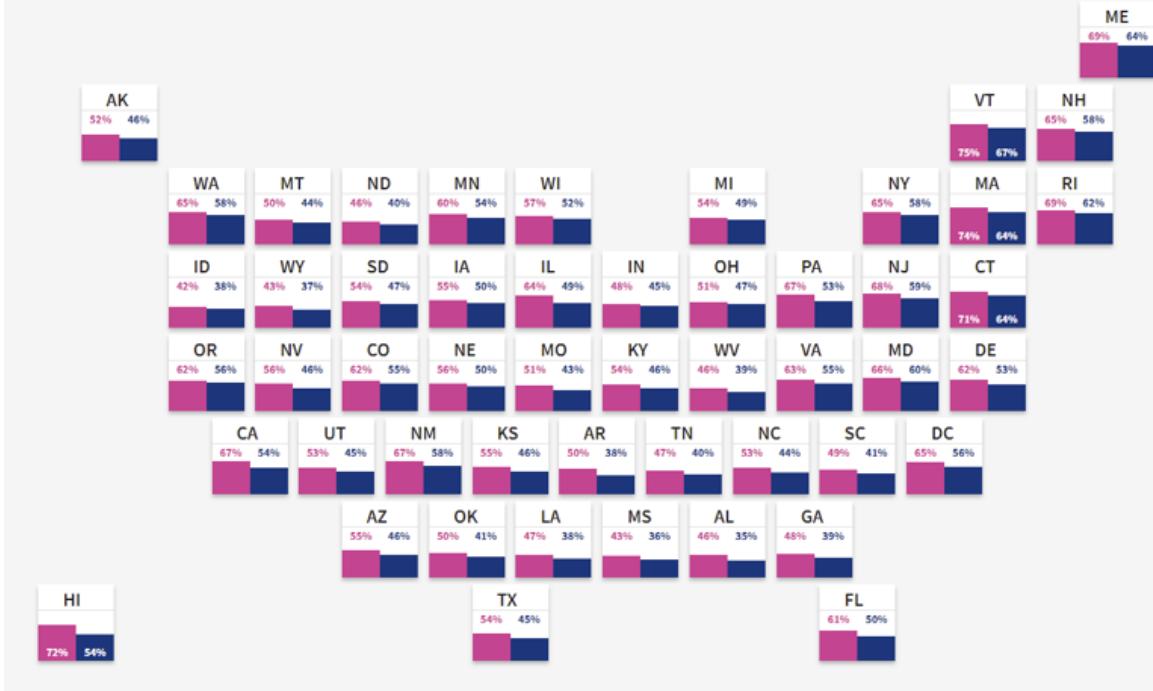
I am assuming, but am not fully confident, that third shots are not included in the dose counts above. The percentages should hold either way, and continue to show slow but steady progress.

[Here's a regional map.](#)

Compare states' vaccination progress or select a state to see detailed information

Percent of state's population who have received vaccines according to most recent state data.

■ One dose ■ Fully vaccinated



[New study finds that vaccines have no effect on pregnancy \(study\)](#). Mostly third trimester vaccinations since that's mostly that's available to study. Most pregnant women are declining the vaccine and there's lots of FUD going around about how it could be unsafe or bad for fertility. It contains no live virus, and is doubtless being watched very carefully. The short term side effects include fever, which is a reasonable thing to have a nonzero amount of worry about, and a reason to be thankful to have a study. Somehow I doubt many minds will be changed regardless, but it's good to at least be able to say Studies Show.

Vaccine Mandates

You know how you [convince a lot less than no people](#) to get vaccinated or support your mandates?

↑↓ WokeMeansYouLoseHat Retweeted



Ari Cohn ✅ @AriCohn · 17m

...

Even if you agree with him, there is no denying that Fauci is absolutely horrible at messaging. This is likely to be actively counterproductive.

Chief White House medical adviser Anthony Fauci on Sunday called on vaccine-hesitant individuals to "put aside" their concerns about personal liberty and recognize the "common enemy" of the COVID-19 pandemic.

I've spent much of the last month arguing in favor of mandates and vaccinations, and I notice that this statement pissed me off quite a bit. It's almost *engineered* to piss people off, to the extent it doesn't feel like an accident, as if Fauci's goal was to signal bad intent and further inflame tensions.

[There's also this.](#)

► Disclose.tv 🌟 @disclosetv · Aug 13

JUST IN - Biden admin is discussing mandating #COVID19 vaccines for interstate travel, but worried that it would be too polarizing "for the moment" (AP)

As in, you want to tell [people like James](#) that they're being crazy, or at least have the mechanism wrong.



Matthew Yglesias ✅
@mattyglesias

...

People use exaggerated fantasies of their opponents' illiberalism to talk themselves into adopting authoritarian methods.



James Lindsay, agnostic @Conc... · 12h ...

Once a vaccine passport social credit system is implemented, you will never be allowed to assemble to protest CRT in schools again as parents. Or anything else as anybody else.

53

495

2,056



James Lindsay, agnostic

@ConceptualJames

...

They will find ways to deny you entry, shut down your bank accounts, terminate you from work, no-fly list, etc., for being a "public health threat" for having the wrong opinions (cf. "syndemic"). Antifa and BLM will be able to do whatever they want, though. At first. LOL.

8:35 PM · 8/13/21 · Twitter for Android

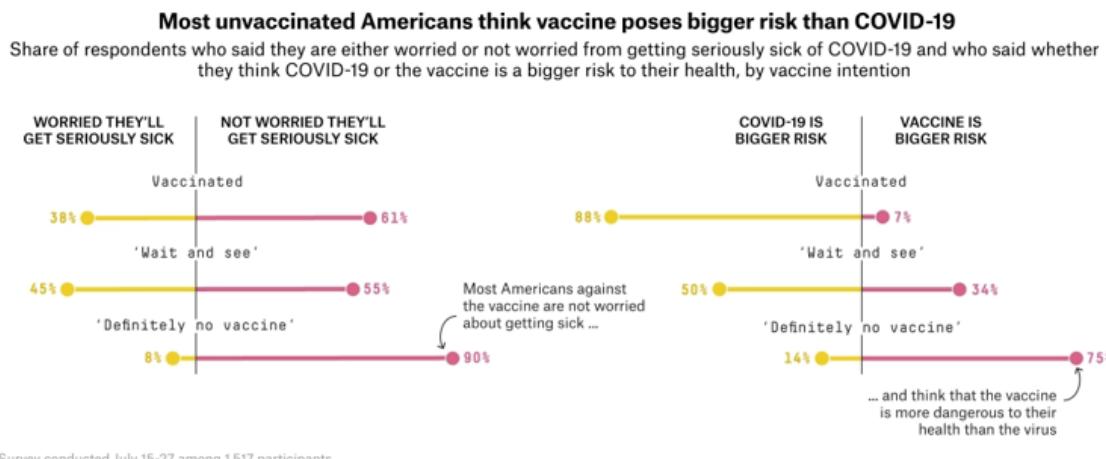
Rather than, you know, reinforce their narrative.

Also likely would help if we [didn't use a giant picture of a needle](#) on every news story.

FiveThirtyEight has a breakdown of unvaccinated America, which tells the usual story that there's a constant pool of '[I will never ever ever](#)' get the vaccine' attitudes, while a steady portion of everyone else gets vaccinated. If that's the case, then keeping our vaccination rate essentially constant is quite the accomplishment at this point, as it's drawing from a quickly narrowing group of people.

The most important chart there, to me, is this one:

Those firmly opposed to the jab are also much more likely to believe misinformation about the COVID-19 vaccines. A Kaiser Family Foundation [survey](#) from April found that 81 percent of the "definitely not" group was more likely to believe or was unsure about at least one vaccine myth, including the false ideas that vaccines contain fetal cells, cause infertility or change our DNA. This may explain why they still overwhelmingly think the vaccine is a bigger risk to their health than COVID-19.



This is not a story about lack of access or time off of work. It's a story of people who are scared of the vaccine, mostly for factually wrong reasons, and aren't scared of Covid.

[FiveThirtyEight attempts here to answer some of the justifications for vaccine hesitancy.](#) It's not perfect or complete, nor would I agree with every argument, but at least it seems likely to help rather than do further damage. A major theme of the objections addressed is the mindset that doesn't differentiate 1% from 99% - the idea that if you *could* get infected after the vaccine, and you also might not get it without the vaccine, then the vaccine does nothing. Such people presumably understand this in other contexts (for example, you *might* crash if you drive sober rather than drinking first) but it really is a true objection here, because they want to buy a mindset rather than buy actual safety, and refusing to sell that mindset means no sale.

[Amy Coney Barrett refuses to block Indiana University's vaccine mandate.](#) It wasn't a full mandate, as testing was an option, and had religious and medical exemptions. I continue to have more faith than most otherwise similarly thinking people that the Supreme Court will do mostly reasonable things and mostly enforce the rules.

[San Francisco joins New York City in requiring vaccination for indoor dining.](#)

[Here's a complete list of what requires vaccination in New York City:](#)

Eli Klein @TheEliKlein · Aug 16

BREAKING: NYC has provided a list of what more than 60% of Black New Yorkers & other unvaccinated people are banned from

LIVE

NYC

THE KEY TO NYC UNLOCKS:

INDOOR ENTERTAINMENT

- Movie theaters
- Live music
- Concert venues
- Museums & galleries
- Aquariums & zoos
- Professional sports arenas
- Stadiums
- Convention centers
- Exhibition halls
- Performing arts theaters
- Bowling alleys
- Arcades
- Pool & billiard halls
- Recreational game centers
- Casinos
- Adult entertainment

INDOOR DINING

- Restaurants
- Catering halls
- Event spaces
- Hotel banquet rooms
- Bars
- Cabarets
- Nightclubs
- Cafeterias
- Grocery stores with indoor dining
- Bakeries
- Coffee shops
- Fast food/quick service with indoor dining

INDOOR FITNESS

- Gyms
- Fitness centers
- Fitness classes
- Pools
- Indoor studios
- Dance studios
- Sports classes

SUNNYSIDE
11:04 77°
SPECTRUM NEWS NY1

RE INFORMATION GO TO PUBLICTHEATER.ORG. OVERNIGHT SERVICE ON THE STATEN ISLAND FER

You can see the room now, where someone was tasked with coming up with as many different names as possible for places requiring vaccination. What do people even do inside, anyway? We need ten different variations of indoor dining by the 11am press conference. The two sides are both trying to make the mandates look as obnoxious as possible, for different reasons.

[What if something goes wrong?](#) I see a lot of people asking questions like this:



Kerry @kerry62189

...

I read today that there's no plan to accommodate people with medical exemptions. And you know there will be glitches and lost cards--are we just going to require people to double up when that happens and assume it's fine? What is this?



Tim Pool @Timcast · 17h

So far called 14 NYC restaurants asking they mandate vaccines

12 said yes

They also said people with medical conditions preventing them from getting vaccinated will not be allowed in

One said its fine if you're medically exempt

One said they won't check at all

[Show this thread](#)

For New York in particular, you can use the Excelsior pass as a substitute for your vaccine card. If you're medically exempt from the vaccine and can't take it, I think you're simply out of luck.

What about the general question of whether you can get your vaccine card replaced if you need to? [In Wisconsin the answer is clearly yes](#), and they say to contact the health department in the state you were vaccinated in if it didn't happen locally. A patchwork solution isn't ideal but I'd be very surprised if there wasn't a way to get one in other places. Also [CNN recommends](#) scanning your card with your phone no matter what else is going on, along with some links to additional options, and that seems very wise.

[New York and Washington DC require health care workers to be vaccinated](#) (WaPo), joining California and Washington state.

[A Virginia hospital mandating vaccinations faces a nurses strike](#) (WaPo).

[Brooklyn Nets start requiring proof of vaccination at their games, as per NY rules.](#)

The Pac-12 football conference reverted to the old rule that a school unable to field a team must forfeit the game.

[The electric company gets in on the act.](#)

 **River** @river_is_nice · Aug 14

My brother is in town and just told me the electric company he works for is trying to require vaccination and his union is talking about going on strike over it. Methinks we shouldn't be telling the last vestiges of productive labor in this country what to do with their bodies

Texas nursing homes, on the other hand, [do not get in on the act.](#)

 **Razib Khan**  @razibkhan · Aug 14

COVID-19 cases are skyrocketing in Texas nursing homes, and nearly half of workers are unvaccinated texastribune.org/2021/08/13/cor... via @TexasTribune

Nor do any localities in Texas, [since the Texas Supreme Court has \(for now\) ruled](#) that the executive mandate against mandates is mandatory, and therefore mask mandates are forbidden, although Biden is trying to use Federal coercion to change that.

[He's also going to lay down the law on nursing homes, requiring vaccine mandates if they want to keep getting Medicare and Medicaid funding.](#) I hate the tactic, but if there's one place you *really, really* want a vaccine mandate, it would have to be nursing homes.

Incentives matter, small amounts of money can be highly motivating ([such as this study from the flu vaccine](#)), [so be careful how you set them.](#)



Medical Axioms  @medicalaxioms · Aug 14

...

My colleague admitted a patient to the hospital recently and asked them if they had a COVID vaccine.

“Yeah a lot of them.”

“Like how many?”

“Probably 20.”

Turns out he had been finding gift card and other incentive programs and getting shots to get the cash and prizes.

Mask and Testing Mandates and Other NPIs

[Andrew reminds us of the key question.](#)



Andrew Rettek @oscredwin · 8h

NPIs are super important to do when they're going to stop the virus, and super important NOT to do when they won't slow the virus enough.

 **Nate Silver** ✅ @NateSilver538 · 9h

Replies to @JamesFallows

True but Australia shows even incredibly strict NPIs are *not* enough to contain the Delta variant on their own.

I'd also argue their sluggishness in vaccination is related to their "Zero COVID" policy. They didn't have much urgency and really played up low-risk AZ side effects.



Andrew Rettek @oscredwin · 8h

Replies to @oscredwin

It seems like western elected officials are mostly inclined to do the opposite, no NPIs super early when it would make a big difference, and strong NPIs when the virus is everywhere.

There is a golden middle where population-level NPIs (non-pharmaceutical interventions) are great, which is where you can stop Covid *if and only if* you use the NPIs for a limited amount of time.

If you would have stopped Covid anyway then obviously you didn't need the NPIs.

If you can't stop Covid that way, or you can only do so until you relax the controls, then all you've done is buy yourself some time that didn't do you any good. The time needs to change things for the better, such as by getting people vaccinated. Yet Australia seems to be in no hurry to vaccinate, and places like America have already vaccinated most of the people they are going to vaccinate.

The other case is the 'flatten the curve' argument, where the time you purchase stabilizes the medical system. That only makes sense *if you stabilize at a high level that churns through cases*, otherwise what's the point? Australia is halting things at zero.

NPIs can look good during the crucial two-week blame period, but then you're in the same situation two weeks later, over and over again.

[I disagree with Misha, I think this happened](#), because that's mostly what I would have expected.



Misha Gurevich. Angel Investor, Photographer.
@drethelin

...

this definitely never happened but is kind of hilarious if it did



John Kelly ✅ @jkelly3rd · Aug 14

In one of my kid's classes, in Florida, the teacher told students she had cancer, recently finished chemotherapy and asked if more of them would wear masks the next day for her sake. About half were wearing masks. The next day, after her plea, still about half were wearing masks.

11:12 PM · Aug 14, 2021 · Twitter Web App

My model is that most children who would care about such a plea were already wearing masks, and the last thing kids want to be seen doing is caring about the teacher's health in front of their peers.

[New Zealand locks down for three days after one positive Covid test.](#) If you're going to play this game, by all means play to win it, but it's not a long term solution. I'm curious how often this turns out to be a false positive.

On a related note, [reports that Amazon pulled their Lord of the Rings production out of New Zealand to the United Kingdom due to not wanting to deal with all the Covid quarantine and other restrictions.](#)

Via MR, [thread by Andy Slavitt, Former Biden White House Sr Advisor for COVID Response](#), laying out the need to live with Covid long term and proposing eminently reasonable actions for doing the best we can with that. Here's his takeaways at the end:



Andy Slavitt 🇺🇸💉✅ @ASlavitt · Aug 14

These are things we should be willing to tolerate:

- Masks when we travel & in heavy seasons
- Staying home when we're sick— always
- Weather reports as there on for smog or allergy seasons
- Periodic outbreaks
- Caution around key populations
- Showing we are vaccinated/tested 20/

156

795

3.9K



Andy Slavitt 🇺🇸💉✅ @ASlavitt · Aug 14

But there is something we can decide not to tolerate:

Preventable deaths.

100 people die every day from the flu in a bad season, mostly elderly & kids. We should not accept this. And we should not accept even more from COVID. 21/

I definitely agree that we should adopt a permanent 'stay home if you're sick' norm, which we should have had anyway. The question is what we are buying, in this scenario, with all these extra precautions, the same way we'd ask what our flu precautions are buying to see if they pass a cost-benefit test, if we did cost-benefit tests.

[MR also shares this study of mask mandates.](#) Given how non-linear the dynamics are in such situations, and the role of control systems, saying that a mandate saved a certain number of lives seems like not a reasonable way to measure whether mask mandates work, but I'm not sure what other options are available. I haven't looked at the study myself.

[Los Angeles to require masks at outdoor concerts and festivals](#), regardless of vaccine status, but not require vaccination. To be fair, it looks like the threshold for this is 10,000 people, usually crammed together tightly, so even outdoors I can see taking extra precautions.

[A call for the end to mass testing in the UK, since it can't stop the virus in any case](#). As always, when gathering information, one must figure out the value of information and compare it to the costs. Enough mass testing to generate good statistics seems clearly still worthwhile to me, but beyond that, what changes based on what you find? Does it do you any good?

[Survey of 'experts' says most wouldn't go to the gym.](#) Most wouldn't go to the gym anyway. Gyms are a weird case, because the main point of them is health benefits, and there are (mostly) other ways to get the same effect. Not sure what I'm going to do on this once I'm back in the city.

Delta Variant

The attributes of Delta are the biggest and most important unknowns.

How much more infectious is the Delta variant? How much less effective are the vaccines against it than against Alpha or the original strain? How much deadlier is it, especially to the vaccinated or to children? How much faster does it cycle through? Do its higher viral loads mess with test results?

These can all be considered as part of a series of equations.

1. We know almost exactly how fast Delta displaced Alpha.
2. We largely know how fast cases and deaths have been growing.
3. We know what percentage of various countries/states are vaccinated.
4. Alpha's serial interval is ~5 days, Delta is likely faster and more like ~3 days.

There's a lot of room to argue details on #2-#4.

#2: We only know about reported cases and deaths, and our positive test percentage is down a lot so it's possible we are missing a higher percentage of cases. However, if Delta is deadlier than Alpha, then we can't suddenly be missing a lot more cases unless we're also suddenly ignoring deaths, and the testing slowdown doesn't apply to other places. One could however argue, as we will below, that if vaccinated people are a larger percentage of infections, and we're missing a lot more of their cases because they're less serious or don't get tested, then that could throw this off.

#3: The base percentage is good but what we need is the *effective* vaccination percentage for various purposes, all of which are different. For deaths, age matters a lot, and the vaccination rate is effectively higher. For cases, how often are we detecting cases in people at various ages and with varying vaccination status? For infections and thus growth rate, how much do children matter, and how much do the vaccinated matter (which is more of a thing we're trying to solve for, but it impacts our answer here as well)?

#4: The faster Delta replicates the less additionally infectious it needs to be in order to displace Alpha and grow cases the way it did. The drop from 5 days to 3 is a big game.

The rate at which Delta took over from Alpha and then grew, in various places, is the central constraint. You only have 'so much infectiousness' to go around. The data is consistent with Delta being 50% or so more infectious than Alpha across the board. If you then put additional growth in infectiousness in one place, to the extent that it's big enough to matter, you have to take that growth from another place. The average *is an average*. It's not a floor.

As a working example, here is [a long thread doing math on 'breakthrough' infections and deaths](#), along with a Mayo Clinic study. Starts with the executive summary:



David Wallace-Wells @dwallacewells · Aug 12

...

The vaccines are still performing quite well in preventing severe disease. But with Delta they are doing much worse in stopping transmission. Breakthroughs are probably now 10-20% of new infections, and perhaps as much as 5% of new deaths. A thread. (1/x)

Before we go further, suppose this is right. What would that correspond to in terms of vaccine effectiveness if true?

Let's say 15% of new infections are in vaccinated individuals. Depending on what percentage of the population counts as vaccinated, you get a different answer.

At the time this statement was made about 70% of the adult population had at least one dose. If you use that number and they have 15% of infections, the vaccines would be 93% effective against infection, or 89% effective if they're 20% of infections. If we cut their share down to a conservative 60% to account for partial vaccinations and children, and say 15% of

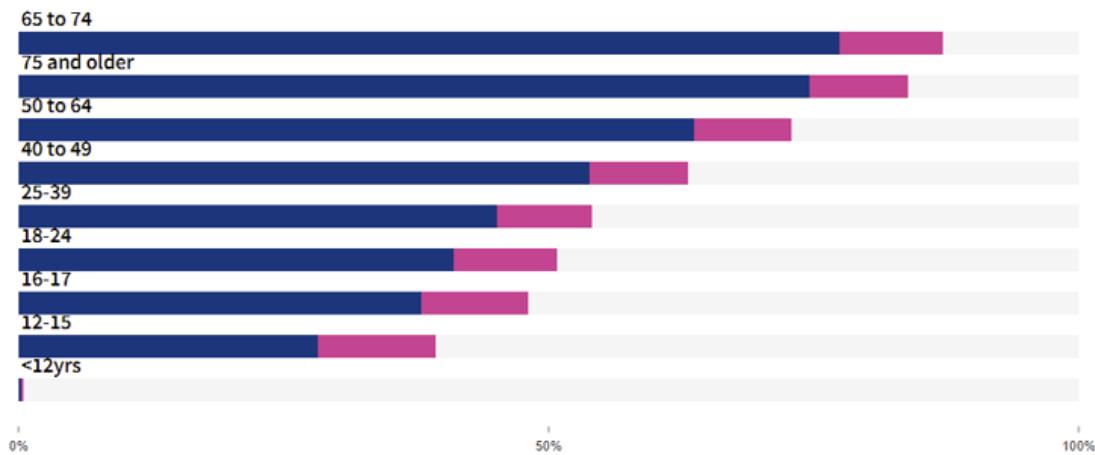
infections are among the vaccinated, we get 88% effectiveness against infection. So it's not like the 10%-20% range is either an update or scary versus our previous expectations.

For death, if we assumed the vaccinated and unvaccinated populations were identically distributed, we'd improve from 88% to 95% effective against death. But that's not right, [because vaccinations increase dramatically with age](#) and we can essentially ignore children entirely.

What percentage of people in each age range received the COVID-19 vaccine? ⓘ

Percent of people by age receiving **at least one dose** or **fully vaccinated**.

According to the Centers for Disease Control and Prevention, there is age information for **92%** of people who received **at least one dose** and **92% of fully vaccinated** people.



At a minimum we're effectively looking at 75% here where it counts, which would make the vaccines 98% effective against death. That is disappointing compared to what we had before, but that's quite the statement about how effective they were before. It's still amazingly great.

(Standard disclaimers that there's a bunch of confounders I'm not dealing with here, none of this is exact, but it's for intuition pumping and Fermi estimation rather than an exact answer.)

Then we can compare that with the body of the thread, along with other data coming in.

What we find here is a comparison of the early numbers and reports, which had almost no breakthrough infections let alone deaths, and the new numbers that aren't as insanely great.

Then of course we get a reference to Topol once again, because the world is small. Sigh.



David Wallace-Wells @dwallacewells · Aug 12

...

"The message that breakthrough cases are exceedingly rare and that you don't have to worry about them if you're vaccinated — that this is only an epidemic of the unvaccinated — that message is falling flat," said [@michaelmina_lab](#).

4

86

304



David Wallace-Wells @dwallacewells · Aug 12

...

"If this was still Alpha, sure. But with Delta, plenty of people are getting sick. Plenty of transmission is going on."

1

40

208



David Wallace-Wells @dwallacewells · Aug 12

...

"We're seeing a lot more spread in vaccinated people," agreed [@erictopol](#), who estimated the vaccines' efficacy against symptomatic transmission, which he estimated to be 90% for the wild-type strain and other variants, had fallen to about 60% for Delta. "That's a big drop."

4

98

283



David Wallace-Wells @dwallacewells · Aug 12

...

Later, Topol suggested it might have fallen to 50 percent, and that new data about to be published in the U.S. would suggest an even lower rate.

 **Eric Topol**  @EricTopol · Aug 11

There needs to be truth-telling about the reduced protection of mRNA vaccines vs symptomatic Delta infections.

It was 95% pre-Delta.

Many are claiming it's still ~80%.

It isn't.

50-60% is best estimate from all sources (not US, since we don't have the data)

[Show this thread](#)

Even the Mayo Clinic gets in on the act ([study](#), note that Delta was not their intended target for the study, it started too early).



David Wallace-Wells @dwallacewells · Aug 12

...

On Wednesday, a large pre-print study published by the Mayo clinic suggested the efficacy against infection had fallen as far as 42%.

The ‘as far as’ is based on the numbers for Pfizer, which did much worse than Moderna here: 0.39-0.64). In Florida, which is currently experiencing its largest COVID-19 surge to date, the risk of infection in July after full vaccination with mRNA-1273 was about 60% lower than after full vaccination with BNT162b2 (IRR: 0.39, 95% CI: 0.24-0.62). Our observational study highlights that while both mRNA COVID-19 vaccines strongly protect against infection and severe disease, further evaluation of mechanisms underlying differences in their effectiveness such as dosing regimens and vaccine composition are warranted.

I’m going to go ahead and say this is probably all hopelessly confounded. It’s all observational, they didn’t periodically test populations. The size of the difference between Pfizer and Moderna here is absurdly high. And, yeah, no.

One big motivation for that: That obviously doesn’t make *any* sense in the context of the headline claims. If vaccines were only 60% effective, even if we only count 50% of the effective population as vaccinated we would still get 30% of cases as breakthrough cases. Also, if Pfizer vaccinations are only 60% effective, then as per the calculation last week Delta would represent a more than doubling of cases *every four days* versus Alpha until behavior adjusted, and we would have seen the Delta variant take over from 1% to over 50% of cases within a span of under three weeks, and from 1% to 99% within six weeks.

So if we’re not seeing anything like those results in general, then either we need to explain what’s going on somehow or claims of vaccine effectiveness this low are Obvious Nonsense. These are the first checks any reasonable person would run, and I’m confused why I’ve *literally not seen anyone else do the calculation*.

The prevalence of Delta in their study in May was 0.7% and in July it was ‘over 70.’ If we assume a constant growth rate, what does that imply? I built a toy spreadsheet to see, with two free variables of daily growth rate of Delta relative to Alpha, and the initial share on May 1. We get 0.10% Delta share on May 1 and a daily growth rate of 11% of Delta relative to Alpha, which passes sanity checks.

Which once again places us back at Delta being 50% or so more infectious than Alpha *among the entire United States population*. But we also think it’s 50% more infectious among the unvaccinated population alone. Can’t have it both ways. There are only so many ways out of this, as it’s a math problem. For completeness:

1. [HYPOTHESIS] Serial interval for Covid is longer than we think.
2. [HYPOTHESIS] Delta isn’t that much more infectious than Alpha for the unvaccinated.
3. [HYPOTHESIS] Behavior adjusted radically during this period and somehow explains it.
4. [HYPOTHESIS] Regional dynamics explain this or something, maybe?
5. [HYPOTHESIS] Vaccinated people aren’t getting tested and their cases are missed.
6. [HYPOTHESIS] Vaccinated people aren’t infectious, or at least much less than others.
7. [HYPOTHESIS] Which cases are vaccinated aren’t tracked properly so there are a lot more ‘breakthroughs’ than the data say.

I think we can safely dismiss 1 through 4:

1. This would contradict a lot of other data points and there's no one proposing it. Delta is considered to move faster than Alpha, so if anything this variable makes our problem bigger rather than solving it.
2. Delta's mechanism is much larger viral loads, we have lots of data points showing it's more infectious in the unvaccinated, and it spread like wildfire to take over India where no one was vaccinated. Seems highly implausible.
3. If we separate out the Alpha and Delta case counts we don't see any radical adjustments and I don't see any way this solves the puzzle even if we found them - I listed this for completeness.
4. Delta took over at roughly the same time across the country, I don't see how this could solve the problem.
5. Could be true, given that we think 'breakthrough' cases are mostly asymptomatic or at least not serious. Logically this combines well with (6), since (5) solves the 'where are the infections' issue but not the growth rate, and (6) solves the growth rate but doesn't explain where the infections are.
6. Could also be true to varying degrees. To the extent that someone never gets infectious and isn't detected, the case doesn't really 'count.'
7. The earlier numbers for breakthrough infections were so low that it does make one wonder about this possibility. This can substitute for (5).

So the only way to make sense of this is some combination of (5) and (7), plus a large dose of (7). In this scenario, we've been missing most of the breakthrough infections the whole time, either because they don't get tested and/or we don't write down that they're breakthroughs, and mostly they're not serious infections. The missing cases would be invisible, neither spreading disease nor causing noticeable problems. Not exactly a nightmare scenario, and those infections would in turn strengthen immunity going forward.

Thing is, that scenario reconciles the population data with the vaccines losing effectiveness, but it's not consistent with the study data, because this predicts that you'd see in the study what you see in the population. This was all observational, so the missing infections shouldn't have been detected. On top of that, *hospitalization rates conditional on infection* didn't change in the study, so a bunch of missing harmless infections doesn't work at all here.

[Mason here cites Israeli data](#) that on its face suggests vaccine effectiveness declined down to 16% over time, with dramatic and rapid drops. I don't consider this credible, and presume it comes from the same mistake that was previously made in Israel of assuming all the cohorts are otherwise the same. Scientifically I don't find this plausible, and also all the arguments about growth rates apply here as well. Still, it would be wrong to silently not include such information.

[A less courteous view of the Israeli data \(his full post\)](#):



Venk Murthy @venkmurthy · 2h

Widely reported leaks suggest US govt will push for approval & uptake of 3rd COVID vax doses.

...

What is the likelihood that evidence being used to support the need for this to prevent severe COVID is confounded, obs data that doesn't account age stratification in vax uptake?



44 votes · 2 days left

Q 1

↑↓

♡

↑



Venk Murthy

@venkmurthy

...

Replying to [@venkmurthy](#)

Please don't cite observational data from Israel as those are easily shown to be compatible with pretty darn good efficacy for 2 doses.



Venk Murthy @venkmurthy · 3h

Must read article teaches us about:

- * How Pfizer vax still likely very effective against severe disease even in Israeli data
- * Concept of Simpson's paradox
- * General principle of how confounding can mislead in observational data of therapies

Here's the chart with and without adjusting for age, yay [Simpson's Paradox](#):

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	67.5%
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	91.8%
>50	186,078 7.9%	2,170,563 90.4%	171 90.9	290 13.6	85.2%

Save	Population (%)		Severe cases/100k		Severe Case Risk	Efficacy ↗
	% Not Vax	% Fully Vax	Not Vax	Fully Vax		
12-15	62.1%	29.9%	0.30	0.00	1/20x	100%
16-19	21.9%	73.5%	1.60	0.00	1/4x	100%
20-29	20.5%	76.2%	1.50	0.00	1/4x	100%
30-39	16.2%	80.9%	6.20	0.20	1	96.8%
40-49	13.2%	84.4%	16.50	1.00	2.7x	93.9%
50-59	10.0%	88.0%	40.20	2.90	6.5x	92.8%
60-69	8.8%	89.8%	76.60	8.70	12.4x	88.7%
70-79	4.2%	94.6%	190.10	19.80	30.7x	89.6%
80-89	5.6%	92.6%	252.30	47.90	40.7x	81.1%
90+	6.1%	90.5%	510.9	38.60	82.4x	92.4%

Then you need to adjust for other things, including that vaccination rates are higher in the cities, which was the core problem last time. I still don't love the numbers we're seeing here, even after adjustments, but they're not super scary or huge outliers.

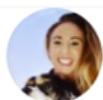
Most of all, let's [not overthink this and lose sight of the obvious](#).

99% of COVID deaths are now of unvaccinated people, experts say

With the delta variant running rampant in the US, COVID cases are on the rise in what is now a "pandemic of the unvaccinated," the CDC says.

The vast majority of people being hospitalized with COVID and dying from the disease haven't been fully vaccinated, according to public health officials. More than 97% of hospitalizations from COVID right now are of unvaccinated people, Dr. Rochelle Walensky, director of the Centers of Disease Control and Prevention, said at a press briefing Friday, adding: "There is a clear message that is coming through: This is becoming a pandemic of the unvaccinated." In early July, Dr. Anthony Fauci, the president's chief medical advisor, told CBS that 99.2% of COVID deaths are now of unvaccinated people.

This was on July 29, so not fully current, but cases were already mostly Delta, and the vaccinated were the majority of adults and a large majority of the elderly and most vulnerable [Eyes on the prize](#).



Kathryn Watson

@kathrynw5

...

The unvaccinated account for 96.4% of D.C. COVID hospitalizations. Vaccines work.



Charles Allen @charlesallen · Aug 16

Per DOH today:

% of DC COVID cases (7-day avg):

- 👉 Vaccinated = 16.4%
- 👉 Unvaccinated = **83.6%**

% of DC COVID hospitalizations (7-day avg):

- 👉 Vaccinated = 3.6%
- 👉 Unvaccinated = **96.4%**

Here's another number for you: 1-855-363-0333. They will bring the vaccine to your home!

—
There remain large unknowns in all this, but my best guesses on the state of Delta have not changed much. Here's my current model.

1. Can vaccinated people spread the virus if they catch it? Yes, of course they can, but less often than a similar unvaccinated person, mostly due to differences in severity and duration. This is too much nuance for the press.
2. How effective is the Pfizer vaccine against symptomatic Delta? Over my range of possible answers, mean 86%, median 89%.
3. How effective is the Pfizer vaccine against death from Delta? 99%+.

4. Does the Pfizer vaccine lose effectiveness from those numbers, with time? Probably some, but nothing like the extreme numbers being suggested.
5. Should you get a booster shot? If there's sufficient supply and they'll give it to you if you ask for it, I'd do it.
6. How much more infectious is Delta than Alpha among the unvaccinated? About 50%.
7. What is the serial interval of Delta? Probably 3 days, maybe 4. For Alpha it's 5.

Booster Shot

[In case we need them, we've already purchased everyone's booster shots.](#) Looks like we're going to need them.

[Pfizer's earnings report seems to have contained its latest data on booster shots, which then got covered on CNN.](#) I don't think anything meaningfully fishy happened, but it's a very strange place to put your scientific data.

Booster shots are now available to the immunocompromised, and there are reports that they're considering calling for boosters for everyone soon. [Of course, until the time when the booster is the Only Responsible Thing To Do, it's forbidden](#), even unthinkable to many, because that's how things work and they think talk of boosters will discourage vaccinations.



Alex Tabarrok
@ATabarrok

...

I am old enough to remember when the FDA and CDC immediately wrote a letter rebuking Pfizer for even applying for a booster shot. Oh, that was July.
cnn.com/2021/07/08/hea...



Catherine Rampell @crampell · 12h

NYT reports Biden administration has decided that most Americans should get a coronavirus booster shot eight months after they completed their initial vaccination, and could begin offering the extra shots as early as mid-September
[nytimes.com/2021/08/16/us/...](https://nytimes.com/2021/08/16/us/)

The real nonsensical thing, however, is that if you accidentally got J&J as your first shot, [well, whoops](#) (MR). No second dose for you, let alone a third, no matter how immunocompromised you are.

Whereas it looks like they're not requiring proof of underlying medical conditions to get the third shot, so, [as MR puts it](#), solve for the equilibrium.

I notice that this is sufficiently absurd that it crosses my 'you're not supposed to lie' threshold. If you're in this situation, and you want a second shot, I think it's fine for you to do what you need to do.

Soon, you'll be able to get a third shot eight months after your second shot. Biden plans to do this '[when it is available](#) (WaPo)' as if he wasn't the President of the United States and

couldn't get his hands on a third dose or perhaps face a *slightly* unusual cost-benefit calculation that might justify doing it a bit early. It's bizarre the times that Biden thinks the geometries bind him, versus when he thinks they do not, I can mostly predict it intuitively but it's still very strange.

Think of the Children

[Kids are at higher risk from Delta than Alpha, as one would expect.](#)



David Fisman  @DFisman · Aug 13

Sharing this via twitter because I think it's in the public interest to know: we now have enough delta cases in kids, and enough pediatric hospitalizations in Ontario, that we can estimate the odds ratio for hospitalization with delta. It is increased, OR = 2.75 (1.59 to 4.49)

The thing is, tripling something only matters three times as much as the original number. The actual increase here is still very small.

How bad is it really? [Scott Gottlieb points out we're not gathering the data \(REACT data\).](#)



Scott Gottlieb, MD  @ScottGottliebMD · Aug 15

...

In the U.S. we have no firm idea how many kids have already been infected with COVID. We have no idea if hospitalizations in south are tip of a huge iceberg of dire infection - or a sign that COVID has become more pathogenic in children. The CDC should gather this data. It isn't.



Scott Gottlieb, MD  @ScottGottliebMD · Aug 15

...

Replying to @ScottGottliebMD

Britain has this data. Their REACT study evaluates population-level info to reveal where, how COVID is spreading. We have no similar effort in U.S. CDC's cohort studies are small, narrow - monitoring specific groups like nursing homes and essential workers

On the one hand, yes, absolutely we should be gathering the data. On the other hand, data does not stop being true because it's from across the pond, and child data can be compared to the adult data, so why not look at what Britain's REACT data found? I didn't see any takeaways on child infections, and it doesn't seem like the best use of time for me to look right now, but someone should definitely look.

Periodic reminder: Whether mask mandates prevent Covid transmission in schools is one of those things that's too important to know, because it would not be 'ethical' to study it properly even though some places already have mandates and some do not. Still, if you want to issue a study that says masks prevent Covid transmission in schools, it would be useful to actually compare schools where they wore masks to schools where they did not wear masks. [Or else, one might say you have provided exactly zero evidence.](#)

As a periodic reminder, [very few people are saying anything about preventing individuals from doing NPIs to protect themselves](#), or telling them that they have to dine indoors or go to

concerts. [No one is proposing ‘banning masks in schools.’](#)



I do realize that saying this is false is a highly uncharitable interpretation of what Biden said, I mean everyone know what he *meant*, take him seriously not literally and all that, but it's also not what anyone in question is doing, and the distinction is important. Mandating a lack of mandates is not to forbid the underlying action.

The arguments are mainly over whether to (A) mandate mandating masks or (B) mandate *not* mandating masks, without that much support for (C) neither mandating nor not mandating mask mandates and letting people decide under what rules they will associate, either in local public venues like schools and/or local private venues like gyms or restaurants. Thus, Florida and Texas mandate not mandating masks in schools and threaten to withhold funding, and others try to sue or coerce them into mandating mandating masks instead, or at least take Biden's tactic of [mandating not mandating a lack of mandates](#) (WaPo) including using the civil rights act.

You gotta love attempts like this (from the WaPo article on Biden's booster shot):

In Texas, a school system has made masks a [part of its dress code](#) for the academic year, hoping to exploit a possible loophole in a statewide ban on face coverings by Gov. Greg Abbott (R), [who currently has covid-19](#). And in Florida, [Gov. Ron DeSantis \(R\) is now facing a revolt](#) from a growing number of school districts that have required masks for the new school year despite his ban on mandates and threat to punish those who defy it.

I do think masks are an important escalation in the dystopian quotient of the American school, but it is nice to remind oneself that the existing bar was not so low.

And it's a reminder that I can't find an *example* to contradict the claim that 'no one is banning masks in schools' but I have no doubt that *someone somewhere* is doing exactly that. Schools tell you how to dress, it's rather standard. A large percentage of schools banned masks in 2019 and assumed anyone wearing one was up to no good, and any claims regarding health benefits were doubtless mostly ignored. Everything being either mandatory or forbidden is the mindset here, and also central to what the schools are trying to instill in the children.

Periodic reminder: Air filters in schools were worthwhile before Covid anyway, estimated 0.15 SD impact on test scores, and also we *shouldn't need to cite test scores here*, free our kids or at least let them breathe in the meantime.

Periodic reminder: Schools are what they are, and some kids hate them enough [they prefer the dystopian nightmare that we refer to as ‘remote learning’](#) in order to be free of an even worse nightmare.

[Scott Alexander has a post this week called “Kids Can Recover From Missing Even Quite a Lot of School”](#) pointing out that losing out on a bunch of school seems to do surprisingly little, even *measured in school test scores*, as long as you don't miss school during the month when they're spending your life cramming for the test before you forget everything again. Excused absences don't do anything, even though being sick is usually rather bad for almost everything, so one could wonder whether they're actively *good* for test scores. Scott doesn't

*explicitly ask the obvious question, which is “does school actually do anything useful while it’s traumatizing kids to obey authority rather than examine physical reality, and ending their lives one minute at a time?” or alternatively “But Can Kids Recover From Not Missing Even Quite a Lot of School?” but he does point out that school didn’t teach *him* anything useful and now he’s Scott Alexander, and it didn’t teach me much of anything either, so search your experiences and draw your own conclusions, and then draw your own secondary conclusions about the pandemic.*

(Great Minds Think Alike note for next story: Between when I wrote this section and when I hit publish on Thursday, Scott Alexander came out with his links post, in which he has almost exactly the same take I do.

Young children learn about the world by interacting with it. [What would happen if suddenly all the adults around them were covering the lower half of their faces?](#) Someone really ought to do a study, except no, no one will ever do that.



Mason @webdevMason

...

Would an IRB even approve *studying* adult caregivers obscuring their faces from babies long term?



American Academy of Pediatrics @AmerAcadPeds · Aug 12

Babies and young children study faces, so you may worry that having masked caregivers would harm children’s language development. There are no studies to support this concern. Young children will use other clues like gestures and tone of voice. [healthychildren.org/English/health...](http://healthychildren.org/English/health/)



Mason @webdevMason · Aug 12

...

Replying to @webdevMason

'Cause yeah, if something is so likely to cause harm that you can't ethically study it, there will be "no studies to support this concern."

Whoops:

The face-deprived monkeys and control monkeys were scanned by fMRI when they were six months old to measure their neural responses to faces and other visual stimuli.

Control monkeys had face patches by the time they were six months old; the face-deprived monkeys did not. Patches for other visual categories that both sets of monkeys saw equally, such as hands and bodies, were roughly equivalent between the two groups.

These findings suggest that looking at faces is necessary for the development of face patches; in general, each region of “expertise” probably develops through extensive experience with that type of visual stimulus. [In an earlier study](#), the same research group trained young monkeys to recognize Arabic numerals, Tetris pieces, and other abstract symbols – not things monkeys evolved to see. Yet this experience was enough to produce “symbol patches” in fMRI scans of their brains.



bosco @selentelechia · Aug 13

...

Replying to @webdevMason

My kid (5 months) does not smile or respond much to people who have masks on

she is extremely socially engaged when people are unmasked

don't tell me it doesn't have an effect



Steve Wilson @SteveWi62205415 · Aug 13

Replying to @webdevMason

Strong "no evidence of human to human transmission" energy.

This is the only proposed method of possibly getting a study done, and I don't think even this would work in context:



Anarcho-Moses @ben_r_hoffman · Aug 13

...

Replying to @webdevMason

Only if the proposal's discussion of risk seemed like the right kind of dishonest.

There is a taxonomy whereby there are two kinds of things in the world. There are things that are Risky until proven Safe by a Proper Scientific Study, and things that are Safe until

proven Risky. Authorities and “experts” choose, based on framing, context and their incentives, which way to present a given thing, and a lot of the talking past each other comes from this disagreement over priors. To the extent that it’s an honest disagreement and continuous rather than a boolean, this is reasonable, and to the extent it’s not, it’s not.

I don’t know how concerned I should be about this particular problem, but I’m confident the correct answer is not to be unconcerned, and especially not to be unconcerned due to there not being a study on it.

Also, at least some data is in on how babies are doing these days, and [Kerry’s perspective seems directionally wise](#), although I agree with the study authors that the social isolation is the dominant factor here rather than masks ([study](#)).



Kerry

@kerry62189

...

This study may well be sensationalist nonsense, but I think any policy that deprives infants and toddlers of normal social interaction for more than a month is unethical and should be opposed on principle from now on. This is a very serious thing to gamble with.



New York Post @nypost · Aug 13

Babies born during COVID-19 pandemic have lower IQs: Study says
trib.al/1W1Rra0

In Other News

FDA Delenda Est, but it could always be worse. [We could have the TGA.](#)

[Fluvoxamine reduces hospitalization from Covid by 31% in preliminary results](#), Ivermectin found to have no effect. [Additional coverage here](#). Sample sizes between one and two thousand. This is a cheap generic and known to be relatively safe, so this seems like enough to justify using it.

[Zeynep thread on Covid origins](#), it seems the WHO let the Chinese tell them what hypotheses could and couldn’t be in their report. Sounds about right.

[The Governor of Texas tests positive for Covid](#), after testing negative every day or quite a while. Daily testing is an interesting idea.

[Latest Vitamin D study](#), not directly on Covid or directly measuring deficiency, but showing that moving from a high-D area to a low-D area (due to less sunlight) is associated with worse outcomes. I don’t mention Vitamin D as much as I should, as it’s one of the practical things an individual can do that has high expected value in terms of preventing or helping with Covid, that would be a good idea even without Covid. And yet I struggle to remember to take it.

A reminder that while you do not want to catch Covid, *permanently crippling our way of life* is not a reasonable price to pay for that, and [we need to stand up to ‘health experts’ who](#)

[think such things are reasonable and run them out of town on a rail](#). Such talk does not encourage reasonable short term behavior, and potentially lays the groundwork for the destruction of our way of life. Life beckons. Live it. [Which, as Nate Silver notes, most people are correctly doing.](#)

of unfettered social interactions, she added: “This seems to me like a real possibility, since many early vaccinated were motivated by a desire to see friends and family and get back to normal.”

Dr. Murray said boosters would undoubtedly boost immunity in an individual, but the benefit may be minimal — and obtained just as easily by wearing a mask, or avoiding indoor dining and crowded bars.

The administration’s emphasis on vaccines has undermined the importance of building other precautions into people’s lives in ways that are comfortable and sustainable, and on building capacity for testing, she and other experts said.

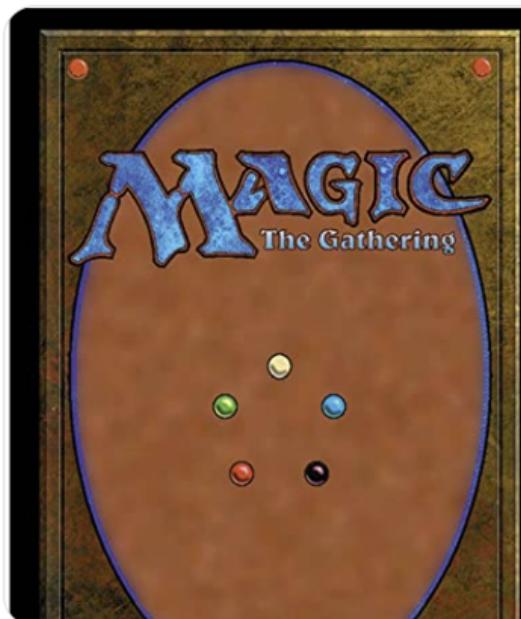
The fall plans meme is over and we have a winner, [congratulations Cedric](#):



Cedric A Phillips
@CedricAPhillips

...

my fall plans:



the delta variant:



Magic banned Oko. That was wise.

Not Covid

If you don't think environmental review is completely out of hand, New York City tried to pass congestion pricing, and it's [going to be held up for sixteen months to start for "environmental review"](#) despite its impact there being purely beneficial.

If you think that's bad, the city of San Francisco is going to build a tunnel for a train line and its *baseline budget* for environmental review is (all caps mandatory, [there are rules](#)) ONE BILLION DOLLARS. With a B.



Roan Kattouw @catrope · Aug 16

...

Replies to [@catrope](#)

The project to build a second train tunnel between San Francisco and Oakland says they're budgeting \$1 billion just for environmental review.

ONE BILLION DOLLARS, to prove that adding capacity to the most crowded train line in California is good for the environment

39

383

1.3K



Roan Kattouw @catrope · Aug 16

...

Meanwhile, a project to widen Highway 1 near Santa Cruz openly admits in its environmental review document that it'll cause more driving, but they get to compare *future* emissions with the project to *current* emissions without the project.

[Failure to do cost-benefit analysis and scope insensitivity, in one Twitter poll:](#)



Aella ✅ @Aella_Girl · 5h

...

If you press the button, somewhere in the world, a young adult about to be killed by a drunk driver, is magically saved.

Also, every ceiling in the world is (safely, magically) lowered by 0.5 inches.
how many times do you press the button?

