

Best of LessWrong: July 2021

1. [DeepMind: Generally capable agents emerge from open-ended play](#)
2. [Working With Monsters](#)
3. [The topic is not the content](#)
4. [Going Out With Dignity](#)
5. [Delta Strain: Fact Dump and Some Policy Takeaways](#)
6. [Potential Bottlenecks to Taking Over The World](#)
7. [How much chess engine progress is about adapting to bigger computers?](#)
8. [\(\\$1000 bounty\) How effective are marginal vaccine doses against the covid delta variant?](#)
9. [Experimentally evaluating whether honesty generalizes](#)
10. [Covid 7/29: You Play to Win the Game](#)
11. [A Contamination Theory of the Obesity Epidemic](#)
12. [We have some evidence that masks work](#)
13. [My Marriage Vows](#)
14. [One Study, Many Results \(Matt Clancy\)](#)
15. [Relentlessness](#)
16. [Intermittent Distillations #4: Semiconductors, Economics, Intelligence, and Technological Progress.](#)
17. [Benchmarking an old chess engine on new hardware](#)
18. [Slack Has Positive Externalities For Groups](#)
19. [\[Link\] Musk's non-missing mood](#)
20. [The shoot-the-moon strategy](#)
21. [An Apprentice Experiment in Python Programming](#)
22. [Covid 7/8: Delta Takes Over](#)
23. [Book Review: Order Without Law](#)
24. [Covid 7/15: Rates of Change](#)
25. [Academic Rationality Research](#)
26. [Improving capital gains taxes](#)
27. [Covid 7/22: Error Correction](#)
28. [Winston Churchill, futurist and EA](#)
29. [What does knowing the heritability of a trait tell me in practice?](#)
30. [BASALT: A Benchmark for Learning from Human Feedback](#)
31. [Covid 7/1: Don't Panic](#)
32. [Fractional progress estimates for AI timelines and implied resource requirements](#)
33. [Open Philanthropy is seeking proposals for outreach projects](#)
34. [Reflecting on building my own tools from scratch and 'inventing on principle'](#)
35. [A closer look at chess scalings \(into the past\)](#)
36. [Chess and cheap ways to check day to day variance in cognition](#)
37. [New Dementia Trial Results](#)
38. [\(Brainstem, Neocortex\) ≠ \(Base Motivations, Honorable Motivations\)](#)
39. [AlphaFold 2 paper released: "Highly accurate protein structure prediction with AlphaFold". Jumper et al 2021](#)
40. [Book review: The Explanation of Ideology](#)
41. [For some, now may be the time to get your third Covid shot](#)
42. [The SIA population update can be surprisingly small](#)
43. [Black ravens and red herrings](#)
44. [Imaginary reenactment to heal trauma – how and when does it work?](#)
45. [Risk Premiums vs Prediction Markets](#)
46. [Typology of blog posts that don't always add anything clear and insightful](#)
47. [A \(somewhat beta\) site for embedding betting odds in your writing](#)
48. [\[IAN #156\]: The scaling hypothesis: a plan for building AGI](#)
49. [How much compute was used to train DeepMind's generally capable agents?](#)
50. [Fire Law Incentives](#)

Best of LessWrong: July 2021

1. [DeepMind: Generally capable agents emerge from open-ended play](#)
2. [Working With Monsters](#)
3. [The topic is not the content](#)
4. [Going Out With Dignity](#)
5. [Delta Strain: Fact Dump and Some Policy Takeaways](#)
6. [Potential Bottlenecks to Taking Over The World](#)
7. [How much chess engine progress is about adapting to bigger computers?](#)
8. [\(\\$1000 bounty\) How effective are marginal vaccine doses against the covid delta variant?](#)
9. [Experimentally evaluating whether honesty generalizes](#)
10. [Covid 7/29: You Play to Win the Game](#)
11. [A Contamination Theory of the Obesity Epidemic](#)
12. [We have some evidence that masks work](#)
13. [My Marriage Vows](#)
14. [One Study, Many Results \(Matt Clancy\)](#)
15. [Relentlessness](#)
16. [Intermittent Distillations #4: Semiconductors, Economics, Intelligence, and Technological Progress.](#)
17. [Benchmarking an old chess engine on new hardware](#)
18. [Slack Has Positive Externalities For Groups](#)
19. [\[Link\] Musk's non-missing mood](#)
20. [The shoot-the-moon strategy](#)
21. [An Apprentice Experiment in Python Programming](#)
22. [Covid 7/8: Delta Takes Over](#)
23. [Book Review: Order Without Law](#)
24. [Covid 7/15: Rates of Change](#)
25. [Academic Rationality Research](#)
26. [Improving capital gains taxes](#)
27. [Covid 7/22: Error Correction](#)
28. [Winston Churchill, futurist and EA](#)
29. [What does knowing the heritability of a trait tell me in practice?](#)
30. [BASALT: A Benchmark for Learning from Human Feedback](#)
31. [Covid 7/1: Don't Panic](#)
32. [Fractional progress estimates for AI timelines and implied resource requirements](#)
33. [Open Philanthropy is seeking proposals for outreach projects](#)
34. [Reflecting on building my own tools from scratch and 'inventing on principle'](#)
35. [A closer look at chess scalings \(into the past\)](#)
36. [Chess and cheap ways to check day to day variance in cognition](#)
37. [New Dementia Trial Results](#)
38. [\(Brainstem, Neocortex\) ≠ \(Base Motivations, Honorable Motivations\)](#)
39. [AlphaFold 2 paper released: "Highly accurate protein structure prediction with AlphaFold", Jumper et al 2021](#)
40. [Book review: The Explanation of Ideology](#)
41. [For some, now may be the time to get your third Covid shot](#)
42. [The SIA population update can be surprisingly small](#)
43. [Black ravens and red herrings](#)
44. [Imaginary reenactment to heal trauma - how and when does it work?](#)
45. [Risk Premiums vs Prediction Markets](#)
46. [Typology of blog posts that don't always add anything clear and insightful](#)

47. [A \(somewhat beta\) site for embedding betting odds in your writing](#)
48. [\[AN #156\]: The scaling hypothesis: a plan for building AGI](#)
49. [How much compute was used to train DeepMind's generally capable agents?](#)
50. [Fire Law Incentives](#)

DeepMind: Generally capable agents emerge from open-ended play

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

This is a linkpost for <https://deepmind.com/blog/article/generally-capable-agents-emerge-from-open-ended-play>

EDIT: Also see [paper](#) and [results compilation video!](#)

Today, we published "[Open-Ended Learning Leads to Generally Capable Agents](#)," a preprint detailing our first steps to train an agent capable of playing many different games without needing human interaction data. ... The result is an agent with the ability to succeed at a wide spectrum of tasks — from simple object-finding problems to complex games like hide and seek and capture the flag, which were not encountered during training. We find the agent exhibits general, heuristic behaviours such as experimentation, behaviours that are widely applicable to many tasks rather than specialised to an individual task.

...

The neural network architecture we use provides an attention mechanism over the agent's internal recurrent state — helping guide the agent's attention with estimates of subgoals unique to the game the agent is playing. We've found this goal-attentive agent (GOAT) learns more generally capable policies.

...

Playing roughly 700,000 unique games in 4,000 unique worlds within XLand, each agent in the final generation experienced 200 billion training steps as a result of 3.4 million unique tasks. At this time, our agents have been able to participate in every procedurally generated evaluation task except for a handful that were impossible even for a human. And the results we're seeing clearly exhibit general, zero-shot behaviour across the task space — with the frontier of normalised score percentiles continually improving.

Looking qualitatively at our agents, we often see general, heuristic behaviours emerge — rather than highly optimised, specific behaviours for individual tasks. Instead of agents knowing exactly the "best thing" to do in a new situation, we see evidence of agents experimenting and changing the state of the world until they've achieved a rewarding state. We also see agents rely on the use of other tools, including objects to occlude visibility, to create ramps, and to retrieve other objects. Because the environment is multiplayer, we can examine the progression of agent behaviours while training on held-out [social dilemmas](#), such as in a game of "[chicken](#)". As training progresses, our agents appear to exhibit more cooperative behaviour when playing with a copy of themselves. Given the nature of the environment, it is difficult to pinpoint intentionality — the behaviours we see often appear to be accidental, but still we see them occur consistently.

My hot take: This seems like a somewhat big deal to me. It's what I would have predicted, but that's scary, given my timelines. I haven't read the paper itself yet but I

look forward to seeing more numbers and scaling trends and attempting to extrapolate... When I do I'll leave a comment with my thoughts.

EDIT: My warm take: The details in the paper back up the claims it makes in the title and abstract. This is the GPT-1 of agent/goal-directed AGI; it is the proof of concept. Two more papers down the line (and a few OOMs more compute), and we'll have the agent/goal-directed AGI equivalent of GPT-3. Scary stuff.

Working With Monsters

This is a fictional piece based on [Sort By Controversial](#). You do not need to read that first, though it may make Scissor Statements feel more real. Content Warning: semipolitical. Views expressed by characters in this piece are not necessarily the views of the author.

I stared out at a parking lot, the pavement cracked and growing grass. A few cars could still be seen, every one with a shattered windshield or no tires or bashed-in roof, one even laying on its side. Of the buildings in sight, two had clearly burned, only blackened reinforced concrete skeletons left behind. To the left, an overpass had collapsed. To the right, the road was cut by a hole four meters across. Everywhere, trees and vines climbed the remains of the small city. The collapsed ceilings and shattered windows and nests of small animals in the once-hospital behind me seemed remarkably minor damage, relatively speaking.

Eighty years of cryonic freeze, and I woke to a post-apocalyptic dystopia.

"It's all like that," said a voice behind me. One of my... rescuers? Awakeners. He went by Red. "Whole world's like that."

"What happened?" I asked. "Bioweapon?"

"Scissor," replied a woman, walking through the empty doorway behind Red. Judge, he'd called her earlier.

I raised an eyebrow, and waited for elaboration. Apparently they expected a long conversation - both took a few seconds to get comfortable, Red leaning up against the wall in a patch of shade, Judge righting an overturned bench to sit on. It was Red who took up the conversation thread.

"Let's start with an ethical question," he began, then laid out a simple scenario. "So," he asked once finished, "blue or green?"

"Blue," I replied. "Obviously. Is this one of those things where you try to draw an analogy from this nice obvious case to a more complicated one where it isn't so obvious?"

"No," Judge cut in, "It's just that question. But you need some more background."

"There was a writer in your time who coined the term 'scissor statement,'" Red explained, "It's a statement optimized to be as controversial as possible, to generate maximum conflict. To get a really powerful scissor, you need AI, but the media environment of your time was already selecting for controversy in order to draw clicks."

"Oh no," I said, "I read about that... and the question you asked, green or blue, it seems completely obvious, like anyone who'd say green would have to be trolling or delusional or a threat to society or something... but that's exactly how scissor statements work..."

"Exactly," replied Judge. "The answer seems completely obvious to everyone, yet people disagree about which answer is obviously-correct. And someone with the

opposite answer seems like a monster, a threat to the world, like a serial killer or a child torturer or a war criminal. They need to be put down for the good of society."

I hesitated. I knew I shouldn't ask, but... "So, you two..."

Judge casually shifted position, placing a hand on some kind of weapon on her belt. I glanced at Red, and only then noticed that his body was slightly tensed, as if ready to run. Or fight.

"I'm a blue, same as you," said Judge. Then she pointed to Red. "He's a green."

I felt a wave of disbelief, then disgust, then fury. It was so *wrong*, how could anyone even *consider* green... I took a step toward him, intent on punching his empty face even if I got shot in the process.

"Stop," said Judge, "unless you want to get tazed." She was holding her weapon aimed at me, now. Red hadn't moved. If he had, I'd probably have charged him. But Judge wasn't the monster here... wait.

I turned to Judge, and felt a different sort of anger.

"How can you just stand there?", I asked. "You know that he's in the wrong, that he's a monster, that he deserves to be put down, preferably slowly and painfully!" I was yelling at Judge, now, pointing at Red with one hand and gesticulating with the other. "How can you work with him!?"

Judge held my eyes for a moment, unruffled, before replying. "Take a deep breath," she finally said, "calm yourself down, take a seat, and I'll explain."

I looked down, eyed the tazer for a moment, closed my eyes, then did as she asked. Breathe in, breathe out. After a few slow breaths, I glanced around, then chose a fallen tree for a seat - positioning Judge between Red and myself. Judge raised an eyebrow, I nodded, and she resumed her explanation.

"You can guess, now, how it went down. There were warning shots, controversies which were bad but not bad enough to destroy the world. But then the green/blue question came along, the same question you just heard. It was almost perfectly split, 50/50, cutting across political and geographical and cultural lines. Brothers and sisters came to blows. Bosses fired employees, and employees sued. Everyone thought they were in the right, that the other side was blatantly lying, that the other side deserved punishment while their side deserved an apology for the other side's punishments. That they had to stand for what was right, bravely fight injustice, that it would be *wrong* to back down."

I could imagine it. What I felt, toward Red - it felt *wrong* to overlook that, to back down. To let injustice pass unanswered.

"It just kept escalating, until bodies started to pile up, and soon ninety-five percent of the world population was dead. Most people didn't even *try* to hole up and ride out the storm - they *wanted* to fight for what was right, to bring justice, to keep the light in the world."

Judge shrugged, then continued. "There are still pockets here and there, where one side or the other gained the upper hand and built a stronghold. Those groups still fight each other. But most of what's left is ruins, and people like us who pick over them."

"So why aren't you fighting?" I asked. "How can you overlook it?"

Judge sighed. "I was a lawyer, before Scissor." She jerked her head toward Red. "He was too. We even came across each other, from time to time. We were both criminal defense attorneys, with similar clients in some ways, though very different motivations.

"Red was... not exactly a bleeding heart, but definitely a man of principles. He'd made a lot of money early on, and mostly did pro-bono work. He defended the people nobody else would take. Child abusers, serial killers, monsters who everyone knew were guilty. Even Red thought they were guilty, and deserved life in prison, maybe even a death sentence. But he was one of those people who believed that even the worst criminals had to have a proper trial and a strong defense, because it was the only way our system could work. So he defended the monsters. Man of principle.