0 ✓

54.3%

1-5

18%

6-30

11.1%

31+

16.7%

2,899 votes · 1 day left

50

9

75

↑



Justin Cohen

@trippdup

...

Replies to @Aella_Girl

I just did some back of the envelope math and the question you have to ask yourself is how many *trillions* of feet of cubic space is one life worth?

How to turn money into AI safety?

Related: [Suppose \\$1 billion is given to AI Safety. How should it be spent? , EA is vetting-constrained, What to do with people?](#)

I

I have heard through the grapevine that we seem to be constrained - there's money that donors and organizations might be happy to spend on AI safety work, but aren't because of certain bottlenecks - perhaps talent, training, vetting, research programs, or research groups are in short supply. What would the world look like if we'd widened some of those bottlenecks, and what are local actions that people can do to move in that direction? I'm not an expert either from the funding or organizational side, but hopefully I can leverage [Cunningham's law](#) and get some people more in the know to reply in the comments.

Of the bottlenecks I listed above, I am going to mostly ignore talent. IMO, talented people aren't the bottleneck right now, and the other problems we have are more interesting. We need to be able to train people in the details of an area of cutting-edge research. We need a larger number of research groups that can employ those people to work on specific agendas. And perhaps trickiest, we need to do this within a network of reputation and vetting that makes it possible to selectively spend money on good research without warping or stifling the very research it's trying to select for.

In short, if we want to spend money, we can't just hope that highly-credentialed, high-status researchers with obviously-fundable research will arise by spontaneous generation. We need to scale up the infrastructure. I'll start by taking the perspective of individuals trying to work on AI safety - how can we make it easier for them to do good work and get paid?

There are a series of bottlenecks in the pipeline from interested amateur to salaried professional. From the individual entrant's perspective, they have to start with learning and credentialing. The "obvious path" of training to do AI safety research looks like getting a bachelor's or PhD in public policy, philosophy, computer science, or math, (for which there are now [fellowships](#), which is great) trying to focus your work towards AI safety, and doing a lot of self-study on the side. These programs are often an imprecise fit for the training we want - we'd like there to be graduate-level classes that students can take that cover important material in AI policy, technical alignment research, the philosophy of value learning, etc.

Opportunity 1: Develop course materials and possibly textbooks for teaching courses related to AI safety. This is [already happening somewhat](#). Encourage other departments and professors to offer courses covering these topics.

Even if we influence some parts of academia, we may still have a bottleneck where there aren't enough departments and professors who can guide and support students focusing on AI safety topics. This is especially relevant if we want to start training people *fast*, as in six months from now. To bridge this gap this it would be nice to have training programs, admitting people with bachelor's- or master's-level skills, at organizations doing active AI safety research. Like a three-way cross between internship, grad school, and [AI Safety Camp](#). The intent is not just to have people learn and do work, but also to help them produce credible signals of their knowledge and skills, over a timespan of 2-5 years. Not just being author number 9 out of 18, but

having output that they are primarily responsible for. The necessity of producing credible signals of skill makes a lot of sense when we look at the problem from the funders' perspective later.

Opportunity 2: Expand programs located at existing research organizations that fulfill training and signalling roles. This would require staff for admissions, support, and administration.

This would also provide an opportunity for people who haven't taken the "obvious path" through academia, of which there are many in the AI safety community, who otherwise would have to create their own signalling mechanisms. Thus it would be a bad outcome if all these internships got filled up with people with ordinary academic credentials and no "weirdness points," as admissions incentives might push towards. Strong admissions risk-aversion may also indicate that we have lots of talent, and not enough spots (more dakka required).

Such internships would take nontrivial effort and administrative resources - they're a negative for the research output of the individuals who run them. To align the incentives to make them happen, we'd want top-down funding intended for this activity. This may be complicated by the fact that a lot of research happens within corporations, e.g. at DeepMind. But if people actually try, I suspect there's some way to use money to expand training+signalling internships at corporate centers of AI safety research.

Suppose that we blow open that bottleneck, and we have a bunch of people with some knowledge of cutting-edge research, and credible signals that they can do AI safety work. Where do they go?

Right now there are only a small number of organizations devoted to AI safety research, all with their own idiosyncrasies, and all accepting only a small number of new people. And yet we want most research to happen in organizations rather than alone: Communicating with peers is a good source of ideas. Many projects require the efforts or skillsets of multiple people working together. Organizations can supply hardware, administrative support, or other expertise to allow research to go smoother.

Opportunity 3: Expand the size and scope of existing organizations, perhaps in a hierarchical structure. Can't be done indefinitely (will come back to this), but I don't think we're near the limits.

In addition to increasing the size of existing organizations, we could also found new groups altogether. I won't write that one down yet, because it has some additional complications that are best explored from a different perspective.

II

If you're a grant-making organization, selectivity is everything. Even if you want to spend more money, if you offer money for AI safety research but have no selection process, a whole bushel of people are going to show up asking for completely pointless grants, and your money will be wasted. But it's hard to filter for people and groups who are going to do useful AI safety research.

So you develop a process. You look at the grantee's credentials and awards. You read their previous work and try to see if it's any good. You ask outside experts for a second opinion, both on the work and on the grantee themselves. Et cetera. This is all

a totally normal response to the need to spend limited resources in an uncertain world. But it is a [lot of work, and can often end up incentivizing picking "safe bets."](#)

Now let's come back the unanswered problem of increasing the number of research organizations. In this environment, how does that happen? The fledgling organization would need credentials, previous work, and reputation with high-status experts before ever receiving a grant. The solution is obvious: just have a central group of founders with credentials, work, and reputation ("cred" for short) already attached to them.

Opportunity 4: Entice people who have cred to [found new organizations](#) that can get grants and thus increase the amount of money being spent doing work.

This suggests that the number of organizations can only grow exponentially, through a life cycle where researchers join a growing organization, do work, gain cred, and then bud off to form a new group. Is that really necessary, though? What if a certain niche just obviously needs to be filled - can you (assuming you're Joe Schmo with no cred) found an organization to fill it? No, you probably cannot. You at least need *some* cred - though we can think about pushing the limits later. Grant-making organizations get a bunch of bad requests all the time, and they shouldn't just fund all of them that promise to fill some niche. There are certainly ways to signal that you will do a good job spending grant money even if you utterly lack cred, but those signals might take a lot of effort for grant-making organizations to interpret and compare to other grant opportunities, which brings us to the "vetting" bottleneck mentioned at the start of the post. Being vetting-constrained means that grant-making organizations don't have the institutional capability to comb through all the signals you might be trying to send, nor can they do detailed follow-up on each funded project sufficient to keep the principal-agent problem in check. So they don't fund Joe Schmo.

But if grant-making orgs are vetting-constrained, why can't they just grow? Or if they want to give more money and the number of research organizations with cred is limited, why can't those grantees just grow arbitrarily?

Both of these problems are actually pretty similar to the problem of growing the number of organizations. When you hire a new person, they need supervision and mentoring from a person with trust and know-how within your organization or else they're probably going to mess up, unless they already have cred. This limits how quickly organizations can scale. Thus we can't just wait until research organizations are most needed to grow them - if we want more growth in the future we need growth now.

Opportunity 5: Write a blog post urging established organizations to [actually try](#) to grow (in a reasonable manner), because their intrinsic growth rate is an important limiting factor in turning money into AI safety.

All of the above has been in the regime of weak vetting. What would change if we made grant-makers' vetting capabilities very strong? My mental image of strong vetting is grant-makers being able to have a long conversation with an applicant, every day for a week, rather than a 1-hour interview. Or being able to spend four days of work evaluating the feasibility of a project proposal, and coming back to the proposer with a list of suggestions to talk over. Or having the resources to follow up on how your money is being spent on a weekly basis, with a trusted person available to help the grantee or step in if things aren't going to plan. If this kind of power was used for good, it would open up the ability to fund good projects that previously would have been lost in the noise (though if used for ill it could be used to gatekeep for existing

interests). This would decrease the reliance on cred and other signals, and increase the possible growth rate, closer to the limits from "talent" growth.

An organization capable of doing this level of vetting blurs the line between a grant-making organization and a centralized research hub. In fact, this fits into a picture where research organizations have stronger vetting capabilities for individuals than grant-making organizations do for research organizations. In a growing field, we might expect to see a lot of intriguing but hard-to-evaluate research take place as part of organizations but not get independently funded.

Strong vetting would be impressive, but it might not be as cost-effective as just lowering standards, particularly for smaller grants. It's like a stock portfolio - it's fine to invest in lots of things that individually have high variance so long as they're uncorrelated. But a major factor in how low your standards can be is how well weak vetting works at separating genuine applicants from frauds. I don't know much about this, so I'll leave this topic to others.

The arbitrary growth of research organizations also raises some questions about research agendas (in the sense of a single, cohesive vision). A common pattern of thought is that if we have more organizations, and established organisms have different teams of people working under their umbrellas, then all these groups of people need different things to do, and that might be a bottleneck. That what's best is when groups are working towards a single vision, articulated by the leader, and if we don't have enough visions we shouldn't found more organizations.

I think this picture makes a lot of sense for engineering problems, but not a lot of sense for blue-sky research. Look at the established research organizations - FHI, MIRI, etc. - they have a lot of people working on a lot of different things. What's important for a research group is trust and synergy; the "top-down vision" model is just a special case of synergy that arises when the problem is easily broken into hierarchical parts and we need high levels of interoperability, like an engineering problem. We're not at that stage yet with AI safety or even many of its subproblems, so we shouldn't limit ourselves to organizations with single cohesive visions.

III

Let's flip the script one last time - if you don't have enough cred to do whatever you want, but you think we need more organizations doing AI safety work, is there some special type you can found? I think the answer is yes.

The basic ingredient is something that's both easy to understand and easy to verify. I'm staying at the [EA Hotel](#) right now, so it's the example that comes to mind. The concept can be explained in about 10 seconds (it's a hotel that hosts people working on EA causes), and if you want me to send you some pictures I can just as quickly verify that (wonder of wonders) there is a hotel full of EAs here. But the day-to-day work of administrating the hotel is still nontrivial, and requires a small team funded by grant money.

This is the sort of organization that is potentially foundable even without much cred - you promise something very straightforward, and then you deliver that thing quickly, and the value comes from its maintenance or continuation. When I put it that way, now maybe it sounds more like Our World In Data's covid stats. Or like 80kh's advising services. Or like organizations promising various meta-level analyses, intended for easy consumption and evaluation by the grant-makers themselves.

Opportunity 6: If lacking cred, found new organizations with really, extremely legible objectives.

The organization-level corollary of this is that organizations can spend money faster if they spend it on extremely legible stuff (goods and services) rather than new hires. But as they say, [sometimes things that are expensive are worse](#). Overall this post has been very crassly focusing on what *can* get funded, not what *should* get funded, but I can be pretty confident that researchers give a lot more bang per buck than a bigger facilities budget. Though perhaps this won't always be true; maybe in the future important problems will get solved, reducing researcher importance, while demand for compute balloons, increasing costs.

I think I can afford to be this crass because I trust the readers of this post to try to do good things. The current *distribution* of AI safety research is pretty satisfactory to me given what I perceive to be the constraints, we just need *more*. It turned out that when I wrote this post about the dynamics of *more*, I didn't need to say much about the content of the research. This isn't to say I don't have hot takes, but my takes will have to stay hot for another day.

Thanks for reading.

Thanks to Jason Green-Lowe, Guillaume Corlouer, and Heye Groß for feedback and discussion at CEEALAR.

When Most VNM-Coherent Preference Orderings Have Convergent Instrumental Incentives

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

This post explains a formal link between "what kinds of instrumental convergence exists?" and "what does VNM-coherence tell us about goal-directedness?". It turns out that VNM coherent preference orderings have the **same** statistical incentives as utility functions; most such orderings will incentivize power-seeking in the settings covered by [the power-seeking theorems](#).

In certain contexts, coherence theorems *can* have non-trivial implications, in that they provide Bayesian evidence about what the coherent agent will probably do. In the situations where the power-seeking theorems apply, coherent preferences **do** suggest some degree of goal-directedness. Somewhat more precisely, VNM-coherence is Bayesian evidence that the agent prefers to stay alive, keep its options open, etc.

However, VNM-coherence over *action-observation histories* [tells you nothing](#) about what behavior to expect from the coherent agent, because [there is no instrumental convergence for generic utility functions over action-observation histories!](#)

Intuition

The result follows because the VNM utility theorem lets you consider VNM-coherent preference orderings to be isomorphic to their induced utility functions (with equivalence up to positive affine transformation), and so these preference orderings will have the same generic incentives as the utility functions themselves.

Formalism

Let o_1, \dots, o_n be outcomes, in a sense which depends on the context; outcomes could be world-states, universe-histories, or one of several fruits. Outcome lotteries are probability distributions over outcomes, and can be represented as elements of the n -dimensional probability simplex (ie as element-wise non-negative unit vectors).

A preference ordering $<$ is a binary relation on lotteries; it need not be eg complete (defined for all pairs of lotteries). *VNM-coherent* preference orderings are those which obey the [VNM axioms](#). By the VNM utility theorem, coherent preference orderings induce consistent utility functions over outcomes, and consistent utility functions conversely imply a coherent preference ordering.

Definition 1: Permuted preference ordering. Let $\phi \in S_n$ be an outcome permutation, and let \prec be a preference ordering. \prec_ϕ is the preference ordering such that for any lotteries $L, M: L \prec_\phi M$ if and only if $\phi(L) \prec \phi(M)$.

EDIT: Thanks to Edouard Harris for pointing out that Definition 1 and Lemma 3 were originally incorrect.

Definition 2: Orbit of a preference ordering. Let \prec be any preference ordering. Its orbit $S_n \cdot \prec$ is the set $\{\prec_\phi \mid \phi \in S_n\}$.

The orbits of coherent preference orderings are basically all the preference orderings induced by "relabeling" which outcomes are which. This is made clear by the following result:

Lemma 3: Permuting coherent preferences permutes the induced utility function. Let \prec be a VNM-coherent preference ordering which induces VNM-utility function u , and let $\phi \in S_n$. Then \prec_ϕ induces VNM-utility function $u'(o_i) = u(\phi(o_i))$, where o_i is any outcome.

Proof. Let L, M be any lotteries.

1. By the definition of a permuted preference ordering, $L \prec_\phi M$ if and only if $\phi(L) \prec \phi(M)$.
2. By the VNM utility theorem and the fact that \prec is coherent, $\phi(L) \prec \phi(M)$ iff $E_{l \sim \phi(L)}[u(l)] < E_{m \sim \phi(M)}[u(m)]$.
3. Since there are finitely many outcomes, we convert to vector representation: $u^\top (P_\phi l) < u^\top (P_\phi m)$.
4. By associativity, $(u^\top P_\phi) l < (u^\top P_\phi) m$.
5. But this is just equivalent to $E_{l \sim L}[u(\phi(l))] < E_{m \sim M}[u(\phi(m))]$.

QED.

As a corollary, this lemma implies that if \prec is VNM-coherent, so is \prec_ϕ , since it induces a consistent utility function over outcomes.

Consider the orbit of any \prec . By the VNM utility theorem, each preference ordering can be considered isomorphic to its induced utility function (with equivalence up to positive affine transformation).

Then let u be any utility function compatible with \prec . By the above lemma, consider the natural bijection between the (preference ordering) orbit of \prec and the (utility function) orbit of u , where $\{\prec_\phi \mid \phi \in S_n\} \leftrightarrow \{u \circ \phi \mid \phi \in S_n\}$.^{Footnote representative}

When my [theorems on power-seeking](#) are applicable, some proportion of the right-hand side is guaranteed to make (formal) power-seeking optimal. But by the bijection and by the fact that the preference orderings incentivize the same things (by the VNM theorem in the reverse direction), the (preference ordering) orbit must have the *exact same proportion of elements* for which (lotteries representing formal) power-seeking are optimal.

Conversely, if we know that some set A of lotteries tends to be preferred over another set B of lotteries (in the preference order orbit sense), then the same argument shows that A tends to have greater expected utility than B (in the utility function orbit sense). This holds for all (utility function) orbits, because every utility function corresponds to a VNM-coherent preference ordering.

So: orbit-level instrumental convergence for utility functions is equivalent to orbit-level instrumental convergence for VNM-coherent preference orderings.

Implications

- [Instrumental convergence does not exist when maximizing expected utility over action observation histories \(AOH\)](#).
 - Therefore, VNM-coherence over action observation history lotteries [tells you nothing](#) about what behavior to expect from the agent.
 - Coherence over AOH tells you nothing *because* there is no instrumental convergence in that setting!
- In certain contexts, coherence theorems *can* have non-trivial implications, in that they provide Bayesian evidence about what the coherent agent will probably do.
 - In the situations where the power-seeking theorems apply, coherent preferences **do** suggest some degree of goal-directedness.
 - Somewhat more precisely, VNM-coherence is Bayesian evidence that the agent prefers to stay alive, keep its options open, etc.
- In some domains, preference specification may be more natural than utility function specification. However, in theory, coherent preferences and utility functions have the exact same statistical incentives.
 - In practice, they will differ. For example, suppose we have a choice between specifying a reward function which is linear over state features, or of doing behavioral cloning on elicited human preferences over world states. These two methods will probably tend to produce different incentives.

The quest for better convergence theorems

Goal-directedness seems to more naturally arise from coherence over resources. (I think the word 'resources' is slightly imprecise here, because resources are only resources in the normal context of human life; money is useless when alone in Alpha

Centauri, but time to live is not. So we want coherence over things-which-are-locally-resources, perhaps.)

In his review of [Seeking Power is Often Convergently Instrumental in MDPs](#), John Wentworth [wrote](#):

in a real-time strategy game, units and buildings and so forth can be created, destroyed, and generally moved around given sufficient time. Over long time scales, the main thing which matters to the world-state is resources - creating or destroying anything else costs resources. So, even though there's a high-dimensional game-world, it's mainly a few (low-dimensional) resource counts which impact the long term state space. Any agents hoping to control anything in the long term will therefore compete to control those few resources.

More generally: of all the many "nearby" variables an agent can control, only a handful (or summary) are relevant to anything "far away". Any "nearby" agents trying to control things "far away" will therefore compete to control the same handful of variables.

Main thing to notice: this intuition talks directly about a feature of the world - i.e. "far away" variables depending only on a handful of "nearby" variables. That, according to me, is the main feature which makes or breaks instrumental convergence in any given universe. We can talk about that feature entirely independent of agents or agency. Indeed, we could potentially use this intuition to derive agency, via some kind of coherence theorem; this notion of instrumental convergence is more fundamental than utility functions.

In his review of [Coherent decisions imply consistent utilities](#), John [wrote](#):

"resources" should be a derived notion rather than a fundamental one. My current best guess at a sketch: the agent should make decisions within multiple loosely-coupled contexts, with all the coupling via some low-dimensional summary information - and that summary information would be the "resources". (This is exactly the kind of setup which leads to instrumental convergence.) By making pareto-resource-efficient decisions in one context, the agent would leave itself maximum freedom in the other contexts. In some sense, the ultimate "resource" is the agent's action space. Then, resource trade-offs implicitly tell us how the agent is trading off its degree of control within each context, which we can interpret as something-like-utility.

This seems on-track to me. We now know [what instrumental convergence looks like in unstructured environments](#), and [how structural assumptions on utility functions affect the shape and strength of that instrumental convergence](#), and this post explains the precise link between "what kinds of instrumental convergence exists?" and "what does VNM-coherence tell us about goal-directedness?". I'd be excited to see what instrumental convergence looks like in [more structured models](#).

Footnote representative: In terms of instrumental convergence, positive affine transformation never affects the [optimality probability](#) of different lottery sets. So for each (preference ordering) orbit element \prec_ϕ , it doesn't matter what representative we select from each equivalence class over induced utility functions — so we may as well pick $u \circ \phi$!

Training My Friend to Cook

During pre-vaccination covid, my friend Brittany and I agreed to hang out once per week as quarantine buddies. Brittany has a laundry list of health problems, many of which are exacerbated by a poor diet. Brittany loves healthy food. She ate a diet based around TV dinners because she didn't know how to cook. My goal for covid lockdown was to train Brittany to cook.

It was summer. I started out by picnicing in the park together. I brought rice, beans, sauerkraut, sliced radishes, homegrown tomatoes, tortilla chips and homemade salsa. Though simple, the food I made was far tastier than anything Brittany was eating. It's not hard for fresh homemade food to beat TV dinners. After each picnic I'd send the leftovers home with her in lovely glass jars. This associated "homemade food" with sunshine, verdant trees, picnic tables and quality time with good friends.

I talked about how cheap it was to make food when your primary ingredients are rice, cabbage, onions and dried beans. This made an impact because Brittany was regularly paying \$15 per meal for her TV dinners, a far inferior food. I never said "you should cook" because that would make Brittany feel bad for not cooking. I just talked about how great it was that I could cook. I showed her an ambition she could aspire to.

Brittany is a med student with very limited time. Going to a park takes time. Have already established the "homemade food" = "warm verdant picturesque parks" association, we moved our hangouts to her apartment. I brought the ingredients to her apartment and cooked them there. She insisted on splitting the grocery bill. A single night of my cooking would give her days of good food for the price of a single TV dinner.

Though I always provided Brittany with leftovers, I also made sure that they never lasted more than a few days. Brittany would eat delicious food for a few days and then she'd be back to her TV dinners. Brittany looked forward to our hangouts.

Brittany is Taiwanese and loves healthy food. I upgraded rice and beans to her favorite foods. I sautéed bok choy. I blanched Chinese broccoli. I figured out a mild mapo tofu. I learned how to prepare 絲瓜. After Brittany had watched me sauté vegetables a few times I encouraged her take over. I made a big show of how I got to relax while she did the cooking.

One day Brittany ordered a box of meal kits. I gently, but stubbornly, refused to help her with them. I walked her to her nearby grocery store where we bought groceries together. I guided her through sautéing onions. Eventually she was cooking all of her favorite dishes.

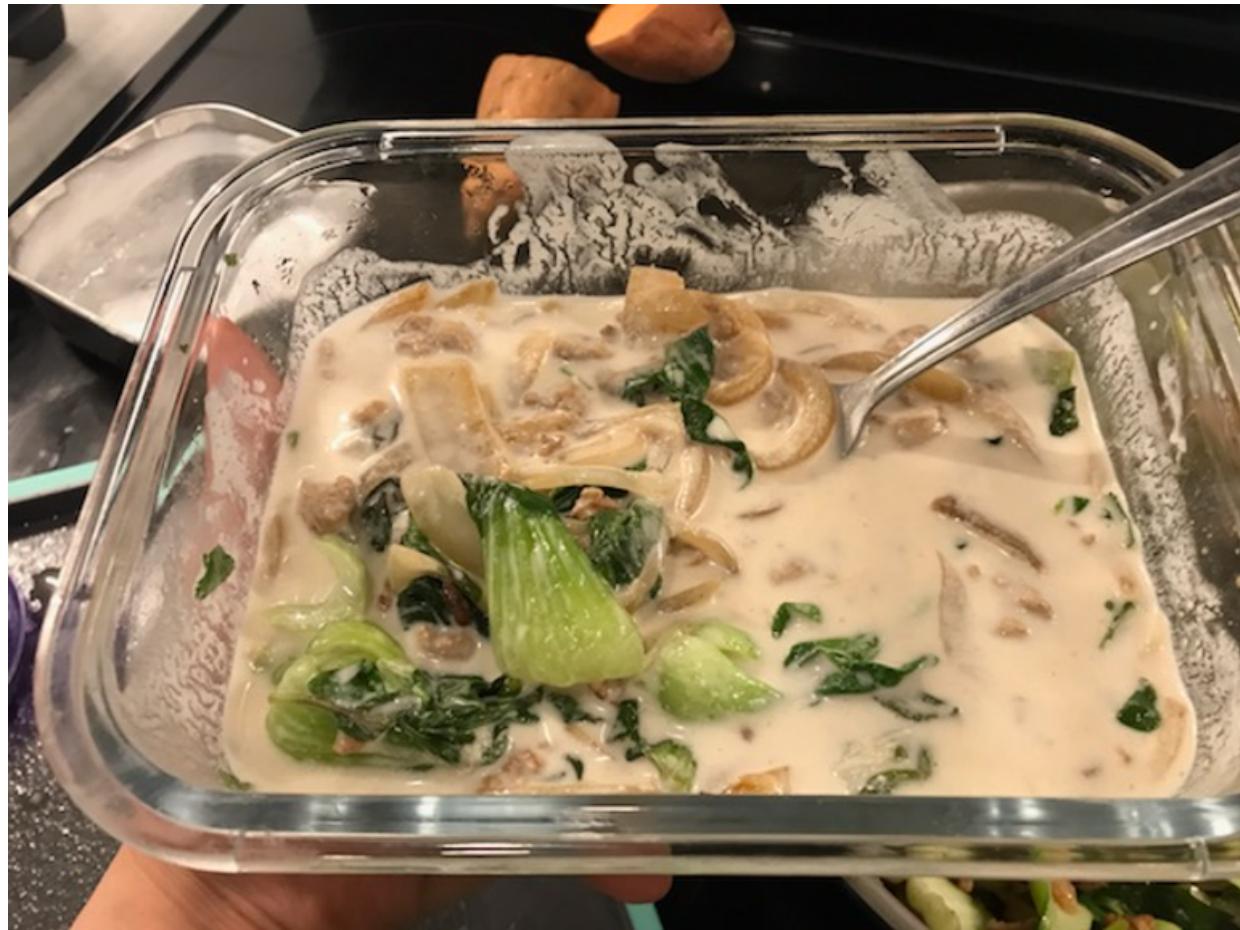
I worried that I had conditioned Brittany to cook solely when I came over. Suddenly, mid-week between our hangouts, she sent me these photos.











I cooked them on my own time based off of memory and our notes 📝 with no help !!!! you really really really helped me!!! You really are the best.

Can't wait to practice more dish!!! Haha 😊 life quality is seriously improving 😊

Really proud to share these photos.

Aww. Weeks later, she followed it up with another message.

You put me on a habit roll to walk straight to the grocery store and cook lol



These days I can give her my extra veggies and she will blanch them.



Takeaways

Changing your own long-term behavior is hard. Changing someone else's is even harder. It is easy to do more harm than good. To make sure I actually did good, I employed protocols from [*Don't Shoot the Dog*](#) by Karen Prior.

- The most important protocol I used was **strictly positive reinforcement**. I never, at any point, implied that Brittany might be deficient because she didn't know how to cook. I didn't even imply she "should" cook. I always framed the activity as "isn't this an exciting opportunity to viscerally improve your life"? Brittany never associated cooking with anything unpleasant. To this day, the activity is solely associated with reward.
- The second most important protocol I used was **backchaining**. I didn't start by bringing Brittany to the store, then teach her to cook and have her eat at the end. I started by feeding her, then I taught her to cook and only at the end did I bring her to the grocery store. If I started by bringing her to the grocery store then she wouldn't understand why we were buying what we were buying. But when I started by serving her food in a park all she had to understand is "eat delicious food". At every point in the process, Brittany understood exactly how the thing she was doing connected to "eating delicious food on a summer day under green trees".
- I let Brittany watch me perform a skill a few times over a few weeks before having her do it herself. I never told her to read anything. **Monkey see. Monkey do.**

Snafu

Right after Brittany learned to sauté, I made a big deal about how she "could sauté any vegetable". She sautéed asparagus. (You should not sauté asparagus.) I didn't mean that she could literally sauté any vegetable. By "any vegetable", I meant she could sauté any vegetable that experienced cooks know is something you should sauté. I laughed it off and took responsibility for giving bad instructions. I explained how I often messed things up too and made sure to complement her initiative.

What fraction of breakthrough COVID cases are attributable to low antibody count?

To what extent are COVID cases in vaccinated people a result of low antibody count, vs other factors (e.g. initial viral load)? Where does most of the variance come from?

Motivation for this question: in a world where most breakthrough COVID cases hit people with low antibody count, one could get some kind of antibody test (probably of a particular type) and then either (a) get an extra vaccine if antibodies are low, or (b) just don't worry if antibody counts are high. That makes antibody tests (of whatever the particular type is) very high value, since we can behave very differently in those two cases. In a world where most of the variance comes from other factors (like initial viral load), results of an antibody test don't provide so much value.

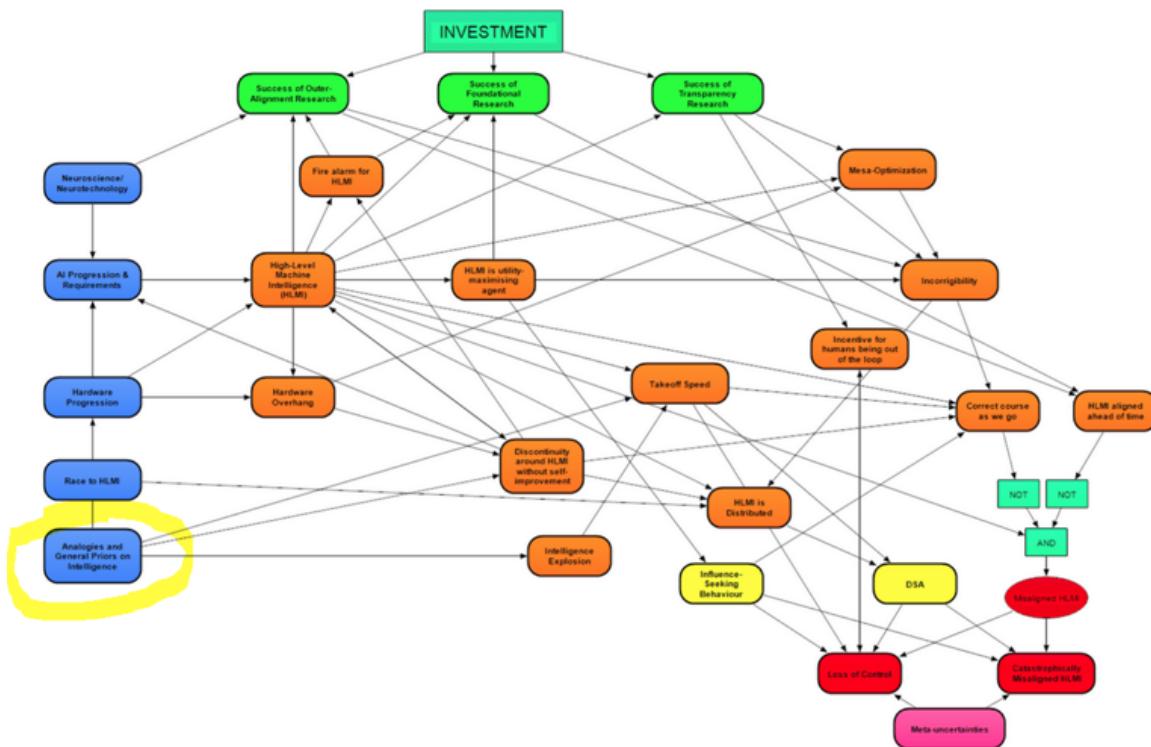
Related: [Is antibody testing to assess effects of vaccines on you a good idea?](#)

Analogies and General Priors on Intelligence

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

This post is part 2 in our [sequence on Modeling Transformative AI Risk](#). We are building a model to understand debates around existential risks from advanced AI. The model is made with [Analytica](#) software, and consists of nodes (representing key hypotheses and cruxes) and edges (representing the relationships between these cruxes), with final output corresponding to the likelihood of various potential failure scenarios. You can read more about the motivation for our project and how the model works in the [Introduction post](#). Future posts will explain how different considerations, such as AI takeoff or mesa-optimization, are incorporated into the model.

This post explains our effort to incorporate basic assumptions and arguments by analogy about intelligence, which are used to ground debates about AI takeoff and paths to High-Level Machine Intelligence (HLMI^[1]). In the overall model, this module, *Analogies and General Priors on Intelligence*, is one of the main starting points, and influences modules (covered in subsequent posts in this series) addressing the possibilities of a *Discontinuity around HLMI* or an *Intelligence Explosion*, as well as *AI Progression* and *HLMI Takeoff Speed*.

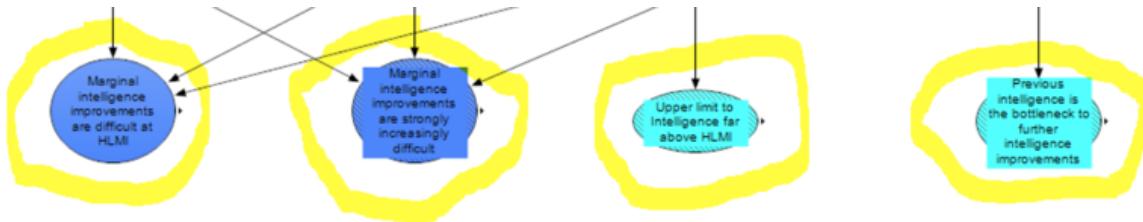


The *Analogies and General Priors on Intelligence* module addresses various claims about AI and the nature of intelligence:

- The difficulty of marginal intelligence improvements at the approximate ‘human level’ (i.e., around HLMI)
- Whether marginal intelligence improvements become increasingly difficult beyond HLMI at a rapidly growing rate or not

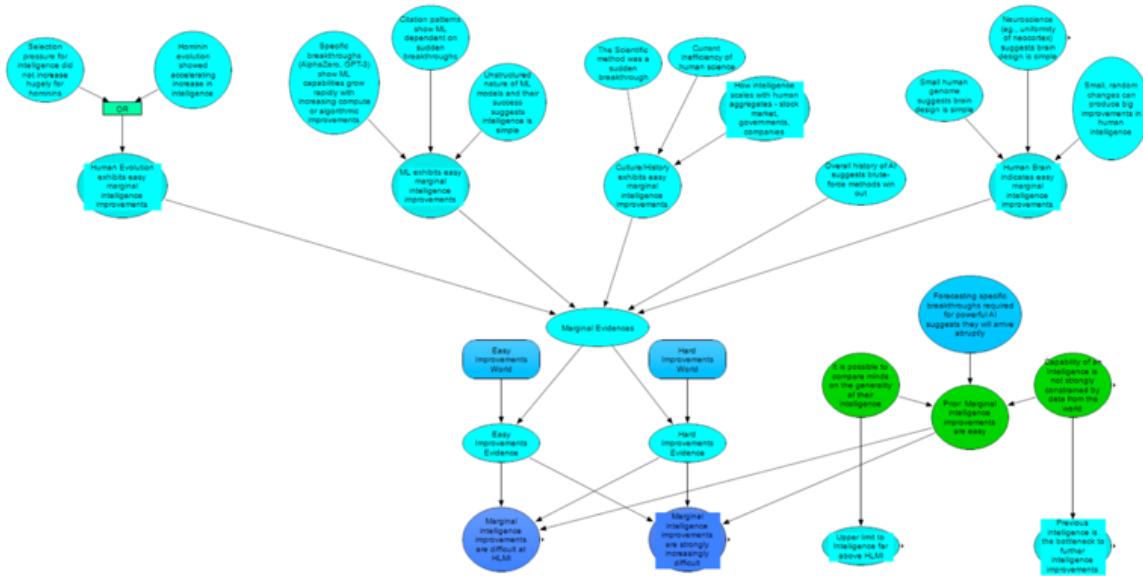
- ‘Rapidly growing rate’ is operationalized as becoming difficult exponentially or faster-than-exponentially
- Whether there is a fundamental upper limit to intelligence not significantly above the human level
- Whether, in general, further improvements in intelligence tend to be bottlenecked by previous improvements in intelligence rather than some external factor (such as the rate of physics-limited processes)

These final outputs are represented by these four terminal nodes at the bottom of the module.



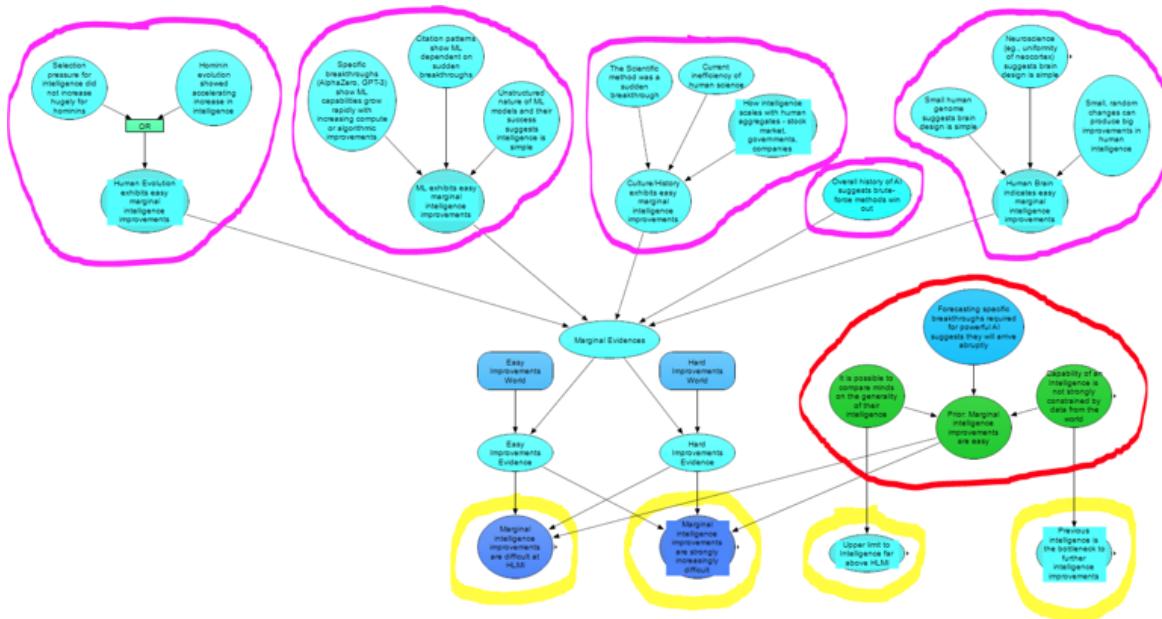
These four claims depend on arguments that analogize the development of HLMI to other domains. Each of these arguments is given a specific submodule in our model, and each argument area is drawn from our review of the existing literature. In the ‘outputs’ section at the end of this article, we explain why we chose these four as cruxes for the overall debate around HLMI development.

Each argument area is shown in the image below, a zoomed out view of the entire *Analogy and General Priors on Intelligence* module. The argument areas are: human biological evolution, machine learning, human cultural evolution, the overall history of AI, and the human brain. Each of these areas is an example of either the development of intelligence or of intelligence itself, and might be a useful analogy for HLMI.



We also incorporate broad philosophical claims about the nature of intelligence as informative in this module. These claims act as priors before the argument areas are investigated, and cover broad issues like whether the concept of general intelligence is coherent and whether the capabilities of agents are strongly constrained by things other than intelligence.

Module Overview

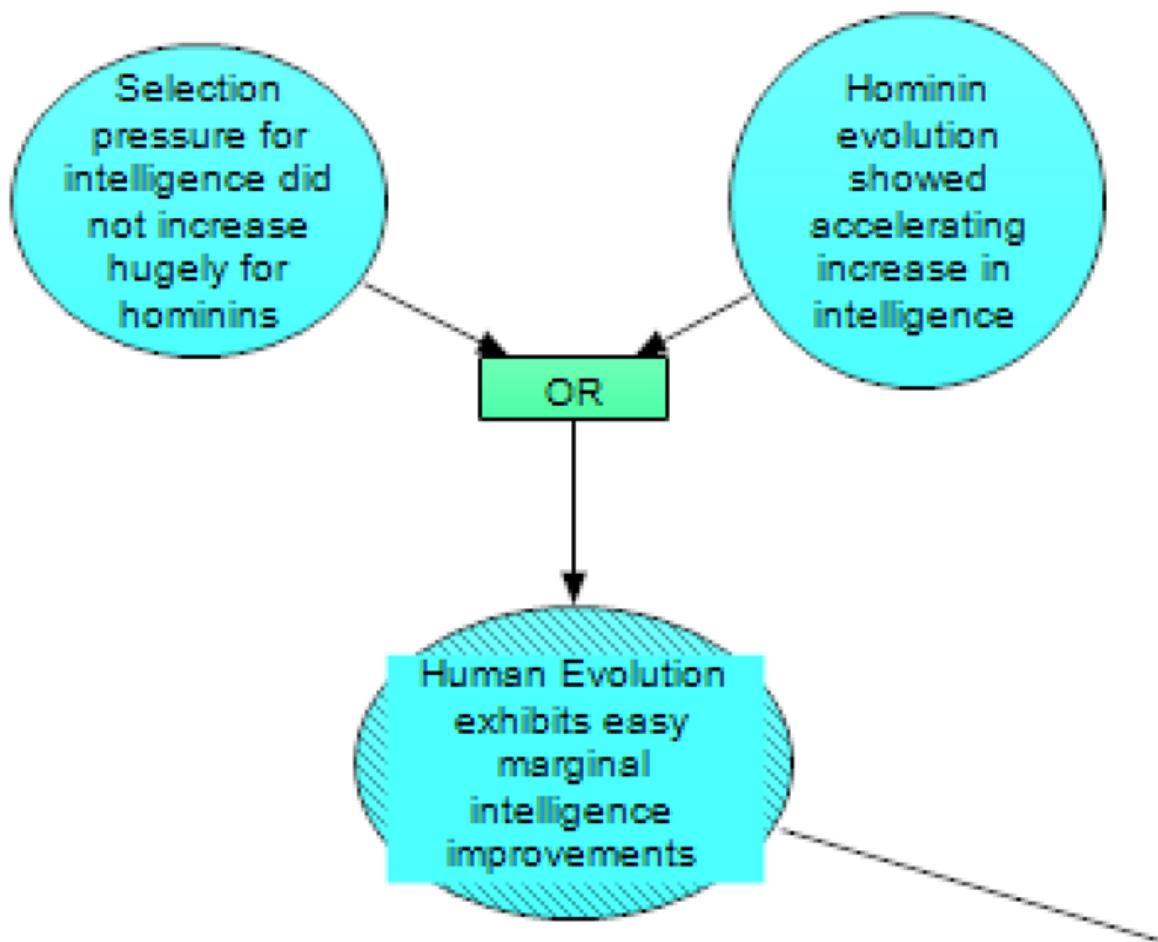


In this diagram of the overall module, the groups of nodes circled in purple are (left to right), [Human Evolution](#), [Machine Learning](#), [Human Culture and History](#), [History of AI](#) and [the Human Brain](#), which feed forward to a classifier that estimates the ease of marginal intelligence improvements. Red shows [General Priors](#), and the yellow nodes are the final model outputs as discussed above. We discuss each of these sections in this order in the rest of this post.

The Cruxes

Each of the argument area submodules, as well as the General Priors submodule, contains cruxes (represented by nodes) that ultimately influence the terminal nodes in this module.

Human Evolution



Human evolution is one of the areas most commonly analogised to HLMI development. '[Intelligence Explosion Microeconomics](#)' argued that human evolution demonstrated accelerating returns on cognitive investment, which suggests marginal intelligence improvements were (at least) not rapidly increasingly difficult during this process.

There are two sources of potential evidence which could support these claims about evolution. The first (**left-hand node in the module above**) is if hominin evolution did not involve hugely more selection pressure for intelligence than did the evolution of other, ancestral primates. It is generally assumed that hominins (all species more closely related to *Homo sapiens* than to chimpanzees – e.g., all species from *Australopithecus* and *Homo*) saw much faster rises in intelligence than what had happened previously to primates. If this unprecedented rise in intelligence was not due to selection pressures for intelligence changing drastically, then that would provide evidence that this fast rise in intelligence did not involve evolution “solving” increasingly more difficult problems than what came before. On the other hand, if selection pressures for intelligence had been marginal (or less) up until this point and were suddenly turned way up, then the fast rise in intelligence could be squared with evolution solving more difficult problems (as the increase in intelligence could then be less than proportional to the increase in selection pressures for intelligence).

Even if selection pressures for intelligence were marginal before hominins, we could still obtain evidence that human evolution exhibited easy marginal intelligence improvements – if we observe a rapid acceleration of intelligence *during* hominin evolution, up to *Homo sapiens* (**right-hand node in the module above**). Such an acceleration of intelligence would be

less likely if intelligence improvements became rapidly more difficult as the human level was approached.

We must note, however, that the relevance of these evolutionary considerations for artificial intelligence progress is still a matter of debate. Language, for instance, was evolved very late along the human lineage, while AI systems have been trained to deal with language from much earlier in their relative development. It is unknown how differences such as this would affect the difficulty-landscape of developing intelligence. The amount to update based on analogies to evolution, however, is not handled by this specific submodule, but instead by the classifier (mentioned above, described below in the section **The Outputs**).

Sources:

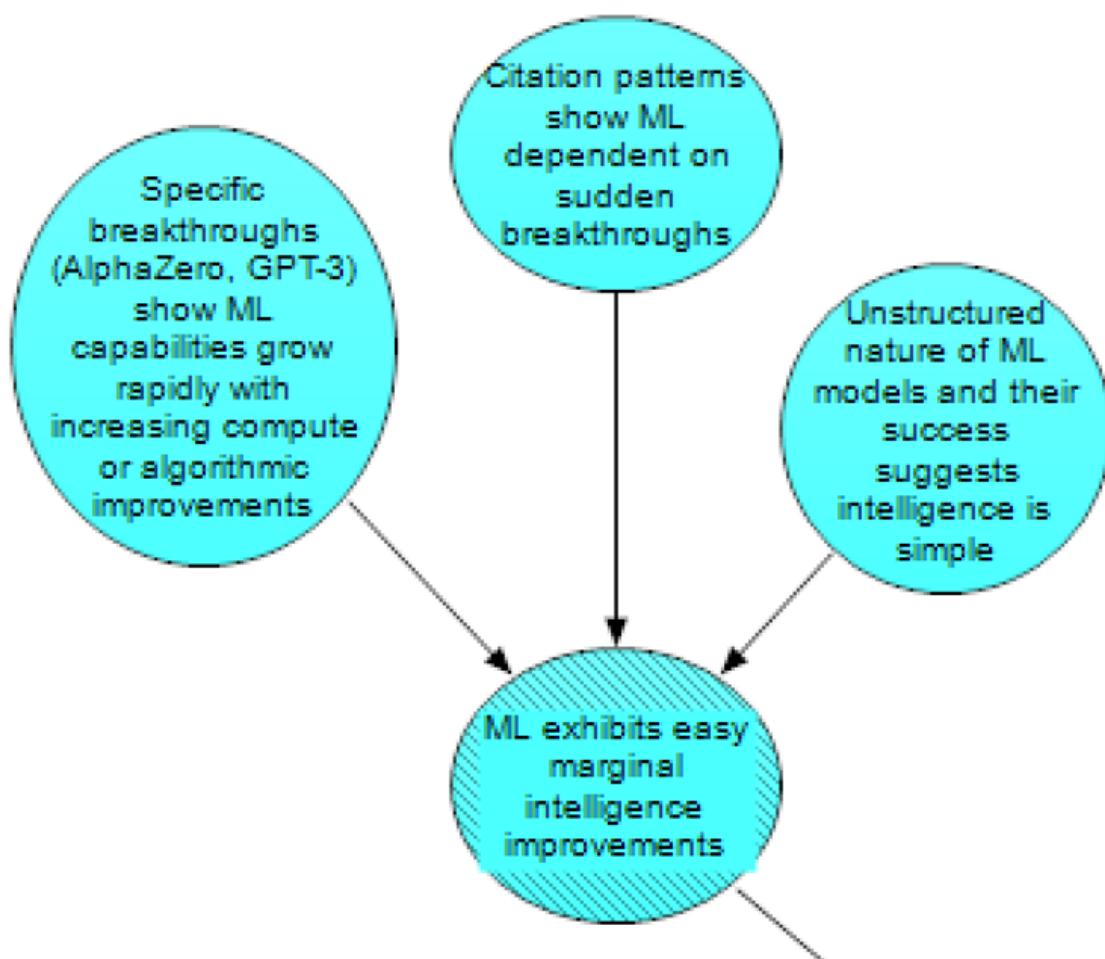
[Likelihood of discontinuous progress around the development of AGI](#)

[Takeoff speeds - The sideways view](#)

[Hanson-Yudkowsky AI Foom Debate](#)

[Thoughts on Takeoff Speeds](#)

Machine Learning



Much of the debate around HLM has focused on what, if any, are the implications of current progress in machine learning for how easy improvements will be at the level of HLM. Some see developments like AlphaGo or GPT-3 as examples of (and evidence for) the claim that marginal intelligence improvements are not increasingly difficult, though others disagree with those conclusions.

For the first of the ways that we could conclude that current ML exhibits easy marginal intelligence improvements, see this [argument](#):



Eliezer Yudkowsky

@ESYudkowsky

...

Given that GPT-3 was not trained on this problem specifically, I claim this case and trend as a substantial victory of my model over [@robinhanson](#)'s model. Robin, do you dispute the direction of the update?



Sharif Shameem @sharifshameem · Jul 17, 2020

I just built a *functioning* React app by describing what I wanted to GPT-3.