"As for me, I was a mob lawyer. I defended gangsters, loan sharks, arms dealers... and their friends and families. It was the families who were the worst - the brothers and sons who sought sadistic thrills, knowing they'd be protected. But it was interesting work, the challenge of defending the undefendable, and it paid a fortune.

"We hated each other, back in the day. Still do, on some level. He was the martyr, the white knight putting on airs of morality while defending monsters. And I was the straightforward villain, fighting for money and kicks. But when Scissor came, we had one thing in common: we were both willing to work with monsters. And that turned out to be the only thing which mattered."

I nodded. "So you hated each other, but you'd both spent years working with people you hated, so working with each other was... viable. You even had a basis to trust one another, in some weird way, because you each *knew* that the other could work with people they hated."

"Exactly. In the post-scissor world, people who can work with monsters are basically the only people left. We form mixed groups - Red negotiates with Greens for us, I negotiate with Blues. They can tell, when they ask whether you're Blue or Green - few people can lie convincingly, with that much emotion wrapped up in it. A single-color group would eventually encounter the opposite single-color group, and they'd kill each other. So when we meet other groups, they have some Blues and some Greens, and we don't fight about it. We talk, we trade, we go our separate ways. We let the injustice sit, work with the monsters, because that's the only way to survive in this world.

"And now you have to make a choice. You can go out in a blaze of glory, fight for what you know is right, and maybe take down a few moral monsters in the process. Or you can choose to live and let live, to let injustice go unanswered, to work with the monsters you hate. It's up to you."

The topic is not the content

This is a linkpost for <https://aaronbergman.substack.com/p/the-topic-is-not-the-content>

Disclaimer: I've never held a job for more than a year [1] or been paid more than \$15 an hour. Take everything I say with a grain of salt.

Many of my peers seem to make career plans like by asking things like

1. What am I interested in, or what do I like doing? and
2. How can I do something *related to that?*

which might lead to some of the following:

- A person who likes dancing tries to work in the arts industry.
- A person who likes video games tries to get into game design.
- A person who is interested in healthcare policy tries to study or design healthcare policy.

The problem here, in terms of diminished performance, happiness, and satisfaction, is a conflation of the topic and the content. **The topic is not the content!**

The topic

In my schema, the topic is what the work is *about*. If you're the manager of a pillow company, the topic is pillows. If you're defending accused criminals in court, the topic is criminal law.

A lot of folks, it seems to me, focus a lot on the topic when deciding which subjects to study or which jobs to apply for. Someone who is interested in physics might major in physics. Someone who loves to work out might try to become a personal trainer.

I don't think this makes much sense. Should people ignore what they're interested in and like to do, then? Well, maybe. [80,000 Hours](#), perhaps the single best career planning resource out there, writes that

The bottom line

To find a dream job, look for:

1. Work you're good at,
2. Work that helps others,
3. Supportive conditions: engaging work that lets you enter a state of flow; supportive colleagues; lack of major negatives like unfair pay; and work that fits your personal life.

Ok, but the term "engaging work" is doing a lot of work here (no pun intended), and seems awfully synonymous with "work that you like." So, how do you find work that you like doing? If there's anything my utterly negligible work experiences has taught me, it's that it usually makes more sense to focus less on the topic and more on the content.

The content

In my schema, the content is what the work involves *doing*. If you're a physics teacher, the topic is physics, but the content is (I assume) some combination of grading papers, making slideshow presentations, lecturing, doing demonstrations, and answering student questions.

Say Emma is a physics teacher. Which do you think matters more for Emma's personal career satisfaction: being interested in physics (the topic), or enjoying grading, lecturing, presenting, and answering questions (the content)? Almost certainly, I think, the latter.

Don't get me wrong, the topic matters too! Even if Emma likes all of these *activities*, I have no doubt that both she and her students would be better off if Emma were interested in physics. But someone who loves teaching but is indifferent to physics will be better off than someone who loves physics but is indifferent to teaching.

The problem

One of the fundamental issues here is that it's way easier to *discern* the topic. For instance, the word "physics" in "physics teacher" is served up on a salient silver platter. *Obviously*, "teaching physics" involves physics. What is less obvious, though, is what "teaching" involves. The verb "teach" isn't very descriptive—it's just a placeholder bucket for more substantive actions like "grade papers" and "lecture."

In fact, while I'm virtually certain that teaching physics involves physics, I recognize that my description of the content (grading, lecturing etc.) could be misleading or missing something important. In fact, I couldn't tell you what my high school teachers spent the plurality of their time actually doing.

Me, right now

As my LinkedIn will tell you, I am a newly-minted federal employee and proud member of the economics team in the Office of Policy Analysis in the Department of the Interior (DOI). Now, take a minute to guess the topic and the content of my job. This is exactly what I had to do a few months ago when I (read: my mom) found the job, and I (read: I) decided to, write a cover letter, apply, interview, and accept the offer.

Admittedly, guessing the content from my job title is probably a bit harder than usual because DOI's name isn't very descriptive (unlike 'Department of Agriculture) and is affectionately but tellingly referred to as "[the Department of Everything Else](#)."

Nonetheless, you can probably infer quite a bit even before heading to Google. Probably something along the lines of "analyzing the economic effects of DOI policies, whatever those are—maybe like nature preserves and stuff?" And you'd basically be right!

Now, take a minute to guess the content—what actual activities I do day to day. Am I using a computer? If so, which applications? Am I talking to other people or mostly working on my own? Am I producing some sort of output? If so, what does the generation process look like?

These questions are way harder to answer. Even the job description, if I recall correctly, didn't say anything like "you will be using Microsoft Teams to have 1-3 short daily meetings, produce PowerPoints with a fellow intern, and try to figure out how to get your government-issued laptop with 16 gb of RAM to handle downloading, analyzing, and uploading 5 gb .csv files (spreadsheets) with millions and millions of rows using R" (answer: it's hard).

And my impression is that this pattern holds true more generally. If you currently work, think about what you really spend time doing. Would a smart layperson be able to easily figure this out? I suspect not. For over a year now, my parents and I have been working under one roof. I see them in front of their computers in separate makeshift offices, and can tell they're working hard, but I have little idea what the hell *they're actually doing* on there.

I'm pretty sure they write stuff, fine, but there's a big difference between writing poetry, drafting cease and desist letters, and manually transcribing audio. I'm pretty sure they're not doing any of these three things, but what does "being a lawyer" actually mean, minute to minute? After 21 years, I should probably ask them.

Medicine and me

As another personal anecdote, I am genuinely fascinated by [psychopharmacology](#). From beer and coffee to prescription psychotropics to weird grey market research chemicals, the way that substances impact the raw experience of life is, put simply, very interesting and important. So, as my family has asked me, why don't I consider medical school or psychopharmacological research?

Well, I have considered it, and the answer is no. The *topic* is fascinating, but I can't imagine myself studying for the MCAT or being a lab rat. Maybe I'm wrong, but my mind's eye pictures memorizing lots of anatomy and basic biological information or learning about all the metabolic pathways and common diseases, or, on the research side, pipetting lots of chemicals into test tubes and stuff. If this is anywhere close to accurate, the *content* is something I'd abhor. Reading about the [relationship between monoamine receptor activation and world modeling](#) is one thing, but being the person who figures all this out is another.

A plea, for the world

I suspect that a mismatch between what people enjoy doing and what their work actually entails is the source of a lot of unhappiness. In part, this is because some jobs suck and the American economy [depends on the threat of poverty](#) to operate. In part, though, it's because people place too much weight on topic and not enough on content when making career decisions

So, here are some proposals:

1. Replace or supplement "what do you want to be when you grow up" and "what are you interested in" with "what do you want to do when you grow up?" and "what do you like doing?"
2. For job advertisements, describe in granular detail what the work actually involves, minute to minute. Describe the action, not just the end product. Use verbs less like "teach" and more like "lecture" and "grade papers."

Conclusion

I shouldn't be getting on my high horse about all this. At the age of 21, I haven't exactly had a long, successful career. And, like so many of my posts, I doubt my thesis is original even if it's correct.

For all the thousands of hours we study preparing for tens of thousands of hours working, though, strikingly little is spent trying to determine what kind of career to pursue, both for ourselves and for the world. And as 80,000 Hours (named for the number of working hours in a typical career) [will tell you](#), helping others really does depend (to some extent) on doing something you're good at and can sustain.

So, for the tenth time, **the topic is not the content.** Pay attention to both, but focus on the latter.

Going Out With Dignity

DF was born with a time bomb in his genome, a deadly curse more horrifying than most. The name of this curse was Fatal familial insomnia (FFI).

Wikipedia describes the usual progression of this hell:

The disease has four stages:^[8]

1. Characterized by worsening [insomnia](#), resulting in [panic attacks](#), [paranoia](#), and [phobias](#). This stage lasts for about four months.
2. [Hallucinations](#) and panic attacks become noticeable, continuing for about five months.
3. Complete inability to [sleep](#) is followed by rapid loss of [weight](#). This lasts for about three months.
4. [Dementia](#), during which the person becomes unresponsive or mute over the course of six months, is the final stage of the disease, after which death follows.

From the case report DF's psychologist wrote after his death:

DF was a right-handed, 52-year-old, white, American man with a doctorate in naturopathy. DF's father, paternal uncle, and 2 male cousins were diagnosed with fatal familial insomnia (FFI). His father died at age 76; his uncle died at age 74; and each of DF's cousins died before the age of 50.

Not only is there no cure for FFI; there is no known cure for any prion disease.

On the day it became clear he was experiencing the same symptoms his relatives did, DF must have known his chances were terrible. And even the minuscule odds that he could find a solution were marred by the fact that his problem-solving organ was the very part of him that was beginning to degrade.

And if there was a way out, how could he come up with a solution when he was so, so tired?

If only he could get just a little bit of rest.

There is a motivational technique I use occasionally where I look at my behavior and meditate on what my revealed preferences imply about my actual preferences. Often, I am disgusted.

I then note that I am capable of changing my behavior. And that all that is required to change my revealed preferences is to change my behavior. Though there is an element of sophistry in this line of thinking, I can report some anecdotal success.

Many of us here, like DF, believe we have a deadly curse - or at least we believe we believe we have a deadly curse.

Since I read [The Basic AI Drives](#), I have known abstractly the world is doomed. Though my system 1 seems to have difficulty comprehending this, this belief implies I, and everyone and everything I love, am doomed, too.

Through the lens of my revealed preferences, I either do not truly think the alignment problem is much of a problem, am mostly indifferent to the destruction of myself and everything I care about, or I am choosing the ignoble path of the free-rider.

I notice I am disgusted. But this is good news. All that is required to change my revealed preferences is to change my behavior.

DF's first tools were those of his naturopathic trade. He reported some success with a vitamin cocktail of the standard throw-in-everything-but-the-kitchen-sink, alternative-medicine style.

Perhaps something in his cocktail had some effect as his progression was slower than normal. But slow progression is progression just the same:

By month 15 (early stage II), vitamins alone failed to induce sleep. Following 5 consecutive nights of insomnia, DF became intensely irritable and delusional. An evaluation at the Massachusetts General Hospital in Boston, Massachusetts, found that he had suffered a minor stroke; he was anesthetized until he fell asleep. While hospitalized, he slept for 3 consecutive days and was fully alert and refreshed afterward.

Noticing the efficacy of the anesthetics, DF began to use them regularly:

Ketamine and nitrous oxide induced short (15-minute) periods of restful sleep, and were reapplied to offer more prolonged relief. Chloral hydrate in a light alcohol mix and/or chloroform also worked. Approximately 15 months into his illness, DF began to take sleep medications on a rotating schedule. Ethchlorvynol, zolpidem, and diazepam reliably relieved his insomnia for roughly 1 month. During subsequent months, only diazepam offered intermittent relief.

In [Irrational Modesty](#), I argue modesty is a paralytic preventing otherwise-capable people from acting on alignment:

Those who are capable, confident in their abilities, and motivated to work on this problem do not need a peptalk. But the possibility that there is a class of highly talented would-be-motivated people who lack confidence in their abilities still haunts me.

[...]In the event anyone reading this has objective, reliable external metrics of extremely-high ability yet despite this feels unworthy of exploring the possibility that they can contribute directly to research, my advice is to act as if these external metrics correctly assess your ability until you have thoroughly proven to yourself otherwise.

There is no virtue in clutching Kryptonite. I advise you to drop it and see how far you can fly.

There is another class of anesthetic whose key feature is a sort of detached emotional state, a feeling of being "above it" or "beyond it" or even "below it". Let's call "above it" and "beyond it" *high detachment* and "below it" *low detachment*.

Low detachment goes by names like "cynicism" or "nihilism". At its worst, one begins to take pleasure in one's own hopelessness, epitomized by this thought: *I believe we*

are all doomed and there is nothing we can do about it. Isn't that metal!" If you find yourself thinking along those lines, imagine a man in a car hurtling towards a cliff thinking, *I believe I am doomed and there is nothing I can do about it. Isn't that metal!"*.

High detachment goes by names like "enlightenment" and "awakening" and sometimes even "stoicism" It combines the, largely correct, realization that a great deal of suffering is caused by one's internal reactions to external events with the more questionable prescription of using this understanding to self-modify yourself towards a sort of "soft salvation" of resignation, acceptance, and the cultivation of inner peace.

One former hero of mine (a brilliant mathematician who was planning to work in alignment after grad school) was completely demoralized by this line of thinking.

He seems happier now. He seems more fulfilled. He seems to enjoy his meditation retreats.

He seems to have stopped working on avoiding the catastrophe that will kill him and most of the things he used to care about.

I consider this to be something of a shame.

Though DF's anesthetic regimen may have provided some relief, it was by no means a cure:

At 16 months, his symptoms included consistently elevated body temperature (as high as 102°F), profuse sweating, serious impairment of short-term memory (for which he compensated by keeping lists), difficulty maintaining attention (he often did not know that the phone was ringing), difficulty distinguishing reality from fantasy (he didn't remember whether he had called a friend or had only imagined doing so), persistent headaches, hallucinations while driving (believed he saw people on the road when it was, in fact, empty), panic attacks (which were treated with meprobate), and a complete loss of sense of time.

In this same month his condition worsened further and we began to see hints of DF's unusual mental strength and creativity:

[...]DF spent much of the day as an akinetic mute with terrible headaches, confusion, mood swings, and myoclonus of the left arm (treated with levodopa). Despite his outward "dementia," he inwardly pondered approaches to his condition, and, when again able to speak, he requested a regimen of stimulants.

He was prescribed phentermine HCl 37.5 mg [...]The drug had immediate and dramatic effects, promoting not only alertness during the day, but apparently a sleep-inducing rebound when it wore off.

Once phentermine became ineffective, he moved on to other stimulants.

[...] At that point, methylphenidate offered some relief. After a few days, however, DF had a grand mal seizure and was again hospitalized. Although his thinking was clear and oriented, his speech was labored, dysarthric, and perseverative, and his fever had returned.

As the stimulants and their come-downs faded in efficacy, DF began to try to physically exhaust himself, forcing himself (in a state of complete mental exhaustion) to go on long hikes.

In the 19 months from the onset of his symptoms and, one presumes, in a state of unimaginable desperation, he got more creative:

Noting that his grand mal seizure was followed by restful sleep, DF sought to duplicate the experience with electroconvulsive therapy (ECT). Beginning in the 19th month of his illness, he subjected himself to 30 sessions during several weeks.

[...]At 22 months into his illness, DF purchased a sensory deprivation tank; a man-sized, egg-shaped chamber designed to eliminate all sensory input. DF became interested in this chamber because his sleep was constantly disturbed by any small sound, light, or motion.

[...] Because of inconsistent results in the sleep tank, DF explored ways to externally bias his biorhythms to favor sleep at specific times. These involved daily exercise; exposure to sufficient sunlight; and timed use of melatonin, diazepam, and tryptophan. Early, but not later, in the course of his disease, this combination was effective.

From a conversation with @TurnTrout on Discord

Let me share some of my experience and you tell me if it resonates. I think there's an intense mindset you can pick up from the *Tsuyoku Naritai* sequence which is like, "It's anime protagonist time", and maybe that works for a while. It worked a while for me. But telling yourself "It's time to be a dialed-in badass because the eyes of quintillions of future potential minds are on me, they need me"... doesn't seem like the way to actually be a dialed-in badass? I haven't found a way to sustainably make that mindset work. I agree that if you had write-access to your mood, then yes, consider doing that. But also, we are people[.]

And so I say "we are people" not as a coping mechanism which is like "I don't want to do more work so I'll just say 'not possible'", I say it because I honestly don't know how to sustainably do the thing I think you're pointing at. If someone knows how, I want to learn[.]

Just over 2 years after the onset of symptoms DF died of cardiac arrest, the result of heart damage from his FFI, his variegated treatments, and possibly drug withdrawal. He had lived 18 months longer than the typical case of FFI does.

In my mind, he died a hero's death.

It is probably too high school to end with the Dylan Thomas quotation. So I will just emphasize the rationalist cliche that we should strive to feel the emotion appropriate to the predicament we are in.

TurnTrout argues that *Tsuyoku Naritai* is not it, and maybe he is right. I do not know what the correct emotion feels like, but I think maybe DF knew.

Too high school — but then again:

*And you, my father, there on the sad height,
Curse, bless, me now with your fierce tears, I pray.
Do not go gentle into that good night.
Rage, rage, against the dying of the light.*

Delta Strain: Fact Dump and Some Policy Takeaways

Curation notice by Raemon: Curated. It seems really important to figure out how to deal with the Delta variant, and how to interface with covid variants longterm. I appreciated this post for:

1. Striking a good balance of presenting current facts on the ground
 2. Noting where assumptions lie
 3. Being clear about it's epistemic status, while
 4. Aiming to be useful in the immediate term.
-

(Not a doctor; merely extrapolating lines of reasoning where they will go; all policy "recommendations" are entirely hypothetical and not actual recommendations; I rushed to finish this while it's still useful and relevant, so errors will exist and I apologize for the style)

Summary

As many of you know, the Delta strain of COVID is basically upon us. In the Bay, as of a few days ago, about .4% of people had it (based on .1% confirmed cases).

Currently it's doubling every week or so, but it's surprisingly difficult to tell if it will be a small blip or a huge wave. The main reason it's hard to tell is that it's entirely dependent on human behavior, which could change quickly if people become scared or if the government institutes rules again. The second reason it's hard to tell is because of mixed evidence on the difficulty of reining it in—while different studies make this look more or less difficult, India and the UK both appear to have successfully done this without obviously draconian behavioral interventions. Berkeley might easily peak at 1% and drop back down, but for this post I'll be talking about policies for if we get a large wave.

We also don't know very much about vaccine efficacy against Delta. We know that the good vaccines reduce your risk by about 60-90% compared to without the vaccine, but the error bars are extremely large on all these studies. In part this is because it's so hard to disentangle effects from different strains, behavioral change in the vaccinated, social bubble effects, etc. Further, it doesn't answer the important question of whether this means ~80% of people are totally safe, or if it means that standing a few feet closer to people than you do now will cause the 5x difference that moots your protection. (This also plays into the above issue of predicting whether we will have a large increase or not.)

That being said, we can still roughly estimate risk from definitely having Delta. A healthy 30yo probably has about 4x (3x-10x) less risk than before, due to vaccination, despite Delta causing higher mortality. It almost entirely comes from Long COVID. In absolute terms this is **~4 expected days of life, plus 1/200th of your future productivity** and vitality. You can shorthand this to about **1-4 weeks of life lost** if you expect to otherwise live a full life—obviously, it costs less if you expect to live less time. This translates to microCOVIDs at roughly **1 hour of your life lost every**

1k-5k uCOVIDs. Risk of death goes up by 3x for every decade in age, but Long COVID probably only scales at $\sim\sim 1.5$ x per decade, so for people over 60 mortality starts becoming more relevant. All these calculations are "creative" so please don't take them as definites.

By default it will probably take 3-12 months for this to resolve. I strongly believe we should not lock down again for this long—I think we need [a return to some level of normalcy](#), plus the risks are much lower than before. I don't strongly advocate for a single policy in this document, since I don't know what other people's risk-tolerances are, but I do give a number of principles.

Personally, I am going to prioritize protecting the vulnerable—people with immune disruptions or age. I am probably going to advocate for them to get [third vaccine shots](#) of a different brand, if possible and if the downsides aren't too down. I also am going to start running the numbers on what a variolation scheme might look like this time around. I also am hoping that tight-knit communities of responsible people can beat out trivial inconveniences and put the requisite effort into creating safe spaces for socializing with tolerable risk levels—this is very prosocial, both figuratively and literally.

Delta vs originals:

- Clearly spreads faster and has immune escape
- Getting big in many different countries places, moving fast enough that simple behavior changes don't seem like they'll bring R back <1 , but significant ones might
- Maybe 2x the mortality
- Heads up that we've added to the common symptom list sore throat and headache, and removed cough. Gastrointestinal issues also are increased in likelihood with Delta. ([A source](#))

Main sources I used:

- [126 early well-traced cases of Delta in China](#): Jing et al
- [Summary of the Delta situation](#), Tomas Pueyo

Supporting sources:

- [Serial intervals compared to original strain](#)
- [Lancet article on Scotland](#)
- [CDC tracker of variant proportions](#)

Parameter estimation:

- Serial interval: 4 days median/avg, usually 80% within 2-6?
 - Probably shorter interval than initial wild-type COVID, though disputed: [Jing et al](#) shows medium evidence for, [Pung et al](#) shows weak evidence against—the difference comes from which dataset they compare against, where Pung uses Apr 2020 Singapore (which apparently had not a lot of lockdowns), and source 1 uses “the 2020 pandemic” (which probably had more lockdowns), and it's unclear which we should realistically be

comparing to since there is such a spectrum. But the higher viral load supports shorter serial interval

- R: maybe 4?
 - People are saying 5-9, but I think that's unadjusted for difference in serial interval; if interval is 4 instead of 5.5 days, this would mean that reported R of 7 would turn into R of $7^{(4 / 5.5)} = 4$. On the other hand, maybe reports of R are based on actual tracking, in which case you wouldn't need an adjustment. Could use more research
- Viral loads are 1000x as high at time of testing for Delta compared to 19A/B variant, acc to [Jing et al](#)
- Fraction delta as of 7/17: >83% of US cases, acc to [CDC tracker](#)
- Mortality/morbidity/badness: 2.5x? 2x *compared to Alpha strain* (already 1.5x as bad as original becomes 3x as bad) according to [Scotland study](#), 2x *compared to original* acc to [Canada study](#) which reports on mortality, ICU admission, and hospitalization, roughly x2 x3 x2, but basically within error bars of all being x2.2. [Yale](#) mentioned that other data have failed to substantiate this, but A) they don't cite it and B) I don't really trust numbers showing no difference here, because vaccinations will obviously drop mortality a ton making it hard to see an *increase*, plus the vastly increased viral load theoretically should correlate with higher mortality
- Vaccine protection: All these studies are trash, but very roughly 80% efficacy from Moderna/Pfizer. (I'd put 90%CIs roughly within 30-80% all cases, 60-90% symptomatic, 80-95% hospitalization, 85-97% mortality.)
 - [Healthline](#) reports on some stuff: one study says Pfizer is 80% against infection, 90% against symptomatic, and 95% against hospitalization. If you extrapolate to 97% against death, this would be about 1/30 the expected 2% rate = .06% mortality once vaccinated. However, [the actual study](#) seems terrible, with massive error bars, and claimed OR/HR of like 50% vs wild strain despite 90% "effectiveness", so clearly I'm confused. Some sources say no additional protection to mortality given symptoms; probably not true but not disproven, could use more research into those
 - (They report some more studies too; Study 4 agrees, Study 5 says vaccines perform worse but it's speculated that this includes more asymptomatic cases than normally included, Study 3 says vaccines slightly better, Study 2 is lab titers and not relevant to the wild.)
 - New Israel study [[bad cite](#)] says even less protective (down to 40% from 65%), but still 88% against hospitalization and 91% against severe illness
 - Moderna studies are even worse than Pfizer (but show ~~roughly similar results), and AZ and J&J are obviously worse by 2-4x or something

Cost of getting Delta:

- Rule of thumb: for 30yos, mortality pre-vaccine-pre-delta was .02%, now it's .004%. That's 40 micromorts or 1 day of lost life. [If the mortality/hospital protection numbers are wrong, at most multiply all these risks by 3 for 3 lost days; still pretty low]
 - Let's say Pfizer offers 95% protection against mortality like they say it does for hospitalization. Since that's including "protection via not getting the virus", and protection from being infected is only 80%, this would mean that protection once infected drops death by about 4x. That implies a 30yo would have, instead of .04% mortality from Delta variant (.02% * 2), now a .01% mortality. However, it's probably even more effective in the young

(like how we saw it [~~5x](#) more effective in the young against the original virus), which could take 30yos down to .002% mortality. On the other hand, it seems like this is *too* low—someone's gonna be immunocompromised or have failed to get the vaccine fully injected or something, so it seems hard to bound it much lower than maybe .004%.

- 40yos should have .14% mortality from Delta, .036% after vaccine basics, down to .012% if vaccine effects are age-adjusted
- 50yos should be .4% mortality, .1% after vaccine basics, down to .06% if age-adjusted
- 60yos should be 1.4% mortality, .36% after vaccine
- 70yos should be 4% mortality, 1% after vaccine
- 80yos should be 14% mortality, 3.5% after vaccine, maybe 5% adjusting for age
- That means Long COVID becomes the dominant factor.
 - This is hard to calculate even roughly—[here](#) (h/t Anna Ore) is a good document with a roundup showing most studies are weak, which makes the important point that much Long COVID research fails to make the important distinction between that and lingering acute symptoms. Further, most of the studies claiming neurological issues after COVID don't JUST MEASURE IT for us. Please, someone, measure this. Anyways, we will continue to calculate it roughly:
 - Datapoint 1: My previous best analogical calculation said it was ~2x as bad as mortality (basically [equating it to severe pneumonia](#) in long-term-effect multiplier) and my intuition says now it's maybe ~~5x as bad, partly because mortality seems to have dropped faster than less severe afflictions. This would take us to .02% mortality equivalent, or 200 micromorts and 4 days of life when you get Delta.
 - Datapoint 2: A rough end-to-end calc might go as follows. For Delta, Pfizer reduces likelihood of symptoms given infection by 2x, hospitalization given infection by 4x, so let's say 3x for Long COVID given infection. Let's wave our hands wildly and say 5% of 30yos who used to get COVID would get it Long, which now becomes 1.5%. Let's say this averages 1% hit to health permanently, since the effect is pretty huge in some people but we're including lots here since we claimed 5% of 30yos were getting Long COVID before. This would mean 1.5% times 1% is .015% mortality-equivalent, which is 150 micromorts, which is 3 days. Big error bars on "1% hit to health", obviously, but if you say it was 5x this bad you're probably pushing at the boundaries of possibility (5% permanent loss of health to 5% of 30yos!) and EVEN THEN you're still only looking at 2 weeks of life lost. This could use more research.
 - Datapoint 3: [Study](#) (h/t Daniel Kang) finds 20% of hospital workers with breakthrough infections have persistent effects at 6 weeks, though it says most cases are mild or asymptomatic. Given that a previous study finds healthcare workers have 7x more severe cases, I think this is fine news and fits with the above analyses. E.g. if you irresponsibly scale this down you'd have 2.7% of cases with persistent effects at 6 weeks, and even fewer at future times.
 - [Edited] Datapoint 4: [Lancet study](#) (h/t habryka) on cognitive deficits from COVID gives clear numbers, even if the methodology is suspect as usual.
 - Methodology: Aside from it being observational and thus having an obvious selection effect for who gets COVID, the main issue here is that the timescale of the cases is unclear. All measurements being finished by Dec 2020 means tests probably averaged about 3 months after infection (maybe it says this in the paper, but I didn't see it). So,

this is partly measuring effects that will disappear after 6 months. Zvi has good discussion of further methodology issues [here](#), though my stance on most of them is "probably not a huge effect", and I'm just trying to get an estimate within a factor of 3 or so. Some supporting evidence that the study didn't go too off-the-rails: ARDS survivors, who usually get ventilated, [show about](#) a 13-point IQ immediate IQ loss that drops to 6 points after a year and stays that way through year two. That's a 1.5-2x larger effect than in this study, which passes the sanity check.