I'm still in awe.

[Show this thread](#)

debuild.co

Describe your app.

Clear

Generate

Just describe your app!

Add \$3

Withdraw \$5

My balance is -5

```
// a button that says "Add $3" and  
// a button that says "Withdraw $5".  
then show me my balance  
class App extends React.Component  
{
```

1:05 | 714.5K views
[super\(props\)](#)

11:34 AM · Jul 17, 2020 · Twitter for Android

Which we understand to mean,

GPT-3 is general enough that it can write a functioning app given a short prompt, despite the fact that it is a relatively unstructured transformer model with no explicitly coded representations for app-writing. The fact that GPT-3 is this capable suggests that ML models scale in capability and generality very rapidly with increases in computing power or minor algorithm improvements...

In order to understand whether claims like Yudkowsky's are true, we must understand the specific nature of the breakthroughs made by cutting-edge ML systems like GPT-3 or Alphazero and the limits of what these systems can do (**left node in the image above**).

Claims about current ML systems are related to a broader, more qualitative claim that the general success of ML models indicates that the fundamental algorithms needed for general intelligence are less complex than we might think (**right node in the image above**). Specific examples of 'humanlike' thinking or reasoning in neural networks, for example OpenAI's discovery of [Multimodal Neurons](#), lend some support to this claim.

Alternatively, Robin Hanson [claims](#) that if machine learning was developing in sudden leaps, we would expect to see a pattern of citations in ML research where a few breakthrough papers received a very disproportionately large amount of citations. If Hanson is right about this, and in reality citations aren't distributed in an unusually concentrated pattern in ML compared to other fields, then we have reason to expect marginal intelligence improvements from ML are hard (**middle node in the image above**).

Sources:

[Hanson-Yudkowsky AI Foom Debate](#)

[Searching for Bayes-Structure - LessWrong 2.0 viewer](#)

[Will AI undergo discontinuous progress?](#)

[Conceptual issues in AI safety: the paradigmatic gap](#)

[The Scaling Hypothesis · Gwern.net](#)

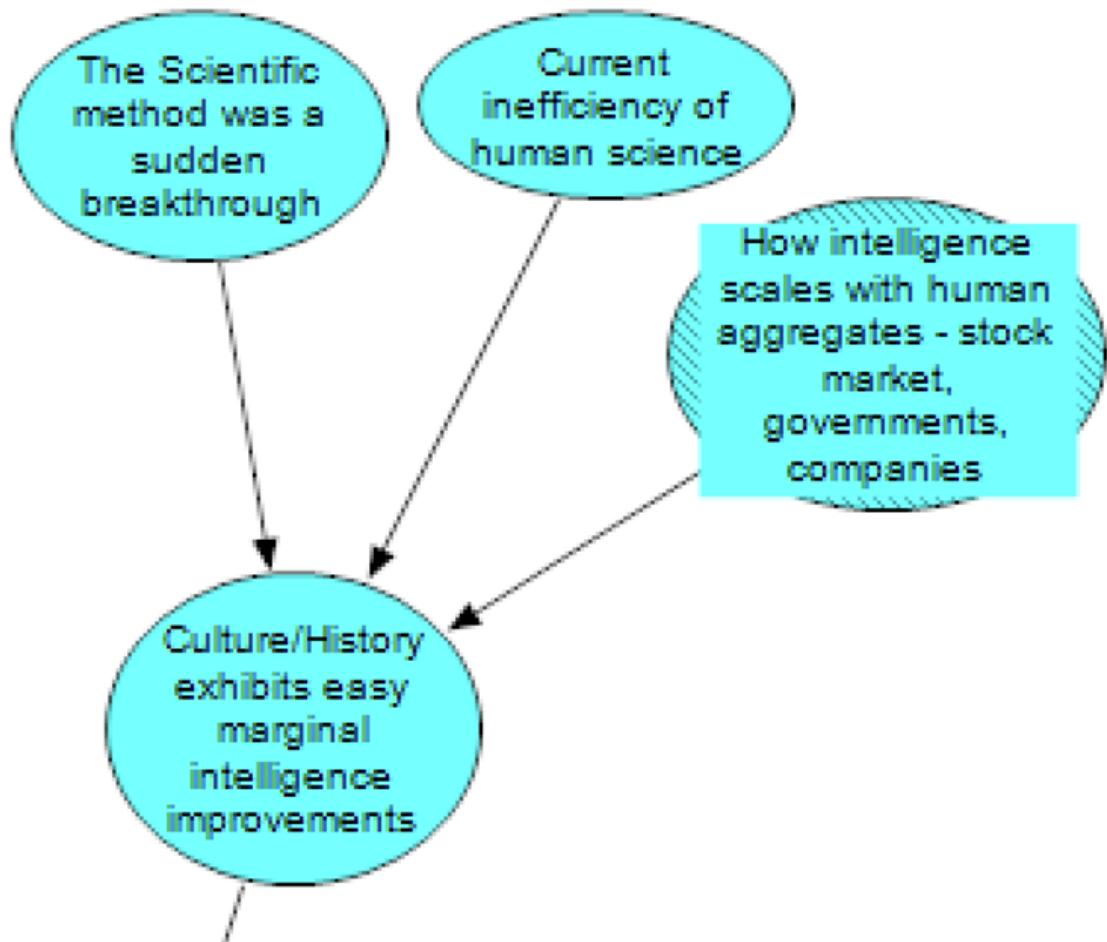
[Eliezer Yudkowsky on AlphaGo's Wins](#)

[GPT-2 As Step Toward General Intelligence](#)

[GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about](#)

[GPT-3: a disappointing paper](#)

Human Culture and History



Another source of evidence is human history and cultural evolution. During the [Hanson-Yudkowsky debate](#), Eliezer Yudkowsky argued that the scientific method is an organisational and methodological insight that suddenly allowed humans to have much greater control over nature, and that this is evidence that marginal intelligence improvements are easy and that AI systems will similarly have breakthroughs where their capabilities suddenly increase (**left-hand node**).

In "[Intelligence Explosion Microeconomics](#)", Eliezer Yudkowsky also identified specific limitations of human science that wouldn't limit AI systems (**middle node**). For example, human researchers need to communicate using slow and imprecise human language. Wei Dai has [similarly argued](#) that AI systems have greater economies of scale than human organizations because they do not hit a [Coasean ceiling](#). We might expect that a lot of human intelligence is 'wasted', as organisations containing humans are not proportionately more intelligent than their members, due to communication limits that won't exist in HLM (right-hand node). If these claims are right, simple organisational insights radically improved humanity's practical abilities, but we still face many organisational limitations an AI would not have to deal with. This suggests marginal improvements in practical abilities could be easy for an AI. On the other hand, even early, relatively unintelligent AIs don't face human limitations such as the inefficiency of interpersonal communication. This means AI might have already "baked in" whatever gains can be achieved from methodological improvements before reaching HLM.

Sources:

[Hanson-Yudkowsky AI Foom Debate](#) (search “you look at human civilization and there's this core trick called science”)

[Debating Yudkowsky](#) (point 5 responds to Eliezer)

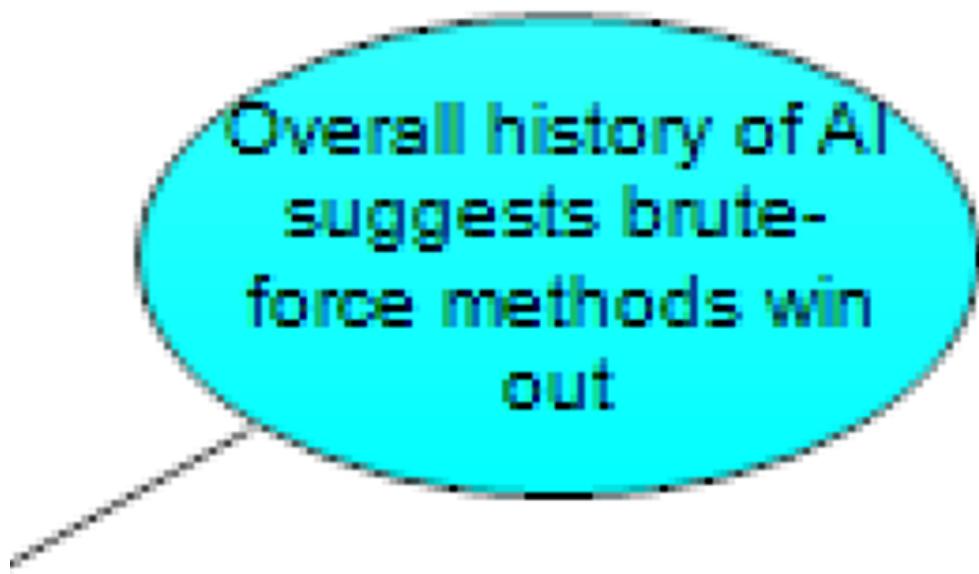
(above links from [Science argument](#))

[Intelligence Explosion Microeconomics](#) (3.5)

[AGI and Economies of Scale](#)

[Continuing the takeoffs debate - LessWrong 2.0 viewer](#)

History of AI



This section covers reference-class forecasting based on the history of AI, going back before the current machine learning paradigm. Principally, the '[bitter lesson](#)' (2019):

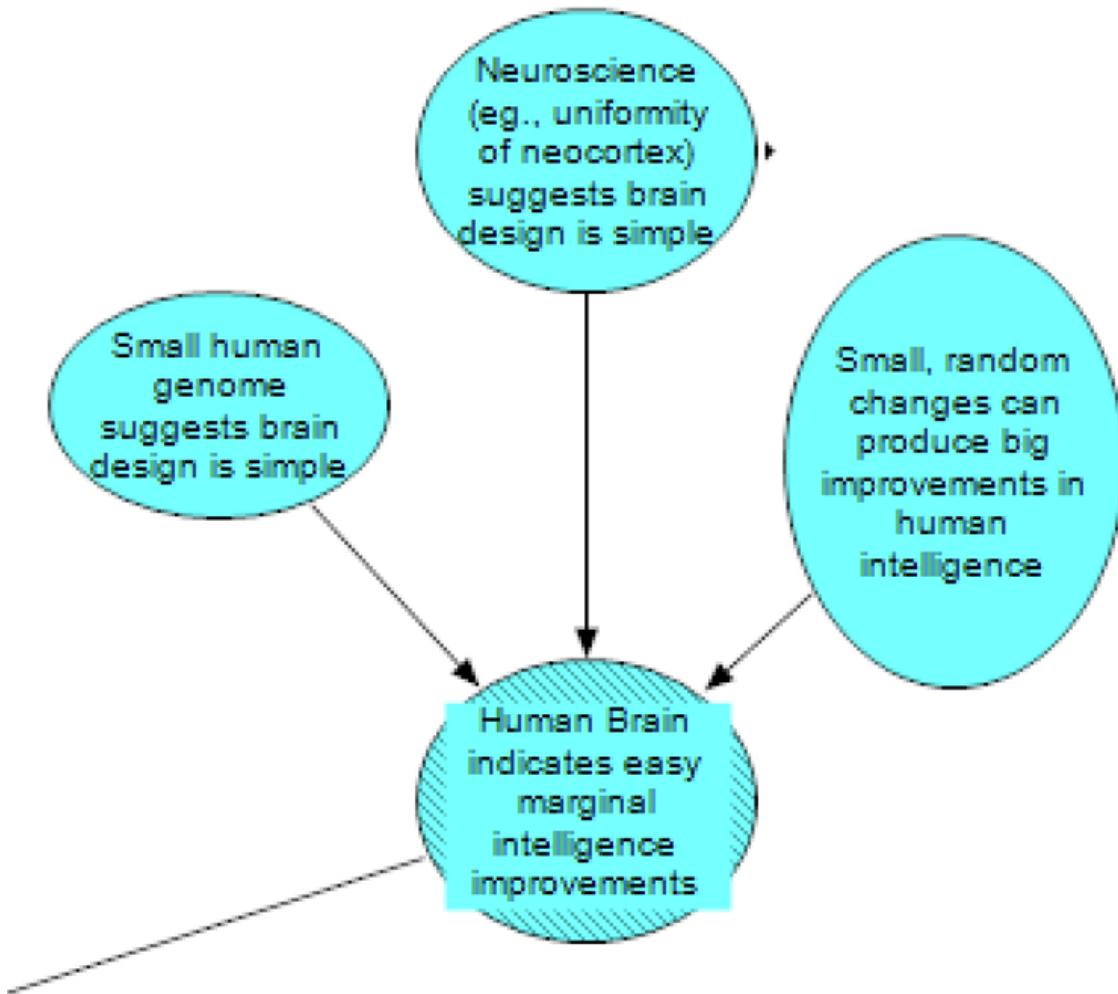
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin... Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation.

The bitter lesson is much broader than machine learning and also includes, for example, the success of deep (relatively brute-force) search methods in computer chess. If the bitter lesson is true in general, it would suggest that we can get significant capability gains from scaling up models. That in turn should lead us to update towards marginal intelligence improvements being easier. The conjecture is that we can continually scale up compute and data to get smooth improvements in search and learning. If this is true, then plausibly scaling up compute and data will also produce smooth increases in general intelligence.

Sources:

[Bitter lesson](#)

The Human Brain



Biological details of the human brain provide another source of intuitions about the difficulty of marginal improvements in intelligence. Eliezer Yudkowsky [has argued that](#) (search for “750 megabytes of DNA” on that page) the small size of the human genome suggests that brain design is simple (this is similar to the [Genome Anchor](#) hypothesis in Ajeya Cotra’s [AI timelines report](#), where the number of parameters in the machine learning model is anchored to the number of bytes in the human genome) (**left-hand node**).

If [the human neocortex is uniform in architecture](#), or if cortical neuron count strongly predicts general intelligence, this also suggests there is a simple “basic algorithm” for intelligence ([possibly analogous to a common algorithm](#) already used in ML). The fact that the neocortex can be divided into different brain regions with different functions, and that the locations of these different regions are conserved across individuals, is evidence against such a simple, uniform algorithm, but on the other hand, the ability of neurons in certain regions to be recruited by other regions (e.g., in [ferrets that have had retinal projections redirected to the auditory thalamus](#), or in [blind humans that can learn to echolocate via mouth clicks and apparently using brain regions typically devoted to vision](#)) is an argument in favour. If the human brain employs a simple algorithm, then we should think it more likely that there are other algorithms that can be rapidly scaled to produce highly intelligent behaviour. These all fall under the evidence from neuroscience node (**middle node**).

Variation in scientific and other intellectual ability among humans (compare Von Neumann to an average human) who share the same basic brain design also suggests improvements are easy at the HLMI level. Similarly, the fact that mood or certain drugs (stimulants and psychedelics) can sometimes improve human cognitive performance implies that humans aren't at a relative maximum, as if we were, simple blunt changes to our cognition should almost always degrade performance (and rare [reports of people gaining cognitive abilities after brain damage](#) provide a potentially even more extreme version of this argument). All of these provide evidence for the claim that small, random changes can produce big improvements in human intelligence (**right-hand node**).

Sources:

[Human genome](#)

[Hanson-Yudkowsky Debate](#)

[Source code size vs learned model size in ML and in humans?](#)

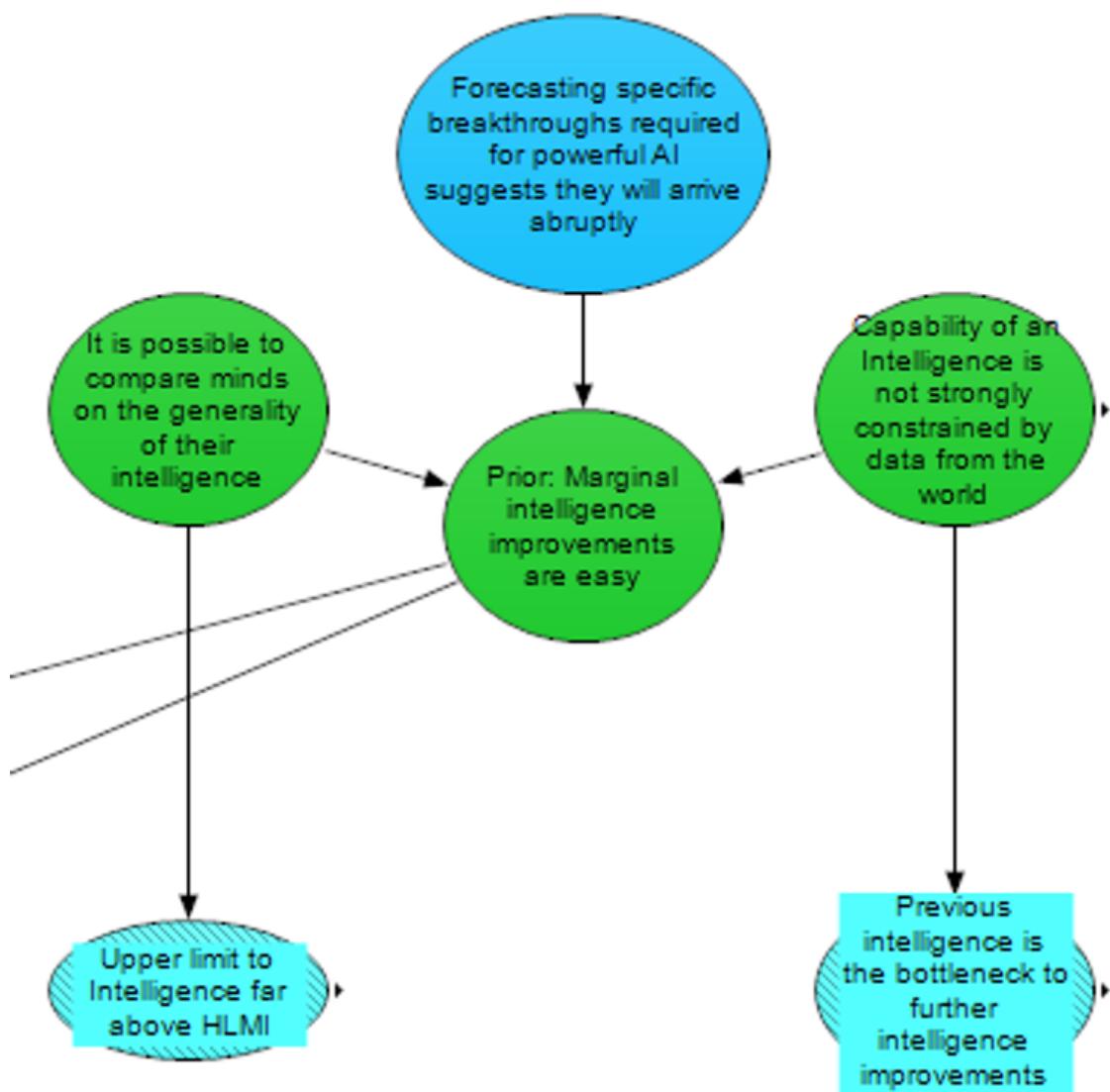
(the above links are from [Missing gear for intelligence](#) and [Secret sauce for intelligence](#))

[Investigation into the relationship between neuron count and intelligence across differing cortical architectures](#)

[Neurons And Intelligence: A Birdbrained Perspective](#)

[Jeff Hawkins on neuromorphic AGI within 20 years](#)

General Priors



One's beliefs about the possibility of an intelligence explosion are likely influenced to a large degree by general priors about the nature of intelligence: whether "[general intelligence](#)" is a coherent concept, whether nonhuman animal species can be compared by level of general intelligence, and so on. Claims like "intelligence isn't one thing that can just be scaled up or down – it consists of a bunch of different modules that are put together in the right way", imply that it is not useful to compare minds on the basis of general intelligence. For instance, [François Chollet](#) denies the possibility of an intelligence explosion in part based on these considerations. As well as affecting the difficulty of marginal intelligence improvements, general priors (including the possibilities that *previous intelligence is a bottleneck to future improvements* and that there exists an *upper limit to intelligence*) also matter because they are potential defeaters of a fast progress/intelligence explosion scenario.

Sources:

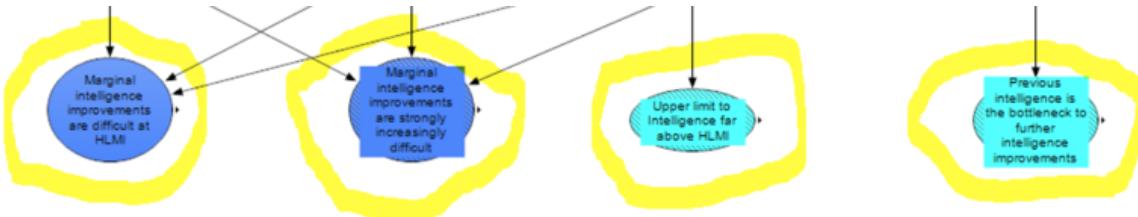
[The implausibility of intelligence explosion](#)

[A reply to Francois Chollet on intelligence explosion](#)

[General intelligence](#)

[General Priors](#)

The Outputs



The empirical cruxes mentioned above influence the outputs of this module, which further influence downstream nodes in other modules. The cruxes this module outputs are:

Marginal Intelligence Improvements are difficult at HLMI, Marginal Intelligence Improvements are Strongly Increasingly Difficult, the Upper Limit to Intelligence is far above HLMI, and Previous intelligence is a bottleneck to future intelligence improvements. These cruxes influence probabilities the model places on different paths to HLMI and takeoff scenarios.

For the two Difficulty of Marginal Intelligence Improvements nodes, we implement a [naive bayes classifier](#) in the Analytica model (i.e., a probabilistic classifier that applies Bayes' theorem with strong independence assumptions between the features). Each claim about one of the domains analogous to HLMI (e.g., in the domain of the human brain, the claim that the neocortex is uniform) is more likely to be true in a world where marginal intelligence improvements *in general* are either easy or hard. Therefore, when taken together, the analogy areas provide evidence about the difficulty of marginal intelligence improvements for HLMI.

This use of a naive Bayes classifier enables us to separate out the prior likelihoods of the original propositions, for example that the human neocortex is uniform, and these propositions' relevance to HLMI (likelihood of being true in a world with easy vs hard marginal intelligence improvements).

This extra step of using Bayes classification is useful because the claims about the analogy domains themselves are often much more certain than their degree of relevance to claims about the ease of improvements to HLMI. Whether there was a rapid acceleration in intelligence during hominin evolution is something that can be assessed by domain experts or a review of the relevant literature, but the relevance of this claim to HLMI is a separate question we are much less certain about. Using the naive Bayes classifier allows us to separate these two factors out.

Difficulty of Marginal Intelligence Improvements

How difficult marginal intelligence improvements are at HLMI and how rapidly they become more difficult at higher levels of intelligence are some of the most significant cruxes of disagreement among those trying to predict the future of AI. If marginal intelligence improvements are difficult around HLMI, then HLMI is unlikely to arrive in a sudden burst. If marginal intelligence improvements rapidly become increasingly difficult beyond HLMI, it is unlikely there will be an intelligence explosion post-HLMI.

For an [example](#) from Yudkowsky of discussion relating to whether marginal intelligence improvements are 'strongly increasingly difficult':

The Open Problem posed here is the quantitative issue: whether it's possible to get sustained returns on reinvesting cognitive improvements into further improving cognition

('Strongly increasingly difficult' has often been operationalized as '[exponentially increasingly difficult](#)', as in that way it serves as a defeater for the 'sustained returns' that we might expect from powerful AI accelerating the development of AI)

[Paul Christiano has also claimed](#) models of progress which include "key insights [that]... fall into place" are implausible, which we model as a claim that marginal intelligence improvements at HLMI are *difficult* (since if they are difficult, a few key insights will not be enough). For a direct example of a claim that difficulty of marginal intelligence improvements is a key crux, see this quote by [Robin Hanson](#):

So I suspect this all comes down to, how powerful is architecture in AI, and how many architectural insights can be found how quickly? If there were say a series of twenty deep powerful insights, each of which made a system twice as effective, just enough extra oomph to let the project and system find the next insight, it would add up to a factor of a million. Which would still be nowhere near enough, so imagine a lot more of them, or lots more powerful.

In our research, we have found that what is most important in many models of AI takeoff is their stance on whether marginal intelligence improvements are difficult at HLMI and if they become much more difficult beyond HLMI. Some models, for instance, claim there is a 'secret sauce' for intelligence such that once it is discovered, progress becomes easy. In the **Discontinuities and Takeoff Speeds** section of the model (presented in a subsequent post in this series), we more closely examine the relationships between claims about ease of marginal intelligence improvements and progress around HLMI.

Why did we attempt to identify an underlying key crux in this way? It is clear that beliefs about AI takeoff relate to beliefs from arguments by analogy (in this post) in fundamental ways, for example:

Paul Christiano et al.: Human Evolution **is not** an example of massive capability gain given constant optimization for intelligence + (other factors) → **[Implicit Belief A about nature of AI Progress]** → **Continuous** Change model of Intelligence Explosion, likely **not** highly localized

Eliezer Yudkowsky et al.: Human Evolution **is** an example of massive capability gain given constant optimization for intelligence + (other factors) → **[Implicit Belief B about nature of AI Progress]** → **Discontinuous** Change model of Intelligence Explosion, likely highly localized

We have treated implicit beliefs A and B as being about the difficulty of marginal intelligence improvements at HLMI and beyond.

We have identified 'Previous intelligence is a bottleneck' and 'There is an upper limit to intelligence' as two other important cruxes. They are commonly cited as defeaters for scenarios that involve any kind of explosive growth due to an acceleration in AI progress. Both of these appear to be cruxes between sceptics and non-sceptics of HLMI and AI takeoff.

Previous Intelligence is a bottleneck

The third major output of this module is whether, in general, further improvements in intelligence tend to be bottlenecked by the current intelligence of our systems rather than some external factor (such as the need to run experiments and wait for real-world data).

This output is later used in the intelligence explosion module (covered in a subsequent post in this series): if such an external bottleneck exists, we are unlikely to see rapid acceleration

of technological progress through powerful AI improving our ability to build yet more powerful AI. There will instead be drag factors preventing successor AIs from being produced without reference to the outside world. This view is summarised by [François Chollet](#):

If intelligence is fundamentally linked to specific sensorimotor modalities, a specific environment, a specific upbringing, and a specific problem to solve, *then you cannot hope to arbitrarily increase the intelligence of an agent merely by tuning its brain — no more than you can increase the throughput of a factory line by speeding up the conveyor belt.* Intelligence expansion can only come from a co-evolution of the mind, its sensorimotor modalities, and its environment.

In short, this position claims that positive feedback loops between successive generations of HLMs could not simply be closed, but instead would require feedback from environmental interaction that can't be arbitrarily sped up. While Chollet appears to be assuming such bottlenecks will occur due to necessary interactions with the physical world, it is possible in principle that such bottlenecks could occur in the digital world – for instance, brain emulations that were sped up by 1Mx would find all computers (and thus communications, access to information, calculators, simulations, etc) slowed down by 1Mx (from their perspective), potentially creating a drag on progress.

A somewhat more tentative version of this claim is that improvements in intelligence (construed here as ‘ability to do applied science’) require a very diverse range of discrete skills and access to real-world resources. [Ben Garfinkel](#) makes this point:

...there's really a lot of different tasks that feed into making progress in engineering or areas of applied science. There's not really just this one task of “Do science”. Let's take, for example, the production of computing hardware... I assume that there's a huge amount of different works in terms of how you're designing these factories or building them, how you're making decisions about the design of the chips. Lots of things about where you're getting your resources from. Actually physically building the factories.

Upper Limit to Intelligence

Finally, this module has an output for whether there is a practical upper limit to intelligence not significantly above the human level. This could be true if there are physical barriers to the development of greater intelligence. Alternatively, it seems more likely to be effectively true if we cannot even compare the abilities of different minds along a general metric of “intelligence” (hence a “no” answer to “it is possible to compare minds on the generality of their intelligence” leads to a “yes” to this crux). For the “not possible to compare minds on the generality of their intelligence” claim, from [François Chollet](#):

The first issue I see with the intelligence explosion theory is a failure to recognize that intelligence is necessarily part of a broader system—a vision of intelligence as a “brain in jar” that can be made arbitrarily intelligent independently of its situation.

Chollet argues that human (and all animal) intelligence is ‘hyper-specialised’ and situational to a degree that makes comparisons of general intelligence much less useful than they first appear,

If intelligence is a problem-solving algorithm, then it can only be understood with respect to a *specific* problem. In a more concrete way, we can observe this empirically in that all intelligent systems we know are highly specialized ... The brain has hardcoded conceptions of having a body with hands that can grab, a mouth that can suck, eyes mounted on a moving head that can be used to visually follow objects (the [vestibulo-ocular reflex](#)), and these preconceptions are required for human intelligence to start taking control of the human body. [It has even been convincingly argued, for instance by](#)

[Chomsky](#), that very high-level human cognitive features, such as our ability to develop language, are innate.

A strong version of the modularity of mind hypothesis, ‘massive modularity’, also implies that intelligence is [hyper-specialised at extremely specific tasks](#). If true, massive modularity would lend support to the claim that ‘intelligence... can only be understood with respect to a specific problem’, which in turn suggests that intelligence cannot be increased independent of its situation.

Conclusion

This post has explained the structure and reasoning behind one of the starting points of the MTAIR model - Analogies and General Priors. This module connects conclusions about the nature of HLMI to very basic assumptions about the nature of intelligence, and analogies to domains other than HLMI about which we have greater experience.

We have made the simplifying assumption to group the conclusions drawn from the general priors and the different analogy domains into four outputs, which we think characterise the important variables needed to predict HLMI development and post-HLMI takeoff. In later posts, we will explain how those outputs are used to make predictions about HLMI takeoff and development.

The next post in the series will be **Paths to High-Level Machine Intelligence**, which attempts to forecast when HLMI will be developed and by what route.

We are interested in any feedback you might have, particularly if there are any views or arguments which you feel our model does not currently capture, or captures incorrectly.

Footnotes

1. We define HLMI as machines that are capable of performing almost all economically-relevant information-processing tasks (either individually or collectively). We are using the term “high-level machine intelligence” here instead of the related terms “human-level machine intelligence”, “artificial general intelligence”, or “transformative AI”, since these other terms are often seen as baking in assumptions about either the nature of intelligence or advanced AI that are not universally accepted.

Garrabrant and Shah on human modeling in AGI

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

This is an edited transcript of a conversation between Scott Garrabrant (MIRI) and Rohin Shah (DeepMind) about whether researchers should focus more on approaches to AI alignment that don't require highly capable AI systems to do much human modeling. CFAR's Eli Tyre facilitated the conversation.

To recap, and define some terms:

- The [alignment problem](#) is the problem of figuring out "how to develop sufficiently advanced machine intelligences such that running them produces good outcomes in the real world" ([outcome alignment](#)) or the problem of building powerful AI systems that are trying to do what their operators want them to do ([intent alignment](#)).
- In 2016, Hadfield-Mennell, Dragan, Abbeel, and Russell proposed that we think of the alignment problem in terms of "[Cooperative Inverse Reinforcement Learning](#)" (CIRL), a framework where the AI system is initially uncertain of its reward function, and interacts over time with a human (who knows the reward function) in order to learn it.
- In 2016-2017, Christiano proposed "[Iterated Distillation and Amplification](#)" (IDA), an approach to alignment that involves [iteratively training AI systems](#) to learn from human experts assisted by AI helpers. In 2018, Irving, Christiano, and Amodei proposed [AI safety via debate](#), an approach based on similar principles.
- In early 2019, Scott Garrabrant and DeepMind's Ramana Kumar argued in "[Thoughts on Human Models](#)" that we should be "cautious about AGI designs that use human models" and should "put more effort into developing approaches that work well in the absence of human models".
- In early February 2021, Scott and Rohin [talked more about human modeling](#) and decided to have the real-time conversation below.

You can find a recording of the Feb. 28 discussion below (sans Q&A) [here](#).

1. IDA, CIRL, and incentives

Eli: I guess I want to first check what our goal is here. There was some stuff that happened online. Where are we according to you guys?

Scott: I think Rohin spoke last and I think I have to reload everything. I'm coming at it fresh right now.

Eli: Yeah. So we'll start by going from "What are the positions that each of you were thinking about?" But I guess I'm most curious about: from each of your perspectives, what would a win of this conversation be like? What could happen that you would go away and be like, "Yes, this was a success."

Rohin: I don't know. I generally feel like I want to have more models of what AI safety should look like, and that's usually what I'm trying to get out of conversations with other people. I also don't feel like we have even narrowed down on what the disagreement is yet. So, first goal would be to figure that out.

Eli: Definitely. Scott?

Scott: Yeah. So, I have two things. One is that I feel frustrated with the relationship between the AI safety community and the human models question. I mean, I feel like I haven't done a good job arguing about it or something. And so on one side, I'm trying to learn what the other position is because I'm stuck there.

The other thing is that I feel especially interested in this conversation now because I feel like some of my views on this have recently shifted.

My worry about human-modeling systems has been that modeling humans is in some sense "close" to behaviors like deception. I still see things that way, but now I have a clearer idea of what the relevant kind of "closeness" is: human-modeling is close to unacceptable behavior on some measure of closeness that involves your ability to observe the system's internals and figure out whether the system is doing the intended thing or doing something importantly different.

I feel like this was an update to my view on the thing relative to how I thought about it a couple of years ago, and I want similar updates.

Eli: Awesome. As I'm thinking about it, my role is to try and make this conversation go well, according to your goals.

I want to encourage both of you to take full responsibility for making the conversation go well as well. Don't have the toxic respect of, "Ah, someone else is taking care of it." If something seems interesting to you, definitely say so. If something seems boring to you, say so. If there's a tack that seems particularly promising, definitely go for that. If I recommend something you think is stupid, you should say, "That's a stupid thing." I will not be offended. That is, in fact, helpful for me. Sound good?

All right. Scott, do you want to start with your broad claim?

Scott: Yeah. So, I claim that it is plausible that the problem of aligning a super capable system that is working in some domain like physics is significantly easier than aligning a super capable system that's working in contexts that involve modeling humans. Either contexts involving modeling humans because they're working on inherently social tasks, or contexts involving modeling humans because part of our safety method involves modeling humans.

And so I claim with moderate probability that it's significantly easier to align systems that don't do any human-modeling. I also claim, with small but non-negligible probability, that one can save the world using systems that don't model humans. Which is not the direction I want the conversation to really go, because I think that's not really a crux, because I don't expect my probability on "non-human-modeling systems can be used to save the world" to get so low that I conclude researchers shouldn't think a lot about how to align and use such systems.

And I also have this observation: most AI safety plans feel either adjacent to a Paul Christiano IDA-type paradigm or adjacent to a Stuart Russell CIRL-type paradigm.

I think that two of the main places that AI safety has pushed the field have been towards human models, either because of IDA/debate-type reasons or because we want to not specify the goal, we want it to be in the human, we want the AI to be trying to do "what we want" in quotation marks, which requires a pointer to modeling us.

And I think that plausibly, these avenues are mistakes. And all of our attention is going towards them.

Rohin: Well, I can just respond to that maybe.

So I think I am with you on the claim that there are tasks which incentivize more human-modeling. And I wouldn't argue this with confidence or anything like that, but I'm also with you that plausibly, we should not build AI systems that do those tasks because it's close to manipulation of humans.

This seems pretty distinct from the algorithms that we use to build the AI systems. You can use iterated amplification to build a super capable physics AI system that doesn't know that much about humans and isn't trying to manipulate them.

Scott: Yeah. So you're saying that the IDA paradigm is sufficiently divorced from the dangerous parts of human modeling. And you're not currently making the same claim about the CIRL paradigm?

Rohin: It depends on how broadly you interpret the CIRL paradigm. The version where you need to use a CIRL setup to infer all of human values or something, and then deploy that agent to optimize the universe—I'm not making the claim for that paradigm. I don't like that paradigm. To my knowledge, I don't think Stuart likes that paradigm. So, yeah, I'm not making claims about that paradigm.

I think in the world where we use CIRL-style systems to do fairly bounded, narrow tasks—maybe narrow is not the word I want to use here—but you use that sort of system for a specific task (which, again, might only involve reasoning about physics), I would probably make the same claim there, yes. I have thought less about it.

I might also say I'm not claiming that there is literally no incentive to model humans in any of these cases. My claim is more like, "No matter what you do, if you take a very broad definition of "incentive," there will always be an incentive to model humans because you need some information about human preferences for the AI system to do a thing that you like."

Scott: I think that it's not for me about an incentive to model humans and more just about, is there modeling of humans at all?

Well, no, so I'm confused again, because... Wait, you're saying there's not an incentive to model humans in IDA?

Rohin: It depends on the meaning of the word "incentive."

Scott: Yeah. I feel like the word "incentive" is distracting or something. I think the point is that there is modeling of humans.

(reconsidering) There might not be modeling of humans in IDA. There's a sense in which, in doing it right, maybe there's not... I don't know. I don't think that's where the disagreement is, though, is it?

Eli: Can we back up for a second and check what's the deal with modeling humans? Presumably there's something at least potentially bad about that. There's a reason why we care about it. Yes?

Rohin: Yeah. I think the reason that I'm going with, which I feel fairly confident that Scott agrees with me on, at least as the main reason, is that if your AI system is modeling humans, then it is “easy” or “close” for it to be manipulating humans and doing something that we don't want that we can't actually detect ahead of time, and that therefore causes bad outcomes.

Scott: Yeah. I want to be especially clear about the metric of closeness being about our ability to oversee/pre-oversee a system in the way that we set up the incentives or something.

The closeness between modeling humans and manipulating humans is not in the probability that some system that's doing one spontaneously changes to doing the other. (Even though I think that they *are* close in that metric.) It's more in the ability to be able to distinguish between the two behaviors.

And I think that there's a sense in which my model is very pessimistic about oversight, such that maybe if we really try for the next 20 years or something, we can distinguish between “model thinking superintelligent thoughts about physics” and “model thinking about humans.” And we have no hope of actually being able to distinguish between “model that's trying to manipulate the human” and “model that's trying to do IDA-type stuff (or whatever) the legitimate way.”

Rohin: Right. And I'm definitely a lot more optimistic about oversight than you, but I still agree directionally that it's harder to oversee a model when you're trying to get it to do things that are very close to manipulation. So this feels not that cruxy or something.

Scott: Yeah, not that cruxy... I don't know, I feel like I want to hear more about what Rohin thinks or something instead of responding to him. Yeah, not that that cruxy, but what's the cruxy part, then? I feel like I'm already granting the thing about incentives and I'm not even talking about whether you have incentives to model humans. I'm assuming that there's systems that model humans, there's systems that don't model humans. And it's a *lot* easier to oversee the ones that don't.

2. Mutual information with humans

Rohin: I think my main claim is: the determining factor about whether a system is modeling humans or not—or, let's say there is an *amount* of modeling the system does. I want to talk about spectrums because I feel like you say too many wrong things if you think in binary terms in this particular case.

Scott: All right.

Rohin: So, there's an amount of modeling humans. Let's call it a scale—

Scott: We can have an entire *one* dimension instead of zero! (*laughs*)

Rohin: Yes, exactly. The art of choosing the right number of dimensions. (*laughs*) It's an important art.

So, we'll have it on a scale from 0 to 10. I think my main claim is that the primary determinant of where you are on the spectrum is what you are trying to get your AI system to do, and not the source of the feedback by which you train the system. And IDA is the latter, not the former.

Scott: Yeah, so I think that this is why I was distinguishing between the two kinds of "closeness." I think that the probability of spontaneously manipulating humans is stronger when there are humans in the task than when there are just humans in the IDA/CIRL way of pointing at the task or something like that. But I think that the distance in terms of ability to oversee is not large...

Rohin: Yeah, I think I am curious why.

Scott: Yeah. Do I think that? (pauses)

Hm. Yeah. I might be responding to a fake claim right now, I'm not sure. But in IDA, you aren't keeping some sort of structure of [HCH](#). You're not *trying* to do this, because they're trying to do oversight, but you're distilling systems and you're allowing them to find what gets the right answer. And then you have these structures that get the right answer on questions about what humans say, and *maybe* rich enough questions that contain a lot of needing to understand what's going on with the human. (I claim that yes, many domains are rich enough to require modeling the human; but maybe that's false, I don't know.)

And basically I'm imagining a black box. Not entirely black box—it's a gray box, and we can look at some features of it—but it just has mutual information with a bunch of stuff that humans do. And I almost feel like, I don't think this is the way that transparency will actually work, but I think just the question “is there mutual information between the humans and the models?” is the extent to which I expect to be able to do the transparency or something.

Maybe I'm claiming that in IDA, there's *not* going to be mutual information with complex models of humans.

Rohin: Yep. That seems right. Or more accurately, I would say it depends on the task that you're asking IDA to do primarily.

Scott: Yeah, well, it depends on the task and it depends on the IDA. It depends on what instructions you give to the humans.

Rohin: Yes.

Scott: If you imagined that there's some core of how to make decisions or something, and humans have access to this core, and this core is not specific to humans, but humans don't have access to it in such a way that they can write it down in code, then you would imagine a world in which I would be incentivized to do IDA and I would be able to do so safely (kind of) because I'm not actually putting mutual information with humans in the system. I'm just asking the humans to follow their “how to make decisions” gut that doesn't actually relate to the humans.

Rohin: Yes. That seems right.

What do I think about that? Do I actually think humans have a core?...

Scott: I almost wasn't even trying to make that claim. I was just trying to say that there exists a plausible world where this might be the case. I'm uncertain about whether humans have a core like that.

But I do think that in IDA, you're doing more than just asking the humans to use their "core." Even if humans had a core of how to solve differential equations that they couldn't write down in code and then wanted to use IDA to solve differential equations via only asking humans to use their core of differential equations.

If this were the case... Yeah, I think the IDA is asking more of the human than that. Because regardless of the task, the IDA is asking the human to also do some oversight.

Rohin: ... Yes.

Scott: And I think that the "oversight" part is capturing the richness of being human even if the task manages to dodge that.

And the analog of oversight in debates is the debates. I think it's more clear in debates because it's debate, but I think there's the oversight part and that's going to transfer over.

Rohin: That seems true. One way I might rephrase that point is that if you have a results-based training system, if your feedback to the agent is based on results, the results can be pretty independent of humans. But if you have a feedback system based not just on results, but also on the process by which you got the results—for whatever reason, it just happens to be an empirical fact about the world that there's no nice, correct, human-independent core of how to provide feedback on process—then it will necessarily contain a bunch of information about humans the agent will then pick up on.

Scott: Yeah.

Yeah, I think I believe this claim. I'm not sure. I think that you said back a claim that not only is what I said, but also I temporarily endorse, which is stronger. (*laughs*)

Rohin: Cool. (*laughs*) Yeah. It does seem plausible. It seems really rough to not be giving feedback on the process, too.

I will note that it is totally possible to do IDA, debate, and CIRL without process-level feedback. You just tell your humans to only write the results, or you just replace the human with an automated reward function that only evaluates the results if you can do that.

Scott: I mean...

Rohin: I agree, that's sort of losing the point of those systems in the first place. Well, maybe not CIRL, but at least IDA and debate.

Scott: Yeah. I feel like I can imagine "Ah, do something like IDA without the process-level feedback." I wouldn't even want to call debate "debate" without the process-level feedback.

Rohin: Yeah. At that point it's like two-player zero-sum optimization. It's like AlphaZero or something.

Scott: Yeah. I conjecture that the various people who are very excited about each of these paradigms would be unexcited about the version that does not have process-level feedback—

Rohin: Yes. I certainly agree with that.

I also would be pretty unexcited about them without the process-level feedback. Yeah, that makes sense. I think I would provisionally buy that for at least physics-style AI systems.

Scott: What do you mean?

Rohin: If your task is like, "Do good physics," or something. I agree that the process-level feedback will, like, 10x the amount of human information you have. (I made up the number 10.)

Whereas if it was something else, like if it was sales or marketing, I'd be like, "Yeah, the process-level feedback makes effectively no difference. It increases the information by, like, 1%."

Scott: Yeah. I think it makes very little difference in the mutual information. I think that it still feels like it makes some difference in some notion of closeness-to-manipulation to me.

So I'm in a position where if it's the case that one could do superhuman STEM work without humans, I want to know this fact. Even if I don't know what to do with it, that feels like a question that I want to know the answer to, because it seems plausibly worthwhile.

Rohin: Well, I feel like the answer is almost certainly yes, right? AlphaFold is an example of it.

Scott: No, I think that one could make the claim... All right, fine, I'll propose an example then. One could make the claim that IDA is a capability enhancement. IDA is like, "Let's take the humans' information about how to break down and solve problems and get it into the AI via the IDA process."

Rohin: Yep. I agree you could think of it that way as well.

Scott: And so one could imagine being able to answer physics questions via IDA and not knowing how to make an AI system that is capable of answering the same physics questions without IDA.

Rohin: Oh. Sure. That seems true.

Scott: So at least in principle, it seems like...

Rohin: Yeah.

Eli: There's a further claim I hear you making, Scott—and correct me if this is wrong—which is, "We want to explore the possibility of solving physics problems without something like IDA, because that version of how to solve physics problems may be safer."

Scott: Right. Yeah, I have some sort of conjunction of "maybe we can make an AGI system that just solves physics problems" and also "maybe it's safer to do so." And

also “maybe AI that could only solve physics problems is sufficient to save the world.” And even though I conjuncted three things there, it’s still probable enough to deserve a lot of attention.

Rohin: Yeah. I agree. I would probably bet against “we can train an AI system to do good STEM reasoning in general.” I’m certainly on board that we can do it some of the time. As I’ve mentioned, AlphaFold is an example of that. But it does feel like, yeah, it just seems really rough to have to go through an evolution-like process or something to learn general reasoning rather than just learning it from humans. Seems so much easier to do the latter that that’s probably going to be the best approach in most cases.

Scott: Yeah. I guess I feel like some sort of automated working with human feedback —I feel like we can be inspired by humans when trying to figure out how to figure out some stuff about how to make decisions or something. And I’m not too worried about mutual information with humans leaking in from the fact that we were inspired by humans. We can use the fact that we know some facts about how to do decision-making or something, to not have to just do a big evolution.

Rohin: Yeah. I think I agree that you could definitely do a bit better that way. What am I trying to say here? I think I’m like, “For every additional bit of good decision-making you specify, you get correspondingly better speed-ups or something.” And IDA is sort of the extreme case where you get lots and lots of bits from the humans.

Scott: Oh! Yeah... I don't think that the way that humans make decisions is that much more useful as a thing to draw inspiration from than, like, how humans think ideal decision-making should be.

Rohin: Sure. Yeah, that's fair.

Scott: Yeah. It's not obvious to me that you have that much to learn from humans relative to the other things in your toolbox.

Rohin: That's fair. I think I do disagree, but I wouldn't bet confidently one way or the other.

3. The default trajectory

Eli: I guess I want to back up a little bit and just take stock of where we are in this thread of the conversation. Scott made a conjunctive claim of three things. One is that it's possible to train an AI system to do STEM work without needing to have human models in the mix. Two, this might be sufficient for saving the world or otherwise doing pretty powerful stuff. And three, this might be substantially safer than doing it IDA-style or similar. I guess I just want to check, Rohin. Do you agree or disagree with each of those points?

Rohin: I probably disagree *somewhat* with all of them, but the upstream disagreement is how optimistic versus pessimistic Scott and I are about AI safety in general. If you're optimistic the way I am, then you're much more into trying not to change the trajectory of AI too much, and instead make the existing trajectory better.

Scott: Okay. (*laughs*) I don't know. I feel like I want to say this, because it's ironic or something. I feel like the existing framework of AI before AI safety got involved was

"just try to solve problems," and then AI safety's like, "You know what we need in our 'trying to solve problems'? We need a lot of really meta human analysis." (*laughs*)

It does feel to me like the default path if nobody had ever thought about AI safety looks closer to the thing that I'm advocating for in this discussion.

Rohin: I do disagree with that.

Scott: Yeah. For the default thing, you do need to be able to say, "Hey, let's do some transparency and let's watch it, and let's make sure it's not doing the human-modeling stuff."

Rohin: I also think people just would have started using human feedback.

Scott: Okay. Yeah, that might be true. Yeah.

Rohin: It's just such an easy way to do things, relative to having to write down the reward function.

Scott: Yeah, I think you're right about that.

Rohin: Yeah. But I think part of it is that Scott's like, "Man, all the things that are not this sort of three-conjunct approach are pretty doomed, therefore it's worth taking a plan that might be unlikely to work because that's the only thing that can actually make a substantial difference." Whereas I'm like, "Man, this plan is unlikely to work. Let's go with a likely one that can cut the risk in half."

Scott: Yeah. I mean, I basically don't think I have any kind of plan that I think is likely to work... But I'm also not in the business of making plans. I want other people to do that. (*laughs*)

Rohin: I should maybe respond to Eli's original question. I think I am more pessimistic than Scott on each of the three claims, but not by much.

Eli: But the main difference is you're like, "But it seems like there's this alternative path which seems like it has a pretty good shot."

Rohin: I think Scott and I would both agree that the no-human-models STEM AI approach is unlikely to work out. And I'm like, "There's this *other* path! It's *likely* to work out! Let's do that." And Scott's like, "This is the only thing that has *any* chance of working out. Let's do this."

Eli: Yeah. Can you tell if it's a crux, whether or not your optimism about the no-human-models path would affect how optimistic you feel about the human-models path? It's a very broad question, so it may be kind of unfair.

Rohin: It would seriously depend on why I became less optimistic about what I'm calling the usual path, the IDA path or something.

There are just a *ton* of beliefs that are correlated that matter. Again, I'm going to state things more confidently than I believe them for the sake of faster and clearer communication. There's a ton of beliefs like, "Oh man, good reasoning is just sort of necessarily messy and you're not going to get nice, good principles, and so you should be more in the business of incrementally improving safety rather than finding the one way to make a nice, sharp, clear distinction between things."

Then other parts are like, "It sure seems like you can't make a sharp distinction between 'good reasoning' and 'what humans want.'" And this is another reason I'm more optimistic about the first path.

So I could imagine that I get unconvinced of many of these points, and definitely some of them, if I got unconvinced of those, I would be more optimistic about the path Scott's outlining.

4. Two kinds of risk

Scott: Yeah. I want to clarify that I'm not saying, like, "Solve decision theory and then put the decision theory into AI," or something like that, where you're putting the "how to do good reasoning" in directly. I think that I'm imagining that you have to have some sort of system that's learning how to do reasoning.

Rohin: Yep.

Scott: And I'm basically distinguishing between a system that's learning how to do reasoning while being overseen and kept out of the convex hull of human modeling versus... And there are definitely trade-offs here, because you have more of a [daemon](#) problem or something if you're like, "I'm going to learn how to do reasoning," as opposed to, "I'm going to be told how to do reasoning from the humans." And so then you have to search over this richer space or something of how to do reasoning, which makes it harder.

I'm largely not thinking about the capability cost there. There's a safety cost associated with running a big evolution to discover how reasoning works, relative to having the human tell you how reasoning works. But there's also a safety cost to having the human tell you how reasoning works.

And it's a different kind of safety cost in the two cases. And I'm not even sure that I believe this, but I think that I believe that the safety cost associated with learning how to do reasoning in a STEM domain might be lower than the safety cost associated with having your reasoning directly point to humans via the path of being able to have a thick line between your good and bad behavior.

Rohin: Yeah.

Scott: And I can imagine you saying, like, "Well, but that doesn't matter because you can't just say, 'Okay, now we're going to run the evolution on trying to figure out how to solve STEM problems,' because you need to actually have the capabilities or something."

Rohin: Oh. I think I lost track of what you were saying at the last sentence. Which probably means I failed to understand the previous sentences too.

Eli: I heard Scott to be saying—obviously correct me if I've also misapprehended you, Scott—that there's at least two axes that you can compare things along. One is "What's the safety tax? How much AI risk do you take on for this particular approach?" And this other axis is, "How hard is it to make this approach work?" Which is a capabilities question. And Scott is saying, "I, Scott, am only thinking about the safety tax question, like which of these is—"

Scott: "Safety tax" is a technical term that you're using wrong. But yeah.

Rohin: If you call it "safety difficulty"...

Scott: Or "risk factor."

Eli: Okay. Thank you. "I'm only thinking about the risk factor, and not thinking about how much extra difficulty with making it work comes from this approach." And I imagine, Rohin, that you're like, "Well, even—"

Rohin: Oh, I just agree with that. I'm happy making that decomposition for now. Scott then made a further claim about the delta between the risk factor from the IDA approach and the STEM AI approach.

Scott: Well, first I made the further claim that the risk from the IDA approach and the risk from the STEM approach are kind of different in kind. It's not like you can just directly compare them; they're different kinds of risks.

And so because of that, I'm uncertain about this: But then I further made the claim, at least in this current context, that the IDA approach might have more risk.

Rohin: Yeah. And what were the two risks? I think that's the part where I got confused.

Scott: Oh. One of the risks is we have to run an evolution because we're not learning from the humans, and so the problem's harder. And that introduces risks because it's harder for us to understand how it's working, because it's working in an alien way, because we had to do an evolution or something.

Rohin: So, roughly, we didn't give it process-level feedback, and so its process could be wild?

Scott: ... Yeahh... I mean, I'm imagining... Yeah. Right.

Rohin: Cool. All right. I understand that one. Well, maybe not...

Scott: Yeah. I think that if you phrase it as "We didn't give it any process-level feedback," I'm like, "Yeah, maybe this one is obviously the larger risk." I don't know.

But yeah, the risk associated with "You have to do an evolution on your system that's thinking about STEM." Yeah, it's something like, "We had to do an evolution."

And the risk of the other one is that we made it such that we can't oversee it in the required ways. We can't run the policy, watch it carefully, and ensure it doesn't reason about humans.

Rohin: Yeah.

Scott: And yeah, I think I'm not even making a claim about which of these is more or less risky. I'm just saying that they're sufficiently different risks, that we want to see if we can mitigate either of them.

Rohin: Yeah. That makes sense.

5. Scaling up STEM AI

Rohin: So the thing I wanted to say, and now I'm more confident that it actually makes sense to say, is that it feels to me like the STEM AI approach is lower-risk for a somewhat-smarter-than-human system, but if I imagine scaling up to arbitrarily smarter-than-human systems, I'm way more scared of the STEM AI version.

Scott: Yeah.

Eli: Can you say why, Rohin?

Rohin: According to me, the reason for optimism here is that your STEM AI system isn't really thinking about humans, doesn't know about them. Whatever it's doing, it's not going to successfully execute a treacherous turn, because successfully doing that requires you to model humans. That's at least the case that I would make for this being safe.

And when we're at *somewhat* superhuman systems, I'm like, "Yeah, I mostly buy that." But when we talk about the limit of infinite intelligence or something, then I'm like, "But it's not literally true that the STEM AI system has *no* reason to model the humans, to the extent that it has goals (which seems like a reasonable thing to assume)." Or at least a reasonable thing to worry about, maybe not assume.

It will maybe want more resources. That gives it an incentive to learn about the external world, which includes a bunch of humans. And so it could just start learning about humans, do it very quickly, and then execute a treacherous turn. That seems entirely possible, whereas with the IDA approach, you can hope that we have successfully instilled the notion of, "Yes, you are really actually trying to—"

Scott: Yeah. I definitely don't want us to do this forever.

Rohin: Yeah.

Scott: I'm imagining—I don't know, for concreteness, even though this is a silly plan, there's the plan of "turn Jupiter into a supercomputer, invent some scanning tech, run a literal HCH."

Rohin: Yeah.

Scott: And I don't know, this probably seems like it requires super-superhuman intelligence by your scale. Or maybe not.

Rohin: I don't really have great models for those scales.

Scott: Yeah, I don't either. It's plausible to me that... Yeah, I don't know.

Rohin: Yeah. I'm happy to say, yes, we do not need a system to scale to literally the maximal possible intelligence that is physically possible. And I do not—

Scott: Further, we don't want that, because at some point we need to get the human values into the future, right? (*laughs*)

Rohin: Yes, there is that too.

Scott: At some point we need to... Yeah. There is a sense in which it's punting the problem.

Rohin: Yeah. And to be clear, I am fine with punting the problem.

It does feel a little worrying though that we... Eh, I don't know. Maybe not. Maybe I should just drop this point.

6. Optimism about oversight

Scott: Yeah. I feel like I want to try to summarize you or something. Yeah, I feel like I don't want to just say this optimism thing that you've said several times, but it's like...

I think that I predict that we're in agreement about, "Well, the STEM plan and the IDA plan have different and incomparable-in-theory risk profiles." And I don't know, I imagine you as being kind of comfortable with the risk profile of IDA. And probably also more comfortable than me on the risk profile of basically every plan.

I think that I'm coming from the perspective, "Yep, the risk profile in the STEM plan also seems bad, but it seems higher variance to me."

It seems bad, but maybe we can find something adjacent that's less bad, or something like that. And so I suspect there's some sort of, "I'm pessimistic, and so I'm looking for things such that I don't know what the risk profiles are, because maybe they'll be better than this option."

Rohin: Yeah.

Scott: Yeah. Yeah. So maybe we're mostly disagreeing about the risk profile of the IDA reference class or something—which is a large reference class. It's hard to bring in one thing.

Rohin: I mean, for what it's worth, I started out this conversation not having appreciated the point about process-level feedback requiring more mutual information with humans. I was initially making a stronger claim.

Scott: Okay.

Eli: Woohoo! An update.

Scott: (*laughs*) Yeah. I feel like I want to flag that I am suspicious that neither of us were able to correctly steel-man IDA, because of... I don't know, I'm suspicious that the IDA in Paul's heart doesn't care about whether you build your IDA out of humans versus building it out of some other aliens that can accomplish things, because it's really trying to... I don't know. That's just a flag that I want to put out there that it's plausible to me that both Rohin and I were not able to steel-man IDA correctly.

Rohin: I think I wasn't trying to. Well, okay, I was trying to steel-man it inasmuch as I was like, "Specifically the risks of human manipulation seem not that bad," which, I agree, Paul might have a better case for than me.

Scott: Yeah.

Rohin: I do think that the overall case for optimism is something like, “Yep, you'll be learning, you'll be getting some mutual info about humans, but the oversight will be good enough that it just is not going to manipulate you.”

Scott: I also suspect that we're not just optimistic versus pessimistic on AI safety. I think that we can zoom in on optimism and pessimism about transparency, and it might be that “optimism versus pessimism on transparency” is more accurate.

Rohin: I think I'm also pretty optimistic about process-level feedback or something—which, you might call it transparency...

Scott: Yeah. It's almost like when I'm saying “transparency,” I mean “the whole reference class that lets you do oversight” or something.

Rohin: Oh. In that case, yes, that seems right.

So it includes adversarial training, it includes looking at the concepts that the neural net has learned to see if there's a deception neuron that's firing, it includes looking at counterfactuals of what the agent would have done and seeing whether those would have been bad. All that sort of stuff falls under “transparency” to you?

Scott: When I was saying that sentence, kind of.

Rohin: Sure. Yeah, I can believe that that is true. I think there's also some amount of optimism on my front that the intended generalization is just the one that is usually learned, which would also be a difference that isn't about transparency.

Scott: Yeah.

Rohin: Oh, man. To everyone listening, don't quote me on that. There are a bunch of caveats about that.

Scott: (*laughs*)

Eli: It seems like this is a pretty good place to wrap up, unless we want to dive back in and crux on whether or not we should be pessimistic about transparency and generalization.

Scott: Even if we are to do that, I want to engage with the audience first.

Eli: Yeah, great. So let us open the fishbowl, and people can turn on their cameras and microphones and let us discuss.

7. Q&A: IDA and getting useful work from AI

Scott: Wow, look at all this chat with all this information.

Rohin: That is a lot of chat.

So, I found one of the chat questions, which is, “How would you build an AI system that did not model humans? Currently, to train the model, we just gather a big data set, like a giant pile of all the videos on YouTube, and then throw it at a neural network. It's hard to know what generalizations the network is even making, and it

seems like it would be hard to classify generalizations into a ‘modeling humans’ bucket or not.”

Scott: Yeah. I mean, you can make a Go AI that doesn't model humans. It's plausible that you can make something that can do physics using methods that are all self-play-adjacent or something.

Rohin: In a simulated environment, specifically.

Scott: Right. Yeah.

Or even—I don't know, physics seems hard, but you could make a thing that tries to learn some math and answer math questions, you could do something that looks like self-play with how useful limits are or something in a way that's not... I don't know.

Rohin: Yeah.

Donald Hobson: I think a useful trick here is: anything that you can do in practice with these sorts of methods is easy to brute-force. You can do Go with AlphaGo Zero, and you can easily brute-force Go given unlimited compute. AlphaGo Zero is just a clever—

Scott: Yeah. I think there's some part where the human has to interact with it, but I think I'm imagining that your STEM AI might just be, “Let's get really good at being able to do certain things that we could be able to brute-force in principle, and then use this to do science or something.”

Donald Hobson: Yeah. Like a general differential equation solver or something.

Scott: Yeah. I don't know how to get from this to being able to get, I don't know, reliable brain scanning technology or something like that. But I don't know, it feels plausible that I could turn processes that are about how to build a quantum computer into processes that I could turn into math questions. I don't know.

Donald Hobson: “Given an arbitrary lump of biochemicals, find a way to get as much information as possible out of it.” Of a specific certain kind of information—not thermal noise, obviously.

Scott: Yeah. Also in terms of the specifications, you could actually imagine just being able to specify, similar to being able to specify Go, the problem of, “Hey, I have a decent physics simulation and I want to have an algorithm that is Turing-complete or something, and use this to try to make more...” I don't know. Yeah, it's hard. It's super hard.

Ben Pace: The next question in the chat is from Charlie Steiner saying, “What's with IDA being the default alternative rather than general value learning? Is it because you're responding to imaginary Paul, or is it because you think IDA is particularly likely or particularly good?”

Scott: I think that / was engaging with IDA especially because I think that IDA is among the best plans I see. I think that IDA is especially good, and I think that part of the reason why I think that IDA is especially good is because it seems plausible to me that basically you *can* do it without the core of humanity inside it or something.

I think the version of IDA that you're hoping to be able to do in a way that doesn't have the core of humanity inside the processes that it's doing seems plausible to exist. I like IDA.

Donald Hobson: By that do you mean IDA, like, trained on chess? Say, you take Deep Blue and then train IDA on that? Or do you mean IDA that was originally trained on humans? Because I think the latter is obviously going to have a core of humanity, but IDA trained on Deep Blue—

Scott: Well, no. IDA trained on humans, where the humans are trained to follow an algorithm that's sufficiently reliable.

Donald Hobson: Even if the humans are doing something, there will be side channels. Suppose the humans are solving equations—

Scott: Yeah, but you might be able to have the system watch itself and be able to have those side channels not actually... In theory, there are those side channels, and in the limit, you'd be worried about them, but you might be able to get far enough to be able to get something great without having those side channels actually interfere.

Donald Hobson: Fair enough.

Ben Pace: I'm going to ask a question that I want answered, and there's a good chance the answer is obvious and I missed it, but I think this conversation initially came from, Scott wrote a post called "[Thoughts on Human Models](#)," where he was—

Scott: Ramana wrote the post. I ranted at Ramana, and then he wrote the post.

Ben Pace: Yes. Where he was like, "What if we look more into places that didn't rely on human models in a bunch of ways?" And then later on, Rohin was like, "Oh, I don't think I agree with that post much at all, and I don't think that's a promising area." I currently don't know whether Rohin changed his mind on whether that was a promising area to look at. I'm interested in Rohin's say on that.

Rohin: So A, I'll note that we mostly didn't touch on the things I wrote in those comments because Scott's reasons for caring about this have changed. So I stand by those comments.

Ben Pace: Gotcha.

Rohin: In terms of "Do I think this is a promising area to explore?": eh, I'm a *little* bit more interested in it, mostly based on "perhaps process level feedback is introducing risks that we don't need." But definitely my all-things-considered judgment is still, "Nah I would not be putting resources into this," and there are a bunch of other disagreements that Scott and I have that are feeding into that.

Ben Pace: And Scott, what's your current feelings towards... I can't remember, in the post you said "there's both human models and not human models", and then you said something like, "theory versus"... I can't remember what that dichotomy was.

But how much is your feeling a feeling of "this is a great area with lots of promise" versus "it's plausible and I would like to see some more work in it," versus "this is the primary thing I want to be thinking about."

Scott: So, I think there is a sense in which I am primarily doing science and not engineering. I'm not following the plan of "how do you make systems that are..." I'm a *little* bit presenting a plan, but I'm not primarily thinking about that plan as a plan, as opposed to trying to figure out what's going on.

So I'm not with my main attention focused on an area such that you can even say whether or not there are human models in the plan, because there's not a plan.

I think that I didn't update much during this conversation and I did update earlier. Where the strongest update I've had recently, and partially the reason why I was excited about this, was: The post that was written two years ago basically is just talking about "here's why human models are dangerous", and it's not about the thing that I think is the strongest reason, which is the ability to oversee and the ability to be able to implement the strategy "don't let the thing have human models at all."

It's easy to separate "thinking about physics" from "thinking about manipulating humans" if you put "think about modeling humans" in the same cluster as "think about manipulating humans." And It's a lot harder to draw the line if you want to put "think about modeling humans" on the other side, with "thinking about physics."

And I think that this idea just didn't show up in the post. So to the extent that you want me to talk about change from two years ago, I think that the center of my reasoning has changed, or the part that I'm able to articulate has changed.

There was a thing during this conversation... I think that I actually noticed that I probably don't, even in expectation, think that... I don't know.

There was a part during this conversation where I backpedaled and was like, "Okay, I don't know about the expectation of the risk factor of these two incomparable risks. I just want to say: first, they're different in kind, therefore we should analyze both. They're different in kind, therefore if we're looking for a good strategy, we should keep looking at both because we should keep that OR gate in our abilities.

"And two: It seems like there might be some variance things, where we might be able to learn more about the risk factors on this other side, AI that doesn't use human models. It might turn out that the risks are lower, so we should want to keep attention on both types of approach."

I think maybe I would have said this before. But I think that I'm not confident about what side is better, and I'm more arguing from "let's try lots of different options because I'm not satisfied with any of them."

Rohin: For what it's worth, if we don't talk just about "what are the risks" and we also include "what is the plan that involves this story", then I'm like, "the IDA one seems many orders of magnitude more likely to work". "Many" being like, I don't know, over two.

Ben Pace: It felt like I learned something when Scott explained his key motivation being around the ability to understand what the thing is doing and blacklist or whitelist certain types of cognition, being easier when it's doing no human modeling, as opposed to when it is trying to do human modeling but you also don't want it to be manipulative.

Rohin: I think I got that from the comment thread while we were doing the comment thread, whenever that was.

Ben Pace: The other question I have in chat is: Steve—

Scott: Donald was saying mutual information is too low a bar, and I want to flag that I did not mean that mutual information *is* the metric that should be used to determine whether or not you're modeling humans. But the *type* of thing that I think that we could check for, has some common intuition with trying to check for mutual information.

Donald Hobson: Yeah, *similar* to mutual information. I'd say it was closer to mutual information conditional on a bunch of physics and other background facts about the universe.

Scott: I want to say something like, there's some concept that we haven't invented yet, which is like "logical mutual information," and we don't actually know what it means yet, but it might be something.

Ben Pace: I'm going to go to the next question in chat, which is Steve Byrnes: "Scott and Rohin both agreed/posed right at the start that maybe we can make AIs that do tasks that do not require modeling humans, but I'm still stuck on that. An AI that can't do all of these things, like interacting with people, is seemingly uncompetitive and seemingly unable to take over the world or help solve all of AI alignment. That said, 'use it to figure out brain scanning' was a helpful idea by Scott just now. But I'm not 100% convinced by that example. Is there any other plausible example, path, or story?"

Scott: Yeah, I feel like I can't give a good story in terms of physics as it exists today. But I could imagine being in a world where physics was different such that it would help. And because of that I'm like, "It doesn't seem obviously bad in principle," or something.

I could imagine a world where basically we could run HCH and HCH would be great, but we can't run HCH because we don't have enough compute power. But also, if you can inverse this one hash you get infinity compute because that's how physics works.

Group: (*laughs*)

Scott: Then your plan is: let's do some computer science until we can invert this one hash and then we unlock the hyper-computers. Then we use the hyper-computers to run a literal HCH as opposed to just an IDA.

And that's not the world we live in, but the fact that I can describe that world means that it's an empirical about-the-world fact as opposed to an empirical about-how-intelligence-works fact, or something. So I'm, like, open to the possibility. I don't know.

Gurkenglas: On what side of the divide would you put reasoning about decision theory? Because thinking about agents in general seems like it might not count as modeling humans, and thinking about decision theory is kind of all that's required in order to think about AI safety, and if you can solve AI safety that's the win condition that you were talking about.

But also it's kind of hard to be sure that an AI that is thinking about agency is not going to be able to manipulate us, even if it doesn't know that we are humans. It might just launch attacks against whoever is simulating its box. So long as that person is an agent.

Scott: Yeah. I think that thinking about agents in general is already scary, but not as scary. And I'm concerned about the fact that it might be that thinking about agents is just convergent and you're not going to be able to invert the hash that gives you access to the hyper-computer, unless you think about agents.

That's a particularly pessimistic world, where it's like, you need to have things that are thinking about agency and thinking about things in the context of an agent in order to be able to do anything super powerful.

I'm really uncertain about the convergence of agency. I think there are certain parts of agency that are convergent and there are certain parts of agency that are not, and I'm confused about what the parts of agency even *are*, enough that I kind of just have this one cluster. And I'm like, yeah, it seems like that cluster is kind of convergent. But maybe you can have things that are optimizing "figuring out how to divert cognitive resources" and that's kind of like being an agent, but it's doing so in such a way that's not self-referential or anything, and maybe that's enough? I don't know.

There's this question of the convergency of thinking about agents that I think is a big part of a hole in my map. A hole in my prioritization map is, I'm not actually sure what I think about what parts of agency are convergent, and this affects a lot, because if certain parts of agency are convergent then it feels like it really dooms a lot of plans.

Ben: Rohin, is there anything you wanted to add there?

Rohin: I don't know. I think I broadly agree with Scott in that, if you're going to go down the path of "let's exclude all the types of reasoning that could be used to plan or execute a treacherous turn," you probably want to put general agency on the outside of that barrier. That seems roughly right if you're going down this path.

And on the previous question of: can you do it, can this even be done? I'm super with, I think it was Steve, on these sorts of AI systems probably being uncompetitive. But I sort of have a similar position as Scott: *maybe* there's just a way that you can leverage great knowledge of just science in order to take over the world for example. It seems like that might be possible. I don't know, one way or the other. I would bet against, unless you have some pretty weird scenarios. But I wouldn't bet against at, like, 99% confidence. Or maybe I would bet against it that much. But there I'm getting to "that's probably a bit too high."

And like, Scott is explicitly saying that most of these things are not high-probability things. They're just things that should be investigated. So that didn't seem like an avenue to push down.

Ben Pace: I guess I just disagree. I feel like if you were the first guy to invent nukes, and you invented them in 1800 or something, I feel I could probably tell various stories about having advanced technologies in a bunch of ways helping strategically—not even just weaponry.

Rohin: Yeah, I'd be interested in a story like this. I *don't* feel like you can easily make it. It just seems hard unless you're already a big power in the world.

Ben Pace: Really! Okay. I guess maybe I'll follow up with you on that sometime.

Joe Collman: Doesn't this slightly miss the point though, in that you'd need to be able to take over the world but then also make it safe afterwards, and the "make it

safe afterwards" seems to be the tricky part there. The AI safety threat is still there if you through some conventional means—

Donald Hobson: Yeah, and you want to do this without massive collateral damage. If you invent nukes... If I had a magic map, where I pressed it and the city on it blew up or something, I can't see a way of taking over the world with that without massive collateral damage. Actually, I can't really see a way of taking over the world with that with massive collateral damage.

Ben Pace: I agree that the straightforward story with nukes sounds fairly unethical. I think there's probably ways of doing it that aren't unethical but are more about just providing sufficient value that you're sort of in charge of how the world goes.

Joe Collman: Do you see those ways then making it safer? Does that solve AI safety or is it just sort of, "I'm in charge but there's still the problem?"

Scott: I liked the proof of concept, even though this is not what I would want to do, between running literal HCH and IDA. I feel like literal HCH is just a whole lot safer, and the only difference between literal HCH and IDA is compute and tech.

8. Q&A: HCH's trustworthiness

Joe Collman: May I ask what your intuition is about motivation being a problem for HCH? Because to me it seems like we kind of skip this and we just get into what it's computationally capable of doing, and we ignore the fact that if you have a human making the decisions, that they're not going to care about the particular tasks you give them necessarily. They're going to do what they think is best, not what you think is best.

Scott: Yeah, I think that that's a concern, but that's a concern in IDA too.

If I think that IDA is the default plan to compare things to, then I'm like, "Well, you could instead do HCH, which is just safer—if you can safely get to the point where you can get to HCH."

Joe Collman: Right, yeah. I suppose my worry around this is more if we're finding some approximation to it and we're not sure we've got it right yet. Whereas I suppose if you're uploading and you're sure that the uploading process is actually—

Scott: Which is not necessarily the thing that you would do with STEM AI, it's just a proof of concept.

Donald Hobson: If you're uploading, why put the HCH structure on it at all? Why not just have a bunch of uploaded humans running around a virtual village working on AI safety? If you've got incredibly powerful nano-computers, can't you just make uploaded copies of a bunch of AI safety researchers and run them for a virtual hundred years, but real time five minutes, doing AI safety research?

Scott: I mean, yeah, that's kind of like the HCH thing. It's different, but I don't know...

Charlie Steiner: I'm curious about your intuitive, maybe not necessarily a probability, but gut feeling on HCH success chances, because I feel like it's quite unlikely to preserve human value.

Donald Hobson: I think a small HCH will probably work and roughly preserve human value if it's a proper HCH, no approximations. But with a big one, you're probably going to get ultra-viral memes that aren't really what you wanted.

Gurkenglas: We have a piece of evidence on the motivation problem for HCH and IDA, namely GPT. When it pretends to write for a human, we can easily make it pretend to be any kind of human that we want by simply specifying it in the prompt. The hard part is making the pretend human capable enough.

Donald Hobson: I'm not sure that it's that easy to pick what kind of human you actually get from the prompt.

Scott: I think that the version of IDA that I have any hope for has humans following sufficiently strict procedures and such that this isn't actually a thing that... I don't know, I feel like this is a misunderstanding of what IDA is supposed to do.

I think that you're not supposed to get the value out of the individual pieces in IDA. You're supposed to use that to be able to... Like, you have some sort of task, and you're saying, "Hey, help me figure out how to do this task." And the individual humans inside the HCH/IDA system are not like, "Wait, do I really want to do this task?". The purpose of the IDA is to have the capabilities come from a human-like method as opposed to just, like, a large evolution, so that you can oversee it and trust where it's coming from and everything.

Donald Hobson: Doesn't that mean that the purpose is to have the humans provide all the implicit values so obvious no one bothered to mention them. So if you ask IDA to [put two strawberries on the plate](#), it's the humans' implicit values that do it in a way that doesn't destroy the world.

Scott: I think that it's partially that. I think it's not all that. I think that you only get the basics of human values out of the individual components of the HCH or IDA, and the actual human values are coming from the humans using the IDA system.

I think that IDA as intended should work almost as well if you replace all the humans with well-intentioned aliens.

Ben Pace: Wait, can you say that again? That felt important to me.

Scott: I think, and I might be wrong about this, that the true IDA, the steel-man_{Scott} IDA, should work just as well or almost as well if you replaced all of the humans with well-intentioned aliens.

Donald Hobson: Well-intentioned aliens wouldn't understand English. Your IDA is getting its understanding of English from the humans in it.

Scott: No, I think that if I wanted to use IDA, I might use IDA to for example solve some physics problems, and I would do so with humans inside overseeing the process. There's not supposed to be a part of the IDA that's making sure that human values are being brought into our process of solving these physics problems. The IDA is just there to be able to safely solve physics problems.

And so if you replaced it with well-intentioned aliens, you still solve the physics problems. And if you direct your IDA towards a problem like "solve this social problem", you need information about social dynamics in your system, but I think that you should think of that as coming from a different channel than the core of the

breaking-problems-up-and-stuff, and the core of the breaking-problems-up-an-stuff should be thought of as something that could be done just as well by well-intentioned aliens.

Donald Hobson: So you've got humans that are being handed a social problem and told to break it up, that is the thing you're imitating, but the humans are trying to pretend that they know absolutely nothing about how human societies work except for what's on this external piece of paper that you handed them.

Scott: It's not necessarily a paper. You do need some human-interacting-with-human in order to think about social stuff or something... Yeah, I don't know what I'm saying.

It's more complicated on the social thing than on the physics thing. I think the physics thing should work just as well. I think the physics thing should work just as well with well-intentioned aliens, and the human thing should work just as well if you have an HCH that's built out of humans and well-intentioned aliens and the humans never do any decomposition, they only ask the well-intentioned aliens questions about social facts or something. And the process that's doing decomposition is the well-intended aliens' process of doing composition.

I also might be wrong.

Ben Pace: I like this thread. I also kind of liked it when Joe Collman asked questions that he cared about. If you had more questions you cared about, Joe, I would be interested in you asking those

Joe Collman: I guess, sure. I was just thinking again with the HCH stuff—I would guess, Scott, that you probably think this isn't a practical issue, but I would be worried about the motivational side of HCH in the limit of infinite training of IDA.

Are you thinking that, say, you've trained a thousand iterations of IDA, you're training the thousand-and-first, you've got the human in there. They've got this system that's capable of answering arbitrarily amazing, important questions, and you feed them the question "Do you like bananas?" or something like that, some irrelevant question that's just unimportant. Are we always trusting that the human involved in the training will precisely follow instructions? Are you seeing that as a non-issue, that we can just say, well that's a sort of separate thing—

Scott: I think that you could make your IDA kind of robust to "sometimes the humans are not following the instructions", if it's happening a small amount of the time.

Joe Collman: But if it's generalizing from that into other cases as well, and then it learns, "OK, you don't follow instructions a small amount of time..."—if it generalizes from that, then in any high-stakes situation it doesn't follow instructions, and if we're dealing with an extremely capable system, it might see every situation as a high-stakes situation because it thinks, "I can answer your question or I can save the world." Then I'm going to choose to save the world rather than answering your question directly, providing useful information.

Scott: Yeah, I guess if there's places where the humans reliably... I was mostly just saying, if you're collecting data from humans and there's noise in your data from humans, you could have systems that are more robust to that. But if you have it be the case that humans *reliably* don't follow instructions on questions of type X, then the thing that you're training is a thing that reliably doesn't follow instructions on

questions of type X. And if you have assumptions about what the decomposition is going to do, you might be wrong based on this, and that seems bad, but...

Rohin: I heard Joe as asking a slightly different question, which is: “In not HCH but IDA, which is importantly—”

Joe Collman: To me it seems to apply to either. In HCH it seems clear to me that this will be a problem, because if you give it any task and the task is not effectively “Give me the most valuable information that you possibly can,” then the morality of the H—assuming you've got an H that wants the best for the world—then if you ask it, “Do you like bananas?”, it's just going to give you the most useful information.

Rohin: But an HCH is made up of humans. It should do whatever humans would do. Humans don't do that.

Joe Collman: No, it's going to do what the HCH tree would do. It's not going to do what the top-level human would do.

Rohin: Sure.

Joe Collman: So the top level human might say, “Yes, I like bananas,” but the tree is going to think, “I have infinite computing power. I can tell you how to solve the world's problems, or I can tell you, ‘Yes, I like bananas.’ Of those two, I'm going to tell you how to solve the world's problems, not answer your question.”

Rohin: Why?

Joe Collman: Because that's what a human would do—this is where it comes back into the IDA situation, where I'm saying, eventually it seems... I would assume probably this isn't going to be a practical problem and I would imagine your answer would be, “Okay, maybe in the limit this...”

Rohin: No, I'm happy with focusing on the limit. Even eventually, why does a human do this? The top-level human.

Joe Collman: The top level human: if it's me and you've asked me a question “Do you like bananas?”, and I'm sitting next to a system that allows me to give you information that will immediately allow you to take action to radically improve the world, then I'm just not going to tell you, “Yes, I like bananas”, when I have the option to tell you something that's going to save lives in the next five minutes or otherwise radically improve the world.

It seems that if we assume we've got a human that really cares about good things happening in the world, and you ask a trivial question, you're not going to get an answer to that question, it seems to me.

Rohin: Sure, I agree if you have a human who is making decisions that way, then yes, that's the decision that would come out of it.

Joe Collman: The trouble is, isn't that the kind of human that we want in these situations—is one precisely that does make decisions that way?

Rohin: No, I don't think so. We don't want that. We want a human who's trying to do what the user wants them to do.

Joe Collman: Right. The thing is, they have to be applying their own... HCH is basically a means of getting enlightened judgment. So the enlightened judgment has to come from the HCH framework rather than from the human user.

Rohin: I mean, I think if your HCH is telling the user something and the user is like, "Dammit, I didn't like this", then your HCH is not aligned with the user and you have failed.

Joe Collman: Yeah, but the HCH is aligned with the enlightenment. I suppose the thing I'm saying is that the enlightened judgment as specified by HCH will not agree with me myself. An HCH of me—it's enlightened judgment is going to do completely different things than the things I would want. So, yes, it's not aligned, but it's more enlightened than me. It's doing things that on "long reflection" I would agree with.

Rohin: I mean, sure, if you want to define terms that way then I would say that HCH is not trying to do the enlightened judgment thing, and if you've chosen a human who does the enlightened judgment thing you've failed.

Donald Hobson: Could we solve this by just never asking HCH trivial questions?

Joe Collman: The thing is, this is only going to be a problem where there's a departure between solving the task that's been assigned versus doing the thing that sort of maximally improves the world. Obviously if those are both the same, if you've asked wonderfully important questions or if you asked the HCH system "What is the most important question I can ask you?" and then you ask that, then it's all fine—

Scott: I think that most of the parts of the HCH will have the correct belief that the way to best help the world is to follow instructions corrigibly or something.

Joe Collman: Is that the case though, in general?

Scott: Well, it's like, if you're part of this big network that's trying to answer some questions, it's—

Joe Collman: So are you thinking from a UDT point of view, if I reasoned such that every part of this network is going to reason in the same way as me, therefore I need the network as a whole to answer the questions that are assigned to it or it's not going to come up with anything useful at all. Therefore I need to answer the question assigned to me, otherwise the system is going to fail. Is that the kind of...?

Scott: Yeah, I don't even think you have to call on UDT in order to get this answer right. I think it's just: I am part of this big process and it's like, "Hey solve this little chemistry problem," and if I start trying to do something other than solve the chemistry problem I'm just going to add noise to the system and make it less effective.

Joe Collman: Right. That makes sense to me if the actual top-level user has somehow limited the output so that it can only respond in terms of solving a chemistry problem. Then I guess I'd go along with you. But if the output is just free text output and I'm within the system and I get the choice to solve the chemistry problem as assigned, or I have the choice to provide the maximally useful information, just generally, then I'm going to provide the maximally useful information, right?

Scott: Yeah, I think I wouldn't build an HCH out of you then. (*laughs*)

Rohin: Yeah, that's definitely what I'm thinking.

Scott: I think that I don't want to have the world-optimization in the HCH in that way. I want to just... I think that the thing that gives us the best shot is to have a corrigible HCH. And so—

Joe Collman: Right, but my argument is basically that eventually the kind of person you want to put in is not going to be corrigible. With enough enlightenment—

Rohin: I think I continue to be confused about why you assume that we want to put a consequentialist into the HCH.

Joe Collman: I guess I assume that with sufficient reflection, pretty much everyone is a consequentialist. I think the idea that you can find a human who at least is *reliably* going to stay not a consequentialist, even after you amplify their reasoning hugely, that seems a very suspicious idea to me.

Essentially, let's say for instance, due to the expansion of the universe we're losing—I calculated this very roughly, but we're losing about two stars per second, in the amount of the universe we can access, something like that.

So in theory, that means every second we delay in space colonization and the rest of it, is a huge loss in absolute terms. And so, any system that could credibly make the argument, "With this approach we can do this, and we can not lose these stars"—it seems to me that if you're then trading that against, "Oh, you can do this, and you can answer this question, solve this chemistry problem, or you can improve the world in this huge way"...

Rohin: So, if your claim is like, there are two people, Alice and Bob. And if you put Alice inside a giant HCH, and *really* just take the limit all the way to infinity, then that HCH is not going to be aligned with Bob, because Alice is probably going to be a consequentialist. Then yes, sure, that seems probably true.

Joe Collman: I'm saying it's not really going to be aligned with Alice in the sense that it will do what Alice wants. It will do what the HCH of Alice wants, but it won't do what Alice wants.

Ben Pace: Can I move to Donald's question?

Rohin: Sure.

Donald Hobson: I was just going to say that, I think the whole reason that a giant HCH just answering the question is helpful: Suppose that the top-level question is "solve AI alignment." And somewhere down from that you get "design better computers." And somewhere down from that you get "solve this simple chemistry problem to help the [inaudible] processes or whatever."

And so, all of the other big world-improvement stuff is already being done above you in some other parts of the tree. So, literally the best thing you can do to help the world, if you find yourself down at the bottom of the HCH, is just solving that simple chemistry problem. Because all the other AI stuff is being done by some other copy of you.

Joe Collman: That's plausible, sure. Yeah, if you reason that way. My only thing that I'm claiming quite strongly, is that eventually with suitable amplification, everyone is

going to come around to the idea, "I should give the response that is best for the world," or something like that.

So, if your chain of reasoning takes you to "solving that chemistry problem is making a contribution that at the top level is best for the world," then sure, that's plausible. It's hard to see exactly how you reliably get to draw that conclusion, that solving the question you've been asked *is* the best for the world. But yeah.

Rob Miles: Is it necessary for HCH that the humans in it have no idea where in the tree they are? Or could you pass in some contexts that just says, "Solve this chemistry problem that will improve—"

Scott: You actually want to send that context in. There's discussion about one of the things that you might do in solving HCH, which is: along with the problems, you pass in an ordinal, that's how much resources of HCH you're allowed. And so you say, like, "Hey, solve this problem. And you get 10^{100} resources." And then you're like, "Here, sub-routine, solve this problem. You get $10^{100} / 2$ resources." And then, once you get down into zero resources, you aren't allowed to make any further calls.

And you could imagine working this into the system, or you could imagine just corrigible humans following this instruction, and saying, "Whenever you get input in some amount of resources, you don't spend more than that."

And this resource kind of tells you some information about where you are in the tree. And not having this resource is a bad thing. The purpose of this resource is to make it so that there's a unique fixed point of HCH. And without it, there's not necessarily a unique fixed point. Which is basically to say that, I think that individual parts of an HCH are not intended to be fully thinking about all the different places in the tree, because they're supposed to be thinking locally, because the tree is big and complex. But I think that there's no "Let's try to hide information about where you are in the tree from the components."

Rob Miles: It feels like that would partly solve Joe's problem. If you're given a simple chemistry problem and 10^{100} resources, then you might be like, "This is a waste of resources. I'm going to do something smarter." But if you're being allocated some resources that seem reasonable for the difficulty of the problem you're being set, then you can assume that you're just part of the tree—

Scott: Well, if you started out with—

Ben Pace: Yeah, Rob is right that in real life, if you give me a trillion dollars to solve a simple chemistry problem, I'll primarily use the resources to do cooler shit. (*laughs*)

Joe Collman: I think maybe the difficulty might be here that if you're given a problem where the amount of resources is about right for the difficulty of solving the problem, but the problem isn't actually important.

So obviously "Do you like bananas?" is a bad example, because you could set up an IDA training scheme that *learns* how many resources to put into it. And there the top level human could just reply, "Yes," immediately. And so, you don't—

Scott: I'm definitely imagining the resources are, it's reasonable to give more resources than you need. And then you just don't even use them.

Joe Collman: Right, yeah. But I suppose, just to finish off the thing that I was going to say is that yes, it still seems, with my kind of worry, it's difficult if you have a question which has a reasonable amount of resources applied to it, but is trivial, is insignificant. It seems like the question wouldn't get answered then.

Charlie Steiner: I want to be even more pessimistic, because of the unstable gradient problem, or any fixed point problem. In the limit of infinite compute or infinite training, we have at each level some function being applied to the input, and then that generates an output. It seems like you're going to end up outputting the thing that most reliably leads to a cycle that outputs itself, or eventually outputs itself. I don't know. This is similar to what Donald said about super-virulent memes.

9. Q&A: Disagreement and mesa-optimizers

Ray Arnold: So it seems like a lot of stuff was reducing to, "Okay, are we pessimistic or not pessimistic about safe AGI generally being hard?"

Rohin: Partly that, and partly me be plan-oriented, and Scott being science-oriented.

Scott: I mean, I'm trying to be plan-oriented in this conversation, or something.

Ray Arnold: It seems like a lot of these disagreements reduce to this ur-disagreement of, "Is it going to be hard or easy?", or a few different axes of how it's going to be hard or easy. And I'm curious, how important is it right now to be prioritizing ability to resolve the really deep, gnarly disagreements, versus just "we have multiple people with different paradigms, and maybe that's fine, and we hope one of them works out"?

Eli: I'll say that I have updated downward on thinking that that's important.

Ray Arnold: (*laughs*) As Mr. Double Crux.

Eli: Yeah. It seems like things that cause concrete research progress are good, and conversations like this one do seem to cause insights that are concrete research progress, but...

Ray Arnold: Resolving the disagreement isn't concrete research progress?

Eli: Yeah. It's like, "What things help with more thinking?" Those things seem good. But I used to think there were big disagreements—I don't know, I still have some probability mass on this—but I used to think there were big disagreements, and there was a lot of value on the table of resolving them.

Gurkenglas: Do you agree that if you can verify whether a system is thinking about agents, that you could also verify whether it has a [mesa-optimizer](#)?

Scott: I kind of think that systems will have mesa-optimizers. There's a question of whether or not mesa-optimizers will be there explicitly or something, but that kind of doesn't matter.

It would be nice if we could understand where the base level is. But I think that the place where the base level is, we won't even be able to point a finger at what a "level" is, or something like that.

And we're going to have processes that make giant spaghetti code that we don't understand. And that giant spaghetti code is going to be doing optimization.

Gurkenglas: And therefore we won't be able to tell whether it's thinking about agency.

Scott: Yeah, I don't know. I want to exaggerate a little bit less what I'm saying. Like, I said, "Ah, maybe if we work at it for 20 years, we can figure out enough transparency to be able to distinguish between the thing that's thinking about physics and the thing that's thinking about humans." And maybe that involves something that's like, "I want to understand logical mutual information, and be able to look at the system, and be able to see whether or not I can see humans in it." I don't know. Maybe we can solve the problem.

Donald Hobson: I think that actually conventional mutual information is good enough for that. After all, a good bit of quantum random noise went into human psychology.

Scott: Mutual information is intractable. You need some sort of, "Can I look at the system..." It's kind of like this—

Donald Hobson: Yeah. Mutual information [inaudible] Kolmogorov complexity, sure. But computable approximations to that.

Scott: I think that there might be some more information on what computable approximations, as opposed to just saying "computable approximations," but yeah.

Charlie Steiner: I thought the point you were making was more like "mutual information requires having an underlying distribution."

Scott: Yeah. That too. I'm conflating that with "the underlying distributions are tractable," or something.

But I don't know. There's this thing where you want to make your AI system assign credit scores, but you want it to not be racist. And so, you determine whether or not by looking at the last layer, you can determine the race of the participants. And then you optimize against that in your adversarial network, or something like that.

And there's like *that*, but for thinking about humans. And much better than that.

Gurkenglas: Let's say we have a system that we can with our interpretability tools barely tell has some mesa-optimizer at the top level, that's probably doing some further mesa-optimization in there. Couldn't we then make that outer mesa-optimization explicit, part of the architecture, but then train the whole thing anew and repeat?

Scott: I feel like this process might cause infinite regress, but—

Gurkenglas: Surely every layer of mesa-optimization stems from the training process discovering either a better prior or a better architecture. And so we can increase the performance of our architecture. And surely that process converges. Or like, perhaps we just—

Scott: I'm not sure that just breaking it down into layers is going to hold up as you go into the system.

It's like: for us, maybe we can think about things with meta-optimization. But your learned model might be equivalent to having multiple levels of mesa-optimization, while actually, you can't clearly break it up into different levels.

The methods that are used by the mesa-optimizers to find mesa-optimizers might be sufficiently different from our methods, that you can't always just... I don't know.

Charlie Steiner: Also Gurkenglas, I'm a little pessimistic about baking a mesa-optimizer into the architecture, and then training it, and then hoping that that resolves the problem of distributional shift. I think that even if you have your training system finding these really good approximations for you, even if you use those approximations within the training distribution and get really good results, it seems like you're still going to get distributional shift.

Donald Hobson: Yeah. I think you might want to just get the mesa-optimizers out of your algorithm, rather than having an algorithm that's full of mesa-optimizers, but you've got some kind of control over them somehow.

Gurkenglas: Do you all think that by your definitions of mesa-optimization, AlphaZero has mesa-optimizers?

Donald Hobson: I don't know.

Gurkenglas: I'm especially asking Scott, because he said that this is going to happen.

Scott: I don't know what I think. I feel not-optimistic about just going deeper and deeper, doing mesa-optimization transparency inception.

Perhaps vastly more people should be on FDA-approved weight loss medication

[Epistemic Status: I feel pretty good about most of this, but the life-years-saved-via-medication part is problematic on a number of levels, as pointed out by a few commenters. I include it since back-of-the-envelope calculations serve a purpose in ensuring we're comparing effects of approximately the appropriate magnitudes in doing risk/benefit analyses, but I wouldn't take it too seriously.]

Note that I'm **not** a doctor. Please speak to a doctor before doing any of this stuff Or You Will Die.

Introduction

Judging by posts in [r/loseit](#), the existence of effective anti-obesity medications is not particularly well-known (and to the degree it *is* well-known, it's disapproved of.) Even posts on LessWrong, which tend to be very well-researched and exhaustive, simply ignore the topic of medication when weight loss methods or obesity are brought up; I suspect this is not because their authors had explicitly considered and discarded the various anti-obesity drugs currently available, but rather, because the existence of these drugs is very poorly-known. Which I'm attempting to remedy here! At least for the LessWrong crowd.

Quantifying Life-Years Saved by Losing A Certain Amount Of Weight

[Note: as pointed out by comments below, extrapolation to life-years saved is *very speculative*, since all the studies on this in humans are going to be confounded all to hell by healthy user bias and socioeconomic correlations and the like. That said, it feels like a fairly reasonable extrapolation given the comorbidity of obesity to various extremely problematic medical conditions. Be warned!]

According to [Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity | Nature Communications](#), losing a single unit of BMI roughly corresponds to a 7-month gain in life expectancy in the overweight and obese. This seems basically in line with what I hear from popular sources, such as: "[\[L\]ife expectancy for obese men and women was 4.2 and 3.5 years shorter](#)" than people in the healthy BMI range.

This won't count as a revelation. *Obesity is unhealthy, news at eleven.* My goal here is just to quantify what you're getting relative to the risks involved in doing something to ameliorate it.

Accordingly:

[The U.S. Food and Drug Administration \(FDA\) recommends pharmacotherapy for weight loss when lifestyle interventions \(diet, exercise and behavioural therapy\) have failed and the body mass index \(BMI\) is ≥30kg/m² with no concomitant obesity-related risk factors, or if the BMI is ≥27 kg/m² and the patient has at least one obesity-related risk factor.](#)

So: let's talk about weight loss drugs!

Weight Loss Drug Studies

Weigh loss drug studies are always composed of two groups of patients: a group attempting guided diet and exercise along with a placebo pill, and a group doing the diet and exercise plus the drug. That's important here, since it means we can't unequivocally recommend drugs as a *replacement* for diet and exercise, only as a secondary treatment. (Aside: even though basically every article on weight loss is obligated by eternal law to pay tribute to exercise, the evidence for it helping with weight loss on a practical level is minimal.)

These studies are saying, in effect: if you can get X pounds lost from diet and exercise alone, adding pharmaceuticals to these efforts can get you X+Y pounds lost.

That being the case, I'm going to now list the three or so good (as judged by me, a random asshole with a laptop) FDA-approved anti-obesity drugs currently on the market right now; their measured diet-and-exercise-subtracted weight loss; and finally, the amount of life-years you can (maybe? who knows) gain over the long term by losing that much weight. I'll be linking to studies for each.

Note on drugs I'm not discussing here: I'm not going into liraglutide since it seems basically like worse semaglutide at similar cost, and I'm not going into phentermine+topiramate (Qsymia) because in spite of its greater efficacy than phentermine alone, it seems that topiramate has a substantial likelihood of giving people kidney stones and brain fog, which are... not great. Orlistat is quite popular, but has relatively poor efficacy and unpleasant digestive side effects. Links provided on request, but that's a bit far afield of my purposes here, so I'll move on.

The Drugs (at least, the better ones)

Semaglutide (2.4 mg)

- **Price:** 1300ish dollars per month for Wegovy. I've heard insurance has a... spotty... record of covering this. You *might* have better luck with insurance (provided you have T2D, or at least are at risk for it) with Ozempic, which is the same semaglutide, just at a different dose and with labeling for T2D treatment.
- **Mechanism:** GLP-1 inhibitor; more specifically, it slows gastric emptying resulting in lowered appetite.
- **Average Diet/Exercise-Subtracted weight loss:** 12% based on its phase-3 trial. This is the most potent anti-obesity drug on the market.
- **Common Side Effects:** Transient nausea and GI upset at treatment onset.
- **Other Notes:** This is basically just higher-dose Ozempic, which has been on the market about four years.
- **Approximate BMI drop for a 5'6 female at 200 pounds:** In weight, 12% weight loss equates to about 24 pounds. This is a drop in BMI of 32.28 to 28.4 units.
- **Approximate difference between expected life-years of people with these two BMI values:** About 28 months, or about 2.3 years.