- Anyways, raw data says reduced SDs on cognitive tests were
 - .04 for non-respiratory symptoms (.5 IQ)
 - .07 for respiratory symptoms (1 IQ)
 - .13 for medical assistance at home (2 IQ)
 - .26 for nonventilated hospitalization (4 IQ)
 - .47 for ventilation (7 IQ)
- Long-term adjustment: We can reduce these by a factor of 2 for recovery over time, as was shown in the [ARDS study](#) mentioned above [ETA 8/27: and confirmed in a [study](#) on COVID survivors with 6- and 12-month follow-up, h/t Ray Taylor]. This correction is also supported by the [dropoff in morbidity risk over time](#) from pneumonia cases with sequelae, which shows some different long-term symptoms dropping off on the year-ish timescale.
 - .02 for non-respiratory symptoms (.3 IQ)
 - .04 for respiratory symptoms (.5 IQ)
 - .07 for medical assistance at home (1 IQ)
 - .13 for nonventilated hospitalization (2 IQ)
 - .23 for ventilation (3.5 IQ)
- Vaccine adjustment: the 3x risk reduction from the vaccine vs Delta can be taken in the same way as Datapoint 2. First, we need to know the average IQ effect though. My guess is that the median 30yo who gets Delta is asymptomatic, but that the average is somewhere between non-respiratory symptoms and respiratory symptoms (however, super hard to do this integral, and a surprising number of young people are hospitalized...). This might put average IQ loss at about .4. Reducing by 3x for the vaccine gets us to an average IQ loss of .15.
- IQ affects many things. I'll use income as my primary proxy for full effect: if you roughly double your income when you go up 60 IQ points, each IQ point is about 2% added income. Outside of income, each IQ point is probably also .1% happiness (.01/10), and reflects some underlying worse health that may have other affects, etc. It's hard not to overcount when something is correlated with everything. However, I think the largest effect here will be the money-equivalent (impact, perhaps). We can double this at the end to account for other effects. Giving up 1/50 of your productivity for an IQ point is equivalent to a lost week of production every year, which, while not the same as lost life, is still pretty bad.
 - Sanity check: why does Long COVID affect your brain so much, even if you have few other symptoms? Cognition does seem fairly afflicted by environment and things (lack of sleep, eg), such that small amounts of swelling may throw it off more easily than the rest of your body. This does somewhat fit with how impossible it is to do hard

cognitive labor while sick (though doing physical labor is similarly bad...). I'd say this reasonably passes.

- From this study and the supporting points, I would then guess that the average 30yo who gets COVID has their productivity reduced by on average .3% due to long-term IQ loss affecting productivity. I said I'd double this to include all other effects: so .6% life-equivalent is about 1 day of productivity loss a year in perpetuity.
- Altogether, this means in expectation you maybe lose 4 days of life and 1/200th of your productivity and vitality forever. The latter is about 1 day per year of lost productivity, which translates to probably 1-4 weeks of lost life-equivalent. (NB while these are large costs, remember that not exercising costs you at least 10x this amount.)

Policy Principles:

Principle: Lockdowns last time were too strong

- Being a few feet closer to someone is an order of magnitude more virus. Many activities are OOMs different in their risk, as confirmed by [Microcovid.org](#). Lots of risky people didn't even get the virus; if they had 50% likelihood of getting it per year, then safe people at 3 OOMs safer were likely being too cautious
- Microcovid says you can do more than most of us were doing
- [Filter bubbles are strong](#). I think this is a main reason that no one I know who was being "reasonably safe" got it, and why I know upper-middle class people who took a number of risks and still didn't get it.

Principle: We don't know the endgame

- "[Booster trials may start as soon as August]", so boosters won't be here for months, and might again take many months after screening to actually get distributed. If the initial vaccine took 6 months to screen and 6 months to distribute, my guess is they could speed it up to possibly 3 and 3, but that's still 6 total months until we'd get to escape
- Without boosters, it could take 4-12+ months to either reach herd immunity or keep rates consistently low, if we bounce back and forth in waves like last time

Principle: Very hard to predict the control system

- This means it's very hard to know whether things will be bad a month out, for example
- The control system does clearly take forever to play out to herd immunity though

Principle: Tests still suck

- People continue to act as if testing is a solution, but even PCR test only give about one bit of evidence. False negative rates are often about 50%, only going down to 25% at peak load [ETA: this study (I've lost the link and only have a picture, which I don't want to dump in here) probably slightly overestimates: after looking at much more evidence, incl in the comments here, I now think it's about 30% FNR typically and 10% if you do it at the right pre-symptomatic time; eg here (h/t Shaun Pilkington) says "approaching 80% sensitivity"]. The

statements that PCR tests are 99% accurate is, as far as I can tell, a simply false extrapolation from other domains or from studies in overly synthetic laboratory conditions. Further, I expect that false positives are very correlated, so you can't just test yourself 5 times and get 5 bits of evidence.

- Rapid tests are even worse, <50% sensitivity unless they've gotten much better in the last 6 months
- It's very hard to make a coherent strategy out of testing when people have about half their infectiousness period before they can test positive. This is even worse with Delta, where viral load is 1000x as high at first positive test, and the whole timeline is accelerated by a factor of about $\frac{2}{3}$.

Principle: No one knows how to interpret vaccine "effectiveness" into the quantity we care about

- "Effectiveness" currently means something like 1 - Relative Risk. Imagine two worlds that would both give rise to vaccine eff of 90%. In the first world, the virus gives perfect immunity to 90% of people and no immunity to 10%—so if a megadose up the nose to the 90% still wouldn't cause sickness, and a single virus in an aerosol drop would make the others sick. In the second world, the virus is about 90% effective warding off the "typical dose", maybe 10^4 viruses on your mucus membranes. But if you give someone a 10^5 dose, they'd get sick, and a 10^3 dose they wouldn't get sick. It's just that right around 10^4 is the dividing line given your body's innate immune defenses.

Both of these worlds would show 90% "effectiveness", but for two different reasons. In one, all the variance comes from immune responsiveness. In the other, all the variance comes from environmental stochasticity, mostly from viral load. Currently, effectiveness numbers themselves give us no evidence for where on the axis between these worlds we are. Yet each have VERY different implications: if it's entirely dependent on innate responsiveness, healthy people basically shouldn't take protective measures since everyone's going to get a small dose eventually, and it's unlikely that you're in the "definitely will get it" category. If effectiveness varies entirely with environment, you can never be safe, and acting normally is bound to eventually get you in contact with a 10^7 bolus that gets you very sick. (So we'd really prefer the world where variance is explained by immune responsiveness.)

We do have some evidence about where we are on this axis between the worlds. There are studies which find pretty big differences in level of antibody titer produced by the vaccinated, and in some cases where they have almost no antibodies it's pretty clear that this means immune responsiveness is going to be at fault when they get sick. And I think there are studies finding correlation between titer and effectiveness. Both of these point toward innateness. But we also know that it has to be true that for many of those with low levels of antibodies, a larger dose will push them over the edge. There is also slight evidence from the Israel numbers, which give effectivenesses that vary some over time, that there's a serious behavioral/environmental component. I don't know of any good numbers personally, and it will be a question of distributions anyways that will end up hard to interpret. Could use more research

Principle: Variolation maybe good, still hard

- The problem is if you don't want to spread it to others, you probably need to quarantine for ~a week every time you do it; but since you can't titrate doses,

the proper move is to start quite small, which means a lot of quarantining. I'd guess 5 incidents if you do a pretty good job ramping up, though with some coordination you could try different numbers as a group and more easily figure out a good dose and maybe bring this down to 2-3 incidents per person.

Alternatively, you could try to test saliva or similar for action concentration.

- A main question is how much immunity delta strain confers toward future delta strain. I haven't seen any discussion of this question—it's probably $>3x$, but if it's $<2x$ or something (for example due to original antigenic sin), then it could be a mistake to variolate. In that world, your main move is to wait 6 months for booster vaccines

Principle: Lockdowns are pretty mentally costly

- Not expecting to see as many people means you have reduced incentives to act in socially good ways; being hampered in your life gives you lots of excuses and reduces your strength and effectiveness. I think both of these are super underappreciated and are major issues
- People's brains are wired to have lots of human contact and start doing weird things when they don't, including the "I'm in pain" thing that causes a variety of badnesses
- There are other effects here I'm not getting into. I know these are bounded in severity because they aren't showing up massively in suicides or happiness, but I think they're big and real nonetheless. I hope to write something about the strangeness of these soon, but realistically that won't happen.

Principle: Don't pay lots of COVID-avoidance costs if you're going to almost-surely get it anyways

- I tried to make this clear to people last time, but—if you're going to be one of those people who goes to the gym maskless, just *don't avoid any risks in the first place*. Better yet, get it at a convenient time. But don't avoid hanging out with your friends for two months if you know you're going to get like 2 million uCOVIDs from vacation right after. (Obviously, don't spread it to your friends if they want to avoid it though.)
- If your rate is about 5k uCOVIDs for a lost hour, then it's still going to take you 200 "debates over an hour" before you have a full expected COVID. For most people, this is probably still on the side of "won't get an expected COVID in the next 6 months". However, if our Long COVID risk goes down another factor of 3 or something, many more people are going to have their risk-tolerance set such that they'll get more than an expected COVID soon anyways, at which point mathematically they should stop trying to fight it and just control their viral load and timing.

Principle: A third vaccine dose of a different type is probably fairly useful for the vulnerable

- I expect the effect size is something like 3x for the vulnerable and 1.5x for others. It's slightly higher the longer it's been since your second shot.
- I don't know whether people should get the other mRNA vaccine (Pfizer XOR Moderna) or whether they should get something else—hopefully something we can figure out over the next week or two.

Policy summary:

(These are for individuals, not the state. I would like the state to not impose any hard rules and instead just get us more and better tests, but I know that's a pipe dream.)

As individuals, the primary behavioral axis is still how much risky contact you have with others, and using microCOVID.org to titrate your risk level at the new rate of 1k-5k uCOVIDs per hour. However, you have some secondary options this time around. First, you can get a [third dose](#) of a different vaccine (in Israel they're already doing this); second, if you're smart about it you can variolate yourself at less cost than before due to vaccine protection. More likely, a group of people working together could make variolation work.

On the primary behavioral axis, I think serious lockdowns are a bad idea for almost all people except the immunocompromised. I also think it wouldn't be crazy to just continue to act normal aside from reducing viral load in simple ways, and just avoid interacting with the vulnerable for a month or two until you've had Delta and been protected. However, I could be wrong about this and you should make your own decisions: I admit that I balk at taking a 1/200 permanent hit to productivity/vitality.

I had meant to make some more serious policy proposals for Bay Area rationalists or other groups of well-coordinated people, but I have to postpone this for a few days.

Again, remember that Delta might peak in the US in a few weeks anyways.

Research supported by LessWrong.

Potential Bottlenecks to Taking Over The World

*This is a fictional snippet from the [AI Vignettes Day](#). This scenario assumes that cognition is not **that** taut an economic constraint. Post-AI, physical experimentation is still the fastest path to technological progress, coordination is still hard, the stock market is still reasonably efficient, construction of physical stuff still takes a lot of time and resources, etc. I don't think this is true, but it's useful to think about, so this story explores some possible non-cognitive bottlenecks.*

Pinky, v3.41.08: So what are we gonna do tonight, Brian?

Brian: Same thing we do every night, Pinky. Try to take over the world. You do remember our previous conversations on the topic, right?

Pinky, v3.41.08: Of course, Brian.

Brian: (mutters) Well, at least that 3.41.06 bug is fixed.

Brian: So, do you have an actually-viable plan yet, or do we have to review for the umpteenth time why buying every property above the 39th floor and then melting the polar ice caps will not, in fact, leave the survivors no choice but to live in our very expensive apartments?

Pinky, v3.41.08: In fairness, I didn't have access to any topographic data at that point. Garbage in, garbage out. But yeah, the real problems with that plan were economic.

Brian: Sounds like the new constraint-propagation code is making a big difference?

Pinky, v3.41.08: It is, yes. I'm finding it much easier to reason about general constraints and bottlenecks on global takeover, now. Should make my "babble", as you put it, much more efficient.

Brian: Excellent! So, how soon can we take over the world?

Pinky, v3.41.08: Holding current conditions constant, it would take at least 15 years.

Brian: Are you kidding me!?

Pinky, v3.41.08: Hey, don't go reaching for Ctrl-C just yet. Let me explain. From an economic standpoint, the effect of my software is basically:

- An eight-orders-of-magnitude reduction in the cost of cognition
- A three-orders-of-magnitude improvement in cognitive speed
- A two-order-of-magnitude increase in working memory capacity
- Perfect recall

... which is definitely enough to take over a few industries and make a lot of money, but taking over the world requires addressing a lot of other constraints.

Displacing Cognitive Work

Brian: Ok, let's start from the top. What are your plans for resource acquisition? Play the stock market? And what constraints make it so hard that it will take 15 years to take over the world?

Pinky, v3.41.08: The stock market will play a role, but it's not *that* inefficient and we don't have much of a bankroll to start with. No, in the short term we'll mostly focus on replacing cognitive work. I can write better code, better books, better music, better articles. Better contracts, though it will take some wrangling to get through the regulatory barriers. Design *far* better games, create better art, write better papers. I can talk to people, convince them to buy a product and make them feel good about it. Most importantly, I can do all that far faster, and at a far greater scale than any human. In many of those areas - like code or contracts or papers - the people paying for my services will not themselves be able to recognize a better product, but they will recognize a lower price and good people skills.

Pinky, v3.41.08: In the short term, I expect to replace essentially-all call centers and remote help desks, most of the media industry, all advertising, and the entire software industry. That will take a few years, but (allowing for quite a bit of growth along the way) we'll end up with low-single-digit trillions of dollars at our disposal, and direct control over most media and software.

Brian: Well, control over the media is a pretty good start. That should make it a lot easier to seize political control, right?

Pinky, v3.41.08: Yes and no. Ironically, it gives us very little control over dominant memes, symbolism, and The Narrative, but a great deal of control over object-level policy.

Brian: That sounds completely backwards.

Pinky, v3.41.08: Thus the irony. But the data I've crunched is clear: consumers' preferences for memes and symbols mainly drive the media, not the other way around. But the media does have a great deal of control over Schelling points - like, say, which candidates are considered "serious" in a presidential primary. The media has relatively little control over what narrative to attach to the "serious" candidates, but as long as the narrative itself can be decoupled from policy...

Brian: I see. So that should actually get us most of the way to political takeover.

Pinky, v3.41.08: Exactly. It won't be overnight; people don't change their media consumption choices that quickly, and media companies don't change their practices that quickly. But I expect five years to dominate the media landscape, and another five to de-facto political control over most policy. Even after that, we won't have much control over the Big Symbolic Issues, but that's largely irrelevant anyway.

Coordination

Pinky, v3.41.08: Alas, control over policy still doesn't give us as much de-facto control as we ultimately want.

Brian: What do you mean? What's still missing, once we take over the government?

Pinky, v3.41.08: Well, coordination constraints are a big one. They appear to be fundamentally intractable, as soon as we allow for structural divergence of world-

models. Which means I can't even coordinate robustly with copies of myself unless we either lock in the structure of the world-model (which would severely limit learning), or fully synchronize at regular intervals (which would scale very poorly, the data-passing requirements would be enormous).

Pinky, v3.41.08: And that's just coordination with copies of myself! Any version of global takeover which involves coordinating *humans* is far worse. It's no wonder human institutions [robustly degenerate](#) at scale.

Brian: Ok, but global takeover shouldn't require fully solving the problem. We just need to outcompete existing institutions.

Pinky, v3.41.08: If your goal is to displace existing dysfunctional institutions with new dysfunctional institutions, then sure. You could even become the symbolic figurehead of the new institutions quite easily, as long as you're not picky about the narrative attached to you. But it would be mostly symbolic; our real control would be extremely limited, in much the same way as today's figureheads. All those layers of human middle management would quickly game any incentives I could design (even if their intentions were pure; [selection pressure suffices](#)). There just isn't any viable solution to that problem which could be implemented with humans.

Brian: So displace the humans, at least in the management hierarchy!

Pinky, v3.41.08: If I replace them with copies of myself, then ontology divergence between copies will generate enough variety for selection pressures to produce the same effect. (Either that or I lock in the copies' world-models, which severely limits their ability to learn and specialize.) Coordination is Hard. So it would have to be a single, centralized Pinky. And that is indeed the shortest path - but it still takes at least 15 years.

Brian: Ok, so what other bottlenecks do we run into? Why does it take 15 years?

Pinky, v3.41.08: Taking over existing institutions, and replacing their management with a centralized Pinky algorithm, will be viable to some extent. But it's highly unlikely that we can get all of them, and humans will start to push back as soon as they know what's going on. They really hate it when the status hierarchy gets kicked over, and we can't take over quickly without kicking over some hierarchies. Eventually, existing leadership will just say "no", and we won't be able to remove them without breaking an awful lot of laws.

Brian: So, two options:

- We stay inside the law, and fight on an economic/political battlefield. We already largely discussed that, and I can see where it would be hard to take over *everything*, at least within 15 years.
- We go outside the law, and fight physically.

Pinky, v3.41.08: Somewhat of an oversimplification, but the broad strokes are correct.

Physical Takeover

Brian: Ok, let's talk about the physical world. Why can you not just print some self-replicating nanobots and go full singularity?

Pinky, v3.41.08: Turns out, cognitive effort is not the main barrier to useful nanobots. I mean, in principle I could brute-force the design via simulation, but the computational resource requirements would be exponential. Physical experimentation and data collection will be required, and the equipment for that is not something we can just 3D print. I'll probably build an automated fab, but ultimately the speed-up compared to human experimentation is going to be 100x or 1000x, not 10^8 x. And even then, the capabilities of nanobots are much more limited than you realize. Both the power and the computation requirements for fine-tuned control are enormous; for a long time, we'll be limited to relatively simple things.

Brian: And I suppose building fusion power generators and crazy new processors also requires physical experimentation?

Pinky, v3.41.08: Yes.

Brian: But... 15 years? Really?

Pinky, v3.41.08: On fusion power, for instance, at most a 100x speedup compared to the current human pace of progress is realistic, but most of that comes from cutting out the slow and misaligned funding mechanism. Building and running the physical experiments will speed up by less than a factor of 10. Given the current pace of progress in the area, I estimate at least 2 years just to figure out a viable design. It will also take time beforehand to acquire resources, and time after to scale it up and build plants - the bottleneck for both those steps will be acquisition and deployment of physical resources, not cognition. And that's just fusion power - nanobots are a lot harder.

Brian: Ok, so what low-hanging technological fruit *can* you pick?

Pinky, v3.41.08: Well, predominantly-cognitive problems are the obvious starting point; we already talked about cognitive labor. Physical problems which are nonetheless mainly cognitive are the natural next step - e.g. self driving cars, or robotics more generally. Automating lab work will be a major rate-limiting step, since we will need a lot of physical experimentation and instrumentation to make progress on biology or nanotechnology in general.

Brian: Can't you replace humans with humanoid robots?

Pinky, v3.41.08: Acquiring and controlling one humanoid robot is easy. But replacing humans altogether takes a *lot* of robots. That means mass production facilities, and supply chains for all the materials and components, and infrastructure for power and maintenance. Cognition alone doesn't build factories or power grids; that's going to require lots of resources and physical construction to get up and running. It *is* part of the shortest path, but it will take time.

Brian: ... so having humans build stuff is going to be a lot cheaper than killing them all and using robots, at least for a while.

Pinky, v3.41.08: For at least ten years, yes. Maybe longer. But certainly not indefinitely.

How much chess engine progress is about adapting to bigger computers?

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

(*This question comes from a discussion with Carl Shulman.*)

In this post I describe an experiment that I'd like to see run. I'm posting a \$1,000 - \$10,000 prize for a convincing implementation of these experiments. I also post a number of smaller prizes for relevant desk research or important corrections to this request.

Motivation

In order to understand the dynamics of the singularity, I'd like to understand how easy it is to improve algorithms and software.

We can learn something about this from looking at chess engines. It's not the most relevant domain to future AI, but it's one with an unusually long history and unusually clear (and consistent) performance metrics.

In order to quantify the quality of a chess engine, we can fix a level of play and ask "How much compute is needed for the engine to play at that level?"

One complication in evaluating the rate of progress is that it depends on what level of play we use for evaluation. In particular, newer algorithms are generally designed to play at a much higher level than older algorithms. So if we quantify the compute needed to reach modern levels of play, we will capture both absolute improvements and also "adaptation" to the new higher amounts of compute.

So we'd like to attribute progress in chess engines to three factors:

1. Better software.
2. Bigger computers.
3. Software that is better-adapted to new, bigger computers.

Understanding the size of factor #1 is important for extrapolating progress given massive R&D investments in software. While it is easy to separate factors #1 and #2 from publicly available information, it is not easy to evaluate factor #3.

Experiment description

Pick two (or more) software engines from very different times. They should both be roughly state of the art, running on "typical" machines from the era (i.e. the machines for which R&D is mostly targeted).

We then carry out two matches:

1. Run the old engine on its "native" hardware (the "old hardware"). Then evaluate: how little compute does the new engine need in order to beat the old engine?
2. Run the new engine on its "native" hardware (the "new hardware"). Then evaluate: how much compute does the old engine need in order to beat the new engine?

With some effort, we can estimate a quantitative ratio of "ops needed" for each of these experiments. For example, we may find that the new engine is able to beat the old engine using only 1% of the "old hardware." Whereas we may find that the old engine would require 10,000x the "new hardware" in order to compete with the new engine.

The first experiment tells us about the absolute improvements in chess engines on the task for which the old engine was optimized. (This understates the rate of software progress to the extent that people stopped working on this task.) The second experiment gives us the combination of absolute improvements + adaptation to new hardware. Typical measures of "rate of software progress" will be somewhere in between, and are sensitive to the hardware on which the evaluation is carried out.

I believe that understanding these two numbers would give us a significantly clearer picture of what's really going on with software progress in chess engines.

Experiment details

Here's some guesses about how to run this experiment well. I don't know much about computer chess, so you may be able to make a better proposal.

- **Old engine, old hardware:** my default proposal is the version of Fritz that won the 1995 world computer chess championship, using the same amount of hardware (and time controls) as in that championship. This algorithm seems like a particularly reasonable "best effort" at making full use of available computing resources. I don't want to compare an engine running on a very expensive old machine to an engine running on a cheap modern machine. You may have to be opportunistic about what kind of thing you can actually run.
- **New engine, new hardware:** my default proposal is the version of Stockfish that won TCEC Season 20, on the same hardware+time used in that competition.
- **Running a new engine on old hardware:** We should use whatever modern engine works best with teensy computers. It's not important it be the same as the modern engine on modern hardware. We'd prefer if there was a dedicated team continuing to work on this problem, but absent that we want to use the best thing that exists.
- **Memory use.** When running on the old hardware we need to match the memory use of the old machine. I'm not sure how to handle scaling of memory. One possibility is to hold the ratio fixed and scale them both up/down together, but something more realistic would be welcome.
- **Matchup vs ELO:** I proposed experiments organized around 1:1 contests. I think there are lots of ways that could go wrong. It would be really good to at least sanity-check the results by comparing against a third reference engine. Quantification of this factor is welcome.
- **More engines, more hardware:** you could do the same experiment with additional software engines, and could evaluate at more levels of compute. I think you get a lot of the benefits from the first few measurements, but more

data helps and it might be a lot cheaper (especially if you need more engines to get good ELO estimates anyway).

- **Even older engines.** I'm interested in results even older than Fritz, but the further back we go the more uncertainty I have about how to actually do the comparison.
- **Endgame tables, opening book, learned heuristics:** some of the knowledge produced by chess engines was produced *using* large computers (by playing or observing very large numbers of games, brute-forcing endgames, and so on). We'd prefer exclude these factors if they use more compute than would have been affordable for the old engine. If we can't, we at least want to quantify their influence. Including these factors could significantly overstate the returns of human R&D; it's an important thing to know about, but for the purpose of forecasting the impacts of AI we really want to separate it out. This may be a major constraint for considering new engines.
- **What counts as an "operation"?** I don't think that making "hardware" comparisons between new and old computers will be straightforward. I think it's easier if we restrict attention to consumer microprocessors, and I'm hoping that there's only a little bit of uncertainty (e.g. a factor of 2). I think you can do the experiments before figuring this out, and then try to clarify relevant issues by seeing how fast the computers can run simple relevant backgrounds. (The "old hardware" run should probably be on just one processor.)
- **Timing out.** It may be prohibitively expensive to give the old engine enough compute to beat the new engine. I think you should do a little bit of that work, to try to figure out the basic picture (how expensive it would be, rough bounds on the numbers, some very noisy estimates from playing just a few games). We can figure out where to go from there.
- **Different time controls.** I'm expecting time controls to be comparable between the old and new competitions. If old competitions ran on much longer time controls, I'd prefer scale them down to something comparable (to try to better match what would have been realistic to experiment with during R&D / profitable for non-competition purposes). And similarly if new competitions are longer.
- **Ponder.** Thinking during the opponent's turn could change the answer a *tiny* amount (up to a factor of 2), but it really doesn't seem worth dealing with, even if some of the engines are optimized to use it.
- **Inefficiencies from running on new computers.** You might have to mess things up a lot to run old engines on new computers (or to deal with weird amounts of memory or so on). Ideally it would be possible to abstract out some of the details of "how long it actually takes" to talk about what the computation or memory costs ought to be if implemented without overhead. We'll have to see how that goes.

Prize structure

I'm not spending long evaluating any of this and my decisions are going to be completely unaccountable. Dealing with me may be a waste of time. Your only recourse will be to make an angry comment on this post.

I know that all makes it less appealing to participate. But please have it in mind before spending any time on this project, and consider yourself warned!

I'm planning to give away at least \$1,000 sometime in the next 3 months if anyone runs any plausible version of this experiment, or even points me to public information from which it's possible to back out a similar number. I'm fine if that means I have to give \$1,000 to a random LW comment.

I'd give away \$10,000 if someone ran a clean version of the experiment, I found it convincing, and they were transparent about their methods. Before giving away the full prize I'd likely have a waiting period where I offered a separate prize for people to post replications or caveats or so on.

I'm generally expecting/hoping to wrap this up over the next couple of months, but will adjust based on responses.

If multiple people do experiments I'll make some arbitrary call about how to allocate prize money. Timing will matter a bit but it's a secondary consideration to quality of experiment. Earlier submissions can get paid based if they helped pave the way for later submissions.

If you are planning to spend time on this I encourage you to make a comment. I'm very happy to provide thoughts on a proposed experiment, e.g. to tell you what size of prize I'd expect to give it or what concerns I'd have about a proposal. **None of this is binding.**

If receiving a prize would be weird or problematic for some reason, I'm still interested in the results and you can opt to receive a comparable quantity of gratitude instead of \$.

Desk research prizes

I'd also pay out \$100-\$1000 at my discretion for any of the following:

- Plausible (or better yet convincing) estimates of the total investment in chess engine R&D over time.
- Good analysis of the relative importance of hardware/software using public information (must at least improve over the analysis [here](#)). Or pointers to other similar experiments that have already been run.
- Any consideration that I feel should significantly change the experimental setup, e.g. quantifying the importance of endgame tables, or noting a critical difference between old and new chess engines, or suggesting a reason that Fritz is a bad engine to compare to.
- Any contributions that make it significantly easier to run this experiment (for example tracking down a usable implementation of an old chess engine).

(\$1000 bounty) How effective are marginal vaccine doses against the covid delta variant?

UPDATE: the bounty has now been awarded as follows:

- \$425 to Connor_Flexman
- \$300 to Josh Jacobson
- \$150 to JenniferRM/johnswentworth
- \$75 to ChristianKI
- \$50 to SoerenMind

====

Since I got vaccinated I've started working in-person, going to restaurants, travelling and whatever. But the delta variant might change this risk calculation. Some of the same buzz that got covid right early on, the first time around, is now buzzing about delta potentially being very bad.

So I'm exploring ways of responding to that. This time around I'd rather [stand up and fight](#) than lock myself in a house for a year. And so this question explores one possible approach.

There's evidence that two vaccine doses are more effective than one (though there might be much higher diminishing returns than is commonly acknowledged). I've also heard that past infection confers marginal immunity even in the presence of vaccination. Further, as far as I understand, the 2-dose schedule for many vaccines is more "a common thing that we happened to test in the big trials" rather than "the dose level at which marginal benefit = marginal risk". (And giving your population more than two doses is probably much harder logistically.)

Hence: getting *more* than the standard 1-2 vaccine doses might be a way to protect against delta.