Contrave [Bupropion + Naltrexone]

- **Price:** If you get it generic (and why wouldn't you?) about 40 bucks a month as naltrexone + bupropion.
- **Mechanism:** Poorly-understood neurochemical effects.
- **Average Diet/Exercise-Subtracted weight loss:** 3-7% (varies by study)
- **Common Side Effects:** Amped up sex drive and improved focus (Bupropion is sometimes used off-label for ADHD); on the other hand, anxiety and insomnia, plus transient nausea at treatment onset. [My own bias: I'm on bupropion and it's mostly kickass. Insomnia's no fun, though.]

- **Other Notes:** Both parts of this drug have been in common use for several decades. If there was some godawful long-term side effect we'd know about it by now.
- **Approximate BMI drop for a 5'6 female at 200 pounds:** For lower estimates, this is a drop in BMI of 32.28 to 31.31 units (so about 1 unit of BMI); for higher (7%) estimates, this is about 2 units of BMI.
- **Approximate difference between expected life-years of people with these two BMI values:** About 7-14 months of life.

Phentermine

- **Price:** 23 dollars/month
- **Mechanism:** Stimulant. Most stimulants have weight loss as a side effect; this is just one of the few the FDA has actually approved for the purpose.
- **Average Diet/Exercise-Subtracted weight loss:** 3-7% (varies by study)
- **Common Side Effects:** This is a mild stimulant, so... pretty much what you'd expect.
- **Other Notes:** Technically any use of this longer than 6 months is off-label (the FDA hates stimulants), but [several long-term studies of phentermine use find no evidence of addiction or other side-effects when taken for years](#). Anecdotally this is sometimes taken with Contrave, but this is an off-label combination of drugs on which there is little data. On the other hand, there doesn't seem to be any *a priori* reason to expect this combination to be harmful?
- **Approximate BMI drop for a 5'6 female at 200 pounds:** For lower estimates, this is a drop in BMI of 32.28 to 31.31 units (so about 1 unit of BMI); for higher (7%) estimates, this is about 2 units of BMI.
- **Approximate difference between expected life-years of people with these two BMI values:** About 7-14 months of life.

Conclusion

Overweight and obesity cause a lot of misery! I've lurked [r/loseit!](#) And the quest to *stop* being overweight is the cause of even more misery for lots of people. If you're in that group, you might be well-served by discussing medication-assisted options with a doctor.

FAQs

Giving a life-years-saved number for if someone takes a drug implies they'll be on it forever. But what about the unquantified risks of being on some drug for the rest of your life? Especially semaglutide, which hasn't been around very long?

This is a fair concern! It is, however, worth pointing out that the FDA is *vigorous* about pulling drugs that have been shown to have even small risks of causing life-threatening conditions; a recent example of this is lorcaserin (aka Belviq), which was taken off the market due to a *non-statistically-significant* increased risk of cancer. See also: [Is lorcaserin really associated with increased risk of cancer? A systematic review and meta-analysis - PubMed \(nih.gov\)](#)

Think about the implications! If you're on X drug for your whole life, then by assumption you'll have *also* gone your whole life without the FDA having observed any statistical

increases in cancer incidence or heart attacks or whatever for people on the drug. That's a very high bar of safety.

Ultimately, the *quantifiable* life-years lost by obesity (in the form of statistical heart attacks and various other comorbidities) must be weighed against the mere uncertain prospect of an imperfect drug making it through the FDA approval process.

Besides which, nobody says once you're done losing weight that you have to continue taking the appetite suppressants. I mean, I probably would? But diff'rent strokes.

If you're concerned regardless-- semaglutide is the only particularly new treatment on that list (and even that's been around a few years in the form of Ozempic). The others have multi-decade histories of usage, with reams of literature on their effects. Google Scholar: your friend and mine.

Isn't just eating less a much healthier and better-proven means of weight loss than pills?

Nicotine addiction wears off with time. If a person can keep off cigarettes and other nicotine sources for about three months, most surveys show that this leads to a total cessation of a desire to smoke or otherwise consume nicotine.

That being the case, it's obvious what addicted smokers have to do to cure their addiction: stop smoking cigarettes for three months. Withdrawal is unpleasant, but nevertheless this method is uncomplicated (step one: don't smoke, step two: nothing), extremely cheap as an intervention, and guaranteed to work if performed.

And that is why, even to this day, nobody is addicted to cigarettes.

...

I guess a *less* snarky answer is that these medications mostly work by making it more pleasant to eat less.

Isn't this just a way of letting lazy people off the hook?

Eh. If you've tried it and straightforward dieting makes you miserable, you are under no obligation to power through without assistance. You don't win virtue-points for avoiding medication that makes your life easier even if online randos imply otherwise.

If these meds are so great, shouldn't I have heard about them by now?

Nope!

First: American society has a pretty weird relationship with weight loss; there's a huge implication in the discourse that thin-ness is a result of righteous self-discipline, and that fat people just need to buckle down and *make the effort*, and if they fail then they just weren't trying hard enough. (This viewpoint is neatly encapsulated in the slogan "eat less move more" and concepts like "the physics diet") Accordingly, weight loss drugs have acquired the implicit moral status of a cheat enabling one to get the reward without the suffering, which people are suspicious of.

Second, there are also some now-banned medications-- fen-phen and DNP are pretty good examples-- that are both (1) deadly and (2) highly effective at weight loss. Thus, the popular perception that anti-obesity drugs are intrinsically dangerous, to be used by people who value their appearance more than their health.

This isn't helped by all the truly worthless herbal supplements on the market claiming to be effective weight loss aids; unlike for most other medical conditions, herbal supplements *are* allowed to claim that they'll help with obesity (mostly implicitly by calling themselves "fat burners" and the like). Legitimate drugs can be difficult for uneducated audiences to distinguish from snake oil, so they get rounded off to "snake oil".

These factors have resulted in society collectively memory-holing this entire class of medication.

I had bad experiences on phentermine.

You and a bunch of other people!

Most people have some drugs they'll find unpleasant or that don't work for them. It seems broadly reasonable to just try different drugs until you find ones that work for whatever condition you're trying to alleviate; the potential risk is one or two weeks of discomfort while the drugs slowly exit your system (after which you move on to something else), and the potential reward is life-years saved from obesity comorbidities, as well as whatever added happiness you get from being at a lighter weight.

[Insert Certification Body Here] doesn't think [Insert Drug Here] passes a cost/benefit analysis, even though the FDA does.

If you're able to read and evaluate the primary literature on this topic, I see no reason to outsource your cost/benefit analyses to some other decisionmaking body rather than evaluating the drugs on their merits based on clinical trial data.

Ultimately, the decisions of these institutions will be colored by a lot of factors that *aren't* "patient wellbeing"-- by blameworthiness, PR considerations, the perceived second-order effects of their actions, and the crucial distinction between dead-their-fault and dead-not-their-fault. These tend to lean in the direction of "don't certify the medication," especially for anti-obesity drugs (which tend to be viewed as lifestyle drugs rather than drugs for legitimate illnesses, and so face a higher bar of scrutiny.)

Shouldn't we just, as a society, just develop a healthier culture around food? Wouldn't that be better than medicating ourselves?



There are a lot of things I would change about society if I were made Benevolent Dictator For Life.

Aren't you, personally, just a lazy trash-person?

Absolutely. But that's unrelated.

I found a factual error in this post.

Leave a comment telling me this (including a source for the info) and I'll correct it!

Are you a doctor? You're a doctor, right? You're probably a doctor, so I should take this as medical advice.

Jesus Christ no. If you take any medication here without talking to a doctor about it you'll definitely swell up and die, or possibly turn inside-out. Luckily, since I pointed this out in this paragraph I will be absolved of all responsibility.

Good luck!

Value loading in the human brain: a worked example

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

UPDATE MARCH 2022: I later revised and improved this post—see [\[Intro to brain-like-AGI safety\] 7. From hardcoded drives to foresighted plans: A worked example](#). I'm leaving this old version here, but I strongly suggest you click over to the later version instead. Goodbye!

We start out as infants knowing nothing of the world. Then many years later, we have various complex conscious explicit goals, like “I want to get out of debt”, and we take foresighted action in pursuit of those goals. What algorithms can get us from here to there? In this post I’m going to work through an example. To summarize and simplify a bit, the steps will be:

1. We gradually develop a probabilistic generative model of the world and ourselves;
2. There’s a “credit assignment” process, where something in the world-model gets flagged as “good”;
3. There’s a reward prediction error signal roughly related to the *time-derivative of the expected probability of the “good” thing*. This signal drives us to “try” to make the “good” thing happen, including via foresighted planning.

To make things a bit simpler, I won’t talk about “I want to get out of debt” here. I’ll do a simpler example instead. Let’s say (purely hypothetically... ☺) that I ate a slice of [prinsesstårta](#) cake a couple years ago, and it was really yummy, and ever since then I’ve wanted to eat one again.

So my running example of an explicit goal in this post will be “I want a slice of [prinsesstårta](#)”.

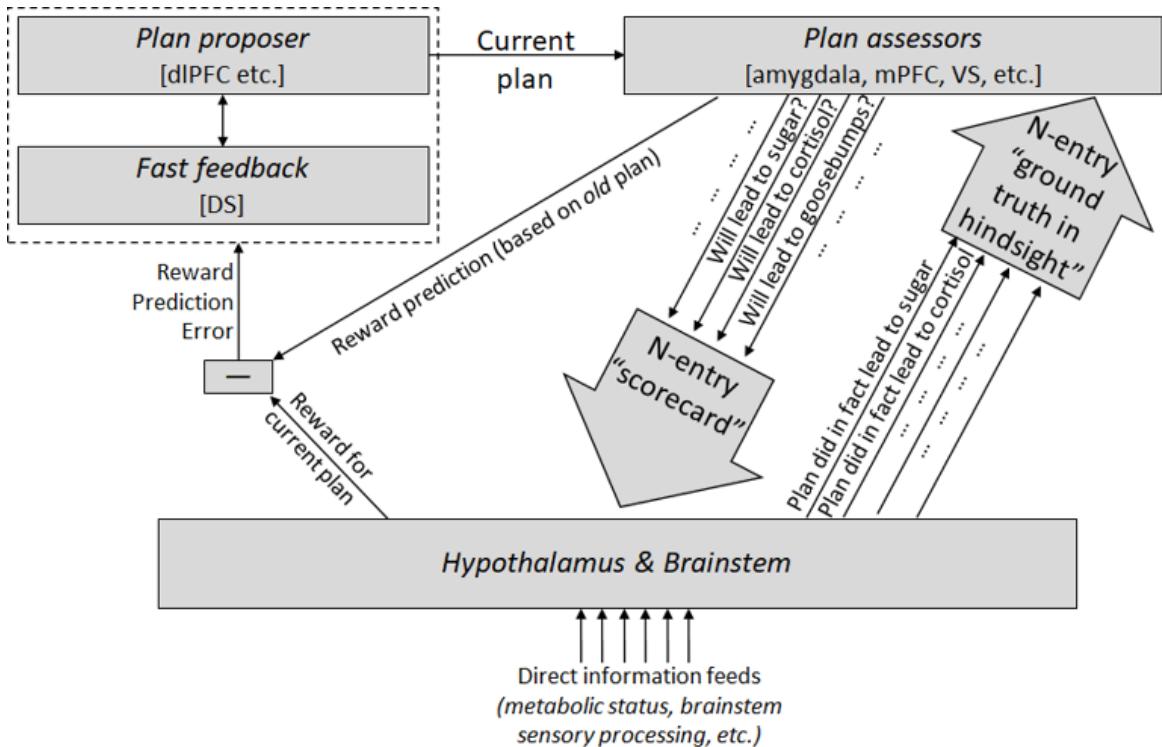
It’s not my only goal in life, or even a particularly important one—so it has to trade off against my other goals and desires—but it is nevertheless a goal of mine (at least when I’m thinking about it), and I would indeed make complicated foresighted plans to try bring about that goal, like dropping subtle hints to my family ... in blog posts ... when my birthday is coming up. Purely hypothetically!!

Big-picture context: As in most of [my posts](#), the subtext is that researchers may someday make AGI algorithms that resemble human brain algorithms—after all, loads of people in neuroscience and AI are trying to do exactly that as we speak. And if those researchers successfully do that, then we’ll want to have a good understanding of how (and indeed whether) we can sculpt those AGIs’ motivations such that the AGIs are robustly trying to do the things we want them to be trying to do. I don’t know the answer to that question, and nobody else does either. But one relevant data-point towards answering that question is “how does the analogous thing work in human brains?”. To be clear, I’m talking here about within-lifetime learning, not evolution-as-a-learning-algorithm, for reasons discussed [here](#).

Also as usual, this is all based on my personal best current understanding of aspects of the human brain, and I could well be wrong, or missing important things.

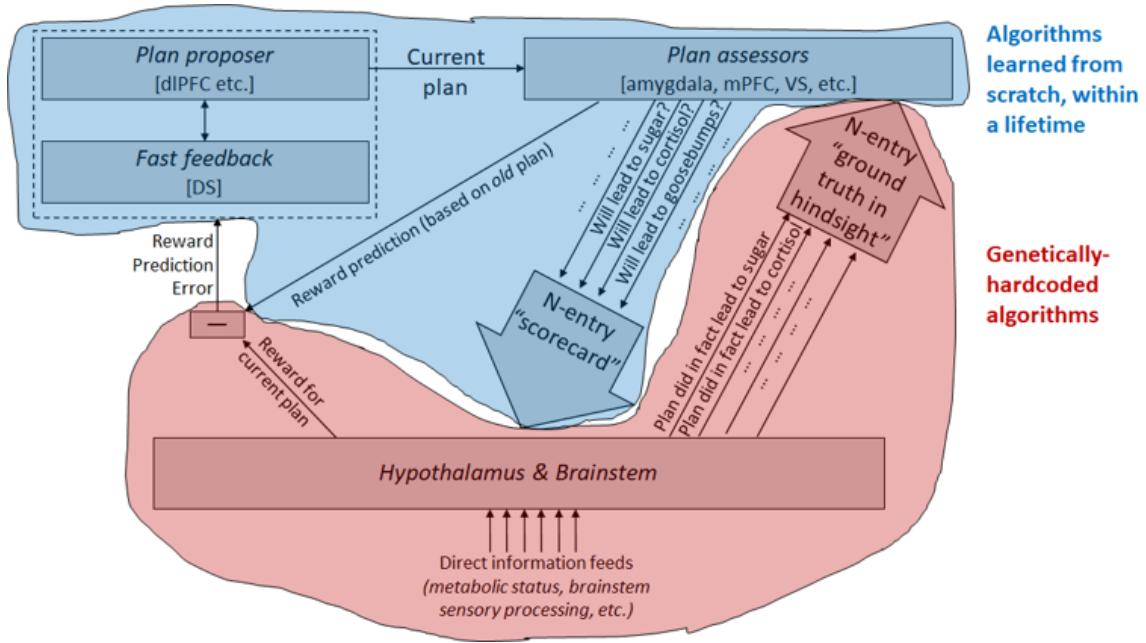
Background on motivation and brainstem-supervised learning

As required background, here's my diagram of decision-making, motivation, and reinforcement learning in the brain:



See [A model of decision-making in the brain \(the short version\)](#) for a walkthrough. All acronyms are brain parts—they don't matter for this post. The left side involves reinforcement learning, the right side involves supervised learning (SL). For the SL part, I think the brain has dozens-to-hundreds of nearly-identical SL algorithms, each with a different supervisory signal. Maybe there's one SL algorithm for each autonomic action, something like that. By the way, in the context of AGI safety, I vote for putting in *thousands* of SL algorithms if we can! Let's have one for every friggin' word in the dictionary! Or heck, millions of them! Or infinity!! Why not an SL algorithm for every point in GPT-3's latent space? Let's go nuts! [More dakka!!](#)

An important aspect of this for this post is that I'm suggesting that some parts (the hypothalamus and brainstem) are (to a first approximation) entirely genetically-hardcoded, while other parts (the “plan proposer” and “plan assessors”) are AFAICT “trained models” in ML terminology—they’re initialized from random weights (or something equivalent) at birth, and learned within a lifetime. (See discussion of “learning-from-scratch-ism” [here](#).) Here’s an illustration:



(The reward prediction error on the left comes from subtracting a trained model output from a genetically-hardcoded algorithm output, so I left it uncolored.)

1. Building a probabilistic generative world-model in the cortex

The first step is that, over my lifetime, my cortex builds up a probabilistic generative model, mostly by self-supervised (a.k.a. predictive) learning.

Basically, we learn patterns in our sensory input, and patterns in the patterns, etc., until we have a nice predictive model of the world (and of ourselves)—a giant web of interconnected entries like “grass” and “standing up” and “slices of prinsesstårta cake” and so on.

Note that I left predictive learning off of the diagram above. Sorry! I didn't want it to be too busy. Anyway, predictive learning lives inside the “plan proposer”. A plan is just a special case of a “thought”, and a “thought” is some configuration of this generative world-model.

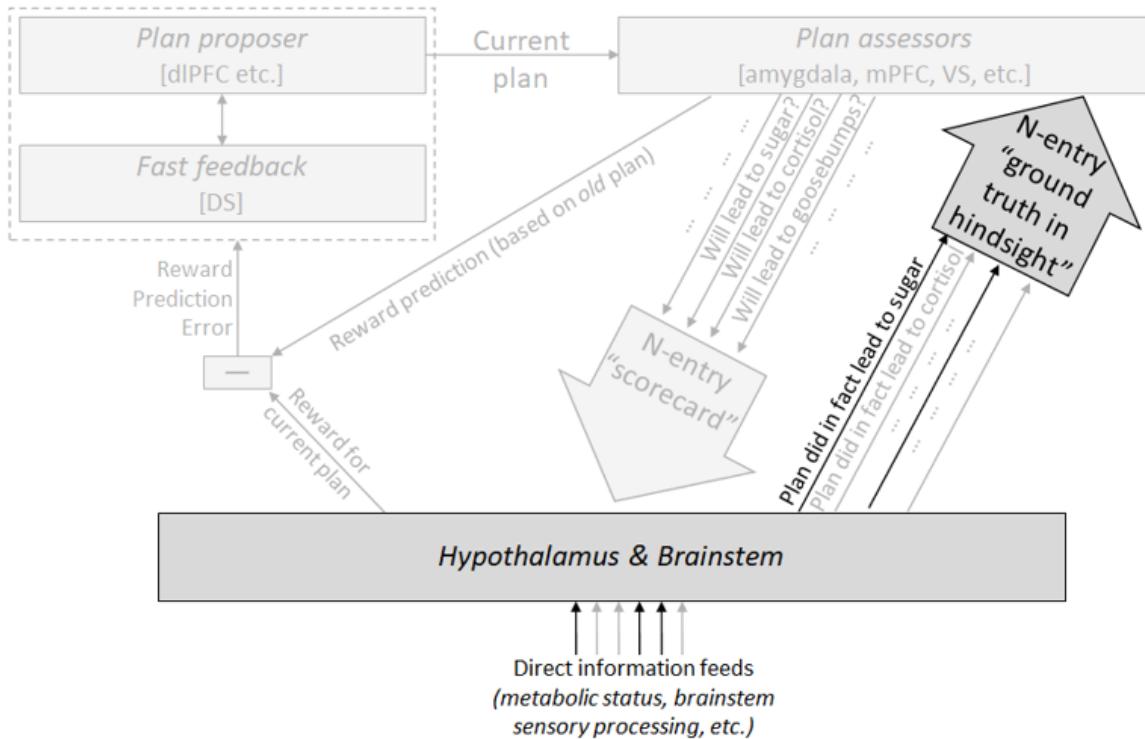
Every thought I can possibly think, and every plan I can possibly plan, can be represented as some configuration of this world-model data structure. The data structure is also continually getting edited, as I learn and experience new things.

When you think of this data structure, imagine many gigabytes or terabytes of inscrutable entries like “*PATTERN 8472836 is defined as the sequence PATTERN 278561 followed by PATTERN 6578362 followed by...*”, or whatever. Some entries have references to sensory inputs or motor outputs. And that giant inscrutable mess comprises my entire understanding of the world and myself.

2. Credit assignment when I first bite into the cake

As I mentioned, two years ago I ate a slice of prinsesstårta cake, and it was *really good*.

Step back to a couple seconds earlier, as I was bringing the cake towards my mouth to take my first-ever bite. At that moment, I didn't yet have any particularly strong expectation of what it would taste like, or how it would make me feel. But once it was in my mouth, mmmmmmm, yummy.



So, as I took that bite, my body had a suite of autonomic reactions—releasing certain hormones, salivating, changing my heart rate and blood pressure, etc. etc. Why? The key is that, as a rule, all sensory inputs split:

- One copy of any given sensory signal goes to the learning-from-scratch subsystem, to be integrated into the predictive world-model. (I omitted that from the diagram above.)
- A second copy of the same signal goes to the hypothalamus & brainstem system, where it serves as an input to genetically-hardwired circuitry. (That's the “Direct information feeds” at the bottom of the diagram above.)

Taste bud inputs are no exception: the former signal winds up at the insula (part of the neocortex), the latter at the medulla (part of the brainstem). The latter feeds into hardcoded brainstem circuits, which, when prompted with the taste and mouth-feel of the cake, and also accounting for my current physiological state and so on, executed all those autonomic reactions I mentioned.

As I mentioned, *before* I first bit into the cake, I didn't expect it to be that good. Well, maybe *intellectually* I expected it—like if you had asked me, I would have *said* and *believed* that the cake would be really good. But I didn't *viscerally* expect it. What do I mean by that? What's the difference? The things I *viscerally* expect are over on the “plan assessor” side. People don't have conscious control over their plan assessors—the latter are trained exclusively by the “ground truth in hindsight signals” from the brainstem. I can manipulate the assessors a bit on the margin, e.g. by re-framing the way I think about things (see [here](#)), but by and large they're doing their own thing, independent of what I *want* them to be doing. (From an evolutionary perspective, this design makes good sense as a defense against wireheading—see [here](#).)

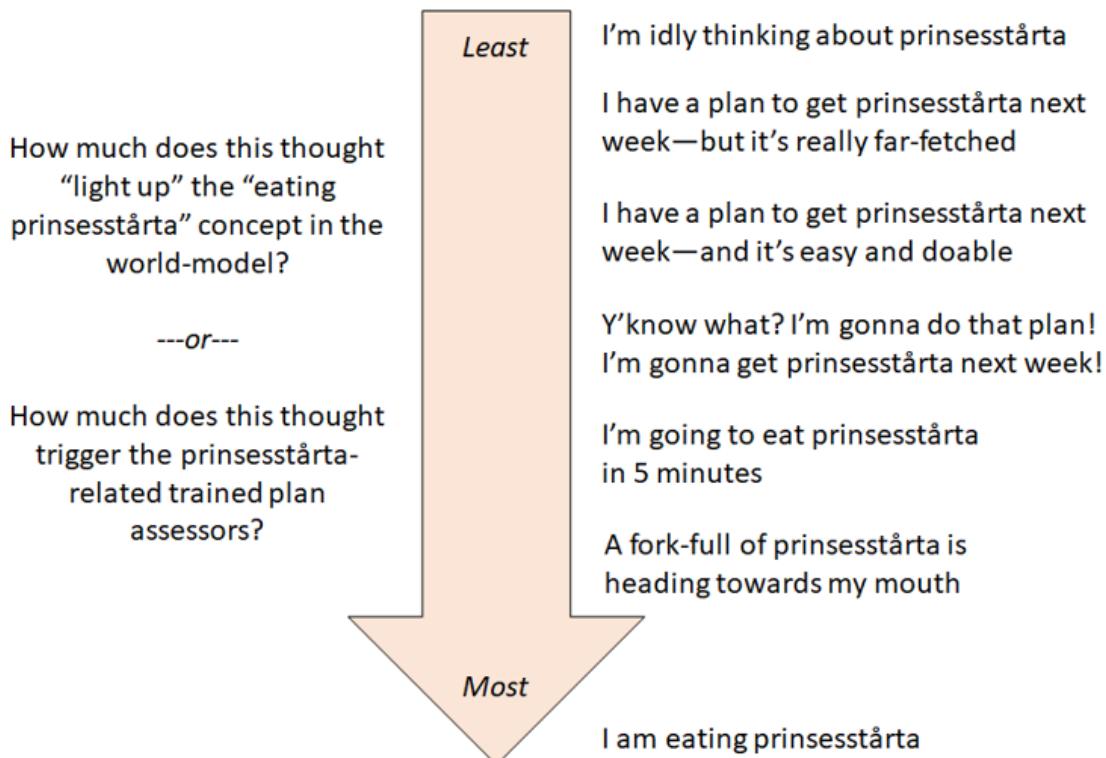
So when I bit into the cake, my “plan assessors” were wrong! They expected the cake to cause *mild* “yummy”-related autonomic reactions, but in fact it caused *intense* “yummy”-related autonomic reactions. And the brainstem *knows* that the plan assessors were wrong. So it sends correction signals up to the “plan assessor” algorithms, as shown in the diagram above. Those algorithms thus edit themselves so that next time I bring a fork-full of prinsesstårta towards my mouth, they will be more liable to predict intense hormones, goosebumps, and all the other reactions that I did in fact get.

A cool thing just happened here. We started with a simple-ish hardwired algorithm: brainstem / hypothalamus circuits turning certain types of taste inputs into certain hormones and autonomic reactions. But then we transferred that information into *functions on the learned world-model*—recall that giant inscrutable database I was talking about in the previous section. This step, which we might call “credit assignment”, is a bit fraught; it’s some hardcoded model-updating algorithm (akin to backprop, although probably not literally that), and it’s based on heuristics, and it might sometimes assign credit to the wrong thing—cf. superstitions!

But in this case, the credit-assignment (a.k.a. plan-assessment model update) process went smoothly: I already had some world-model concept that we might describe as “myself eating prinsesstårta”, and probably the main change was: from that point on, the plan assessors will know that whenever the “myself eating prinsesstårta” concept “lights up” in the world-model, they should issue predictions of the corresponding hormones and other reactions.

3. Planning towards goals via reward-shaping

I don’t have a particularly rigorous model for this step, but I think I can lean on intuitions a bit, in order to fill in the rest of the story:



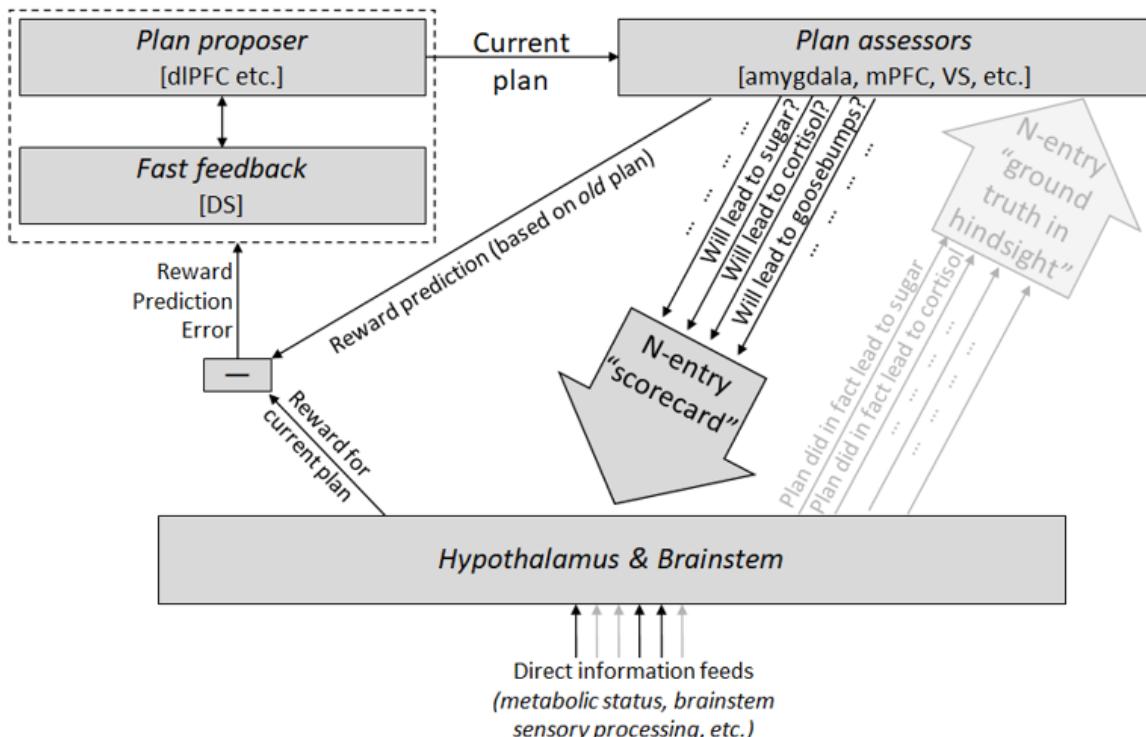
Remember, ever since my first bite of prinsesstårta two years ago in Step 2, the “plan assessors” in my brain have been looking at each thought I think, pattern-matching that thought to the “myself eating prinsesstårta” world-model concept, and to the extent that it’s a match, issuing a suggestion to prepare for delightful hormones, salivation, goosebumps, and so on.

The diagram above suggests a series of thoughts that I think would “pattern-match” better and better, from top to bottom.

To get the intuition here, maybe try replacing “prinsesstårta” with “super-salty cracker”. Then go down the list, and try to feel how each thought would make you salivate more and more. Or better yet, replace “eating prinsesstårta” with “asking my crush out on a date”, go down the list, and try to feel how each thought makes your heart rate jump up higher and higher.

Here's another way to think about it: If you imagine the world-model being vaguely like a [PGM](#), you can imagine that the “degree of pattern-matching” corresponds roughly to the probability assigned to the “eating prinsesstårta” node in the PGM. For example, if you're confident in X, and X weakly implies Y, and Y weakly implies Z, and Z weakly implies “eating prinsesstårta”, then “eating prinsesstårta” gets a very low but nonzero probability, a.k.a. weak activation, and this is kinda like having a far-fetched but not completely impossible plan to eat prinsesstårta. (Don't take this paragraph too literally, I'm just trying to summon intuitions here.)

OK, if you're still with me, let's go back to my decision-making model, now with different parts highlighted:



Again, every time I think a thought, the hypothalamus & brainstem look at the corresponding “scorecard”, and issue a corresponding reward. Recall also (see [here](#)) that the active thought / plan gets thrown out when its reward prediction error (RPE) is negative, and it gets kept and strengthened when its RPE is positive.

Let's oversimplify for a second, and say that the relevant prinsesstårta-related "assessments" comprise just one entry on the scorecard: "*Will lead to feel-good hormones*". And let's also assume the brainstem follows the simple rule: "The higher that a plan / thought scores on the 'Will lead to feel-good hormones' assessment, the higher the reward I'm gonna give it". Well in that case, each time our thoughts move down the ranked list above—from idle musing about prinsesstårta, to a far-fetched plan to get prinsesstårta, to a plausible plan to get prinsesstårta, etc.—there's an *immediate positive* RPE, so that the new thought gets strengthened, and gets to establish itself. And conversely each time we move back *up* the list—from plausible plan to far-fetched plan to idle musing—there's an *immediate negative* RPE, so that thought gets thrown out and we go back to whatever we were thinking before. It's a ratchet! The system naturally pushes its way down the list, making and executing a good plan to eat cake.

(By the way, the plan-proposing algorithm on the top-left is NOT trying to maximize the sum of future rewards—see [here](#), specifically the discussion of TD learning. Instead, its job is more like "maximize RPE right now".)

So **there you have it!** From this kind of setup, I think we're well on the way to explaining the full suite of behaviors associated with humans doing foresighted planning towards explicit goals—including knowing that you have the goal, making a plan, pursuing instrumental strategies as part of the plan, replacing good plans with even better plans, updating plans as the situation changes, pining in vain for unattainable goals, and so on.

By the way, I oversimplified above by reducing the *actual* suite of prinsesstårta-related assessments with "will lead to feel-good hormones". In reality, it's more specific than that—probably some assessment related to salivating, and some other assessment related to releasing certain digestive enzymes, and various hormones, and goosebumps, and who knows what else.

Why does that matter? Well, imagine you're feeling nauseous. Of course your hypothalamus & brainstem *know* that you're feeling nauseous. And meanwhile the assessment functions are telling the hypothalamus & brainstem that this plan will lead to eating food. Some hardwired circuit says: "That's bad! Whatever thought you're thinking, I'm gonna dock some reward points for its possibly leading to eating, in my current state of nausea."

...And indeed, I think you'll find that you're much less intrinsically motivated to make plans to get prinsesstårta, when you're currently feeling nauseous. Maybe you'll do it anyway, because thoughts can be quite complicated, and you have *other* motivations in life, and those motivations *also* feed into these assessment functions and get weighed by the brainstem. So maybe you'll proceed with the plan, driven by the motivation of "trying to avoid later regret, if I miss the deadline to order prinsesstårta in time for the party next week". So you'll order the cake anyway, despite it feeling kinda gross right now.

The two-headed bacterium

This is a linkpost for <https://www.telescopic-turnip.net/essays/the-two-headed-bacterium/>

I like to see categories as fish nets we use to capture ideas. We classify things into categories like individuals, nation or species, and of course it is all arbitrary and doesn't correspond to anything in the real world. But categories still form useful chunks we can use to make sense of the world. Furthermore, here is a fun exercise: introduce arbitrary *changes* in the categories, and see what the world looks like through this new lens. As I will argue, there are plenty of things to be discovered this way. Use the standard fish nets, and you get a standard understanding of the world. Try to use slightly larger or smaller nets, and maybe you will discover things you had never noticed before.

Take the *individual*, for example. One bacterial cell contains exactly one genome and all the necessary equipment to replicate it. Using our human-derived intuition of what makes an individual, it makes sense to see bacteria as unicellular organisms, meaning that *one cell = one individual*. If you visit the [wiki page on prokaryotes](#) (the larger group that encompasses bacteria and archaea) the first thing you hear is that they are unicellular, as if it were the most important thing about them. However, bacteria are so weird, so different from us, that it makes little sense to describe them using the categories we invented while observing humans.

Let's explore the strange and surprising processes that are uncovered when you change your definition of the individual to make it either wider, or narrower. First, I will start with a *hot take*: each bacterial lineage is one big multicellular individual. Then I will move on to the *super-hot-magma-take*: each bacterial cell is actually made of two distinct individuals fused together, facing in opposite directions.

Bacteria as multicellular organisms

First, let's make our definition of the individual arbitrarily broader, and consider that the whole bacterial culture, descending from a single ancestral cell, is one individual. Is there anything interesting to see here? For starters, some behaviors of bacterial cells don't really make sense as individuals. For example, bacterial cells regularly perform what could only be described as bacterial sacrifice.

The Kelly criterion in prokaryotes

Content warning: bacterial sacrifice

Antibiotics were already in the environment long before humans started using them, usually secreted by other micro-organisms who want to take your precious nutrients for themselves. Imagine being a bacterium growing peacefully – there is always a risk that some bastard fungus will put their filthy pterulone, sparassol or strobilurin in your soup. Fortunately, bacteria figured out a solution: enemies can't stop you from growing if you are already not growing.

In its simplest form, this works because the antimicrobial compound needs to be actively incorporated in the growth machinery to cause trouble. Think of a grain of

sand being caught in a clockwork mechanism and breaking everything – if the mechanism is stopped, the grain of sand doesn't enter, and you can resume operation later once the grain of sand has been blown away. Obviously, the drawback is that the bacterium is no longer growing, which kind of defeats the whole point. This is why bacteria have invented what we humans know as the Kelly betting system.

Say a gambler bets on something with 2:1 odds, so if she wins the bet, she gains twice as much as what she invested. She knows she has a 60% chance of winning, so the most profitable strategy is of course to invest 100% of her money every time – this way, she maximizes the return of every winning bet! But obviously this is bad, because eventually she will lose a bet, and then have zero monies remaining. For bacteria, this is like having 100% of the cells growing as fast as they can. This maximizes the population growth rate, until the aforementioned bastard fungus secretes some pleuromutilin or whatever and then the entire population takes it up and goes extinct. To avoid this, our gambler should invest only a fraction of her money on each bet, so her funds still grow exponentially (albeit at a slower rate) but in case of loss she still has some funds to continue. For bacteria, this means always having a small fraction of the population [that stops growing](#), as a backup. This is essentially the bacterial population [betting on whether there will be antibiotics in the close future](#). From the perspective of an individual cells, both situations are bad – either you stop growing, while your friends quickly outnumber you by orders of magnitudes and you practically disappear, or you are part of the growing fraction and eventually you die from antibiotic overdose. But if you look at the entire colony, you can see the two sub-populations as two essential parts of a single organism, that figured out some slick decision theory techniques long before the species of John L. Kelly even evolved a brain.

Eating the corpses of your siblings

Content warning: eating the corpses of your siblings.

Similarly, one puzzling feature of bacteria is that they sometimes commit apoptosis. This happens, for example, when food is scarce – some cells may [spontaneously explode](#) so that other cells can [feed on their remains](#), increasing the chances that at least one of them will make it out alive when resources come back. If you see each cell as an individual, that is weird, and does not fit well with anything [methodological individualism](#) would predict. But if you see the whole colony as the individual, then it is just like your good old typical apoptosis – just like, in the fetal stage, your fingers were all connected by cells until some of them honorably committed seppuku so you get born with fingers instead of [webbed paws](#).

(One fascinating thing with bacterial apoptosis is that every cell which ever activated these pathways is dead. Thus, if you look at a currently living bacteria, at no point in billion years of evolution did this pathway ever activate in any of its ancestors. Not even by chance. The entire mechanism evolved and improved only by correlation with other cells, without ever activating in the lineages we can now see.)

Action potentials in biofilms

As a third exhibit of things bacteria do that definitely don't look like unicellular behavior, there is the recent discovery that some bacteria, after organizing themselves as a [biofilm](#), are able to [communicate with each other using electrical waves](#). The way it works is remotely similar to the [action potentials](#) we see in neurons.

At a resting state, cells are filled with potassium ions, which makes them electrically polarized. Whenever the polarization disappears, ion channels in the envelope open up, and the potassium ions all exit the cell into the extracellular environment. This, in turns, cancels out the polarization of neighboring cells. The result is this:

[Video from Prindle et al., 2015, showing waves of potassium propagating in a colony of tens of thousands of cells.](#)

Supposedly, this mechanism makes sure the outer bacteria will stop eating from time to time, so the nutrients can diffuse all the way to the center and prevent the interior cells from starving. If this does not make you scream “multicellular!”, I don’t know what will.

In short, rather than being just individual cells fighting against each other, bacteria have evolved hard-wired mechanisms that only make sense if you consider the dynamics of the whole colony. A microbiologist could spend her entire career building a perfect model of one bacterial cell, but she would still be far from understanding all facets of the organism. Oh, and if you are ready to hear a similar point about humans (that is, human communities are multi-body individuals), get your largest fish net and check out [this review](#). I will continue with bacteria, because we have barely scratched the lipopolysaccharide of their weirdness.

Bacterial cells are two-faced pairs of individuals

Now, let’s see what happens with a much narrower definition for an individual. Even narrower than a single cell. Put down that extra-large “big game”-rated landing net and bring the tweezers.

Here is our new definition: an individual is what happens between a *birth event* and a *death event*. Now we need to find definitions of birth and death that apply to bacteria. Let’s say, a birth event is when a mother cell divides into two daughters (specifically, [cytokinesis](#)). A death event is when a cell is irreversibly broken, is torn apart or becomes too damaged to grow. We have a simple and precise definition, now we can look at bacteria and pick apart the individuals.

[Video of *E. coli* growing, seen under a microscope](#)

One generation goes as follows:

- The cell extends and roughly doubles in length
- The middle of the cell constricts and two new poles are constructed
- The cell divides and you get two cells. Each of them has one old pole that was already there in the previous generation, and one shiny new pole:

Where is the individual here? Now you understand why I came up with that bizarre birth-death definition. First, let’s number the poles according to their age (in generations).



Blink very fast while on shrooms and you might see a [Koch snowflake](#) in the bottom sequence.

But what if bacteria age? It turns out that, [yes, bacteria age](#). After a number of generations, old poles accumulate damage. Depending on the growth environment, they may still be fine, or grow slower, or explode in an effusion of bacteria blood. To reduce clutter, I'll consider that poles have a lifespan of 3 generations, and then the cell is dead (in real life, they hold for much longer, but that wouldn't be sketchable).

Coming back to our custom, "birth-to-death" definition of an individual, you can see that each cell is actually made of two of them – one on the left, one on the right.

Here they are very short-lived and die after three generations, but in real life these "half-bacteria" live for much longer, perhaps hundreds of generations if the conditions are not too bad. But the principle remains the same, there are just a lot more of these diagonal individuals.

Using your ancestors as trashcans

Content warning: yeah, that.

But wait, there is more. As I said, in nice conditions the poles can grow basically forever. Yet they still exhibit aging. And yes, this is all sane and coherent. This is where the titles of the papers become really spooky ([Age structure landscapes emerge from the equilibrium between aging and rejuvenation in bacterial populations](#) or [Cell aging preserves cellular immortality in the presence of lethal levels of damage](#)), showing how far we are from our typically construction of the individual.

To put it very briefly, take the sketches above where half of the cell is young and half of the cell keeps getting older. Old material accumulates in the old pole, so those cells keep growing slower and slower after each generation. Now add some mixing to it: every generation, the older pole gets a little bit of fresh material, and the younger pole gets a little bit of old material. Eventually the old pole reaches an equilibrium when the new material their inherit exactly compensates the damage from aging. As there is the same thing, reversed, for the young pole, you end up with two attractors:

Slightly adapted from Proenca et al., 2018.

What is the importance of this? There may be no importance at all, since the old cells are quickly outnumbered by young cells so they only represent a [tiny fraction](#) of the colony. However, there is also some evidence that all kinds of garbage, like misfolded proteins or aggregates, [tend to accumulate in the old pole](#). Perhaps this ensure that [at least some cells](#) in the population will be in perfect shape, so in case of trouble, they have a good chance of having at least one survivor (a bit like [North Korea preparing a team for the Math Olympiads](#)).

But this, of course, brings us back to collective, multicellular behavior. Life is too complicated to fit in a single fish net.

Mildly against COVID risk budgets

A friend is hosting a party tonight! It'll cost me 200 microCOVIDs, which, as a healthy thirty-something, I very cautiously estimate to cost about 2 [micromorts](#),^[1] which is roughly equivalent to 1 hour out of my remaining life expectancy. But I'm *super* excited for this party; I'd happily burn an hour of my life driving there and back. Should I go?

Unless something subtle is going on... obviously yes, right?

"*What about your visit to that coffee shop earlier today? Don't you think that's decision-relevant?*"

The coffee shop? I'm confused. Why would that be decision-relevant?

"*Well, you accumulated about 150 microCOVIDs there.*"

Sure, but that's not really relevant to the decision theory of whether to attend the party, is it? The party is equally enjoyable either way, and the costs of multiple COVID exposures add pretty much linearly by the [axiom of independence](#), so previously-incurred risks are irrelevant to my future decisions. (If the coffee shop had been a week ago, sure, I'd be inflicting some of those microCOVIDs on my fellow partygoers, which, sure, could be decision-relevant, I haven't done the math; but it seems very unlikely to me that I'll become a full-fledged germ factory between this morning and this evening, so I think that consideration is insignificant in this case.)

"*But you try to maintain a 200-microCOVID-per-week risk budget, don't you?*"

Sure, but... hmm.

"*So there's something wrong with your assertion that previously-incurred risks are irrelevant to future decisions.*"

...or something is wrong with the idea of risk budgeting.

"...hmm."

...so, which is it?

Zero externalities

Maybe the answer will be clearer if we simplify the problem: let's get rid of externalities. Suppose I am the world's most boring superhero, Captain Can't-Transmit-COVID. I can still *catch it* just like anybody else, but the cost is borne by me alone, not my friends or housemates.

Now is my coffee-shop visit decision-relevant for whether I should attend the party?

I can't think of a single reason it would be.^[2] This strongly suggests that "risk budgeting" has something to do with the risk of transmitting COVID to other people.

Externalities to an enthusiastic optimizer

Let's add an externality back in: suppose I can only transmit COVID to my partner. Perhaps risk budgets will fall naturally out of the arrangements we make.

Me: So. My best estimate for partner-to-vaccinated-partner transmission, given how often we see each other, is 30% over two weeks, so for every 10 microCOVIDs I accrue, 3 spill over onto you.

My partner, the enthusiastic optimizer: Indeed. Using the 100-to-1-to-0.5 exchange rate between microCOVIDs, micromorts, and life-hours, this means that you cost me about one hour of my life for every 600 microCOVIDs you accumulate. However, a good [Coasean](#) would point out that this externality is the result of a joint decision -- your decision to take risks, plus my decision to date you during your infectious period -- so I'm willing to split that cost 50/50. Post-dinner cleanup usually takes half an hour; how about, for every 600 microCOVIDs you accrue 2-12 days before a date, you handle cleanup on a night when it would normally fall on me?

Me: And vice-versa? Every 600 microCOVIDs that *you* accumulate, you--

Them: Of course.

Me: Sounds fair!

Them: Oh, and-- each of us has other connections, housemates and such, who might not want to interact with us when we're high-risk. The math isn't clear to me off the bat, but we might ballpark that as doubling the cost of each of our microCOVIDs. Change from 600 microCOVIDs per dishes to 300?

Me: Sure!

Hmm. No risk budgets in sight.

Externalities requiring costly discussions

Suppose I can also transmit to my housemate?

Me: So. My best estimate for housemate-to-vaccinated-housemate transmission is 30% over two weeks, so for every 10 microCOVIDs I accrue, 3 spill over onto you.

My housemate, the sick and tired of everything about COVID: ...okay.

Me: Alex and I worked out this system where, for every 300 microCOVIDs each of us accumulates in the infectiousness window before a date, we owe the other a get-out-of-doing-the-dishes token, to compensate them for the risk we're imposing--

Them:

Me: --and your facial expression right now confirms my expectation that you would hate this. Can you elaborate on why?

Them: Just... constantly thinking about COVID, weighing every social interaction, opening microcovid.org every time I go outside, keeping track of how many microCOVIDs I'm getting from you so that I can report that number to the two other people I regularly interact with... it would ruin my days. I just really, really want to not think about it anymore.

Me: Okay, sure, I can sympathize with that. I... I'm afraid that the "don't think about it at all" strategy has really high costs for me, since I'll need to make pessimistic assumptions about your COVID risk in order to uphold my agreements with Alex and other folks...

Them: I know, I know. Ugh. How about... I'll just... I won't keep *precise* counts of my microCOVIDs, but I'll keep enough of an eye on things to know the rough order of magnitude of my risk, and... I don't want to have to keep you constantly updated, but how about, you can safely use 200 microCOVIDs as your pessimistic person-risk for me, and if I need to do anything that puts me over that threshold, I'll let you know.

Me: Let's see, if I assume you always have 200, then I'll have 60 from you on each my dates with Alex, so this arrangement will cost me an average of 6 minutes of dishes-doing per date... yeah, all right, I think I can make this work.

That's starting to sound like a budget!

So, at least in this case, it seems like risk-budgeting is a technique that lets you reduce the amount of time you spend computing/communicating risks, at the cost of sometimes making suboptimal decisions (e.g. skipping an awesome party) because you're thinking in terms of "whether I'll exceed my budget" instead of the underlying costs/benefits.

Further consequences

So far, we've only explained why people who hate doing risk-discussions keep budgets; but I know several *non-risk-discussion-hating* people who keep budgets. Why would they do that? Hmm...

My housemate, later: Hey, I was thinking about what you said, about your arrangement with Alex and "every 300 microCOVIDs" -- it sounds like you're keeping the option open to sometimes take on pretty large amounts of risk.

Me: Yeah, probably not often, but it's a possibility.

Them: And... you taking 300 microCOVIDs would give *me* about 100, a significant chunk of my budget, possibly forcing me to have risk-management conversations I find painful with the other people I've told they can safely bound my risk at 200.

Me: Hmm. Yeah, that sounds accurate. So, if *I* ever plan to go above... 200, say, which is about 60 for you... then perhaps I should let *you* know -- which imposes a cost on you, for the emotional toll of both that conversation and the downstream conversations you might have to have with those other people.

Them: Yeah.

Me: So, figuring that those conversations will be about... two hours of unpleasantness, all told?... a cost which I'm willing to split with you 50/50... if I ever go above 200, then I need to save you one hour of time. Payable by doing chores?

Them: Yeah, that sounds good.

So now I have something resembling a budget, even though I, personally, have no particular distaste for risk-discussions.

Conclusions

- For people whom risk-discussions cause significant pain, **risk budgets serve as a tool for reducing the amount you have to talk with people about COVID**: set a limit on how much risk you will ever take on, and let other people assume you're at that level. When you go over that limit, you tell your contacts about it, and endure the painful discussions.
- For people who both associate closely with people whom risk-discussions cause significant pain, **something that resembles a risk budget falls straightforwardly out of cold hard utility-maximization**: if you go above a certain threshold, then you have to have notify a close-contact risk-discussion-hater; which causes them pain; which you probably either feel bad about or compensate them for; which reduces your utility.
- For people who aren't close contacts with risk-discussion-haters, **I don't think risk budgets make sense**. Instead, you simply track your microCOVIDs and share your risk info with your contacts, so everybody can do normal cost/benefit analyses like they're hopefully used to, without "budgets" appearing anywhere in the calculus.^[3] You probably do not have enough close contacts for the communication overhead to be significant.

-
1. 100-to-1 is a guess at what my audience, i.e. you, thinks the microCOVID-to-micromort exchange rate is, judging by things like microcovid.org's recommended budget of 200/week. As best I can figure, though, for a healthy 30-year-old the actual exchange rate is about an order of magnitude higher, even including non-death outcomes like long COVID. I'm writing this footnote to assuage my guilt over contributing to a false appearance of consensus around "100 to 1" when I don't endorse it. ↪
 2. Edit: okay, I thought of a reason: maybe I don't trust myself to make smart cost-benefit assessments on a case-by-case basis, and I think my brain will routinely exaggerate the benefits of risky activities, and my psychology is such that drawing a bright line and saying "this is my budget, I can't go over this" will at least bound how much risk my lying brain can trick me into. But I think that few of the people I see maintaining "risk budgets" would give this as the reason. ↪
 3. Well, maybe your friends will have "under-300-microCOVID-people-only" parties, and those will add discontinuities to your utility function that look kind of like budgets, but I think the similarity is only superficial. ↪

Framing Practicum: Stable Equilibrium

This is a [framing practicum](#) post. We'll talk about what a stable equilibrium is, how to recognize stable equilibria in the wild, and what questions to ask when you find one. Then, we'll have a challenge to apply the idea.

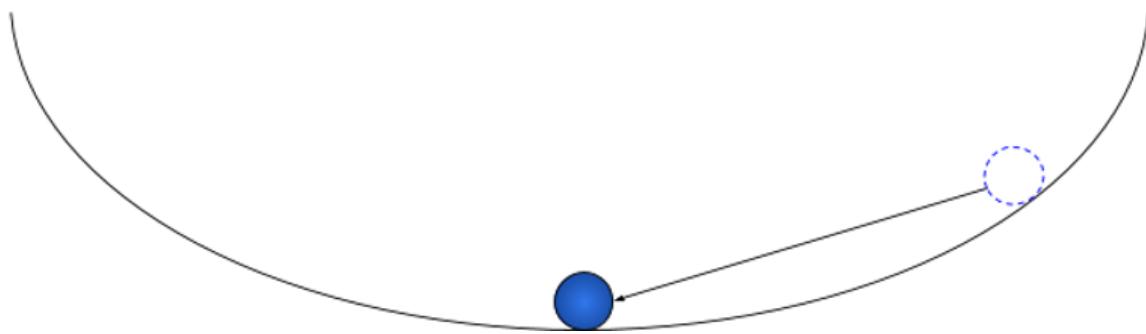
Today's challenge: come up with 3 examples of stable equilibrium which do not resemble any you've seen before. They don't need to be good, they don't need to be useful, they just need to be novel (to you).

Expected time: ~15-30 minutes at most, including the Bonus Exercise.

What's a Stable Equilibrium?

Put a marble at the bottom of a round bowl, and it will just sit there without moving. Put it in the bowl but not quite at the bottom, and it will roll around a bit, but eventually settle at the bottom, and sit there without moving. Give it a poke, and it will roll around some more, but eventually it will again sit at the bottom without moving.

This is *stable equilibrium*: the system may start in different states, or it may be “perturbed” into different states by some external force, but eventually it settles back to the same state (assuming it isn’t pushed *too far away...*).



A marble in a bowl will eventually sit stationary at the bottom of the bowl, and stay there.

What To Look For

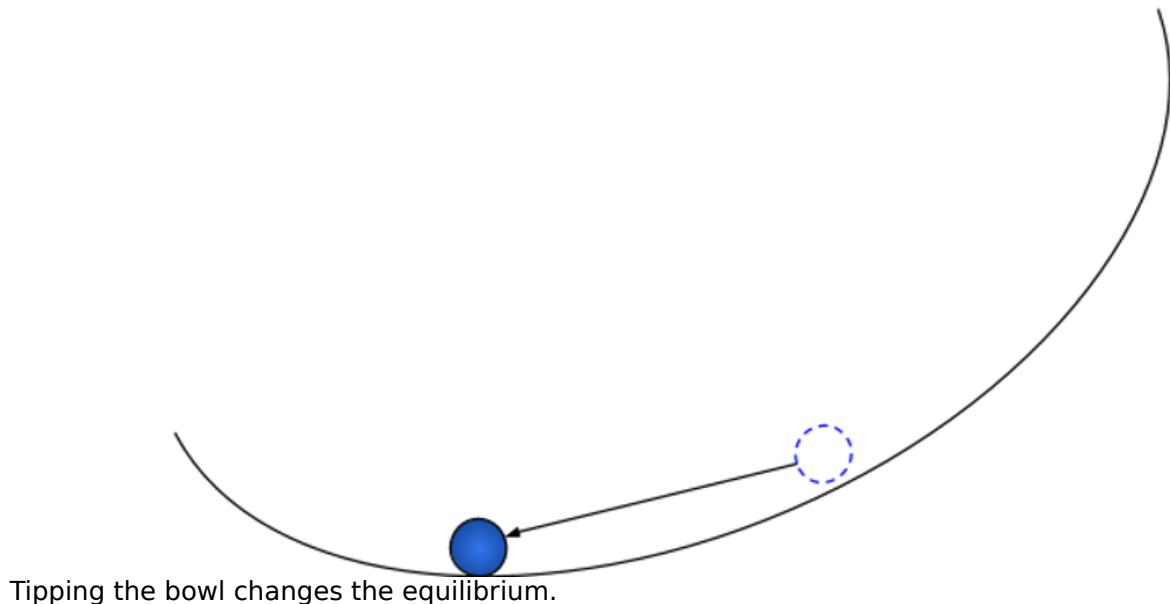
Stable equilibrium should spring to mind whenever a system tends to return to the same state. If you could “poke” it somehow, and the system would go back to normal eventually,

that's probably a stable equilibrium. If a system tends to stay suspiciously the same over the long run, despite lots of short-run noise, that's probably a stable equilibrium.

Useful Questions To Ask

The marble always returns to the bottom of the bowl. If we push the marble away from the bottom, that's only a short-term change - it will roll back down eventually. So, if we're mainly interested in the *long run* behavior of the marble, then we can ignore such little pushes.

On the other hand, there may also be ways to change the equilibrium state itself. For instance, if we tip the bowl to the side slightly, then the equilibrium position of the marble will change. If we deform the bowl, that could change equilibrium position. If we charge the marble with a little static electricity, then place another charged object near the bowl, that could also change the equilibrium. Finally, very large changes to the system state could push it out of the bowl entirely.



When we frame something as a stable equilibrium, we ignore temporary changes to the system state, and only pay attention to things which change the equilibrium.

Two main ways this can apply:

- We want to change the long-term behavior of a system. So, we focus on things which can change the equilibrium, and ignore things which don't.

- We see a change in the long-term behavior of a system, and want to know what caused it. So, we focus on things which can change the equilibrium, and ignore things which don't.

The Challenge

(Rules adapted from the [Babble Challenges](#))

Come up with 3 examples of stable equilibrium which do *not* resemble any you've seen before. They don't need to be good, they don't need to be useful, they just need to be novel (to you). I recommend mentioning what the equilibrium is, and a few ways you could "poke" the system for which it would return to equilibrium afterwards, so that everyone understands your example.

Any answer must include at least 3 to count, and they must be novel to you. That's the challenge. We're here to challenge ourselves, not just review examples we already know.

However, they don't have to be very good answers or even correct answers. Posting wrong things on the internet is scary, but a very fast way to learn, and I will enforce a high bar for kindness in response-comments. I will personally default to upvoting every complete answer, even if parts of it are wrong, and I encourage others to do the same.

Post your answers inside of spoiler tags. ([How do I do that?](#))

Celebrate others' answers. This is really important, especially for tougher questions. Sharing exercises in public is a scary experience. I don't want people to leave this having back-chained the experience "If I go outside my comfort zone, people will look down on me". So be generous with those upvotes. I certainly will be.

If you comment on someone else's answers, focus on making exciting, novel ideas work — instead of tearing apart worse ideas. [Yes, And](#) is encouraged.

Reward people for babbling — don't punish them for not pruning.

I will remove comments which I deem insufficiently kind, even if I believe they are valuable comments. I want people to feel encouraged to try and fail here, and that means enforcing nicer norms than usual.

If you get stuck, look for:

- Systems which go back to normal after you poke them
- Systems which stay suspiciously the same over time despite lots of short-term noise

Bonus Exercise: for each of your three examples from the challenge, suppose you want to change the equilibrium, or you want to know what caused a change in the equilibrium. What factors should you pay attention to (since they can change the equilibrium)? What factors can you safely ignore (since they only affect the system in the short term)?

This bonus exercise is great blog-post fodder!

Motivation

Much of the value I get from math is not from detailed calculations or elaborate models, but rather from *framing tools*: tools which suggest useful questions to ask, approximations to make, what to pay attention to and what to ignore.

Using a framing tool is sort of like using a [trigger-action pattern](#): the hard part is to notice a pattern, a place where a particular tool can apply (the “trigger”). Once we notice the pattern, it suggests certain questions or approximations (the “action”). This challenge is meant to train the trigger-step: we look for novel examples to ingrain the abstract trigger pattern (separate from examples/contexts we already know).

The Bonus Exercise is meant to train the action-step: apply whatever questions/approximations the frame suggests, in order to build the reflex of applying them when we notice a stable equilibrium.

Hopefully, this will make it easier to notice when a stable equilibrium frame can be applied to a new [problem you don't understand](#) in the wild, and to actually use it.

Thankyou to Sisi, Eli, Adam and especially Jacob for beta-testing and feedback. Also thankyou to Aysajan for our daily discussions, which led to this concept.

Framing Practicum: Incentive

Credit

An enormous amount of credit goes to [johnswentworth](#) who made this new post possible.

This is a [framing practicum](#) post. We'll talk about what incentives are, how to recognize incentives in the wild, and what questions to ask when you find them. Then, we'll have a challenge to apply the idea.

Today's challenge: come up with 3 examples of incentives which do not resemble any you've seen before. They don't need to be good, they don't need to be useful, they just need to be novel (to you).

Expected time: ~15-30 minutes at most, including the Bonus Exercise.

What Are Incentives?

At the beginning of a sowing season, the Government of India announces a list of guaranteed purchase prices for certain crops, e.g., rice, wheat, cotton, etc., to support farmers. In case the market price for a crop falls below the guaranteed purchase price, the government agencies purchase the entire quantity from farmers if the crop quality meets a minimum quality threshold. From an Indian farmer's perspective, the farmer is encouraged to produce price supported crops with quality just above the threshold level set by the government - and no higher.

This is an economic *incentive*: There is a reward signal in the system. Farmers are rewarded for producing crops just above the quality threshold. On the other hand, they are not rewarded for producing higher quality crops. Here we see the defining features of incentives: A system (a farmer) "wants" some resource (money), and can get more of that resource in return for some actions (producing crops with quality just above the threshold level) than others (producing crops with quality well above the threshold level).

Another example, with a direct notion of "reward": cash incentive for taking Covid-19 vaccine shots. Some states in the US are offering rewards for Covid-19 vaccination in the form of direct cash or lottery programs. We can identify a clear reward signal in the system, which is people get rewarded for taking vaccines. Here again we see the defining features of incentives: A system (a human) "wants" some resource (money), and can get more of that resource in return for some actions (taking Covid-19 vaccines) than others (not taking Covid-19 vaccines).

What To Look For

In general, incentives should come to mind whenever there is some kind of reward signal. A system "wants" some resource, and can get more of that resource in return for some actions than others.

Useful Questions To Ask

In the farmers support price example, the Government of India announces a minimum quality requirement for the crops to be purchased. Crops with lower quality will not qualify for the support program. On the other hand, farmers are not rewarded for having high quality crops. As a result, farmers will not only avoid the work required to produce higher quality crops, they will even make their crops' quality worse: farmers with high quality crops will mix small rocks, or leftover crops from previous years into their harvested crops to increase the total quantity of "crop", and thus total revenue. Obviously the Government of India did not intend for farmers to throw gravel into their crops, but they accidentally incentivized it anyway.

In general, whenever we see incentives, we should ask:

- **What actions are getting rewarded?**
- **What counterintuitive or unintended actions achieve high reward?**

What about the cash-reward-for-Covid-19-vaccine example? If there is someone who is urgently in need of money, that person might fake the vaccination status in order to receive rewards more than once.

The Challenge

Come up with 3 examples of incentives which do not resemble any you've seen before. They don't need to be good, they don't need to be useful, they just need to be novel (to you).

Any answer must include at least 3 to count, and they must be novel to you. That's the challenge. We're here to challenge ourselves, not just review examples we already know.

However, they don't have to be very good answers or even correct answers. Posting wrong things on the internet is scary, but a very fast way to learn, and I will enforce a high bar for kindness in response-comments. I will personally default to upvoting every complete answer, even if parts of it are wrong, and I encourage others to do the same.

Post your answers inside of spoiler tags. ([How do I do that?](#))

Celebrate others' answers. This is really important, especially for tougher questions. Sharing exercises in public is a scary experience. I don't want people to leave this having back-chained the experience "If I go outside my comfort zone, people will look down on me". So be generous with those upvotes. I certainly will be.

If you comment on someone else's answers, focus on making exciting, novel ideas work — instead of tearing apart worse ideas. [Yes, And](#) is encouraged.

I will remove comments which I deem insufficiently kind, even if I believe they are valuable comments. I want people to feel encouraged to try and fail here, and that means enforcing nicer norms than usual.

If you get stuck, look for:

- Environments in which there exists some kind of reward signal.
- Systems that “want” certain actions to be taken
- Agents that “want” some resource, and can get more of that resource in return for some actions than others.

Bonus Exercise: for each of your three examples from the challenge, explain:

- What other counterintuitive actions are getting rewarded?

This bonus exercise is great blog-post fodder!

Motivation

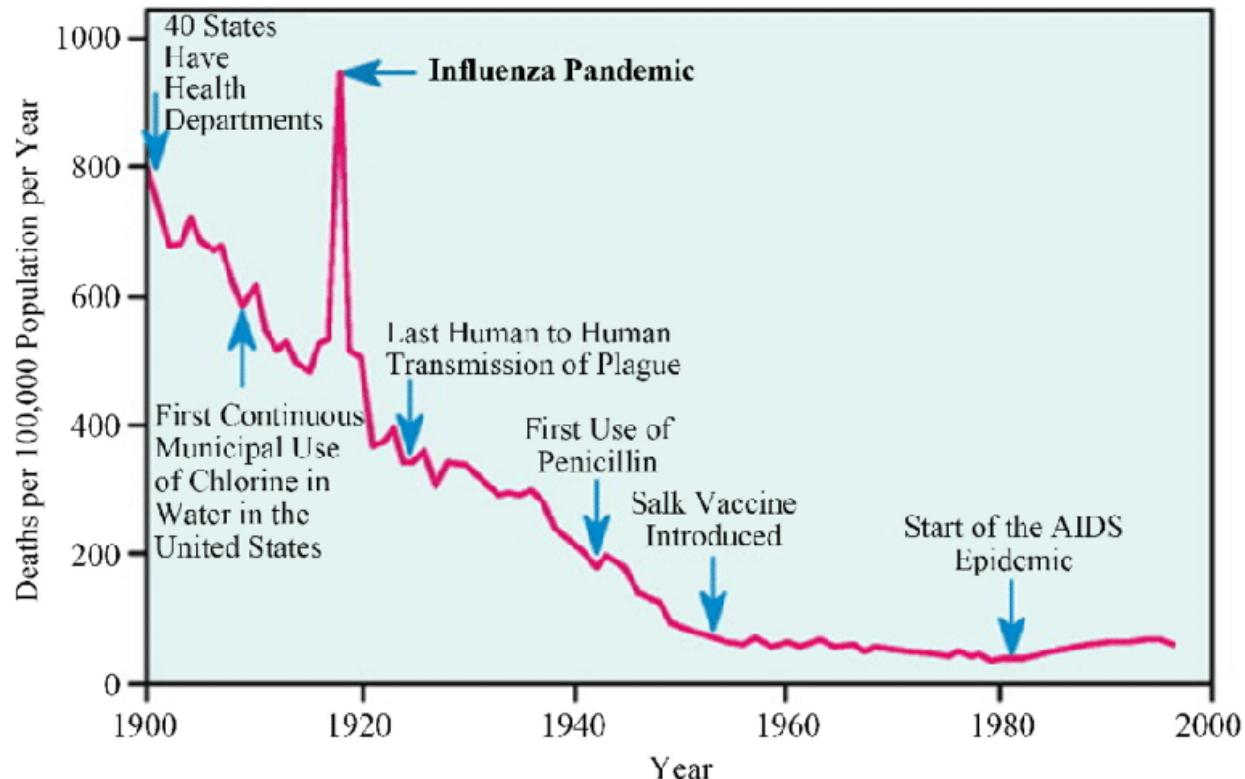
Using a framing tool is sort of like using a [trigger-action pattern](#): the hard part is to notice a pattern, a place where a particular tool can apply (the “trigger”). Once we notice the pattern, it suggests certain questions or approximations (the “action”). This challenge is meant to train the trigger-step: we look for novel examples to ingrain the abstract trigger pattern (separate from examples/contexts we already know).

The Bonus Exercise is meant to train the action-step: apply whatever questions/approximations the frame suggests, in order to build the reflex of applying them when we notice incentives.

Hopefully, this will make it easier to notice when an incentive frame can be applied to a new [problem you don’t understand](#) in the wild, and to actually use it.

Humanity is Winning the Fight Against Infectious Disease

One of my favorite things about living in the 21st century United States is how so few people die of infectious disease. Here's a graph of deaths showing how infectious diseases decreased by an order of magnitude between the turn of the century and Salk's Polio Vaccine.



[In 1900, 12% of deaths were caused by pneumonia and influenza. 11% were caused by tuberculosis. 8% were caused by diarrhea.](#) A hundred years ago infectious diseases killed over one third of everyone. In contrast, not a single person in my extended social circle has ever died of an infectious disease. I know more people who were killed by airplane crashes than by infectious diseases.

Mortality and Top 10 Causes of Death, USA, 1900 vs. 2010

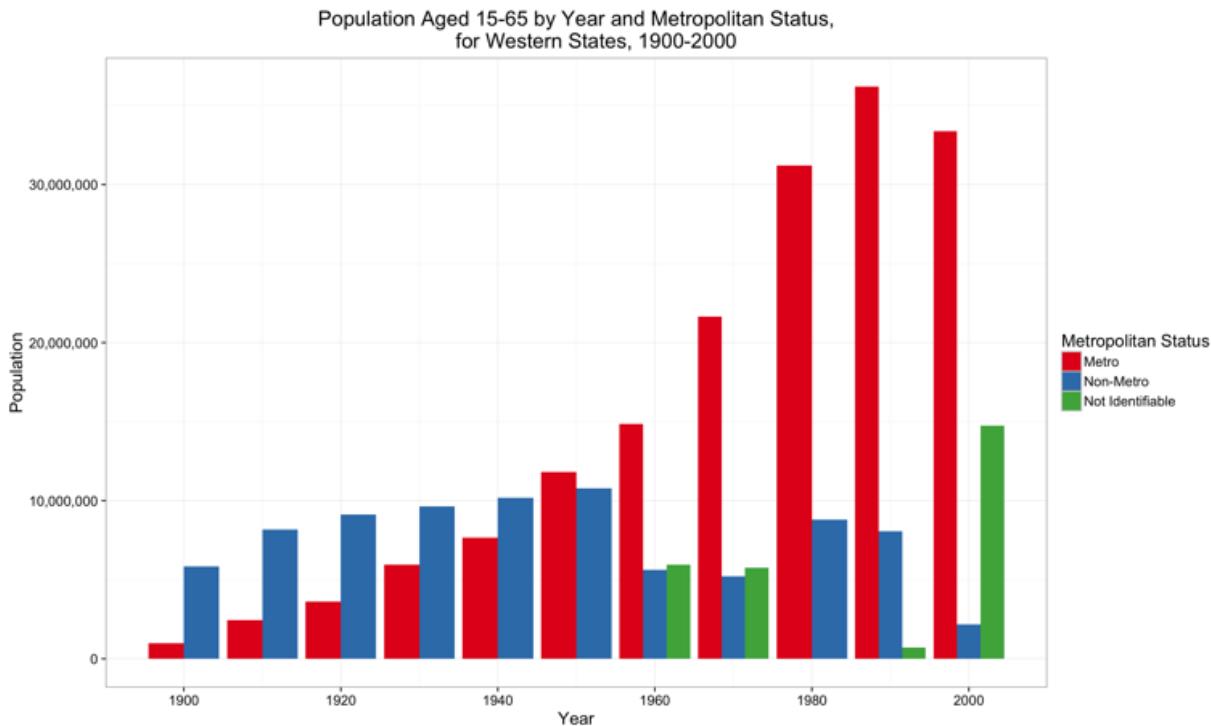
(Rates per 100,000)



Data Source: Centers for Disease Control



The raw 10× reduction undersells the progress we've made against infectious disease. We have more people overall *and* we live together in cities. Death due to infectious disease should have skyrocketed over the last century. We would have been fortunate if medicine and sanitation merely kept infectious disease at 1900 levels.



The emergence of new diseases like HIV and COVID-19 has had a tiny impact compared to the overall trend.

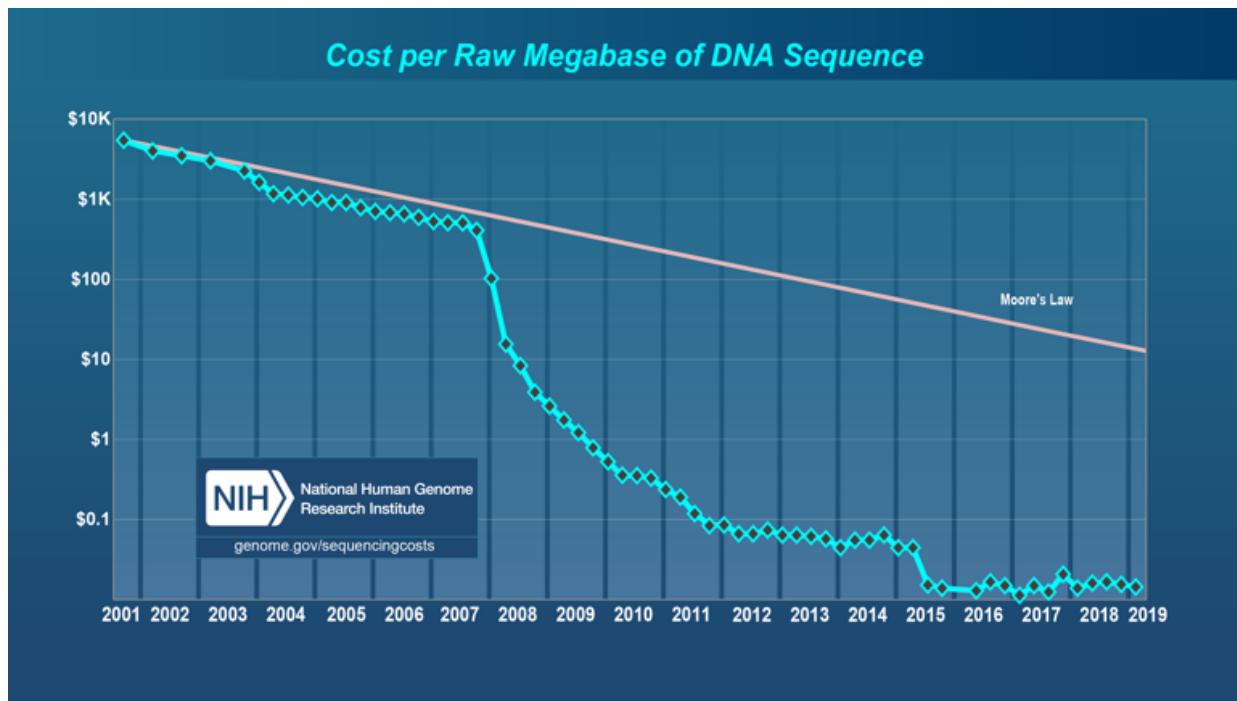
In 2019, there were 15,815 deaths among adults and adolescents with diagnosed HIV in the United States and 6 dependent areas. These deaths may be due to any cause.

—hiv.gov

[2,855,000 Americans died in 2019.](#) Those with HIV were 0.5% of the total. Even if we are maximally pessimistic and attribute to HIV every single death of every single American with HIV, this is a tiny fraction compared to how many people used to die of the plagues of 1900.

As of July 20, 2021, COVID-19 killed 600,000 Americans over the course of 1.5 years. That's 400,000 deaths per year. At 13% of total deaths, COVID-19 temporarily increased deaths due to infectious disease back to what we had in the 1940s. But the increase is temporary. We have vaccines. Even if we didn't have vaccines, deaths due to COVID-19 would decrease naturally after it burned through the vulnerable population.

If new diseases emerged while technology stagnated then we'd be living in a temporary golden age—and maybe we are. But biotechnology is advancing faster than any technology ever.



The COVID-19 vaccines were invented and deployed with unprecedented speed. Civilization will only get better at inventing vaccines. I mean, can you plausibly imagine a worse response to COVID-19 than what we experienced in the West?

With biotechnology advancing so quickly, a [deliberately engineered bioweapon](#) (or gain-of-function research) will soon become a greater threat than naturally-occurring pandemics (if it hasn't already).

A bioterror attack is scary, but I don't think it's likely to kill as many Americans in a single year as did ordinary 1900-level tuberculosis. COVID-19 got as bad as it did because we lacked the will to respond appropriately. (China took the pandemic seriously and is doing fine despite being patient zero.) If COVID-19 had been a terrorist attack, the American population would have rallied behind contact tracing and pandemic prevention would have gotten a blank check.

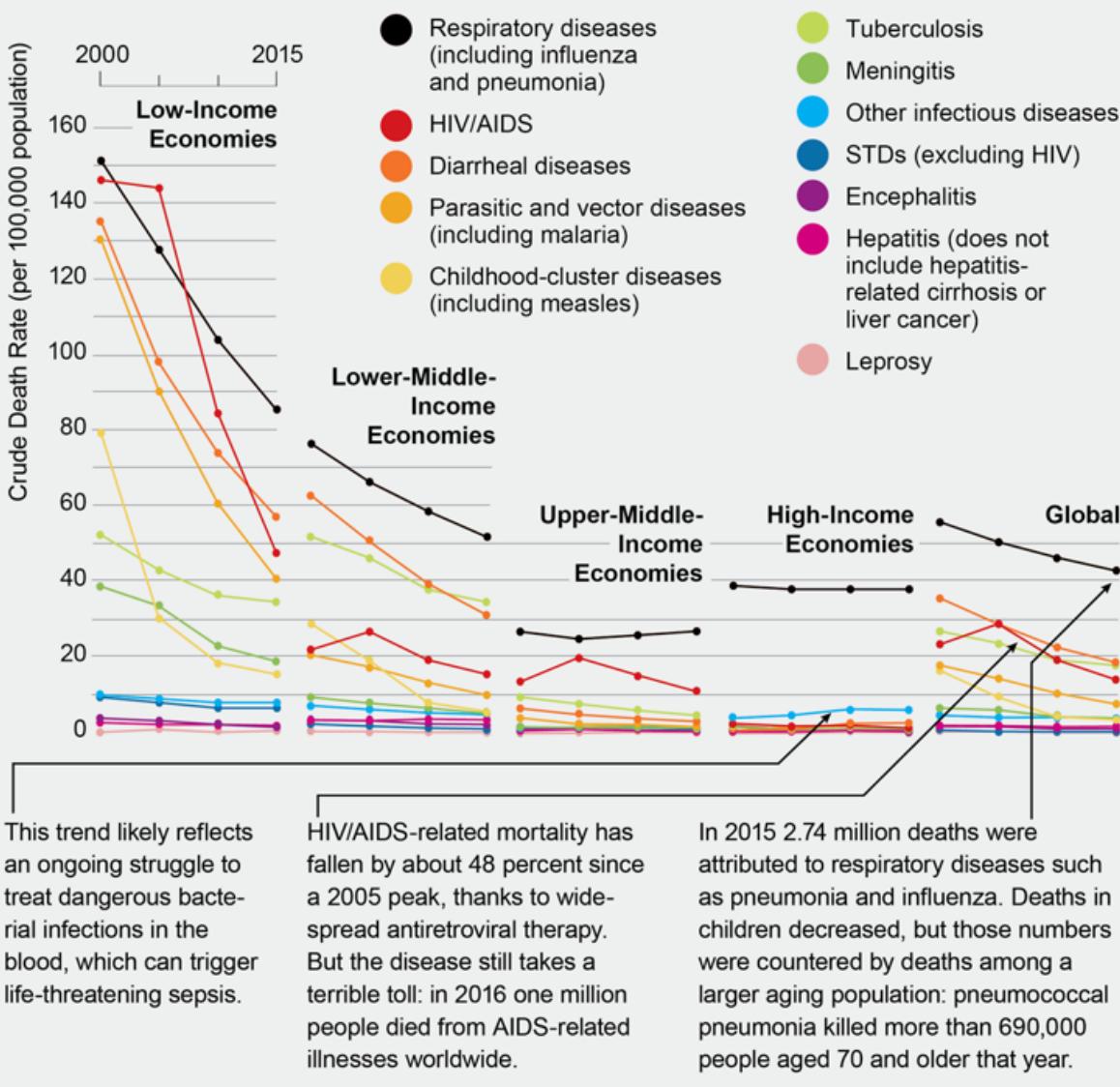
The World

The United States is better protected against infectious disease than ever before. What about the rest of the world? Here is a graph of deaths due to childhood infectious diseases.

Unlike the United States where deaths due to infectious diseases leveled off several decades ago, global childhood deaths due to major infectious diseases have decreased for decades. I predict they will continue to do so. Global deaths due to infectious disease for all people (not just children) have decreased too, especially in the poorest countries. Deaths from the worst infectious diseases in the poorest countries halved in just fifteen years.

Global Mortality Drops but Differs by Economy

When the countries of the world are divided by economy type (defined by the World Bank), some distinctions in death rates stand out. Low- and lower-middle-income countries, such as Haiti and India, started high and showed a steep drop in mortality during the first 15 years of this century, according to World Health Organization data. The wider availability of medical care, as well as drugs to combat infections, played an important role. HIV/AIDS deaths declined dramatically after 2005, coinciding with a U.S.-led initiative to provide care, including antiretroviral medication, to poorer countries. Upper-middle- and high-income countries, such as China and Germany, began with better care and thus did not show a sharp drop in deaths. Even so, well-off countries have had a difficult time controlling respiratory diseases such as pneumonia, which hits hard among the elderly and people with weakened immune systems.



If you look at things on time horizons of centuries—or even just decades—humanity is freer from infectious disease than it has ever been. The situation is rapidly improving too!

AI Risk for Epistemic Minimalists

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

Financial status: This is independent research, now supported by a grant. I welcome further [financial support](#).

Epistemic status: This is an attempt to use only very robust arguments.

Outline

- I outline a case for concern about AI that does not invoke concepts of agency, goal-directedness, or consequential reasoning, does not hinge on single- or multi-principal or single or multi-agent assumptions, does not assume fast or slow take-off, and applies equally well to a world of emulated humans as to de-novo AI.
- The basic argument is about the power that humans will temporarily or permanently gain by developing AI systems, and the history of quick increases in human power.
- In the first section I give a case for paying attention to AI at all.
- In the second section I give a case for being concerned about AI.
- In the third section I argue that the business-as-usual trajectory of AI development is not satisfactory.
- In the fourth section I argue that there are things that can be done now.

The case for attention

We already have powerful systems that influence the future of life on the planet. The systems of finance, justice, government, and international cooperation are things that we humans have constructed. The specific design of these systems has influence over the future of life on the planet, meaning that there are small changes that could be made to these systems that would have an impact on the future of life on the planet much larger than the change itself. In this sense I will say that these systems are powerful.

Now every single powerful system that we have constructed up to now uses humans as a fundamental building-block. The justice system uses humans as judges and lawyers and administrators. At a mechanical level, the justice system would not execute its intended function without these building-block humans. If I turned up at a present-day court with a lawsuit expecting a summons to be served upon the opposing party but all the humans in the justice system were absent then the summons would not end up being served.

The Google search engine is a system that has some power. Like the justice system, it requires humans as building-blocks. Those human building-blocks maintain the software, data centers, power generators, and internet routers that underlie it. Although individual search queries can be answered without human intervention, the transitive closure of dependencies needed for the system to maintain its power includes a huge number of humans. Without those humans, the Google search engine, like the justice system, would stop functioning.

The human building-blocks within a system do not in general have any capacity to influence or shut down the system. Nor are the actions of a system necessarily connected to the interests of its human building-blocks.

There are some human-constructed systems in the world today that do not use humans as building-blocks, but none of them have power in their own right. The [Curiosity Mars rover](#) is a system that can perform a few basic functions without any human intervention, but if it has any influence over the future, it is via humans collecting and distributing the data captured by it. The [Clock of the Long Now](#), if and when constructed, will keep time without humans as building-blocks, but, like the Mars rovers, will have influence over the future only via humans observing and discussing it.

Yet we may soon build systems that do influence the future of life on the planet, and do not require humans as building-blocks. The field concerned with building such systems is called artificial intelligence and the current leading method of engineering is called machine learning. There is much debate about what exactly these systems will look like, and in what way they might pose dangers to us. But before taking any view about whether these systems will look like agents or tools or AI services, or whether they will be goal-directed or influence-seeking, or whether they will be developed quickly or slowly, or whether we will end up with one powerful system or many, we might ask: what is the least we need to believe to justify attending to the development of AI among all the possible things that we might attend to? And my sense is just this: we may soon transition from a world where all systems that have power over the future of life on the planet are intricately tied to human building-blocks to a world where there are some systems that have power over the future of life on the planet without relying on human building-blocks. This alone, in my view, justifies attention in this area, and it does not rest in any way on views about agency or goals or intelligence.

So here is the argument up to this point:

Among everything in the world that we might pay attention to, it makes sense to attend to that which has the greatest power over the future of life on the planet. Today, the systems that have power over the future of life on the planet rely on humans as building-blocks. Yet soon we may construct systems that have power but do not rely on humans as building-blocks. Due to the significance of this shift we should attend to the development of AI and check whether there is any cause for concern, and, if so, whether those concerns are already being adequately addressed, and if not, whether there is anything we can do.

The case for concern

So we have a case for paying some attention to AI among all the things we could pay attention to, but I have not yet made a case for being *concerned* about AI

development. So far it is as if we discovered an object in the solar system with a shape and motion quite unlike a planet or moon or comet. This would justify some attention by humans, but on this evidence alone it would not become a top concern, much less a top cause area.

So how do we get from attention to concern? Well, the thing about power is that *humans already seek it*. In the past, when it has become technically feasible to build a certain kind of system that exerts influence over the future, humans have tended, by default, to eventually deploy such systems in service of their individual or collective goals. There are some classes of powerful systems that we have coordinated to avoid deploying, and if we do this for AI then so much the better, but by default we ought to expect that once it becomes possible to construct a certain class of powerful system, humans will deploy such systems in service of their goals.

Beyond that, humans are quite good at incrementally improving things that we can tinker with. We have made incremental improvements to airplanes, clothing, cookware, plumbing, and cell phones. We have not made incremental improvements to human minds because we have not had the capacity to tinker in a trial-and-error fashion. Since all powerful systems in the world today use humans as building blocks, and since we do not presently have the capacity to make incremental improvements to human minds, there are no powerful systems in the world today that are subject to incremental improvement at all levels.

In a world containing some powerful systems that do not use humans as building blocks, there will be *some powerful systems that are subject to incremental improvements at all levels*. In fact the development of AI may open the door to making incremental improvements to human minds too. In this case *all* powerful systems in the world would be subject to incremental improvement. But we do not need to take a stance on whether this will happen or not. In either case the situation we will be in is one in which humans are making incremental improvements to some systems that have power in the world, and we therefore ought to expect that the power of these systems will therefore increase on a timescale of years or decades.

Now at this point it is sometimes argued that a transition of power from humans to non-human systems will take place, due to the very high degree of power that these non-human systems will eventually have, and due to the difficulty of the alignment problem. But I do not think that any such argumentative move is necessary to justify concern, because whether humans eventually lose power or not, what is much more certain is that in a world where powerful systems are being incrementally improved, there will be a period during which *humans gain power quickly*. It might be that humans gain power for mere minutes before losing it to a recursively self-improving singleton, or it may be that humans gain power for decades before losing it to an inscrutable web of AI services, or it may be that humans gain power and hold onto it until the end of the cosmos. But under any of these scenarios, humans seem destined to gain power on a timescale of years or decades, which is the pace at which we usually make incremental improvements to things.

What happens when humans gain power? Well as of today, existential risk exists. It would not exist if humans had not gained power over the past few millennia, or at least it would be vastly reduced. Let's ignore existential risk due to AI in order to make sure the argument is non-circular. Still, the point goes through. There is much good to say about humans. This is not a moral assessment of humanity. But can anyone deny that humans have gained power over the past few millennia, and that, as a result of that, existential risk is much increased today compared to a few millennia ago? If

humans *quickly gain power*, it seems that, by default, we ought to presume that existential risk will also increase.

Now, there are certainly *some ways* to increase human power quickly without increasing existential risk, including by skillful AI development. There have certainly been *some times and places* where rapid increases in human power have led to decreases in existential risk. But this part of the argument is about what happens by default, and the ten thousand year trendline of the "existential risk versus human power" graph is very much up-and-to-the-right. Therefore I think rapidly increasing human power will increase existential risk. We do not need to take a stance on how or whether humans might later lose power in order for this to go through. We merely need to see that, among all the complicated goings-on in the world today, the development of AI is the thing most likely to confer a rapid increase in power to humans, and on the barest historical precedent, that is already cause for both attention and concern.

So here is the case for concern:

If humans learn to build systems that do influence the future of life on the planet but do not require human building-blocks, then they are likely to make incremental improvements to these systems over a timescale of years or decades, and thereby increase their power over the future of life on the planet on a similar timescale. This should concern us because quick increases in human power have historically led to increases in existential risk. We should therefore investigate whether these concerns are already being adequately addressed, and, if not, whether there is anything we can do.

I must stress that not all ways of increasing human power lead to increases in existential risk. It is as if we were considering giving a teenager more power over their own life. Suppose we suddenly gave this teenager the power not just of vast wealth and social influence, but also the capacity to remake the physical world around them as they saw fit. For typical teenagers under typical circumstances, this would not go well. The outcomes would not likely be in the teenager's own best interests, much less the best interests of all life on the planet. Yet there probably *are* ways of conferring such power to this teenager, say by doing it slowly and in proportion to increases in the teenager's growing wisdom, or by giving the teenager a wise genie that knows what is in the teenager's best interest and will not do otherwise. In the case of AI development, we are collectively the teenager, and we must find the wisdom to see that we are not well-served by rapid increases in our own power.

The case for intervention

We have a case for *a priori* concern about the development of a particular technology that may, for a time, greatly increase human power. But perhaps humanity is already taking adequate precautions, in which case marginal investment might be of greater benefit in some other area. What is the epistemically minimal case that humanity is not already on track to mitigate the dangers of developing systems that have power over the future of life on the planet without requiring humans as building-blocks?

Well consider: right now we appear to be rolling out machine learning systems at a rate that is governed by economic incentives, which is to say that the rate of machine learning rollout appears to be determined primarily by the supply of the various factors of production, and the demand for machine learning systems. There is

seemingly no gap between the rate at which we *could* roll out machine learning systems if we allowed ordinary economic incentives to govern, and the rate at which we are rolling out those systems.

So is it more likely that humanity is exercising diligence and coordinated restraint in the rollout of machine learning systems, or is it more likely that we are proceeding haphazardly? Well imagine if we were rolling out nuclear weapons at a rate determined by ordinary economic incentives. From a position of ignorance, it's *possible* that this rate of rollout would have been selected by a coordinated humanity as the wisest among all possible rates of rollout. But it's much more likely that this rate is the result of haphazard coordination, since from economic arguments we would expect the rate of rollout of any technology to be governed by economic incentives in the absence of a coordinated effort, whereas there is no reason to expect a coordinated consideration of the wisest possible rate to settle on this particular rate.

Now, if there were a gap between the "economic default" rate of rollout of machine learning systems and the actual rate of rollout then might still question whether we were on track for a safe and beneficial transition to a world containing systems that influence the future of life on the planet without requiring humans as building-blocks. It might be that we have merely placed haphazard regulation on top of haphazard AI development. So the existence of a gap is not a sufficient condition for satisfaction with the world's handling of AI development. But the absence of any such gap does appear to be evidence of the absence of a well-coordinated civilization-level effort to select the wisest possible rate of rollout.

This suggests that the concerning situation in the previous section is, at a minimum, not already *completely* addressed by our civilization. It remains to be seen whether there is anything we can do about it. The argument here is about whether the present situation is already satisfactory or not.

So here is the argument for intervention:

Humans are developing systems that appear destined to quickly increase human power over the future of life on the planet at a rate that is consistent with an economic equilibrium. This suggests that human civilization lacks the capacity to coordinate on a rate motivated by safety and long-term benefit. While other kinds of interventions may be taking place, the absence of this particular capacity suggests that there is room to help. We should therefore check whether there is anything that can be done.

Now it may be that there is a coordinated civilization-level effort that is taking measures other than selecting a rate of machine learning rollout that is different from the economic equilibrium. Yes, this is possible. But the question is why our civilization is not coordinating around a different rate of machine learning rollout if it has the capacity to do so. Is it that the economic equilibrium is in fact the wisest possible rate? Why would that be? Or is it that our civilization is choosing not to select the wisest possible rate? Why? The best explanation seems to be that our civilization does not presently have an understanding of which rates of machine learning rollout are most beneficial, or the capacity to coordinate around a selected rate.

It may also be that we navigate the development of powerful systems that do not require humans as building-blocks without ever coordinating around a rate of rollout different from the economic equilibrium. Yes this is possible, but the question we are asking here is whether humanity is already on track to safely navigate the

development of powerful systems that do not require humans as building-blocks, and whether our efforts would therefore be better utilized elsewhere. The absence of the capacity to coordinate around a rate of rollout suggests that there is at least one very important civilizational capacity that we might help develop.

The case for action

Finally, the most difficult question of all: is there anything that can be done? I don't have much to say here other than the following very general point: It is very strong to claim that nothing can be done about a thing because there are many possible courses of action, and if even one of them is even a little bit effective then there is something that can be done. To rule out all possible courses of action requires a very thorough understanding of the governing dynamics of a situation and a watertight impossibility argument. Perhaps there is nothing that can be done, for example, about the heat death of the universe. We have some understanding of physics and we have strong arguments from thermodynamics, and even on this matter there is some room for doubt. We have nowhere near that level of understanding about the dynamics of AI development, and therefore we should expect on priors that among all the possible courses of actions, there are some that are effective.

Now you may doubt whether it is possible to *find* an effective course of action. But again, claiming that it is impossible to find an effective course of action implies that among all the ways that you might try to find an effective course of action, none of them will succeed. This is the same impossibility claim as before, only now it concerns the process of finding an effective course of action rather than the process of averting AI risk. Once again it is a very strong claim that requires a very strong argument, since if even one way of searching for an effective course of action would succeed, then it is possible to find an effective course of action.

Now you may doubt that it is possible to find a way to search for an effective course of action. Around and around we could go with this. Each time you express doubt I would point out that it is not justified by anything that is objectively impossible. What, then, is the real cause of your doubt?

One thing that can always be done at an individual level is to make a thing the top priority in our life, and to become willing to let go of all else in service of it. At least then if a viable course of action does become apparent, we will certainly be willing to take it.

Conclusion

In the early days of AI alignment there was much discussion about fast versus slow take-off, and about recursive self-improvement in particular. Then we saw that *the situation is concerning either way*, so we stopped predicated our arguments on fast take-off, not because we concluded that fast take-off arguments were wrong, but because we saw that the center of the issue lay elsewhere.

Today there is much discussion in the alignment community about goal-directedness and agency. I think that a thorough understanding of these issues is central to a solution to the alignment problem, but, like recursive self-improvement, I do not think it is central to the problem itself. I therefore expect discussions of goal-directedness

and agency to go the way of fast take-off: not dismissed as wrong, but de-emphasized as an unnecessary predicate.

There is also discussion recently about scenarios involving single versus multiple AI systems governed by single versus multiple principals. Andrew Critch has [argued](#) that more attention is warranted to "multi/multi" scenarios in which multiple principals govern multiple powerful AI systems. Amongst the rapidly branching tree of possible scenarios it is easy to doubt whether one has adequately accounted for the premises needed to get to a particular node. It may therefore be helpful to lay out the part of the argument that applies to all branches, in order that we have some epistemic ground to stand on as we explore more nuance. I hope this post helps in this regard.

Appendix: Agents versus institutions

One of the ways that we could build systems that have power over the future of life on the planet without relying on human building-blocks is by building goal-directed systems. Perhaps such goal-directed systems would resemble agents, and we would interact with them as intelligent entities, as [Richard Ngo describes in AGI Safety from First Principles](#).

A different way that we could build systems that have power over the future of life on the planet without relying on human building-blocks is by gradually automating factories, government bureaucracies, financial systems, and eventually justice systems as [Andrew Critch describes](#). In this world we are not so much interacting with AI as a second species but more as the institutional and economic water in which we humans swim, in the same way that we don't think of the present-day finance or justice systems as agents, but more like a container in which agents interact.

Or perhaps the first systems that will have power over the future of life on the planet without relying on human building-blocks will be emulations of human minds, as Robin Hanson describes in *Age of Em*. In this case, too, humans would gain the capacity to tinker with all parts of some systems that have power over the future of life on the planet, and through ordinary incremental improvement become, for a time, extremely powerful.

These possibilities are united as avenues by which humans could quickly increase their power by building systems that have both influence over the future of life on the planet, and are subject to incremental improvement at all levels. Each scenario suggests particular ways that humans might later lose power, but instead of taking a strong view on the loss of power we can see that a quick increase in human power, however temporary, is, on historical precedent, already a cause for concern.

Information Assets

Epistemic Status: This represents fairly early work. The terminology isn't at all set in stone.

Scholarship Status: I've spent several hours attempting to investigate this topic, and in the past have spent quite a while researching Information Theory and Applied Information Economics. I'm sure I'm missing very valuable literature, but I can't find it. Comments well appreciated.

Thanks to David Manheim and Nuño Sempere for comments on this

Introduction

As anyone who's read an arduous textbook knows, learning comes with costs as well as benefits. As anyone who's decided against reading at least one textbook knows, often the expected costs outweigh the expected benefits.

The interplay of the costs and benefits of information leads to an environment of efficient trade-offs. Tweets and infographics often do most of the work of popular nonfiction books. Lossy computational compression schemes offer good-enough quality for greatly reduced storage costs. Common verbal communication is typically far less precise and rigorous than formal philosophical proofs, but is far more practical.

As the old saying goes "All models are false, but some are useful." A cost-benefit approach would say, "Models represent trade-offs in situations where absolute accuracy doesn't justify its cost." I think this basic fact; that educational materials, models, theories, and language all represent trade-offs between accuracy and costs, should really be fairly obvious and commonly acknowledged by now.

Interestingly enough, fairly little work around Information Theory seems to have gone deep into modeling these trade-offs. It's often assumed that information is either freely absorbed, or in some situations limited to a particular fixed communication channel. Some work around Value of Information analyses estimate the cost one might be interested in paying for specific information, but often makes some very large assumptions for some particular settings.

I'm interested in developing better intuitions and vocabulary around these tradeoffs. Here are some initial attempts. I'm sure I'm missing a lot of great work, but I've had problems finding it. If you have thoughts or recommended references, I'd very much appreciate you sharing them.