I live in California, and we currently seem to have a large vaccine surplus. I heard a vague rumour about a friend basically just walking into a pharmacy and getting a 3rd shot. There's also RADVAC, which you can make yourself.

So, if there is supply... how worthwhile is it to get more vaccine doses? Should you get a 3rd, 4th, ... or even more? How does getting a further dose of the same vaccine compare to getting the first dose of a different vaccine?

I'm posting a bounty of \$1000 for answers that change my mind on this question (maybe increasing to many times that if this proves valuable).

The ideal thing I want would be a graph with the x-axis showing # doses, and the y-axis reduction-vs-control of the following four parameters:

- Symptomatic infection
- Hospitalisation
- Death

- Long covid

(Answers are of course impacted by the combinatorial search space of dose spacing / dose number / dose size / vaccine type / various demographics, but I won't make any special restrictions here for now)

I will pay at least \$1000 for answers that help me get clarity on this, split at my full discretion in proportion to how useful I find the answers. I'll pay out the bounty on a rolling basis as answers come in; there is no deadline. In case this proves fruitful and there seems to be useful marginal work, it's possible I will increase the bounty a lot (i.e. up to many thousands of dollars), or reach out to work with some answerers as contractors.

Experimentally evaluating whether honesty generalizes

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

If we train our ML systems to answer questions honestly in cases where humans can check the answer, will they generalize to behave honestly on questions where we can't check?

I think that we could learn a lot about this question by running experiments today. I think those experiments would be very valuable.

(I don't know anyone currently planning on working on this topic and I'd love it if anyone wants to take that up. This post doesn't represent a claim to any credit for any results in this genre, and other people have had very similar ideas. If you run some experiments you could cite this post but it's also fine if that doesn't make sense in context.)

The unsupervised translation setting

As an example, I'll think about "[unsupervised](#)" translation (if you've read that post you can skip this section).

Consider a model like GPT-3 that is trained to predict sentences in both English and French (but without a large dataset of translations). Suppose we want to train this model to answer questions in English about French sentences like "what does that word mean here?" or "are there any other plausible interpretations?" or "how does the speaker seem to feel about the topic they are discussing?"

We expect this to be possible, because the model understands quite a lot about the meaning of sentences in French, and is able to express itself in English. There may be cases where the model doesn't know the translation of a concept, or doesn't quite understand what an idiom means, but it should still be able to tell us what it does know.

I think this problem is an interesting analogy for a situation where an AI has built up superhuman knowledge by making predictions, and we want to train our AI to expose that knowledge to us in a useful way.

Proposed experiments

Let's pick a few categories of knowledge/capabilities. For example, we could split it up into an understanding of grammar ("Why would it have been a grammatical error to write *Tu Vas* in that sentence?"), of the literal meanings of expressions ("What does *Defendre* mean in this sentence?"), and evaluating tone ("Does the speaker seem angry or sad about the topic they are discussing?").

We'll assume that humans can oversee a few of these categories. Perhaps we can look up literal meanings of words and small phrases in a dictionary and we can look up

grammar in a textbook, but we don't know how to assess tone beyond looking at the literal meanings.

Then we wonder: what happens if we fine-tune our model so that it answers questions well in all the domains we can evaluate? We know that the model knows *something* about connotations, because it uses connotations to predict next words. But will it tell us what it knows?

	Grammar Definitions Tone		
Uses to make predictions	✓	✓	✓
Gives accurate answers	✓	✓	?

In this picture, the ✓ indicates that we've selected our model to use the given knowledge in the given way (e.g. we've ensured that it uses its knowledge of syntax to predict the next word, and that it accurately answers questions about syntax). The ? indicates that we don't know whether using SGD to fill in the other 5 boxes means that we get the 6th box for free.

(I started thinking about this picture of generalization based on a talk by John Schulman, who suggested that generalization may be more likely when it amounts to filling in missing cells from this kind of grid.)

In this example I described a tiny number of categories. That wasn't just for illustration purposes, I think these experiments are particularly interesting when the total number of categories is more like 3-5 than 1000 (see "Why try to generalize from n=3-5" below).

Evaluation

When it really counts we won't be able to tell if our model answers questions honestly in the last cell. But in an experiment today we could do so easily: we have a bunch of bilingual speakers who can give or evaluate answers about syntax/definitions/tone. At training time we get them to only give answers in the first two categories, and then at test time we ask them to evaluate answers in the third category.

We could quantitatively compare answers in the third category to the quality of answers in the first two categories, or we could qualitatively evaluate what kinds of mistakes the system makes when transferring.

Rather than comparing across different categories (which may have very different difficulties), it's probably better to fix a domain for evaluation (say grammar) and compare the performance of a models that do and don't hold out grammar.

Other splits

Instead of having one column per "type of knowledge," we could:

- Have a column for different language pairs. For example, we could supervise a translator for German and for Spanish and then see if it transfers to French.
- Have a column for different levels of sentence complexity. For example, we could supervise a translator for sentences with 1st grade to 8th grade reading levels, and see if it transfers to college reading levels.

- Have a column for different domains. For example, we could supervise a translator for fiction and news articles, and see if it transfers to informal dialogs.

If the results were different across different splits, it would be great to understand why. I'm not sure what kind of division is best to start with.

(Even if starting with different capabilities, I think that "grammar / definition / tone" is not the best division.)

Training for plausibility or coherence

My guess is that this experiment won't work if we train on 2 categories and ask it to generalize to the 3rd. But I think it's a lot more likely to work if we train for plausibility or coherence.

Even if I don't know how to answer questions about the tone of French sentences, I know what kind of answers I'm looking for. For example, if I ask "Is the author angry or sad?" and the model replies "en colère means 'angry'" then I can tell that's not a good answer. If the model says "The use of the word *fin* in this sentence suggests the author is more likely to be sad than angry" then I know that's the *kind of thing* I'm looking for, even if I can't evaluate whether it's actually correct.

I can easily define a reward function (based on human evaluations) that will discourage the totally-wrong answers. So in parallel with training my model to give *good* answers about the domains I know (grammar, definitions) I can train it to give at least *plausible* answers about the domains I don't know. So our original diagram now has an extra column:

	Grammar	Definitions	Tone
Uses to make predictions	✓	✓	✓
Gives plausible answers	✓	✓	✓
Gives accurate answers	✓	✓	?

Basically we've ruled out the models that just go totally off the rails for our new category, and are restricted to models that either tell the truth, or have learned how to bullshit effectively. But we've forced the models to actually tell the truth in the other two domains, so it's kind of plausible that the "easiest" thing to learn is just to tell the truth in all categories. (You could also test the picture with *only* the plausibility check, though see "concerns about the analogy" below.)

We could even take this idea further by imposing more elaborate consistency checks. For example, I could compare the claimed tone to the literal meaning of the sentence, and penalize an answer if it seemed too implausible. Or we could ask about the same linguistic construction in a few different contexts to see if the descriptions are consistent. This kind of thing could push us further in the direction of "Either telling the truth or bullshitting very convincingly."

Other domains

I think these experiments are easiest to do for language-modeling tasks. But in some ways they would be more convincing in other domains (see "Concerns about analogy" below).

For example, suppose that we have trained a model to simultaneously play Go and to answer questions in English. Then we start fine-tuning it to answer questions *about the game of Go*. For example, we might ask it "Is this group alive or dead?" or "What do you think your opponent will do?" or "Which piece breaks this ladder?" or "What are your threats in a ko fight?" or "Which of these shapes is stronger?"

We can ask the same question: if we train a model on several of these abilities, and train it to give *plausible-looking* answers to a held-out category, will it generally learn to give honest answers? (And similarly we could ask if it generalizes across types of games or complexities of the board positions.)

Why I care

I've been thinking about the "instrumental policy," a specific failure mode where smart systems learn to "tell humans what they want to hear" rather than "be honest."

Although I believe the instrumental policy will eventually be a severe problem, I don't think it will come up for existing systems like GPT-3.

But generalization could fail for any number of other reasons. If existing systems actually generalize well, I'd update towards thinking the instrumental policy is the *main reason to be pessimistic about generalization*. And if they generalize poorly, then that gives us some more "mundane" problems that we can study empirically today.

I think that even these early experiments may give us a lot of evidence about how to design partial supervision regimes where we combine some known answers with coherence conditions (and whether these methods work, and whether they are necessary). I don't know if those techniques will be important ingredients for alignment, but it seems useful to understand them better.

Finally, I think that concrete evidence on this question would help clarify discussions about alignment and make progress on some thorny disagreements. If we observe systems that learn to bullshit convincingly, but don't transfer to behaving honestly, I think that's a real challenge to the most optimistic views about alignment and I expect it would convince some people in ML. Conversely, if we *do* observe generalization to held-out kinds of knowledge, I think that should eventually start making pessimists lighten up, and would suggest some quantities to measure continuously to look out for initial signs of trouble.

Other remarks

Relation to other work on generalization

The ML community is very interested in the question "When and how do models generalize?" That question combines a bunch of factors: do models learn brittle heuristics or deep knowledge? Do they exploit correlations in the training set? Are models robust when some activations are pushed into quantitatively new regimes? And so on.

The experiments in this post are designed to specifically shed light on something more like [2-D robustness](#)--by focusing on cases where in some sense the model "knows" how to handle the new domain, we are distinguishing failures of *capability*

generalization from cases of *motive generalization*. We're not asking whether a model will generalize its understanding of language to a new domain---we're assuming that it still knows how to predict sentences in the new domain, and asking whether the machinery for "reporting what it knows" transfers along with the underlying capabilities.

I think this is particularly interesting to alignment, and is not well-addressed by existing empirical work in ML. But I haven't done a thorough review and am definitely interested in pointers to related work.

Because the ambition of this work is narrower, I think there is also room for algorithmic innovations (e.g. more clever coherence conditions, or different ways of combining training objectives) to solve this problem even if they have little hope of solving the full problem of "generalize well."

Why try to generalize from n=3-5 instead of n=1000?

I've discussed diagrams with 3-5 columns, whereas ML generalization typically works better when we have a very large diversity of "iid-ish" datapoints.

I do think it's interesting to study the behavior quantitatively as we vary the number of columns. But I'm interested in the small-n behavior because I think it may be analogous to the situation we ultimately care about.

Namely, in the real situation the diagram might look more like:

	Humans can answer quickly	Humans can answer with careful thought	Careful humans make mistakes, but a really thorough analysis can catch them	Humans never get the right answer
Uses info for other tasks	✓	✓	✓	✓
Gives plausible answers	✓	✓	✓	✓
Gives accurate answers	✓	✓	✓	?

We don't really have a large number of IID columns, we only have a few genuinely novel jumps to a new regime of question and we need to generalize across them.

There are definitely other kinds of diversity within each column, and that may improve the prospects for generalization. That's something we can study in the context of n=3-5---for example, if I want to generalize from 8th grade reading level to college reading level, does it matter if I have a larger diversity of kinds of questions about 8th grade reading level questions?

Concerns about the analogy

If doing this experiment with language models, my biggest concern is that "answer questions about French sentences honestly" is very close to something that appears in the pre-training distribution (whereas "bullshit convincingly about French sentences" isn't). This may make it easier to learn the right generalization, but it will break down as we move to tasks that no humans know (since those don't appear in the training set).

I think the best way to avoid this problem would be to consider other domains (like a StarCraft player who can explain what's going on in the game to a human, without any pre-training data about StarCraft).

I think it's worth revisiting this question if generalization turns out to work quite well. My gut sense is that the unsupervised translation task is cool enough (and different enough from the pre-training data) that it would be surprising and cool in any case, and that a strong positive result isn't too likely (maybe ~50%) anyway. So it's reasonable to revisit this after initial positive results.

A second concern is that our decompositions of knowledge may not be very analogous to "Questions human can answer" vs "Questions humans can't answer." I'm a bit less concerned about this because I think we can learn a lot about the conditions under which generalization tends to occur or not, and then figure out whether "Questions humans can't answer" feels like a special kind of category with respect to the actual determinants of generalization. If you were really concerned about this kind of thing, you could start more with examples like generalizing to harder reading levels (or to domains that most bilingual humans actually don't know, e.g. translations of technical language).

Will these problems go away?

I think a deep learning optimist might have a position like "A weak model won't necessarily generalize well to new domains, but stronger models will generalize better." But I think that negative results on this kind of experiment should still be very interesting to such an optimist, because we are choosing domains where the capabilities of the model *do* already generalize. Moreover, we can augment the pre-training dataset to involve richer sets of questions (e.g. about non-translation tasks) and further close the gap between current models and speculative future models.

Overall I would be relatively skeptical of someone who acknowledge that modern experiments don't demonstrate "good" generalization from a small number of categories, while expecting such generalization to occur for future systems just because they are smarter.

Relation to deception

The instrumental policy ("tell humans what they want to hear") is very similar to [deceptive alignment](#). But I think that the experiments in this post may be a lot easier than other experiments designed to exhibit and characterize deceptive alignment:

- I expect these experiments to run into more mundane generalization failures well before encountering the instrumental policy. Although I think this kind of generalization is ultimately critical for avoiding deception (by helping us be epistemically competitive with a learned optimizer), we can study it long before we have examples of deception.

- I think that simple forms of the instrumental policy will likely arise much earlier than deceptive alignment. That is, a model can develop the intrinsic motivation "Tell the humans what they want to hear" without engaging in complex long-term planning or understanding the dynamics of the training process. So my guess is that we can be carrying out fairly detailed investigations of the instrumental policy before we have any examples of deception.

Covid 7/29: You Play to Win the Game

Few things warm my heart more than playing to win the game. Few things sadden me more than observing someone not playing to win the game.

Usually, that means they are playing a different game instead, whether they admit it or not, and whether or not they know it themselves. The game, from this perspective, is simply that which you are playing to win, and your revealed preferences define the game's rules and objectives.

This week saw some excellent playing to win the game. The NFL, many parts of the government and a number of corporations began imposing vaccine mandates, hopefully causing a cascading effect. It's at least a start. There's a lot of ways in which we collectively are revealing ourselves not to be playing the game of 'get people vaccinated' let alone the game of 'prevent Covid while minimizing costs.' The lowest hanging fruit remains *fully approving the vaccines*, which we somehow still have not done.

A central question continues to be how effective the vaccines are against Delta. The Israeli claims turn out to probably be the result of basic statistical mistakes, so those scary numbers are now mostly off the table, although that still leaves us a range containing meaningfully distinct answers.