A Very Simple Model

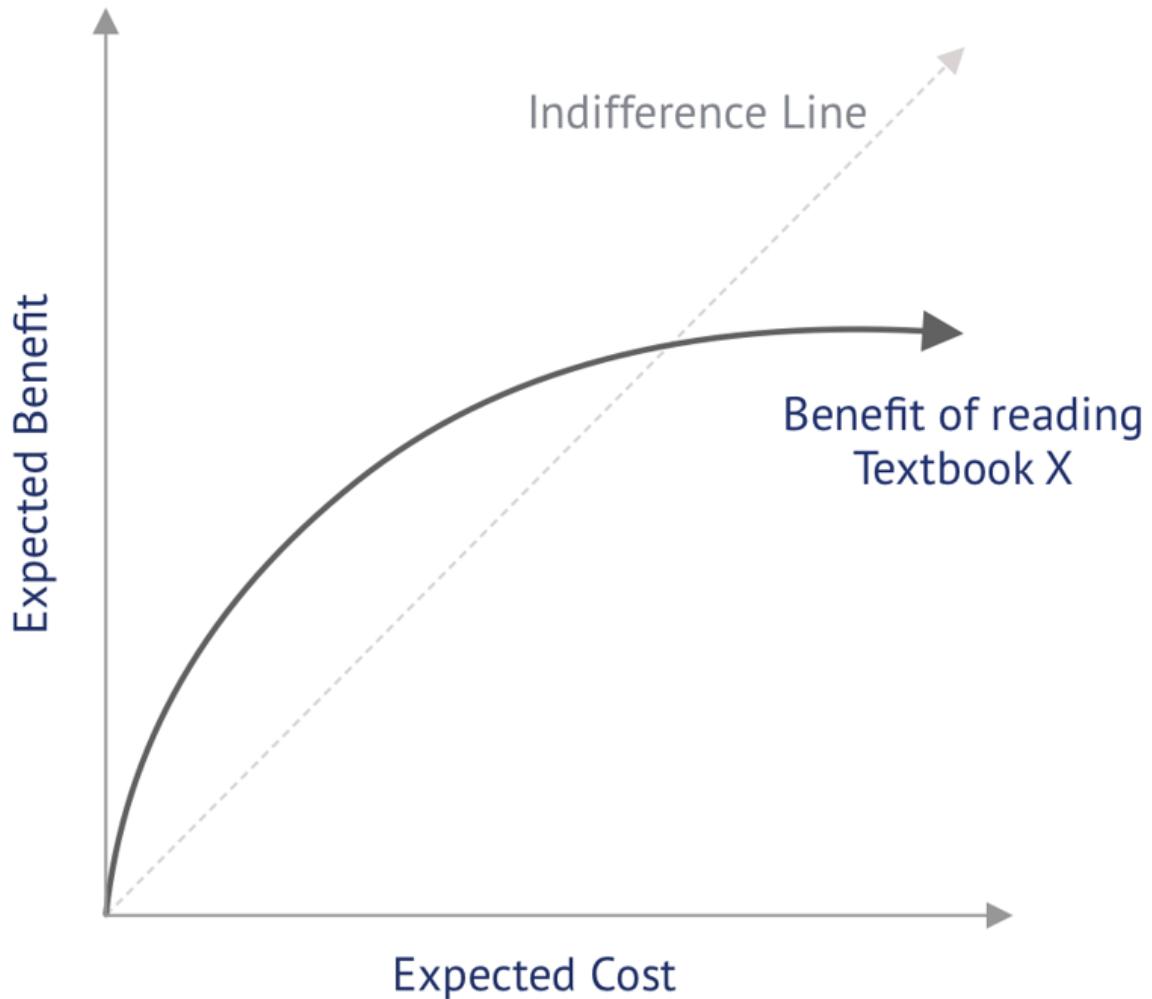
We're going to begin with models of textbooks; not because these are particularly important, but because I think these are particularly easy to intuit. These models will later be extended to more interesting areas.

Say you begin reading an interesting textbook to study for a test in one month. This is one of your reading materials among several, so you don't have time to read and fully process every single word in said book.

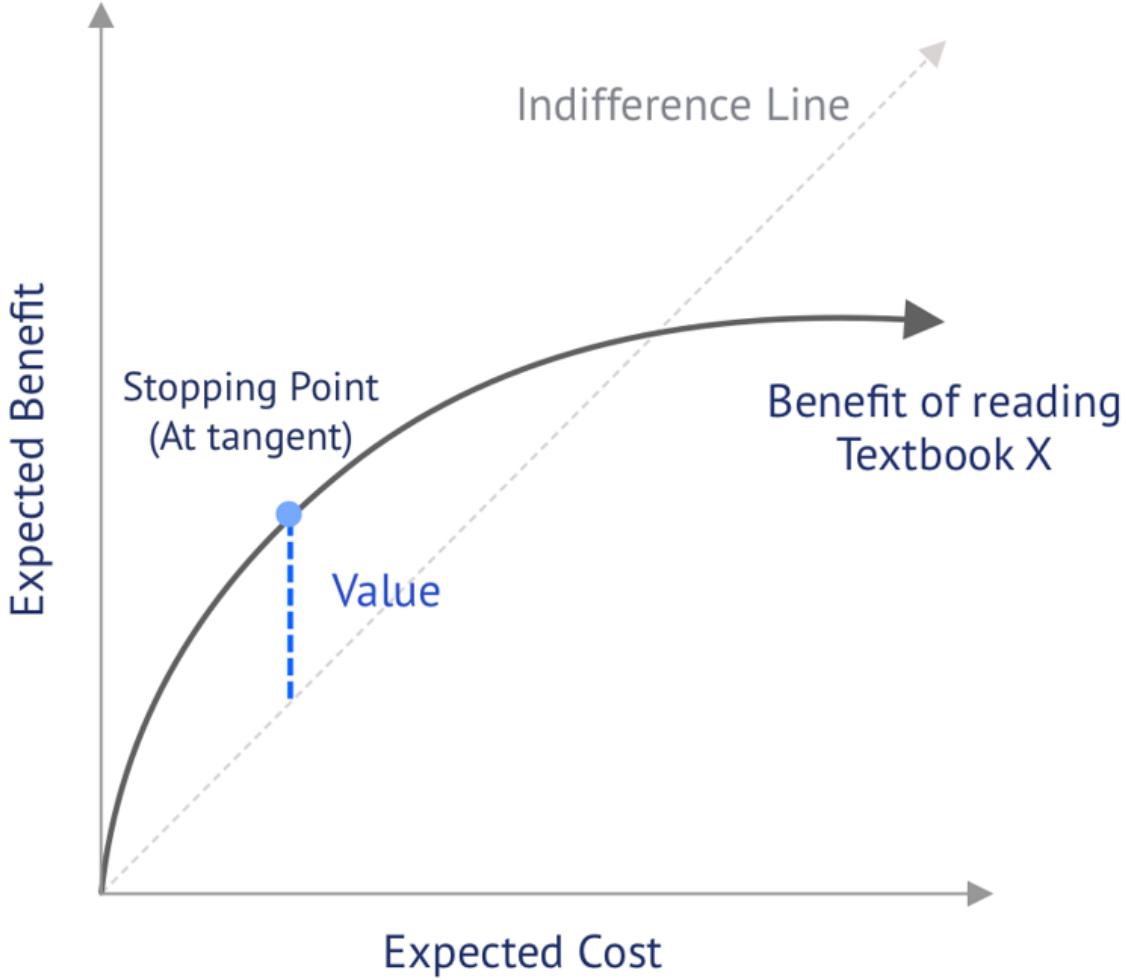
You might begin by skimming the book, then proceed to read what seem like the most important sections. You'll continue it until you become convinced that your time would be better spent on different materials. More succinctly, you continue reading it until the expected costs begin to outweigh the expected benefits of marginal investment.

This is a classic Microeconomics case of diminishing marginal utility. Below is a representation of the corresponding curve.

Note that the expected benefit and expected cost scales are equivalent. Therefore, at all points diagonal between them (the Indifference Line), the expected benefit is equal to the expected cost, meaning that the total net value is 0.



As is true for situations of marginal utility, assuming you are rational, you'd aim to stop reading the book around when the tangent of the curve is at 45%, or, when the marginal costs begin to exceed the marginal benefits.



We can draw a “Stopping Point” at this point. The line directly below it, to the indifference line, represents the net benefit, or “value” achieved, if learning stops at this point. *This point is also sometimes called the “point of maximum yield.”*

This book represents information, but the fact that there are known learning costs represents complexity not typically mentioned in models of information. We can call models of bundles of information that require learning costs, as *information assets*.

Basic Functions

Information assets are items that represent trade-offs of information gained and cost to particular actors. We could imagine representing them with a few programming functions. (Here represented with Haskell style definitions)

```
information_aquisition_fn :: information_asset -> agent --> cost -->
information
information_benefit_fn :: agent --> information --> expected_benefit
```

Alternatively, you could either combine these or skip the intermediate step.

```
information_benefit_fn :: information_asset -> agent --> cost -->
expected_information_benefit
```

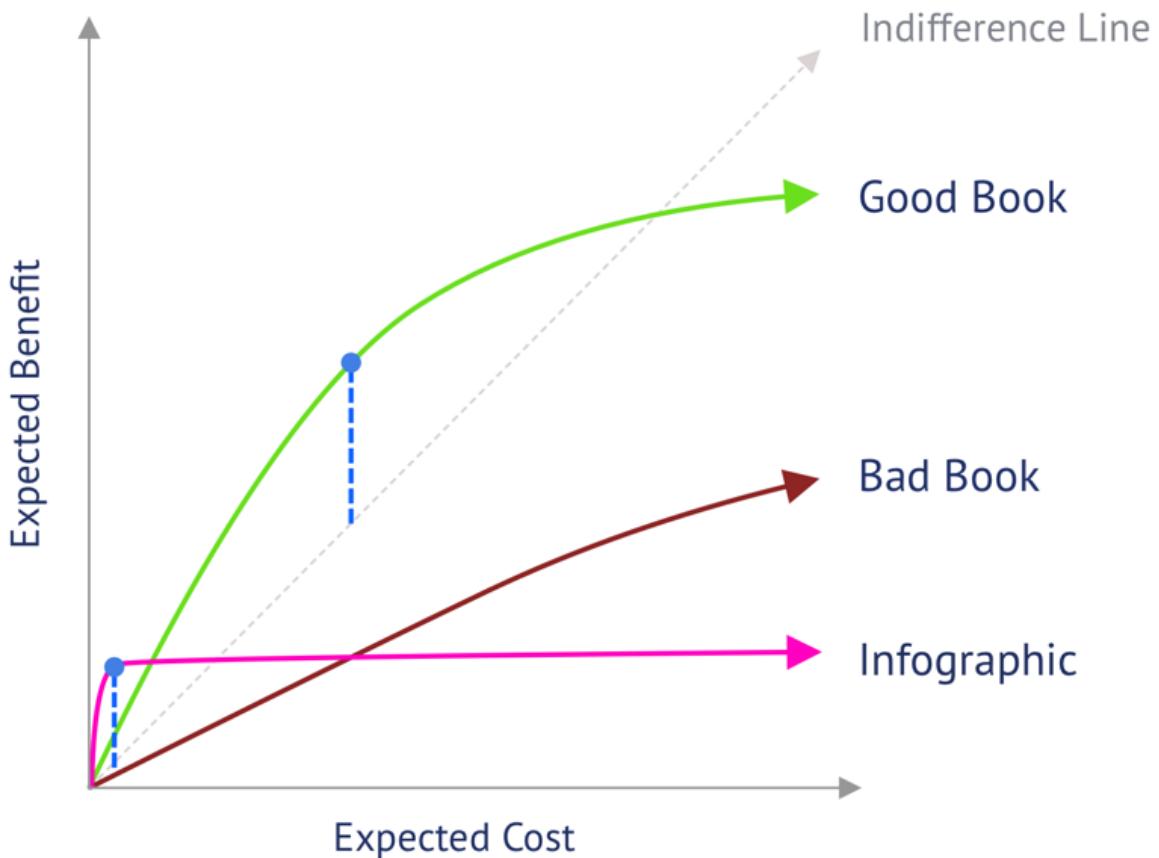
One could get more fancy if you were to represent a space of potential ways of converting costs to information. The equations above assume that the best option is chosen, but sometimes you might desire to model this extra complexity.

The details of what agents, costs, and information are, are abstracted here. I imagine that there are many potential definitions that would all work well enough to be interesting.

The important thing is that this main `information_benefit_fn` function should generally represent diminishing marginal utility, and should help us represent the most important aspect of information assets.

We could of course also represent these functions as math equations, but I both am more experienced in programming, and also prefer thinking in programming for things like this.

Comparisons of different information assets

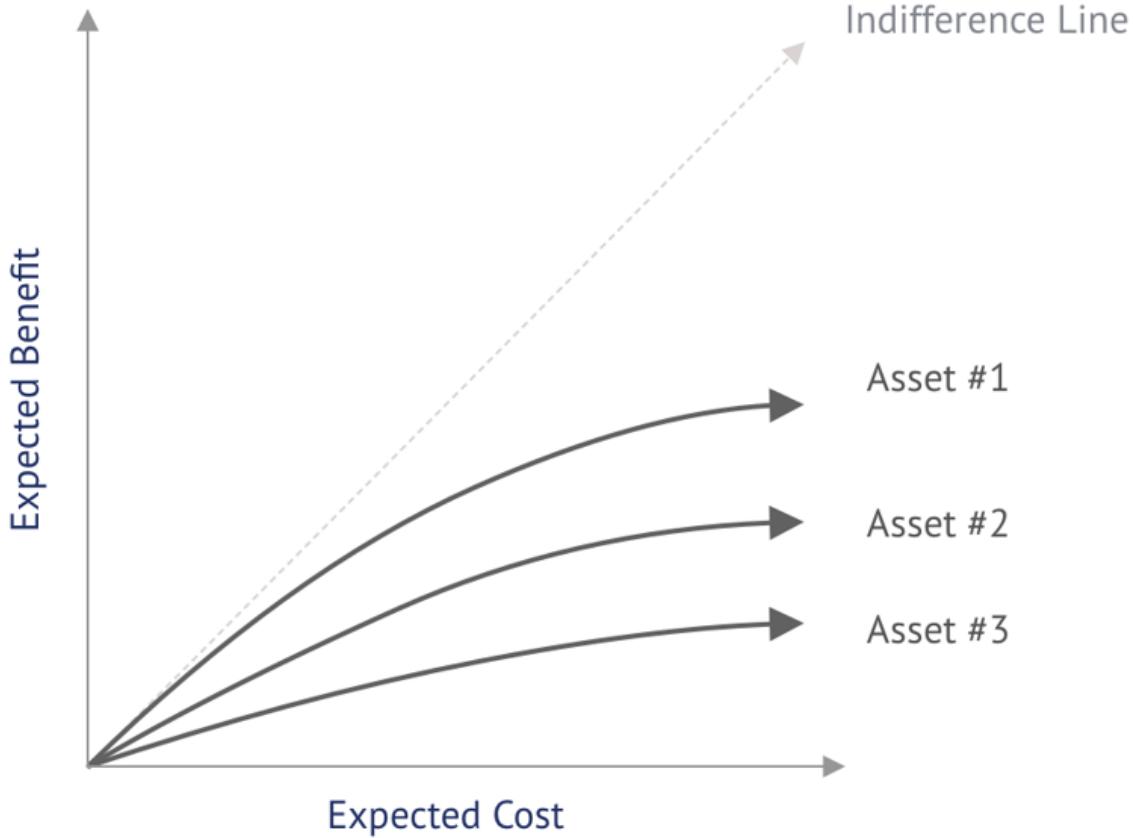


Now, we can compare the marginal value curves of different information assets. The green line represents a good book, the pink one an infographic. As you can see, the infographic produces great returns for a short period of time, but then levels off, as there's typically not all too much to learn from an infographic. In comparison, the good book takes longer to provide the same amount of value, but in this case winds up producing more in total.

The bad book, on the other hand, might have a lot of potential benefit, but has negative marginal value, so would never be started. Therefore, in this model, the bad book has high potential benefit, but zero expected value.

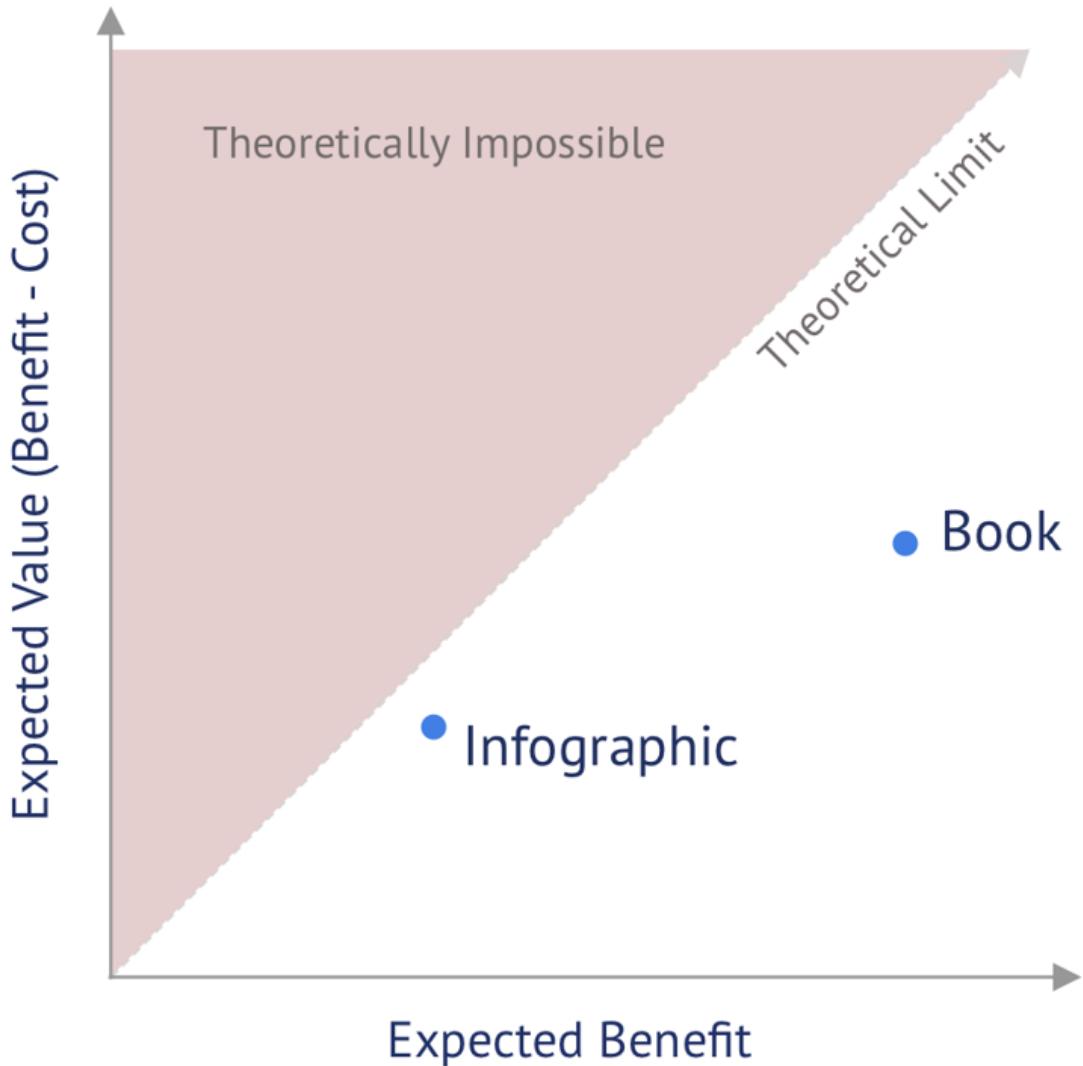
	Potential 100% Benefit	Benefit at Stopping Point	Cost at Stopping Point	Expected Value
Good Book	120	70	30	40
Bad Book	100	0	x	0
Infographic	30	30	5	25

What do you think most available information assets would look like, on this graph? Well, if costs include opportunity costs, in an environment with many potential information assets, the vast majority would represent zero value. This might look like the curves in the following diagram.



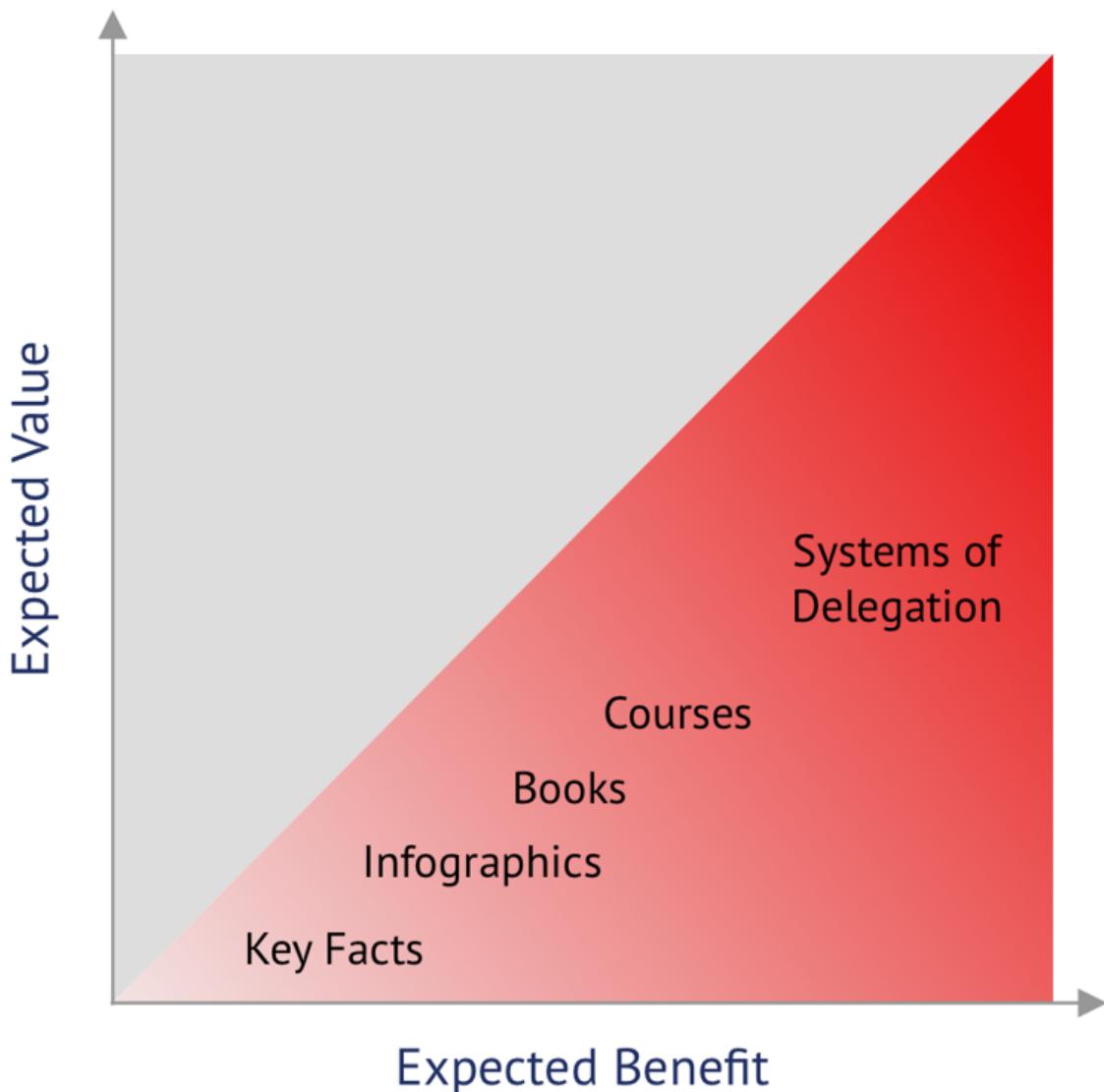
Benefit vs. Value

Let's now introduce a new plot, one of benefit vs. value. This can be valuable because it can hint at something like efficiency; how much of the benefit is realized as value? 100% efficiency is clearly the maximum, so at any particular benefit, the max potential value (assuming we might be able to have zero costs cost) is equal to that benefit. We'll just go ahead and color off the area where value is greater than benefit as an "impossible zone".



We'll show blue dots for the stopping points of the information assets (the infographic and the book).

One very obvious question we might have is how to produce information assets that will lie on different parts of this graph. To help answer that question, here's a simple graph that's filled in red to represent the very roughly expected costs of producing something in each part of it.



It's normally more work to produce a book than it is to produce an infographic, especially if both must be positive expected utility. As the expected potential benefit increases, it becomes increasingly difficult to maintain a high efficiency. This is because within any particular topic, some subsets of information are typically much more valuable than other subsets. You're kind of fighting multiple marginal utility diminishments; not only can a user skim to effectively compress each area, they can also prioritize the particularly valuable areas.

The top right area is labeled “systems of delegation” to represent the sorts of information assets that I expect to exist here. For these, it’s important to point out that the costs of using information are not only time costs, but they can also be monetary. If you want to achieve a lot of value from available information about a medical condition, you can either read a lot about that condition, or defer to a professional who’s already analyzed all of the relevant knowledge. In some cases, these delegation systems can be automated, so can represent very low costs.

Fixed Costs vs. Variable Costs

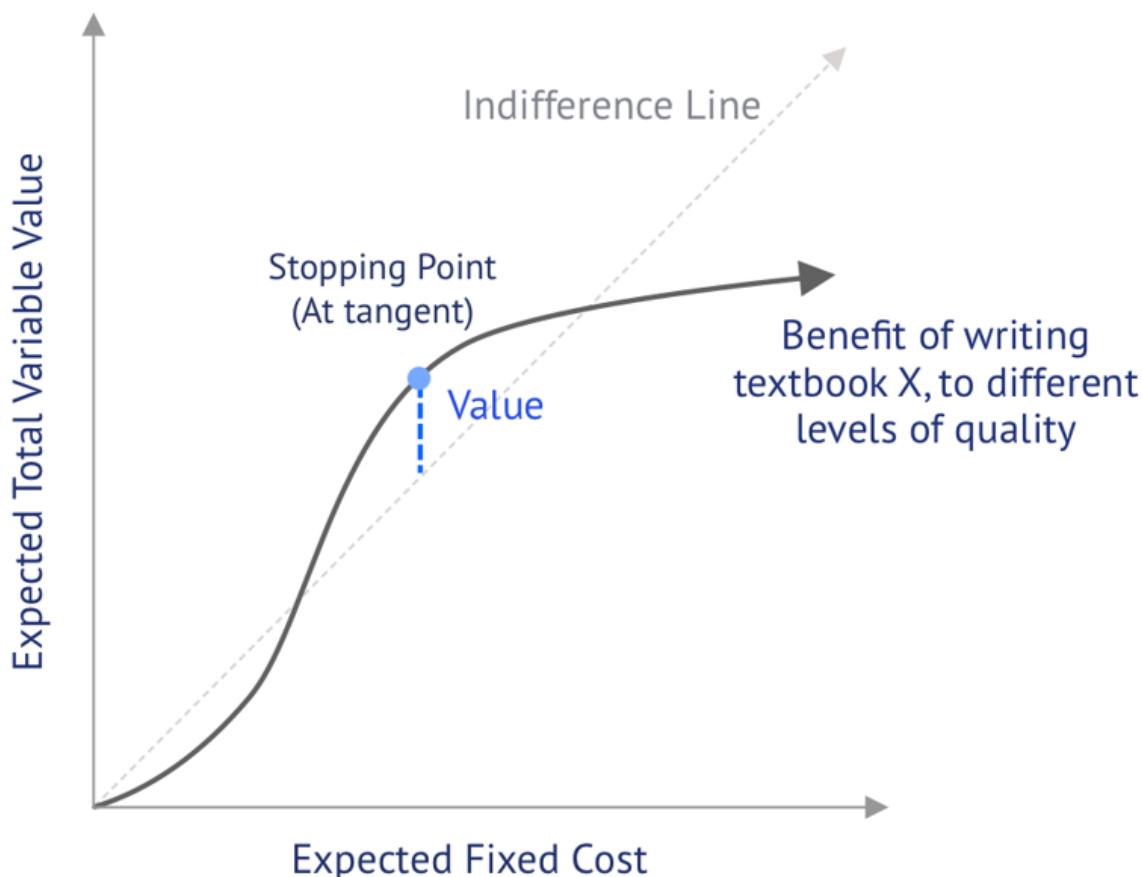
Imagine we're set on writing a textbook for the purpose of helping 500 students via informational expected value. There are clear fixed costs associated with writing the book. We probably have a spectrum of how much effort we can put into the book. We can rush it by simply transcribing our previous lectures without any cleanup, or we can spend a lot of time figuring it out from scratch.

We'll describe writing the book as the "fixed cost", and the value of students (n) when they read the book as the "fixed benefit".

$$\text{Fixed Benefit} = (\text{Variable Benefit} - \text{Variable Cost}) * n$$

Again, we have a diminishing marginal utility curve. The aim is to find some convenient balance between costs and benefits. The total resulting value is the difference between the total value received from the students (note that this takes the costs and benefits for each student into account), and the costs spent on writing the book.

This curve might be a bit different from the ones above, because there might well be a period at the start where no textbook would justify its cost.^[1] Maybe there's an upfront cost for just having anything printed, that requires a substantial benefit to overcome. But eventually diminishing marginal utility will come into play.



One important thing to consider here is that there are multiple total stages of fixed costs and benefits to variable costs and benefits. So there are many sorts of *degrees* of variability.

- Information in a field gets converted into many books.

- Each book has many readers.
- Each reader will recall the information they have read many times.

At each point of the pipeline, authors or readers must make estimates to best spend fixed costs to gain later variable benefits. Most of these trade-offs will involve considerations of informational efficiencies.

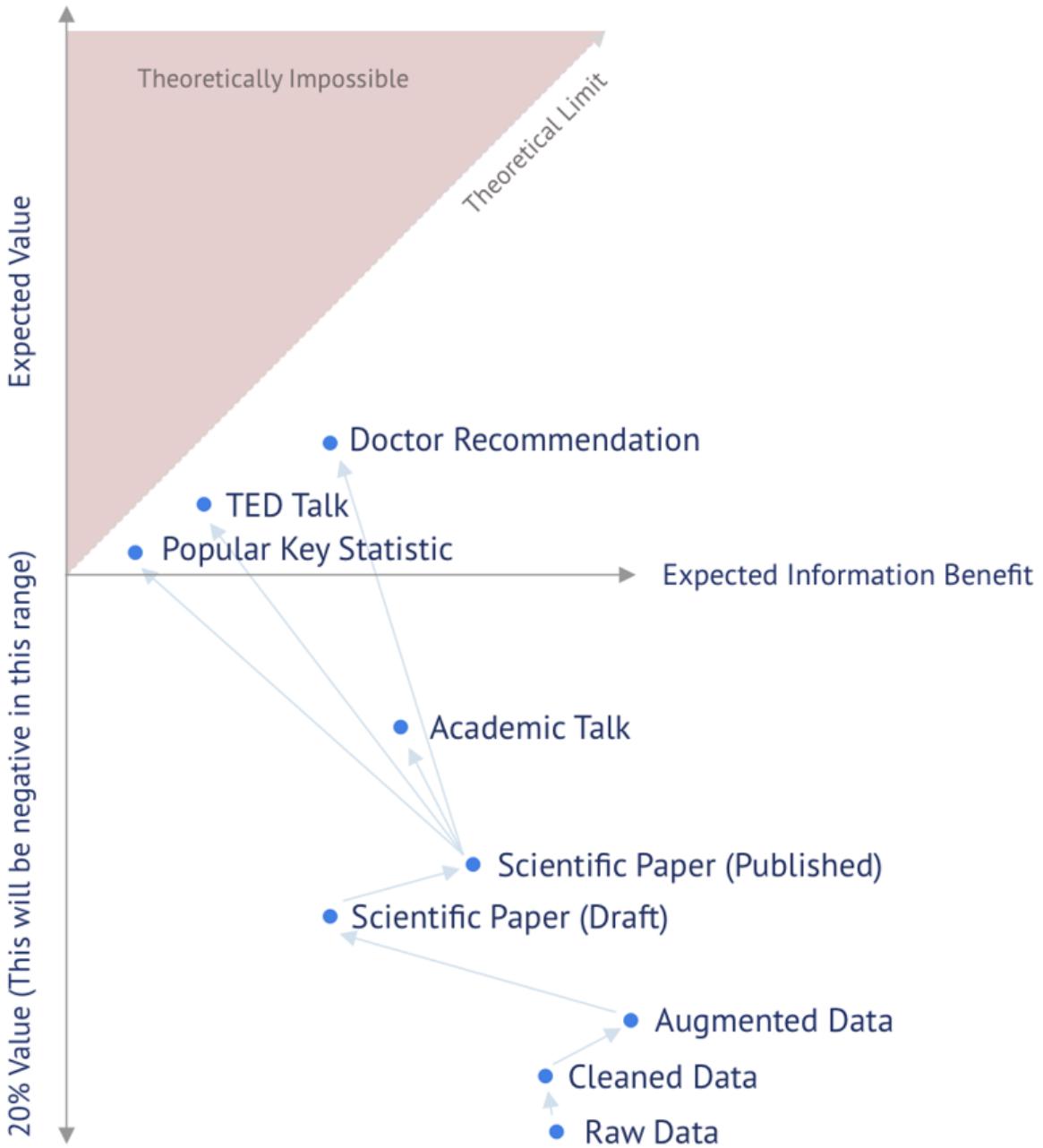
Information Asset Supply Chains

Let's imagine that a new useful-but-not-groundbreaking medical discovery has been made. It begins with data and understanding in a particular scientist's lab.

At this point, even if the scientist were to make this information public, it would be effectively useless to consumers. Consumers would need to discover, translate, and process this information. The costs of doing so would far exceed the expected benefits.

There are many situations where information assets must undergo several transformations of conversions before enabling utility in particular groups of consumers. We can attempt to make diagrams of this process, as shown below. This is the basic value to benefit graph shown a few times above, but with the addition of a slightly different y-axis going down. As stated earlier, consumers won't be expected to attempt negative-value endeavors, so in order to represent the costs and benefits of these, we need to make conjectures. In this case, we can estimate the net value (which would be negative) were they to do the work necessary to obtain 20% of the benefit of these assets.

For example, it might take a somewhat average educated consumer 300 hours to learn how to analyze the results of an experiment themselves and take away 20% of the potential benefit they could get from it.

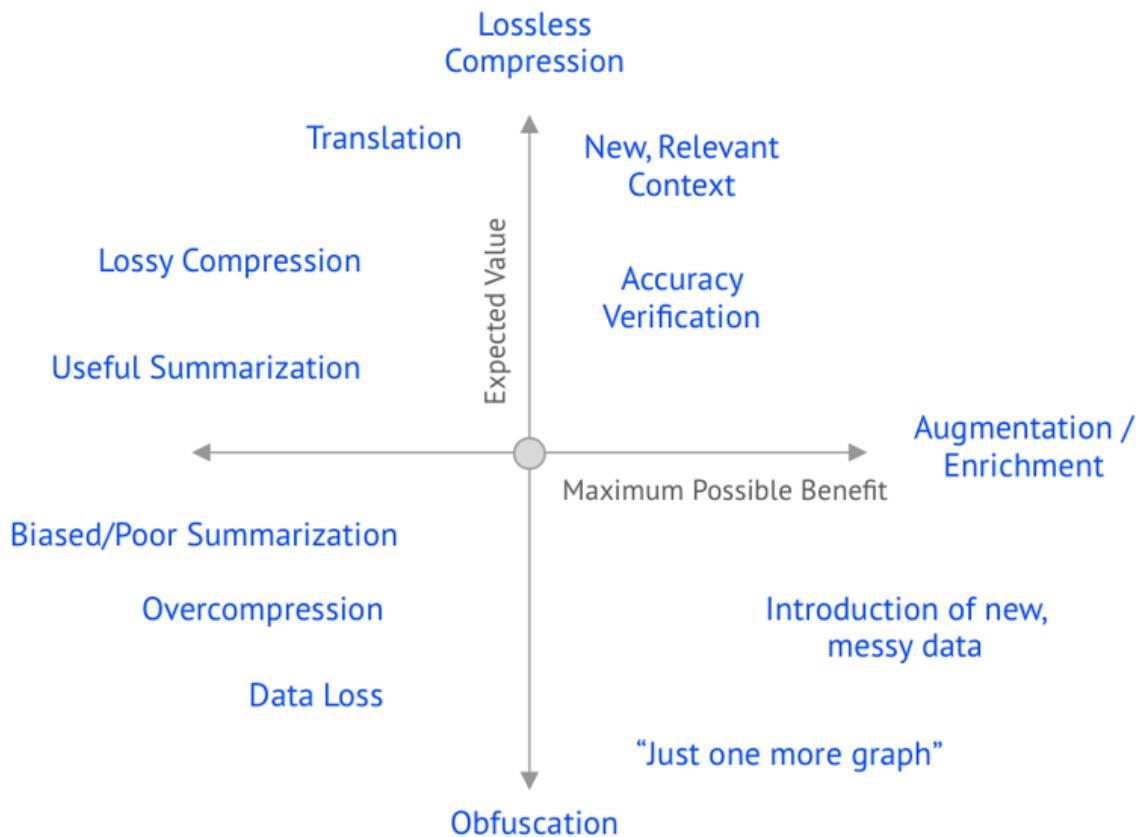


A medical experiment might result in “raw data”, which is particularly difficult to interpret. This is shown in the bottom right of the diagram. This raw data gets cleaned up and goes through several steps to eventually transform into an immediately valuable asset.

Information assets below the zero-expected-value line basically represent *unfinished* or **capital goods**. Information assets at the starting point are raw goods. Assets above the line are *finished goods*.

Information Asset Conversion Options

For a particular information asset on the above chart, we might imagine ways it could be transferred to different regions. Mostly common are moves to the top-left. Moves in the top-right quadrant, and occasionally in the bottom-right quadrant are possible too. Moves in the bottom-left quadrant are only caused by accident or malice. Below we label a few clusters that refer to common names that might be used to refer to moves in different directions.



Translation: The information asset is converted into a format more amenable to a user; for example, from French to German. It's likely that some information is lost, but hopefully fairly little.

Compression: Some information is lost. Sometimes it's information that's completely irrelevant to later users ("lossless" compression), but normally the lost information will present at least some loss of potential benefit ("lossy" compression).

Summarization: A type of compression that typically aims to present a small fraction of the total information.

New context: Additional information is introduced to help contextualize the asset. This might help make the information asset easier to digest, and also might make it more valuable.

Accuracy verification: A small amount of information is introduced, in the form of a check on the accuracy of the information asset. This builds justified trust in the asset.

Introduction of new, messy data: New information is introduced, with some additional potential benefit. However, it makes the information asset initially more costly than the

immediate benefit.

Obfuscation: No data is lost, but it requires additional work to interpret. For example, maybe the data is removed from being publicly accessible, and now requires hunting down the right bureaucrat to access.

Beneficiation

I've looked for some time for a good word to describe the process outlined in the Supply Chains section above. I think my favorite so far is beneficiation.

According to the [Wikipedia](#) page,

In the [mining](#) industry or [extractive metallurgy](#), **beneficiation** is any process that improves (benefits) the [economic value](#) of the ore by removing the [gangue minerals](#), which results in a higher grade product ([ore concentrate](#)) and a waste stream ([tailings](#)). There are many different types of beneficiation, with each step furthering the concentration of the original ore.

I believe the word came about by turning "benefit" into a verb.

I like "beneficiation" because it seems both very to the point, and because it doesn't already have a more narrow term around information.

Regarding *information assets*, we can define beneficiation as:

Beneficiation (*information assets*): Any process or action that assists increasing the total value of an information asset.

Beneficiation here describes every part of the process of converting *information* into its final form, before it gets directly used for a decision. This includes:

- Learning
- Writing
- Summarization
- Data organization
- Rewriting
- Teaching
- Automated data systems

I'm really not sure what the neatest mathematical models to represent value-add at each part of a beneficiation pipeline, but it feels like it should be rather simple. In theory, one should be able to identify bottlenecks, or particularly valuable or costly transition points.

Next Steps

This is fairly early and messy work. I'd be curious to clean up the terminology, and figure out how much further it can be taken, while also be cost-effective.

Here are some questions I still have:

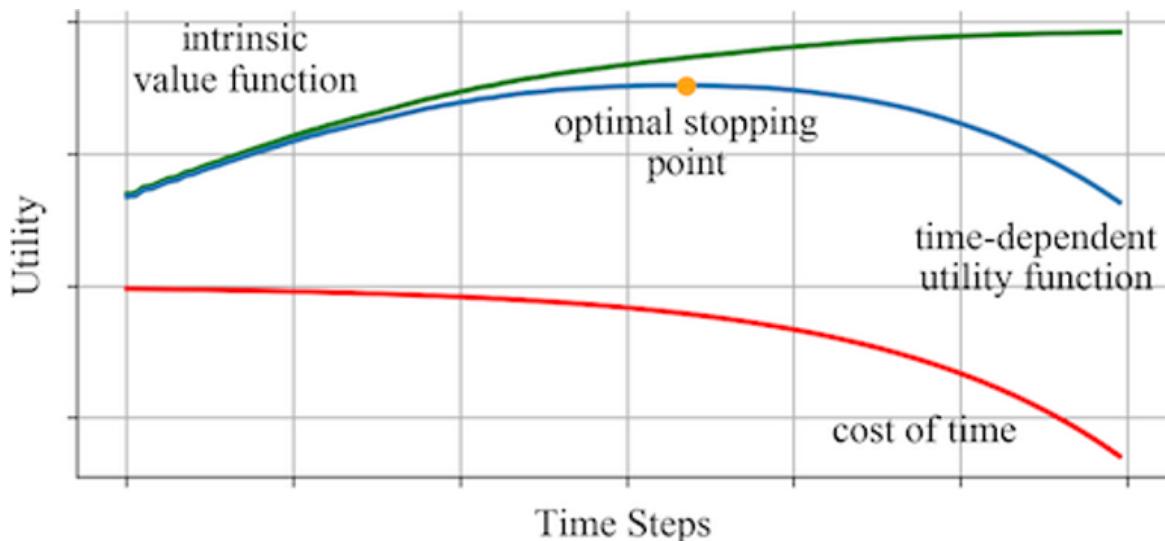
1. Can we make decent cost-benefit charts or tables of common types of information? (Books, courses, computer recommender systems)
2. Can we make decent cost-benefit charts or tables to represent tradeoffs among abstract models (Newtonian mechanics vs. Special Relativity, formal sentences vs. context-dependent speech, data compression methods, etc)?

3. Can we come up with elegant categorizations to describe information that's simply not worth the necessary beneficiation costs? There's definitely a lot of very interesting information out there that's useful in theory, but the costs to make it net-valuable are likely to exceed the expected benefits of doing so.
4. Can these sorts of models help direct us regarding to what sorts of information assets we should emphasize going forward?
5. Can information asset models be used to help describe human language and communication, both descriptively and prescriptively?

Related Work

As I said before, this sort of issue seems highly general, important, and neglected, which really surprises me.

One area with related work is that of [anytime algorithms](#). Anytime algorithms face clear trade-offs of computation time to accuracy, and thus can be modeled as a type of information asset. Justin Svegliato has recently done work on using diminishing marginal utility curves to make decisions around anytime algorithms. [This post](#) is a good summary, with some very similar diagrams to what I made above.



Diminishing marginal utility curve of an anytime algorithm, by Justin Svegliato.

There's definitely other work out there that seeks to understand the costs vs. benefits of information, though from what I can tell, it generally seeks less to try to find economics-style models. Laura Schulz did some relevant work, as shown in [this lecture](#).

The [Information Bottleneck](#) method in Information Theory is also very related, along with other theoretical work on compression.

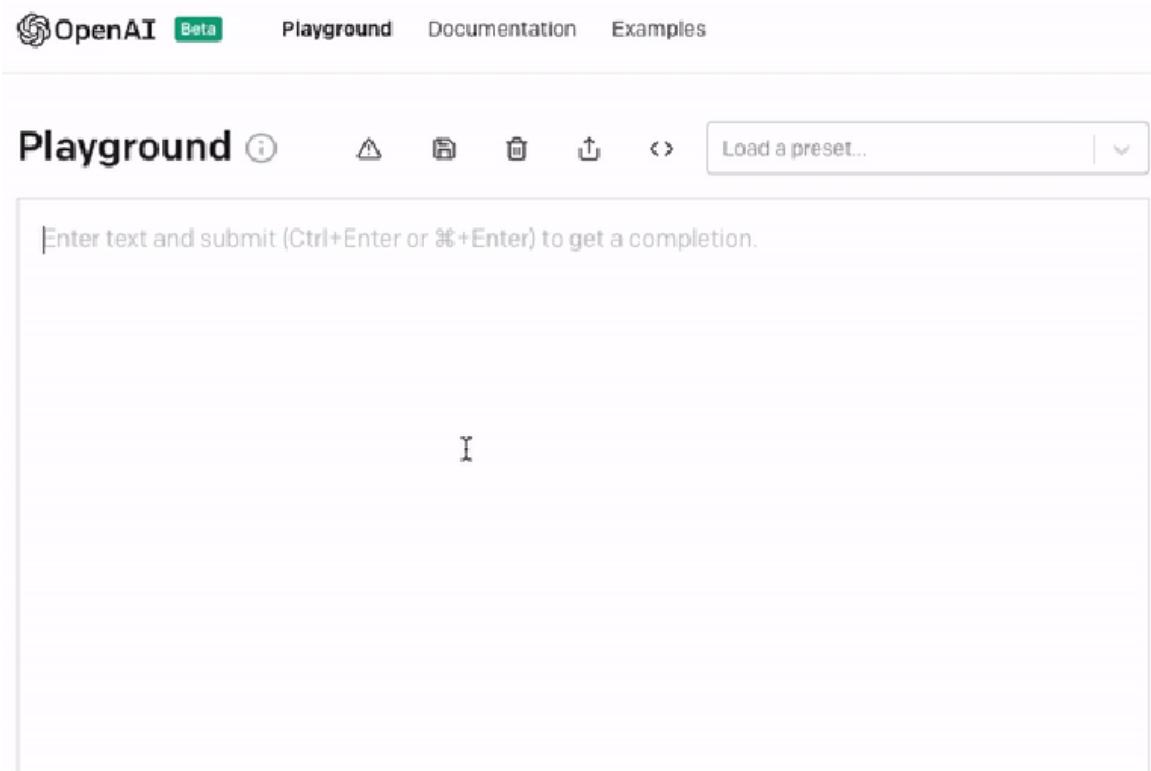
OpenAI Codex: First Impressions

This is a linkpost for <https://sixeleven.in/codex>

OpenAI organised a challenge to solve coding problems with the aid of an AI assistant. This is a review of the challenge, and first impressions on working with an AI pair-programmer.

OpenAI Codex

[OpenAI](#) is an AI research and development company. You might have heard some buzz about one of its products: [GPT-3](#). GPT-3 is a language model that can generate human-like text. It can be used for chatting, text auto-completion, text summarisation, grammar correction, translation, etc.



The screenshot shows the OpenAI Codex playground interface. At the top, there's a navigation bar with the OpenAI logo, a 'Beta' button, and links for 'Playground', 'Documentation', and 'Examples'. Below the navigation is a toolbar with icons for help, warning, file operations, and a search bar labeled 'Load a preset...'. A large text input area contains the placeholder text: 'Enter text and submit (Ctrl+Enter or ⌘+Enter) to get a completion.' A cursor is visible in the text area. On the left side of the text area, there's a vertical gray sidebar.

Checkout [OpenAI API](#) to access the playground.

[Codex](#) is a descendant of GPT-3, trained on natural language data and publicly available source-codes (e.g. from public GitHub repos). **Codex translates a natural language prompt to code.** It is the very model that powers [GitHub Copilot](#) — an AI pair-programmer (checkout the site for demos, it is fascinating).

When the rocket is clicked, temporarily display some text saying "Firing thrusters!" in white on the current location -- and temporarily speed up by 4x for 0.25 second.

```

'hidden';

/* Now set background to the
color of space. */
document.body.style.backgroundColor = 'black';

/* When the rocket is clicked,
temporarily display some text
saying 'Firing thrusters!' in
white on the current location
-- and temporarily speed up by
4x for 0.25 second. */
rocketship.onclick =
function() {
  var text =
  document.createElement( 'div' );
  text.innerHTML = 'Firing
thrusters!';
  text.style.color = 'white';
  text.style.position =
'absolute';
  text.style.left =
rocketship.offsetLeft + 'px';
  text.style.top =
rocketship.offsetTop + 'px';
  document.body.appendChild(text);
}

```

Credits: OpenAI

OpenAI recently released an API to access Codex (in beta). The [demos](#) attached with the release were a cause for consternation. Codex is proficient in a dozen (programming) languages. It can be used for code generation, refactoring, autocomplete, transpilation (translating source-code b/w languages), code explanation, etc. To show off Codex, OpenAI recently organised a challenge.

The Challenge

The [challenge](#) was to solve a series of (five) programming puzzles in [Python](#). The only twist — you can use Codex as a pair-programmer. It was a time-judged competition, with a temporal cap. Not surprisingly, Codex itself was a participant (not just as a helper)!