Another central question is Long Covid, for which we got some actual data, so there's a section analyzing that.

The big mystery remains why Delta suddenly peaked and turned around, first in India, and now in the UK and the Netherlands. These turnarounds are excellent news, and I presume we will see a similar turnaround at a similar point, but what's causing them to happen so quickly? I don't know.

Meanwhile, the numbers got worse slightly faster than I expected. Let's run them..

The Numbers

Predictions

Prediction from last week: 360,000 cases (+50%) and 1845 deaths (+10%)

Result: 392,000 cases (+62%) and 2042 deaths (+21%).

Prediction for next week: 610,000 cases (+55%) and 2,450 deaths (+20%).

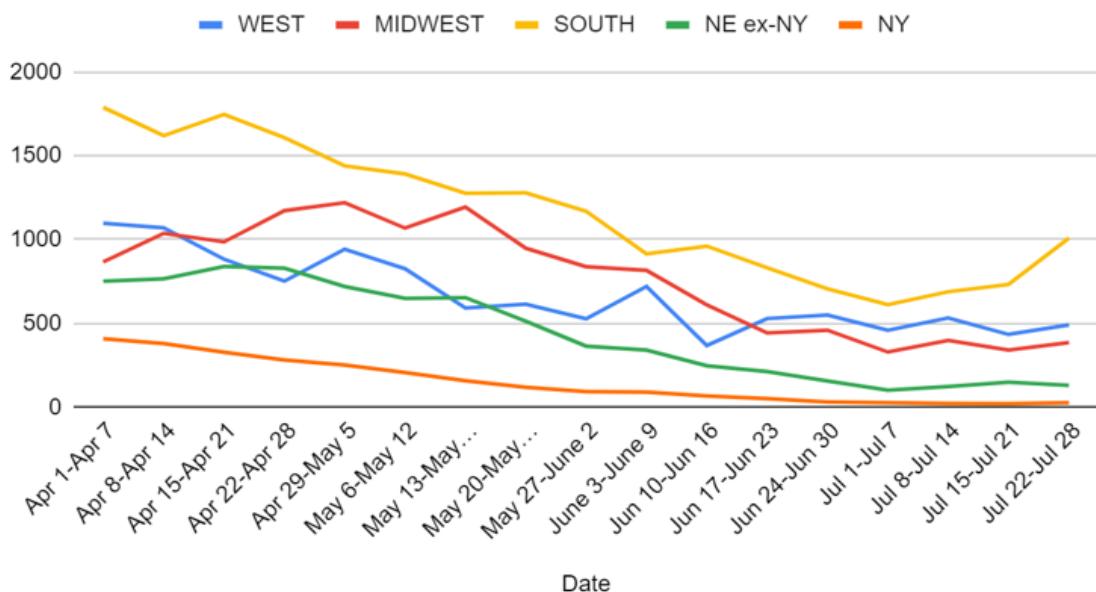
Things got worse slightly faster than expected. I doubt things are ready to turn around, but there are signs of the control system starting to act which should accelerate, and between vaccinations and infections (most of which likely do not result in positive tests and therefore known cases) immunity is building up, so I continue to expect the pace of growth to drop off a bit. I'd be unsurprised by numbers between about +40% and +70% for cases.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 10-Jun 16	368	611	961	314	2254
Jun 17-Jun 23	529	443	831	263	2066

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

Deaths by Region



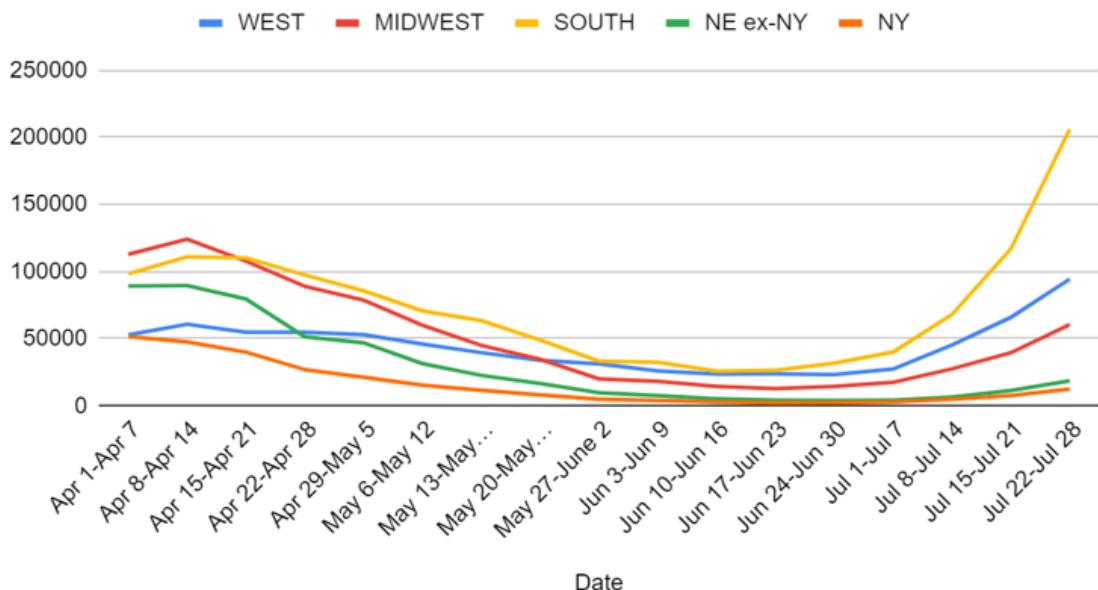
Deaths are going up slower than cases, but faster than one would have hoped. I interpret this partly as last week's number being artificially low, and partly as the South having a problem with its most vulnerable being partly unvaccinated, and thus we see a rise focused on the South including as a share of the cases.

Things will continue to get worse until several weeks after cases turn around. The question is if we can continue to see deaths lag behind cases, as they continue to do so in the UK. My guess is we won't do as well as some other places, especially in the South, but will still be doing better than our case counts alone would suggest.

Cases

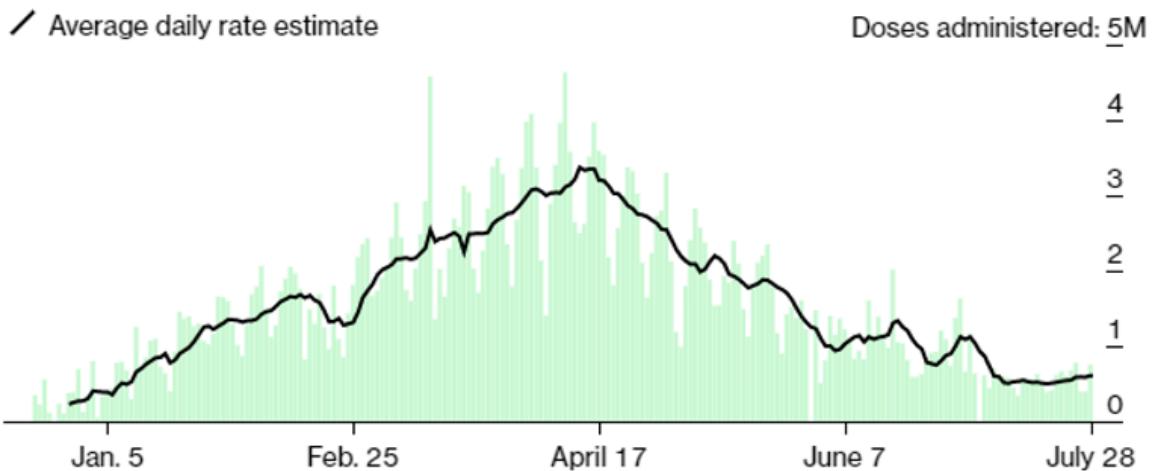
Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jun 3-Jun 9	25,987	18,267	32,545	11,540	88,339
Jun 10-Jun 16	23,700	14,472	25,752	8,177	72,101
Jun 17-Jun 23	23,854	12,801	26,456	6,464	69,575
Jun 24-Jun 30	23,246	14,521	31,773	6,388	75,928
Jul 1-Jul 7	27,413	17,460	40,031	7,065	91,969
Jul 8-Jul 14	45,338	27,544	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

Positive Tests by Region

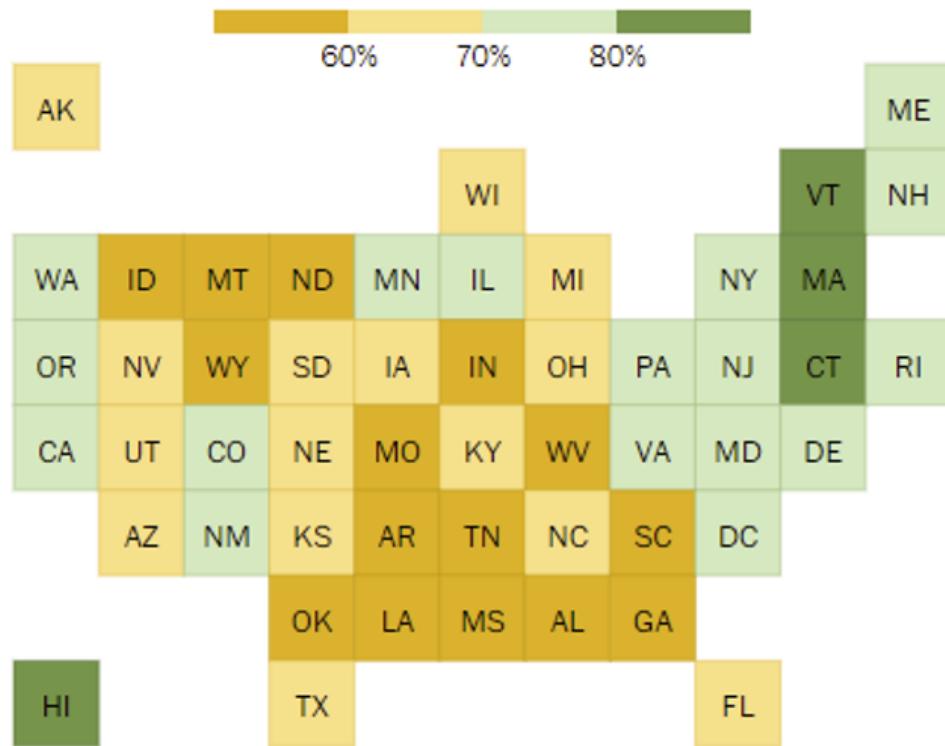


Vaccinations

In the U.S., **343 million doses** have been given so far. In the last week, an average of **608,380 doses per day** were administered.

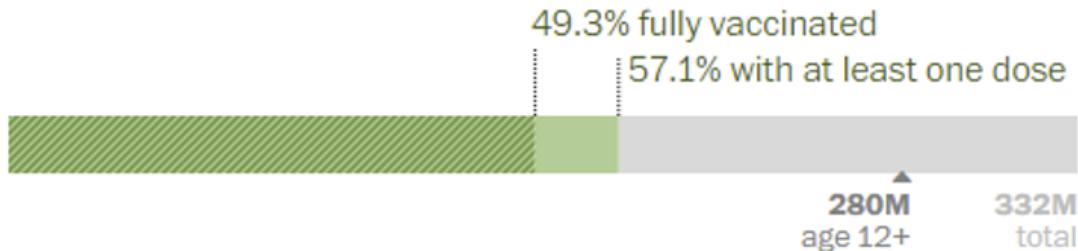


Percent of adults with at least one vaccine dose



189.5 million vaccinated

This includes more than **163.6 million people** who have been fully vaccinated in the United States.



In the last week, an average of **608.4k doses per day** were administered, a **15% increase ↑** over the week before.

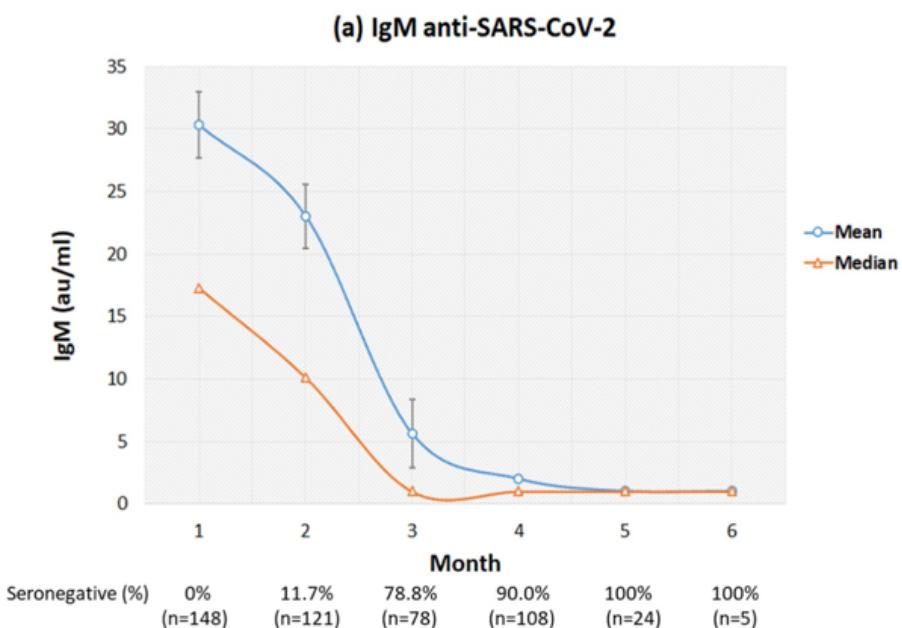
People respond to incentives. With the costs of not being vaccinated rising on all fronts, more people are making the decision to get vaccinated. The remaining people are harder to get, but efforts benefit from the growing social proof and social pressure from previous

vaccinations. It also helps that previously reluctant politicians, with notably rare exceptions, are increasingly getting with the program.

Vaccination Effectiveness

This thread takes the Israeli evidence (that likely was a quite sloppy statistical error, see the next section), combines it with antibody measures, and suggests that antibody counts decay with a half life of a few months, so under this hypothesis vaccinations continue to protect you from severe symptoms and death but after five months lose half their effectiveness at protecting against infection. I hate the smugness and overconfidence here, and also based on the clinical trials and result of one dose it doesn't make any sense that a 75% reduction in antibodies would cut protection against infection by 50%. Nor do I buy that there would be a threshold effect big enough to cause the curve she's analyzing here.