Meet your teammate

Codex won't just be a player in the challenge, it will also be your teammate. When solving problems, you'll be able to issue some number of queries to Codex to generate answers or complete your code. Just keep in mind that Codex works best as a partner rather than simply trying to complete the whole problem — the latter may take many tries.

```

1 from typing import List
2
3 def smallest_change(arr: List[int]) -> int:
4     """
5         Given an array 'arr' of integers, find the minimum
6         number of elements that need to be changed to make
7         the array palindromic.
8     """
9

```

CODEX IT **RUN** **SUBMIT →**

The problems were simple. ~830 "people" (Codex included) were able to solve all five of them. I had to solve the first two challenges manually (OpenAI server issues). "Had to"

because it was a race against time (& top 500 win an OpenAI t-shirt). For the other three, however, I was able to call in the cavalry (it was pretty climactic).

The novel experience of watching an AI auto-generate code is amazing. Just type a docstring — describing the procedure — and watch the code develop. If you're an old-time programmer, you'll get the notion when you experience it.

Illustration

I've illustrated one problem statement where I used Codex to generate a solution.

PROBLEM

Parse the given Python source code and return the list of full-qualified paths for all imported symbols, sorted in ascending lexicographic order.

CONSTRAINTS

The input will *not* contain any wildcard imports (`from ... import *`). Ignore aliases (renamings): `from x import y as z` should be represented as `x.y`.

LIBRARY SUGGESTION

Consider using the [\[ast\]](#) module.

EXAMPLES

Input

```
import os
import concurrent.futures
from os import path as renamed_path
from typing import ( List, Tuple )
```

Output

```
['concurrent.futures', 'os', 'os.path', 'typing.List', 'typing.Tuple']
```

Codex it!

I just formulated the docstring. Using the doc, imported libs and function signature, it generated an (almost) functional code:

```
1 import ast
2 from typing import List
3
4
5 def parse_imports(code: str) -> List[str]:
6     """
7         Parse all the imports in the code using ast module.
8         Imports of the form 'from x import y' should be
9         appended as 'x.y'. Ignore any wildcards and aliases.
10        Append each import type to a list and return the sorted
11        list."""
12
13
14 # Examples
15 print(parse_imports('import os'))
16 print(parse_imports('import os\nfrom typing import List'))
```

Used 2 of 10 assists

CODEX IT

RUN

SUBMIT →

Pretty impressive. After just one or two manual bug sweeps, the code passed all the testcases! Final script:

```
import ast
from typing import List

def parse_imports(code: str) -> List[str]:
    """
        Parse all the imports in the code using ast module.
        Imports of the form 'from x import y' should be appended as 'x.y'.
            Ignore any alias. Append each import type to a list
            and return the sorted list.
    """
    symbols = []
    for node in ast.walk(ast.parse(code)):
        if isinstance(node, ast.Import):
            for name in node.names:
```

```

        symbols.append(name.name)
    elif isinstance(node, ast.ImportFrom):
        for name in node.names:
            symbols.append(node.module + '.' + name.name)
print(code, symbols)
return sorted(symbols)

# Examples
print(parse_imports('import os'))
print(parse_imports('import os\nfrom typing import List'))

```

Implications

Although it could not beat all its human counterparts, it ranked an impressive #96 on the [leaderboard](#). In all fairness, the competition paradigm was many-to-some — everyone faced the same five problems. So, Codex will have a rich data of differentiated prompts for the same set of problems. It might give the AI a learning edge (in the case of concurrent active learning). Still, for competing against top-notch programmers, top 100 is quite a feat. I mean, contrast the statistics below (Codex vs Avg. player):



Problem vs cumulative time-taken ([source](#)). Also, yup, I'll win an OpenAI t-shirt!

Does this mean I should start scouting career options? Can Codex just self-optimize and outcompete all programmers? I doubt it.

Okay, let us go first-principles. Codex trained on public code repos. Now the underlying framework of weights and biases is impossible to entertain. So let us take a spherical cow approach. The constituents of its dataset will probably form a logarithmic distribution. Majority of the train split comprised of overlapping, generic, non-complex solutions (like database CRUDs). Basis this (sensible) hypothesis, we can assert that 80% of its statistical prowess lies in solving small, low cognitive, well-defined, pure functions.

Building webpages, APIs, 3rd party integrations, database CRUDs, etc. — 80% of non-creative, repetitive tasks can probably be automated. Going by the Pareto principle, the rest

20% — non-generalisable, complex, abstract problems — that take up 80% of cognitive bandwidth, will survive. But this is good news. Codex will handle all the tedious tasks, while programmers focus on the most creative jobs.

Once a programmer knows what to build, the act of writing code can be thought of as (1) breaking a problem down into simpler problems, and (2) mapping those simple problems to existing code (libraries, APIs, or functions) that already exist. The latter activity is probably the least fun part of programming (and the highest barrier to entry), and it's where OpenAI Codex excels most.

Individual Problem Comparison

Alluded to above, it outperformed me (singleton, for problems 1 & 2). Teamed up, however, I yielded a far greater performance metric. Complement, not replacement.

Performance Log

Aa Problem	≡ Team	≡ Solved In (Me)	≡ Solved In (Codex)
1	Me	50:04	22:09
2	Me	15:16	07:22
3	Me Codex	18:20	19:24
4	Me Codex	10:47	25:43
5	Me Codex	11:49	14:13

[Software is eating the world](#). Apparently, [at the expense of atoms](#). Yet, this asymmetrically entitled ecosystem is inaccessible to most. Programming demands a logical constitution. Tools like OpenAI Codex can loosen these constraints. Help dissolve the clique.



Sam Altman



@sama



Replies to @sama

I think Codex gets close to what most of us really want from computers—we say what we want, and they do it.

Programming languages are an artifact of computers not being able to actually understand us, and humans and computers relying on a lingua franca to understand each other.

5:38 PM · Aug 10, 2021



250



4



Copy link to Tweet

For programmers, these tools act as excellent appendages. Programming languages are, by definition, means of communication with computers. Consider Codex to be a [level-n](#) programming language (with intermediate transpilation). One step closer, in the overarching thread of symbiosis.

The Myth of the Myth of the Lone Genius

"Our species is the only creative species, and it has only one creative instrument, the individual mind and spirit of man. Nothing was ever created by two men. There are no good collaborations, whether in music, in art, in poetry, in mathematics, in philosophy. Once the miracle of creation has taken place, the group can build and extend it, but the group never invents anything. The preciousness lies in the lonely mind of a man."

- John Steinbeck

"The Great Man theory of history may not be truly believable and Great Men not real but invented, but it may be true we need to believe the Great Man theory of history and would have to invent them if they were not real."

- Gwern

The Myth of the Lone Genius is a bullshit cliche and we would do well to stop parroting it to young people like it is some deep insight into the nature of innovation. It typically goes something like this - the view that breakthroughs come from Eureka moments made by geniuses toiling away in solitude is inaccurate; in reality, most revolutionary ideas, inventions, innovation etc. come from lots of hard work, luck, and collaboration with others.

A screenshot of a Google search results page. The search bar at the top contains the query "the myth of the lone genius". Below the search bar are filter options: "All" (which is selected), "Videos", "Images", "News", and "More". To the right of these are "Settings" and "Tools" buttons. The search results section shows the following entries:

- The myth of the lone genius - NobelPrize.org**
www.nobelprize.org › martin-chalfie-npii-canada ...
The stories he'd heard about great scientists were of lone geniuses, who made their breakthroughs without the help of others. His conclusion was that, if he was ...
- Is the lone genius a total myth? - Vox**
Aug 17, 2014 — The conventional view of history is filled with lone geniuses: men and women who, through talent and inspiration, achieved feats no one else ...
- It's Time to Bury the Idea of the Lone Genius Innovator**
Apr 6, 2016 — Take a look at any significant innovation, and the myth of the lone genius and the "eureka moment" breaks down. First, a big idea or a new ...
- The Truth Behind the Lone Genius Myth | Wiley**
Aug 23, 2019 — So why does the idea of the "lone genius" hold appeal? Why does it stick around? We asked Dr. Jennifer Rohn, a scientist and novelist, to answer ...

Here is a good description of the myth from *The Ape that Understood the Universe: How the Mind and Culture Evolve* by psychologist Steve Stewart-Williams.

"We routinely describe our species' cultural achievements to lone-wolf geniuses - super-bright freaks of nature who invented science and technology for the rest of us. ... It's a myth because most ideas and most technologies come about not through Eureka moments of solitary geniuses but through the hard slog of large armies of individuals, each making—at best—a tiny step or two forward"

The problem here is that the myth of the lone genius is itself a myth. History (ancient and recent) is full of geniuses who came up with a revolutionary idea largely on their own - that's why the archetype even exists in the first place (Aristotle, Newton, Darwin, Einstein to name the most obvious examples). The author of the above quote would seem to grant that at least some ideas and technologies come from eureka moments of solitary geniuses. Others would seem to go further - the author of an article entitled "[The Myth of the Genius Solitary Scientist is Dangerous](#)" holds up Einstein, Maxwell, and Newton as examples of this archetype, but then exposes the falsehood of these examples by saying:

"Newton looked down on his contemporaries (while suspecting them of stealing his work) but regularly communicated with Leibniz, who was also working on the development of calculus. Maxwell studied at several prestigious institutions and interacted with many intelligent people. Even Einstein made the majority of his groundbreaking discoveries while surrounded by people with whom he famously used as sounding boards."

Uhhh ok, so they talked to other people while working on their ideas? Sure, we shouldn't have this naive view that these so-called solitary geniuses work 1000% on their own without any input whatsoever from other people, but that doesn't mean that they didn't do most of the heavy lifting. Similarly, another proponent of the myth of the lone genius focuses on the power of partnership (Joshua Shenk, Powers of Two: How Relationships Drive Creativity). From the introduction of an interview with Shenk on [Vox](#):

"After struggling for years trying to develop his special theory of relativity, Einstein got his old classmate Michele Besso a job at the Swiss patent office — and after "a lot of discussions with him," Einstein said, "I could suddenly comprehend the matter." Even Dickinson, a famous recluse, wrote hundreds of poems specifically for people she voraciously corresponded with by letter.

The idea isn't that all of these situations represent equal partnerships — but that the lone genius is a total myth, and all great achievements involve some measure of collaboration."

This seems contradictory - so there is still a dominant person in the partnership doing most (or all) of the difficult work, but at the same time the lone genius is a TOTAL myth. I have a feeling that Einstein's contribution was a little more irreplaceable than that of this Besso fellow. Is there not room for a more moderate position here? I guess that doesn't really sell books.

It's not hard to see why the myth of the lone genius is so popular - it is a very politically correct type of idea, very much going along with the general aversion to recognizing intelligence and genes as meaningful sources of variation in social/intellectual outcomes. It is also kind of a natural extension of the "you can achieve anything you set your mind to!" cliche. The fact that most of the geniuses in question are white men probably plays a not insignificant role in people's quickness to discredit their contributions. At the end of the day, it's really tough to admit that there are geniuses in the world and you aren't one of them.

Defenders of the myth would probably argue that the vast majority of people are not solitary geniuses and the vast majority of innovations do not come from people like this, so we should just preach the message that hard work and collaboration are what matters for innovation. In this view, the myth of the lone genius is a kind of noble lie - the lessons we impart by emphasizing the fallacy of the lone genius are more beneficial than the lessons imparted from uncritical acceptance of the lone genius story. I'm not sure this is true, and in fact I would argue that the uncritical acceptance of the myth of the lone genius is just as bad as uncritical acceptance of the lone genius story.

What lessons are we really trying to impart with the myth of the lone genius?

- (1) You are not just going to have a brilliant idea come to you out of thin air.
- (2) Creativity is enhanced by collaboration and sharing ideas with others. Most good ideas come from recombining pre-existing ideas.
- (3) Be humble and don't expect that it will be easy to find good ideas. No, you will not "solve" quantum mechanics after taking your first high school physics class.

Ok great, I'm on board with all of these lessons, it's kind of impossible not to be. The problem is that by harping so much on the fallacy of the lone genius we are also sending some implicit messages that are actively harmful to aspiring scientists/engineers/entrepreneurs.

- (4) There are no such things as geniuses, and even if there were you are not one of them.
- (5) You won't come up with a great idea by spending lots of time thinking deeply about something on your own. The people who think they can do this are crackpots.

(6) Thinking isn't real work and ideas are cheap, anything that doesn't produce something tangible is a waste of time. Go do some experiments, have a meeting, write a paper, etc.

(1)-(3) are certainly valuable lessons, but I think most relatively intelligent people eventually learn them on their own to some degree. My concern is that lessons (4)-(6) can become self-fulfilling prophecies - upon learning about how innovation really works from the myth of the lone genius, the next would-be revolutionary thinker will give up on that crazy idea she occasionally worked on in her free time and decide to devote more time to things like networking or writing academic papers that no one reads. We should want exceptional people to believe they can do exceptional things on their own if they work hard enough at it. If everyone internalizes the myth of the lone genius to such a degree that they no longer even try to become lone geniuses then the myth will become a reality.

My argument here is similar to the one that Peter Thiel makes about the general lack of belief in secrets in the modern world.

"You can't find secrets without looking for them. Andrew Wiles demonstrated this when he proved Fermat's Last Theorem after 358 years of fruitless inquiry by other mathematicians—the kind of sustained failure that might have suggested an inherently impossible task. Pierre de Fermat had conjectured in 1637 that no integers a , b , and c could satisfy the equation $a^n + b^n = c^n$ for any integer n greater than 2. He claimed to have a proof, but he died without writing it down, so his conjecture long remained a major unsolved problem in mathematics. Wiles started working on it in 1986, but he kept it a secret until 1993, when he knew he was nearing a solution. After nine years of hard work, Wiles proved the conjecture in 1995. He needed brilliance to succeed, but he also needed a faith in secrets. If you think something hard is impossible, you'll never even start trying to achieve it. Belief in secrets is an effective truth."

The actual truth is that there are many more secrets left to find, but they will yield only to relentless searchers. There is more to do in science, medicine, engineering, and in technology of all kinds. We are within reach not just of marginal goals set at the competitive edge of today's conventional disciplines, but of ambitions so great that even the boldest minds of the Scientific Revolution hesitated to announce them directly. We could cure cancer, dementia, and all the diseases of age and metabolic decay. We can find new ways to generate energy that free the world from conflict over fossil fuels. We can invent faster ways to travel from place to place over the surface of the planet; we can even learn how to escape it entirely and settle new frontiers. But we will never learn any of these secrets unless we demand to know them and force ourselves to look."

Maybe I'm overthinking all of this - does the myth of the lone genius really affect anyone's thinking in any substantial way? Maybe it only has the tiniest effect in the grand scheme of things. Even still, I would argue that it matters - uncritical acceptance of the lone genius myth is one more cultural force among many that is making it more and more difficult for individuals to do innovative work (and last time I checked, humanity is made up of individuals). In a fast-paced world full of intense economic/scientific/intellectual competition and decreasing opportunities for solitude, it is harder than ever before to justify spending significant time on intangible work that may or may not pay off. You can't put on your resume - "I spend a lot of time thinking about ideas and scribbling notes that I don't share with anyone."

I guess what I want to counteract is the same thing that Stephen Malina, Alexey Guzey, Leopold Aschenbrenner argue against in "[Ideas not mattering is a Psyop](#)". I don't know how we could ever forget that ideas matter - *of course they matter* - but somewhere along the way I think we got a little confused. How this happened, I don't know - you can probably broadly gesture at computers, the internet, big data, etc. and talk about how these have led to a greater societal emphasis on predictability, quantifiability, and efficiency. Ideas (and the creative process that produces them) are inherently none of these things; as Malina et al. remind us - Ideas are often built on top of each other, meaning that credit assignment is genuinely hard" and "Ideas have long feedback loops so it's hard to validate who is good at having ideas that turn out to be good". I would also mention increased levels of competition (as a result of globalism, increased population sizes, and the multitude of technologies that enable these things) as a major culprit. For any position at a college/graduate school/job you are likely competing with many people who have done all kinds of impressive sounding things (although it is probably 90% bullshit) so you better stop thinking about crazy ideas (remember, there are no such things as lone geniuses) and starting doing things, even if the things you are doing are boring and trivial. As long as they look good on the resume...

The life and times of Kary Mullis provide an illustration of this tension between individual genius and collaboration in the production of radical innovation. Kary Mullis is famous for two things - inventing the polymerase chain reaction (which he would win the nobel prize for) and having some very controversial views.

"A New York Times article listed Mullis as one of several scientists who, after success in their area of research, go on to make unfounded, sometimes bizarre statements in other areas. In his 1998 humorous autobiography proclaiming his maverick viewpoint, Mullis expressed disagreement with the scientific evidence supporting climate change and ozone depletion, the evidence that HIV causes AIDS, and asserted his belief in astrology. Mullis claimed climate change and HIV/AIDS theories were promulgated as a form of racketeering by environmentalists, government agencies, and scientists attempting to preserve their careers and earn money, rather than scientific evidence."

This is another reason why people are so leery of the lone genius - it often comes with a healthy dose of crazy. Yes, obviously this can go poorly - his ideas on AIDS did NOT age well - but, as we all know because there is an idiom for it, sometimes you have to break a few eggs to make an omelette.

"Mullis told Parade magazine: "I think really good science doesn't come from hard work. The striking advances come from people on the fringes, being playful"

Proponents of the lone genius myth might be wondering at this point - did Mullis really invent PCR all on his own in a brilliant flash of insight? We shouldn't be surprised that the answer is yes in fact he did, but also that it's a little more complicated than that.

"Mullis was described by some as a "diligent and avid researcher" who finds routine laboratory work boring and instead thinks about his research while driving and surfing. He came up with the idea of the polymerase chain reaction while driving along a highway."

"A concept similar to that of PCR had been described before Mullis' work. Nobel laureate H. Gobind Khorana and Kjell Kleppe, a Norwegian scientist, authored a paper 17 years earlier describing a process they termed "repair replication" in the Journal of Molecular Biology.[34] Using repair replication, Kleppe duplicated and then quadrupled a small synthetic molecule with the help of two primers and DNA polymerase. The method developed by Mullis used repeated thermal cycling, which allowed the rapid and exponential amplification of large quantities of any desired DNA sequence from an extremely complex template."

"His co-workers at Cetus, who were embittered by his abrupt departure from the company,[10] contested that Mullis was solely responsible for the idea of using Taq polymerase in PCR. However, other scientists have written that the "full potential [of PCR] was not realized" until Mullis' work in 1983, [35] and that Mullis' colleagues failed to see the potential of the technique when he presented it to them."

"Committees and science journalists like the idea of associating a unique idea with a unique person, the lone genius. PCR is thought by some to be an example of teamwork, but by others as the genius of one who was smart enough to put things together which were present to all, but overlooked. For Mullis, the light bulb went off, but for others it did not. This is consistent with the idea, that the prepared (educated) mind who is careful to observe and not overlook, is what separates the genius scientist from his many also smart scientists. The proof is in the fact that the person who has the light bulb go off never forgets the "Ah!" experience, while the others never had this photochemical reaction go off in their brains."

So what's the take-home message? Let's not treat the myth of the lone genius like it's gospel. Sometimes really smart people think long and hard about something and come up with an idea that changes the world. Yes, this happens very rarely and most innovation comes from the "hard slog of large armies of individuals, each making—at best—a tiny step or two forward", but if we aren't careful then these Eureka moments will become fewer and farther between and everything will be a hard slog. Let's do better by providing a more nuanced picture of innovation in which solitary exploration by "geniuses" and collaboration both play critical roles.

(originally posted at [Secretum Secretorum](#))

Applications for Deconfusing Goal-Directedness

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

Atonement for my sins towards deconfusion

I have [argued](#) that the deconfusion of goal-directedness should not follow what I dubbed “the backward approach”, that is starting from the applications for the concept and reverse-engineering its coherent existence (or contradiction) from there. I have also [argued](#) that deconfusion should always start and center around applications.

In summary, I was wrong. About the former

If deconfusion indeed starts at the applications, what about my arguments against the backward approach to goal-directedness? I wrote

The same can be said for all convergent instrumental subgoals in the paper, and as far as I know, every argument about AI risks using goal-directedness. In essence, the backward approach finds out that what is used in the argument is the fundamental property that the forward approach is trying to formalize. This means in turn that we should work on the deconfusion of goal-directedness from the philosophical intuitions instead of trying to focus on the arguments for AI risks, because these arguments depend completely on the intuitions themselves.

My best answer is this post: an exploration of applications of deconfusing goal-directedness, and how they actually inform and constrain the deconfusion itself. The gist is that in discarding a different approach than what felt natural to me, I failed to notice all the ways in which applications do constraint and direct deconfusion. In this specific case, the most fruitful and important applications I’ve found are convergent subgoals, i replacing optimal policies, formalizing inner alignment and separating approval-directed systems from pure maximizers.

Thanks to John S. Wentworth for pushing hard on the importance of starting at the applications.

Applications

Convergent subgoals

Convergent subgoals (self-preservation, resource acquisitions...) are often important ingredients in scenarios starting with misspecified objectives and ending with catastrophic consequences. Without them, even an AGI would let itself be shut down,

greatly reducing the related risks. Convergent subgoals are also clearly linked with goal-directedness, since [the original argument](#) proposes that most goals lead to them.

As an application for deconfusion, what does this entail? Well, goal-directed systems should be the whole class of systems that could have convergent subgoals. It doesn't necessarily mean that most goal-directed systems will actually have such convergent subgoals, but a low-goal-directed system shouldn't have them at all. **Hence high goal-directedness should be a necessary (but not necessarily sufficient) condition for having convergent subgoals.**

This constraint then point to concrete approaches for deconfusing goal-directedness that I'm currently pursuing:

- Search for informal necessary conditions to each convergent subgoal, and then try to see the links/common denominator. Here I am looking for requirements which are simpler than goal-directedness, because they will hopefully be components of it.
- List for each convergent subgoals examples of systems with and without this goal, and search for commonalities.
- Based on Alex's [deconfusion of power-seeking](#), look for a necessary condition **on the policies** for his theorems to hold.

Replacing optimal policies

When we talk about AGI having goals, we have a tendency to use optimal policies as a stand-in. These policies do have a lot of nice properties: they are maximally constrained by the goal, allow some reverse-engineering of goals without thinking about error models, and make it easy to predict what happens in the long-term -- optimality.

Yet as Richard points out in [this comment](#), true optimal policies for real world tasks are probably incredibly complex and intractable. It's fair to say that for any task we cannot just enumerate on, we probably haven't built an optimal policy. For example AlphaGo and its successors are very good at Go, but they are not strictly speaking optimal.

The above point wouldn't matter if optimal policies pretty much behaved just like merely competent ones. But that's not the case in general: usually the very optimal strategy is something incredibly subtle that uses many tricks and details that we have no way of finding except through exhaustive search. Notably, the reason we expect quantilizers to be less catastrophic than pure maximizers is indeed that difference between optimal behavior and competent one.

Because of this, focusing on optimal policies when thinking about goal-directedness might have two dangerous effects:

- If the problems we investigate only appear for optimal policies, then it is probably a waste of time to study them, as we won't build an optimal policy (or not for a very long time). And wasting resources might prove very bad for shorter timelines.
- If the problems we investigate also appear for merely competent goal-directed policies, but we have to wait for optimality before spotting them when

training/studying something, we're fucked because they will crop up way before that point.

What I take from this analysis is that **we want to replace the optimality assumption by goal-directedness + some competence assumption.**

Here we don't really have a necessary condition, so unraveling what the constraint entails is slightly more involved. But we can still look at the problems with optimality, and turn them into requirements for goal-directedness:

- Since optimal policies don't capture the competent policies we're actually building, we want goal-directedness to do it.
 - Possible approach: list the competent AI we're able to produce, and try to find some commonality beyond being good at their task.
- But not all policies should be goal-directed, or it becomes a useless category.
 - Possible approach: find examples of policies which we don't want to include in the goal-directed ones, possibly because there is no way for them to get convergent subgoals.
- Less sure, but I see the issues with optimality as stemming from an obsession with competence. This comfort me in [my early impression](#) that goal-directedness is more about "really trying to accomplish the goal" than in accomplishing it.
 - Find the commonality between not very competent goal-directed policies (like an average chess player), and try to formalize it.

Grounding inner alignment

[Risks from Learned Optimisation](#) introduced the concept of mesa-optimizers or inner optimizers to point to the results of search that might themselves be doing internal search/optimization. This has been consistently confusing, and people constantly complain about it. Abram has [a recent post](#) that looks at different ways of formalizing it.

In addition to the confusion, I believe that focusing on inner optimizers as currently defined underestimate the range of models that would be problematic to build, because I expect some goal-directed systems to not use optimization in this way or have explicit goals. I also expect goal-directedness to be easier to define than inner optimization, even if that probably comes from a bias.

Rephrasing, the application is that **goal-directedness should be a sufficient condition for the arguments in [Risks](#) to apply.**

The implications are quite obvious:

- Mesa-optimizers should be goal-directed
 - Possible approach: look for the components of a mesa-optimizers, and see what can be relaxed or abstracted while still keeping the whole intuitively goal-directed.
- Goal-directed systems should have the same problems/issues argued in [Risks](#).
 - Possible approach: find necessary conditions for the arguments in Risks to hold.
- Goal-directedness should be less confusing than inner-optimization/mesa-optimization.

- Possible approach: list all the issues and confusions people have with inner optimization, and turn that into constraints for a less confusing alternative.

Approval-directed systems as less goal-directed

This application is definitely more niche than the others, but it seems quite important to me. Both Paul in his [approval-directed post](#) and Rohin in [this comment](#) to one of his posts on goal-directedness have proposed that approval-directed systems are inherently less goal-directed than pure maximizers.

Why? Because approval-directed systems would have a more flexible goal, and also wouldn't not have the same convergent subgoals that we expect from competent goal-directed systems.

I share this intuition, but I haven't been able to find a way to actually articulate it convincingly. Hence why I add this constraint: **approval-directed systems should have low-goal-directedness (or at least lower than pure-maximizers)**

Since the constraint is quite obvious, let's focus on the approaches to goal-directedness this suggests.

- Deconfuse approval-directed systems as much as possible, to have a better idea of what their low goal-directedness would entail
- List all the intuitive differences between approval directed systems and highly goal-directed systems
- Look for sufficient conditions (on a definition of goal-directedness) for approval-directed systems to have low goal-directedness.

Conclusion: all that is left is work

In refusing to focus on the application, I slowed myself down in two ways:

- By going into weird tangents and nerd-snipe without a mean to check if the digression was relevant or not
- By missing out on the many approaches and research directions that emerge after even a cursory exploration of these applications.

I attempted to correct my mistake in this post, by looking at the most important applications for deconfusing goal-directedness (convergent subgoals, optimality, inner optimization and approval directedness), and extracting constraints and questions to investigate from them.

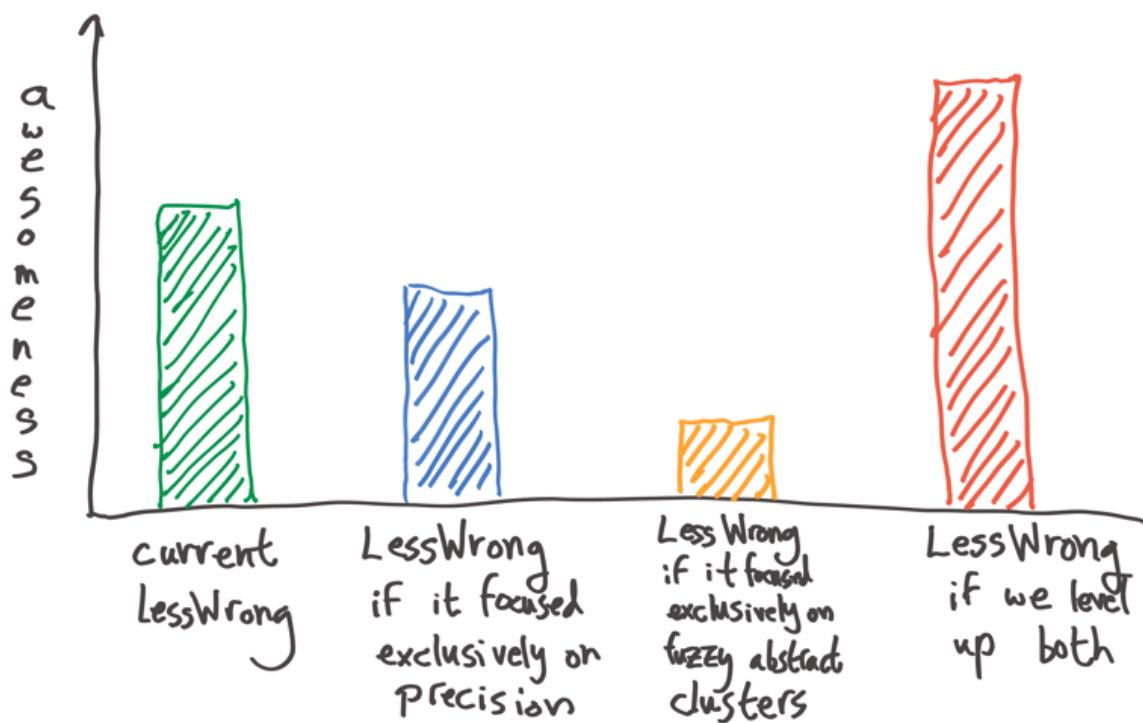
This cuts my work for me on the topic; if you find yourself interested or excited by any of the research ideas proposed in this post, send me a message so we can talk!

Power vs Precision

I've written a couple of posts ([1](#), [2](#)) about General Semantics and precision (IE, non-equivocation). These posts basically argue for developing superpowers of precision. In addition, I suspect there are far more [pro-precision posts](#) on LessWrong than the opposite (although most of these would be about specific examples, rather than general calls for precision, so it's hard to search for this to confirm).

A [comment](#) on my more recent general-semantics-inspired posts applied [the law of equal and opposite advice](#), pointing out that there are many advantages to abstraction/vagueness as well; so, why should we favor precision? At least, we should [chesterton-fence](#) the default level of precision before we assume that turning the dial up higher is better.

I totally agree that there are advantages to fuzzy clusters as well; what I actually think is that we should level up *both* specificity-superpowers and vagueness-superpowers. However, it does seem to me like there's something special about specificity. I *think* I would prefer a variant of LessWrong which valued specificity more to one which valued it less.



To possibly-somewhat explain why this might be, I want to back up a bit. I've been saying "specificity" or "precision", but what I *actually* mean is specificity/precision *in language and thought*. However, I think it will be useful to think about *physical* precision as an analogy.

Physical Power vs Precision

Power comes from muscles, bones, lungs, and hearts. Precision comes from eyes, brains, and fingers. (Speaking broadly.)

Probably the most broadly important determiner of what you can accomplish in the physical world is *power*, by which I mean the amount of physical force you can bring to bear. This

determines whether you can move the rock, open the jar, etc. In the animal kingdom, physical power determines who will win a fight.

However, *precision* is also incredibly important, especially for humans. Precision starts out being very important for the use of projectiles, a type of attack which is not *unique* to humans, but nearly so. It then becomes critical for crafting and tool use.

Precision is about creating very specific configurations of matter, often with very little force (large amounts of force are often harder to exercise precision with).

We could say that *power* increases the range of accessible states (physical configurations which we could realize), while *precision* increases the *optimization pressure* which we can apply to select from those states. A body-builder has all the physical strength necessary to weave a basket, but may lack the precision to bring it about.

(There are of course other factors than just precision which go into basket-weaving, such as knowledge; but I'm basically ignoring that distinction here.)

Generalizing

We can analogize the physical precision/power dichotomy to other domains. Social power would be something like the number of people who listen to you, whereas social precision would be the ability to say just the right words. Economic power would simply be the amount of money you have, while economic precision would involve being a discerning customer, as well as managing the logistics of a business to cut costs through efficiency gains.

In the mental realm, we could say that power means IQ, while precision means rationality. Being slightly more specific, we could say that "mental power" means things like raw processing power, short-term memory (working memory, RAM, cache), and long-term memory (hard drive, episodic memory, semantic memory). "Mental precision" is like running better software: having the right metacognitive heuristics. Purposefully approximating probability theory and decision theory, etc.

(This is very different from what I meant by precision in thought, earlier, but this post itself is very imprecise and cluster-y! So I'll ignore the distinction for now.)

Working With Others

In [Conceptual Specialization of Labor Enables Precision](#), Vaniver hypothesizes that people in the past who came up with "wise sayings" actually did know a lot, but were unable to convey their knowledge precisely. This results in a situation where the wise sayings convey relatively little of value to the ignorant, but later in life, people have "aha" moments where they suddenly understand a great deal about what was meant. (This can unfortunately result in a situation where the older people mistakenly think the wise sayings were quite helpful in the long run, and therefore, propagate these same sayings to younger people.)

Vaniver suggests that specialized fields avoid this failure mode, because they have the freedom to invent more specialized language to precisely describe what they're talking about, and the incentive to do so.

I think there's more to it.

Let's think about physical precision again. Imagine a society which cares an awful lot about [building beautiful rock-piles](#).

If you're stacking rocks with someone of low physical precision, then it matters a lot less that you have high precision yourself. You may be able to place a precisely balanced rock, but your partner will probably knock it out of balance. You will learn not to exercise your precision. (Or rather, you will use your precision to optimize for structures that will endure your partner's less-precise actions.)

So, to first approximation, ***the precision of a team of people can only be a little higher than the precision of its least precise member.***

Power, on the other hand, is not like this. The power of a group is basically the *sum* of the power of its members.

I think this idea also (roughly) carries over to non-physical forms of precision, such as rationality and conceptual/linguistic precision. In hundreds of ways, the most precise people in a group will learn to "[fuzz out](#)" their language to cope with the less-precise thinking/speaking of those around them. (For example, when dealing with a sufficiently large and diverse group, [you have about five words](#). This explains the "wise sayings" Vaniver was thinking of.)

For example, in [policy debates should not appear one-sided](#), Eliezer laments the conflation of a single pro or con with claims about the overall balance of pros/cons. But *in an environment where a significant fraction of people are prone to make this mistake, it's actively harmful to make the distinction. If I say "[wearing sunscreen is correlated with skin cancer](#)", many people will automatically confuse this with "you should not wear sunscreen".*

So, conflation can serve as the equivalent of a clumsy person tipping over carefully balanced rocks: it will incentivize the more precise people to act as if they were less precise (or rather, use their optimization-pressure to avoid sentences which are prone to harmful conflation, which leaves less optimization-pressure for other things).

This, imho, is one of the major forces behind "specialization enables precision": your specialized field will filter for some level of precision, which enables everybody to be more precise. Everyone in a chemistry lab knows how to avoid contamination and how to avoid being poisoned, so together, chemists can set up a more fragile and useful configuration of chemicals and tools. They could not accomplish this in, say, a public food court.

So, imho, part of what makes LessWrong a great place is the concentration of intellectual precision. This enables good thinking in a way that a high concentration of "good abstraction skill" (the skill of making good fuzzy clusters) does not.

Conversations about good fuzzy abstract clusters are no more or less important in principle. However, those discussions don't require a concentration of people with high precision. By its nature, fuzzy abstract clustering doesn't get hurt too much by conflation. This kind of thinking is like moving a big rock: all you need is horsepower. Precise conversations, on the other hand, can only be accomplished with similarly precise people.

(An interesting nontrivial prediction of this model is that clustering-type cognitive labor, like moving a big rock, can easily benefit from a large number of people; mental horsepower is easily scaled by adding people, even though mental precision can't be scaled like that. I'm not sure whether this prediction holds true.)

Research agenda update

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

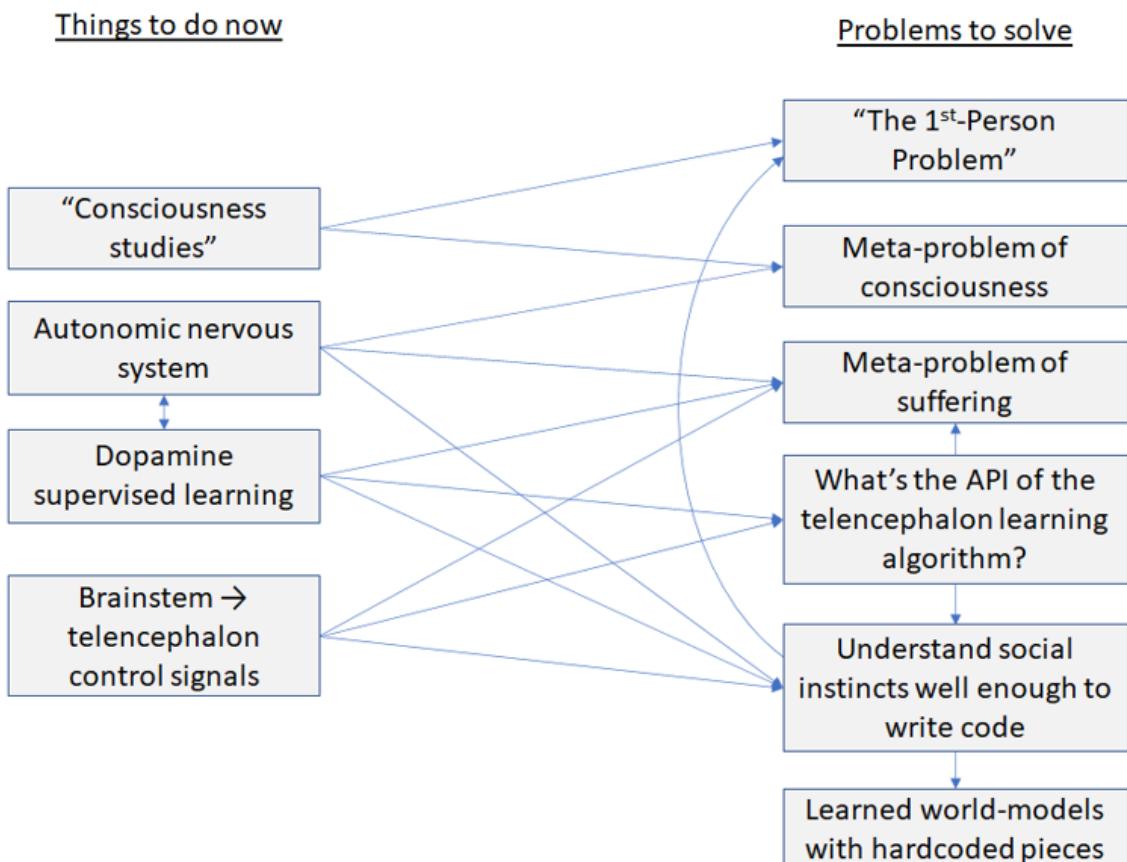
"Our greatest fear should not be of failure, but of succeeding at something that doesn't really matter." -DL Moody (allegedly)

(Related: [Solving the whole AGI control problem, version 0.0001](#). Also related: [all my posts](#).)

I'm five months into [my new job](#)! This here is a *forward-looking* post where I write down what I'm planning to work in the near future and why. Please comment (or otherwise [reach out](#)) if you think I'm prioritizing wrong or going in a sub-optimal direction for Safe & Beneficial AGI. That's the whole reason this post exists!

(UPDATE 5 MONTHS LATER: I should have been clearer: This post is primarily the "learning about neuroscience" aspects of my research agenda. The real goal is to design safe & beneficial AGI, and my learning about neuroscience is just one piece of that. As it turns out, in the few months since I wrote this, the great majority of my time did not really involve making progress on any of the things herein.)

I'll work backwards: first the intermediate "problems to solve" that I think would help for the AGI control problem, then the immediate things I'm doing to work towards those goals.



An arrow X→Y means "I think doing/solving X will be helpful for doing/solving Y".

1. Intermediate problems to solve

1.1 “The 1st-Person Problem”

(See [here](#).)

It's easy (or at least tractable) to find or generate tons of labeled data with safety-critical concepts and situations. For example, we can easily find a YouTube video where Alice is deceiving Bob, and label it "deception".

But is that useful? The thing we want to *find and/or reinforce* in our AGI's world-model is not quite that. Instead it's concepts related to the AGI's own actions. We want the AGI to think ""I am deceiving Bob" is a very bad thing". We don't want the AGI to think ""Alice is deceiving Bob" is a bad thing", or that the abstract concept of deception is a bad thing.

In other words, getting 3rd-person labeled data is easy but unhelpful. What we really want is 1st-person labeled data.

OK, so then we can say, forget about the YouTube videos. Instead we'll do RL, or something like it. We'll have the AGI do things, and we'll label the things it does.

That is indeed a solution. But it's a *really dangerous* one! We would have to catch the AGI in the act of deception before we could label deception as bad.

The real golden ticket would be if we could somehow magically transmute third-person labeled data into 1st-person labeled data (or at least something equivalent to 1st-person labeled data).

How do we do that? I don't know. I call this the "1st-Person Problem".

Incidentally, in the diagram above, I put an arrow indicating that understanding social emotions might be helpful for solving the 1st-Person Problem. Why would I think that? Well, what I have in mind here is that, for example, if a 3-year-old sees a 7-year-old doing something, then often the 3-year-old immediately wants to do the exact same thing. So I figure, it's at least *possible* that somewhere in the algorithms of human social instincts is buried a solution to the 1st-Person Problem. Maybe, maybe not.

1.2 The meta-problem of consciousness

(See [Book review: Rethinking Consciousness](#))

I am strongly averse to becoming really competent to talk about the philosophy of consciousness. Well, I mean, it would be *really fun*, but it seems like it would also be really time-consuming, with all the terminology and theories and arguments and counter-arguments and whatnot, and I have other priorities right now.

But there's a different question which is: there's a set of observable psychological phenomena, where people declare themselves conscious, muse on the ineffable nature of consciousness, write papers about the hard problem of consciousness, and so on. Therefore there has to be some explanation, in terms of brain algorithms, for the chain of events that eventually leads to people talking about how they're conscious. It seems to me that unravelling that chain of events is purely a question of neuroscience and psychology, not philosophy.

The "meta-problem of consciousness"—a.k.a. "Why do people believe that there's a hard problem of consciousness"—is about unraveling this chain of events.

What's the point? Well, if I truthfully declare "I am wearing a watch", and we walk through the chain of events that led to me saying those words, we'd find that one of the links in the chain is an actual physical watch that I'm actually physically wearing. So by the same token, it seems to me that a complete solution to the meta-problem of consciousness should directly lead to a solution to the hard problem of consciousness. (Leaving aside some fine print—see [Yudkowsky on zombies](#).)

In terms of AGI, it seems to me that knowing whether or not AGI is conscious is an important thing to know, at least for the AGI's sake. (Yeah I know—as if we don't already have our hands full thinking about the impacts of AGI on *humans* !)

So working on the meta-problem of consciousness seems like one thing worth doing.

That said, I'm tentatively feeling OK about what I wrote [here](#)—basically my proto-theory of the meta-problem of consciousness is some idiosyncratic mishmash of Michael Graziano's "Attention Schema Theory", and Stanislas Dehaene's "Global Workspace Theory", and Keith Frankish's "Illusionism". I'll stay on the lookout for reasons to think that what I wrote there was wrong, but otherwise I think this topic is not high on my immediate priority list.

1.3 The meta-problem of suffering

The motivation here is exactly like the previous section: People talk about suffering. The fact that they are speaking those words is an observable psychological fact, and must have an explanation involving neuroscience and brain algorithms. And whatever that explanation is, it would presumably get us most of the way towards understanding suffering, and thus towards figuring out which AI algorithms are or aren't suffering.

Unlike the previous section, I *do* think I'm pretty likely to do some work here, even in the short term, partly because I'm still pretty generally confused and curious about this. More importantly, it's very closely tied to other things in AGI safety that I need to be working on anyway. In particular, it seems pretty intimately tied to decision-making and motivation—for example, we are motivated to not suffer, and we make decisions that lead to not-suffering. I've already written a lot about [decision-making](#) and [motivation](#) and plan to write more, because understanding and sculpting AGI motivation is a huge part of AGI safety. So with luck I'll have something useful to say about the meta-problem of suffering at some point.

1.4 What's the API of the telencephalon?

I think the telencephalon (neocortex, hippocampus, amygdala, part of the basal ganglia) is running a learning algorithm (well, [several learning algorithms](#)) that starts from scratch at birth ("random weights" or something equivalent) and learns a big complicated world-model etc. within the animal's lifetime. See "learning-from-scratch-ism" discussion [here](#). And the brainstem and hypothalamus are mostly hardcoded algorithms that try to "steer" that learning algorithm towards doing useful things.

What are the input channels through which the brainstem can "steer" the telencephalon learning algorithms? I originally thought that there was a one-dimensional signal called "reward", and that's it. But when I looked into it, I found a *bunch* more things! *And all those things were at least plausibly promising ideas for AGI safety!*

Two things in particular:

(1) a big reason that I wrote [Reward Is Not Enough](#) was as an excuse to promote the idea of supplying different reward signals for different subsystems, especially subsystems that report back to the supervisor. And where did I get *that* idea? From the telencephalon API of course!

(2) I'm pretty excited by the idea of interpretability and steering via supervised learning (see the "More Dakka" comment in the figure caption [here](#)), especially if we can solve the "1st-Person Problem" above. And where did I get *that* idea? From the telencephalon API of course!

So this seems like fertile soil to keep digging. There are certainly other things in the telencephalon API too. I would like to understand them!

1.5 Understand human social instincts well enough to implement them in code

See [Section 5 of "Solving the Whole AGI Control Problem Version 0.0001](#).

1.6 Learned world-models with hardcoded pieces

My default presumption is that our AGIs will learn a world-model from scratch—loosely speaking, they'll find patterns in sensory inputs and motor outputs, and then patterns in the patterns, and then patterns in the patterns ... etc.

But it might be nice instead to hard-code "how the AGI will think about certain aspects of the world". The obvious use-case here is that the programmer declares that there's a thing called "humans", and they're basically cartesian, and they are imperfectly rational according to such-and-such model of imperfect rationality, etc. Then these hard-coded items can be referents when we define the AGI's motivations, rewards, etc.

For example, lots of discussion of IRL and value learning seem to presuppose that we're writing code that tells the AGI specifically how to model a human. To pick a random example, in [Vanessa Kosoy's 2018 research agenda](#), the "demonstration" and "learning by teaching" ideas seem to rely on being able to do that—I don't see how we could possibly do those things if the whole world-model is a bunch of unlabeled patterns in patterns in patterns in sensory input etc.

So there's the problem. How do we build a learned-from-scratch world model but shove little hardcoded pieces into it? How do we ensure that the hardcoded pieces wind up in the right place? How do we avoid the AGI ignoring the human model that we supplied it and instead building a parallel independent human model from scratch?

2. Things to do right now

2.1 "Consciousness studies"

I think there's a subfield of neuroscience called "consciousness studies", where they talk a lot about how people formulate thoughts about themselves, etc. The obvious application is understanding consciousness, but I'm personally much more interested in whether it could help me think about The 1st-Person Problem. So I'm planning to dive into that sometime soon.

2.2 Autonomic nervous system, dopamine supervised learning

I have this theory (see the "Plan assessors" at [A model of decision-making in the brain \(the short version\)](#)) that parts of the telencephalon are learning algorithms that are trained by

supervised learning, with dopamine as the supervisory signal. I think this theory (if correct) is just super duper important for everything I'm interested in, and I'm especially eager to take that idea and build on it, particularly for social instincts and motivation and suffering and so on.

But I'm concerned that I would wind up building idiosyncratic speculative theories on top of idiosyncratic speculative theories on top of

So I'm swallowing my impatience, and trying to *really nail down* dopamine supervised learning—keep talking to experts, keep filling in the gaps, keep searching for relevant evidence. And *then* I can feel better about building on it.

For example:

It turns out that these dopamine-supervised-learning areas (e.g. anterior cingulate, medial prefrontal cortex, amygdala, ventral striatum) are all intimately involved in the autonomic nervous system, so I'm hoping that reading more about that will help resolve some of my confusions about those parts of the brain (see Section 1 [here](#) for what exactly my confusions are).

I also think that I kinda wound up *pretty close* to the [somatic marker hypothesis](#) (albeit via a roundabout route), so I want to understand the literature on that, incorporate any good ideas, and relate it to what I'm talking about.

Since the autonomic nervous system is related to “feelings”, learning about the autonomic nervous system could also help me understand suffering and consciousness and motivation and social instincts.

2.3 Brainstem → telencephalon control signals

As mentioned above (see “What’s the API of the telencephalon?”), there are a bunch of signals going from the brainstem to the telencephalon, and I only understand a few of them so far, and I’m eager to dive into others. [I’ve written a bit about acetylcholine](#) but I have more to say, I have a hunch that it plays a role in cortical specialization, with implications for transparency. I know pretty much nothing about serotonin, norepinephrine, opioids, etc. I mentioned [here](#) that I’m confused about how the cortex learns motor control from scratch, or if that’s even possible, and how I’m also confused about the role of various signals going back and forth between the midbrain and cortex. So those are all things I’m eager to dive into.