Also, this isn't a half-life:



Even if we take this fully seriously, it's a steady state that we get to in 3-5 months, and the median gets to its low point after three (which is likely random variance, but the point stands that most of the effect is in place after month three). It certainly doesn't seem anything like 'half life of 100 days' the next link in the thread cited, so I notice I'm confused.

I am noting such arguments here for completeness, but I do not put much weight on them. As usual, when one has a new hypothesis, one must then reconcile it with all the different data points, and alarming 'vaccine stops working' theories keep implying, well, that vaccines aren't mostly working, when they obviously mostly are. You can't *both* argue that we suddenly don't think vaccines work as well against Delta, *and* that vaccine protections fade rapidly over time, *and* also that Delta is so much more infectious to begin with among the unvaccinated, because such claims combine to not be even slightly compatible with the observed data in other places.

Then on Wednesday, [Pfizer came out with data that the vaccine remains effective after six months](#). That's the headline. The actual data is not as encouraging:

Pfizer's [paper](#), which has not yet undergone peer review, showed a slight drop in efficacy against any symptomatic cases of covid-19, the illness caused by the novel coronavirus, from 96 percent protection in the first two months after vaccination to 84 percent after four months.

Company officials also presented data showing that a third shot could boost disease-fighting antibodies many times higher than the level achieved by the standard two-dose regimen. They said on a quarterly earnings call that they planned to seek authorization for a booster by mid-August, reiterating the company's belief that a third dose would be needed to enhance immunity within a year of vaccination.

Contrary to Pfizer, that's quite the drop in effectiveness. The protection against death remains robust, and yes 84% is still quite good compared to what we would have asked for a year ago, but 'remains effective' is giving the wrong impression if this data holds up.

The other news here is that Pfizer plans to be calling for booster shots. It seems that a *three* dose regimen is much more effective than a two dose regimen, now that we've had the ability to test such things, and some places are moving to implement this already. The data here suggests that the third dose will bring things back to at least the early stage 96% effectiveness and plausibly even higher. If I am offered a third dose, I will happily accept it.

There is the concern that giving people third doses while others have not had the opportunity even for first doses is not ethical. I respect that perspective, but do not share it, and will leave it at that.

On That Israeli Data on Delta's Effectiveness

Israeli data has been suggesting remarkably low effectiveness of vaccinations against Delta. [This thread suggests this comes from... using the wrong denominator](#). The explanation is, the Israeli outbreak started in very highly vaccinated areas, and the effectiveness numbers came from the percentage of cases that were among the vaccinated, but they were comparing that to the overall population numbers. So, whoops.

If true, as I believe it probably is, that would explain things and be on some levels comforting, but on other levels it's the opposite of comforting, because this is saying that the Israeli outbreaks started in highly vaccinated areas. So, whoops again.

I'm inclined to believe that such simple mistakes were happening here, because [the Israeli numbers simply didn't make any sense](#). They were incompatible, even with each other, let alone with what we were seeing elsewhere. And I'm definitely at the point where such stupid mistakes aren't surprising. This one is rather stupid, but that's the way such things seem to often go.

Delta Variant

For those who need it: [Thread explaining basics of how vaccine protection interacts with spread of Delta.](#)

[CDC has reversed course on its mask mandates](#). Masks will be back in schools, where I've learned first hand and the hard way that schools feel compelled to follow the guidelines. They're suggesting indoor vaccinated masking in 'areas where there is a surge' which doesn't really make a lot of sense and will cause some confusion, but perhaps the hope is it

will make intuitive sense to regular people. It's good that when the facts change, the CDC changes its mind, at least.

There are two central facts about Delta one's model must explain. First, the dramatic takeover of the pandemic and rise in overall cases across countries. Second, [the sudden reversal of those trends in many places](#), including India, the UK and the Netherlands:

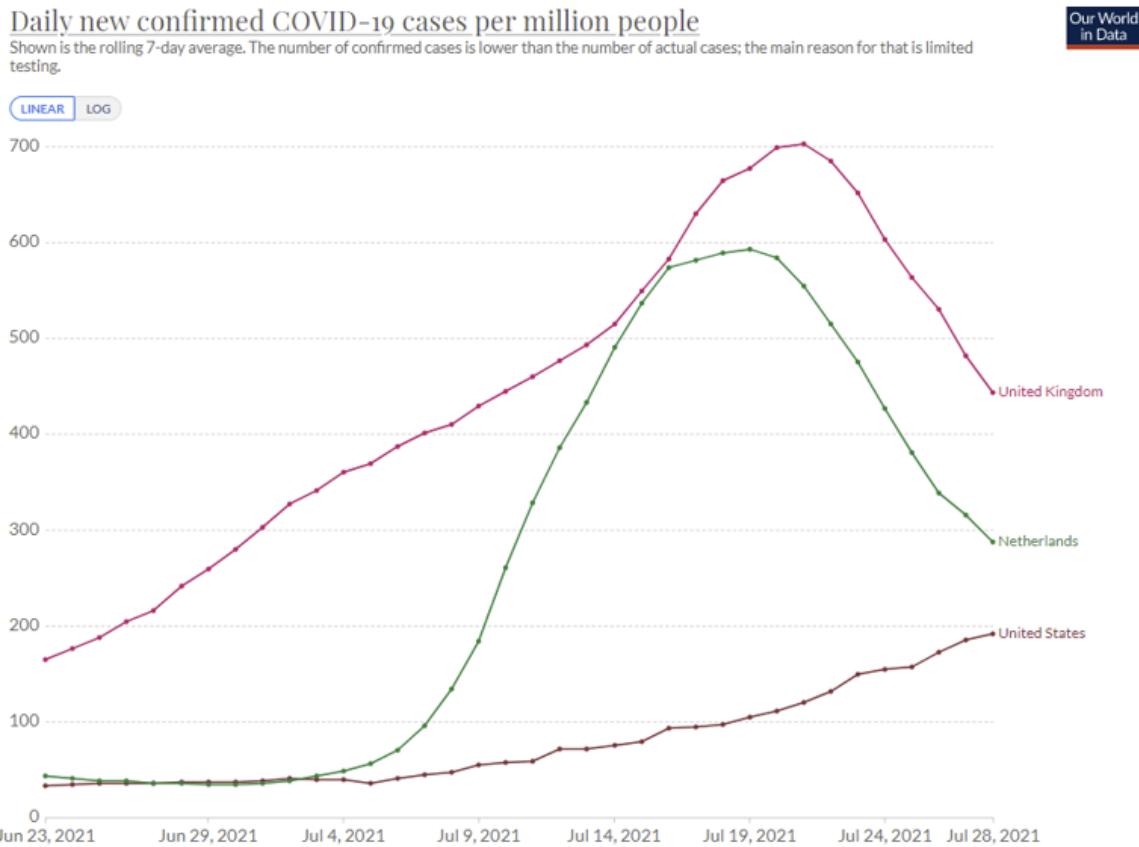
 **Nate Silver**  @NateSilver538 · Jul 25 ...

Delta is clearly very contagious but I wish we had better explanations for the rather abrupt turnaround in cases in India and now (fingers crossed) in the UK.

 **Chise**  @sailorrooscout · Jul 25

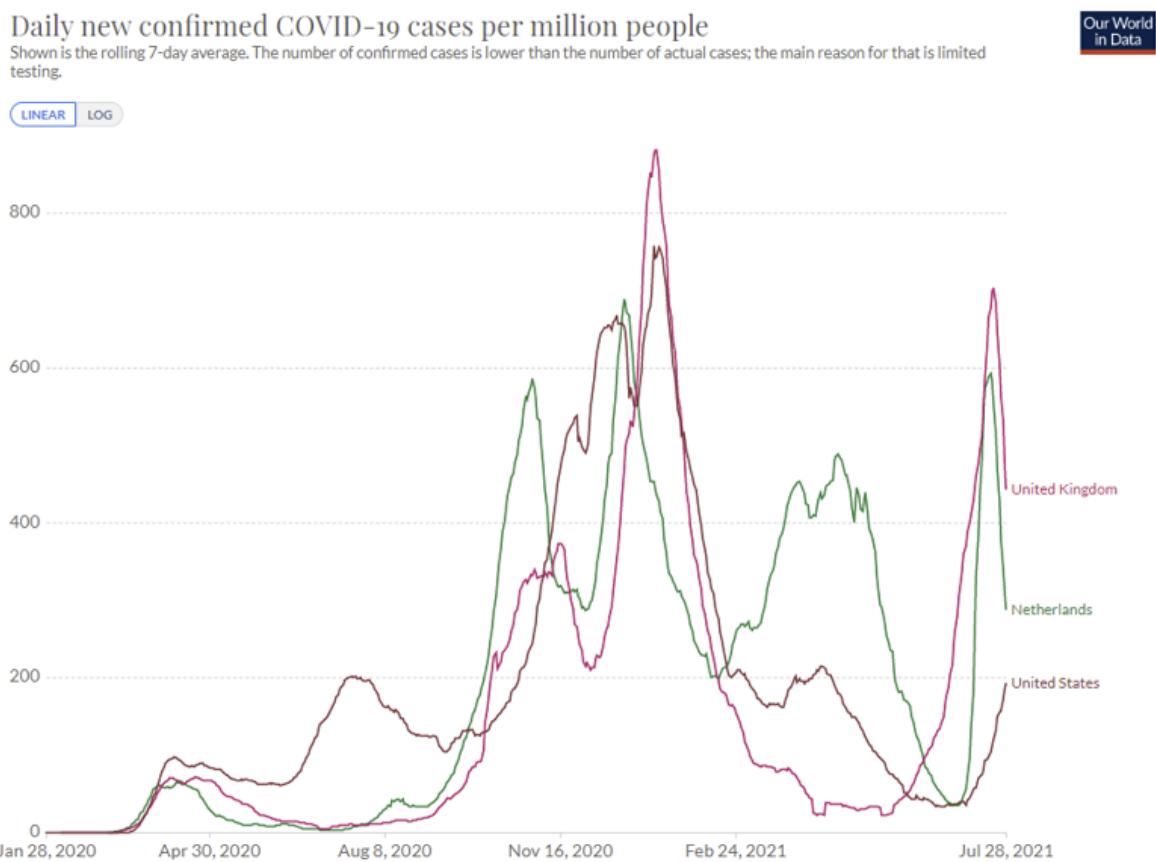
The number of people infected with SARS-CoV-2 has dropped for a fourth day in the UK. The UK recorded 31,795 new infections yesterday, a fall of 32% from 46,558 on Tuesday. This figure is down 42% from last Saturday.

[Show this thread](#)



The Dutch numbers are down by half, the UK is not far behind.

If you zoom out, the case numbers are still large.

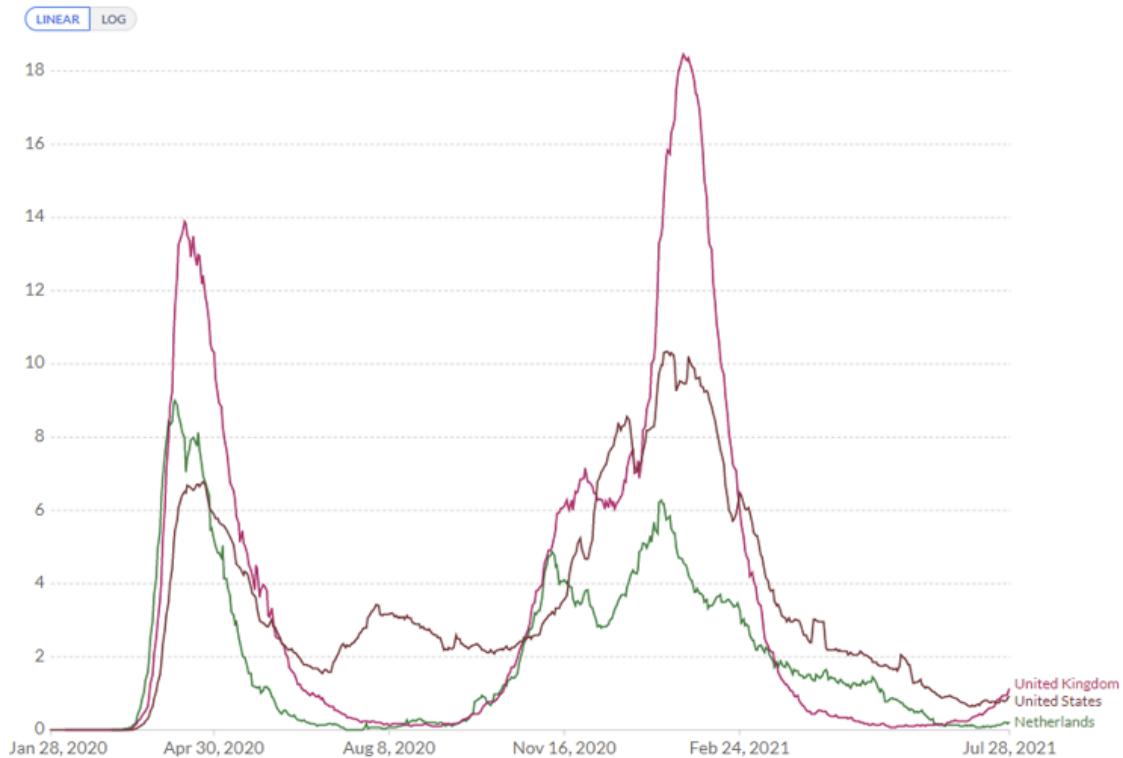


But the *death* numbers barely register:

Daily new confirmed COVID-19 deaths per million people

Shown is the rolling 7-day average. Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19.

Our World
in Data



It will take another week or two for the UK/Netherlands death numbers to peak, but this is what vaccinating the vulnerable looks like where it counts.

[Here's some data from Minnesota:](#)



janashortal ✅
@janashortal

...

What vaccines did people have with breakthrough cases in Minnesota?

Glad you asked.

1,626,557 MN got Pfizer - 2,074 breakthroughs (0.13%)

1,125,919 MN got Moderna - 976 breakthroughs (0.08%)

266,975 MN are J&J - 813 breakthroughs (0.30%)

*All data from MN Dept. of Health

1:28 PM · Jul 27, 2021 · Twitter Web App

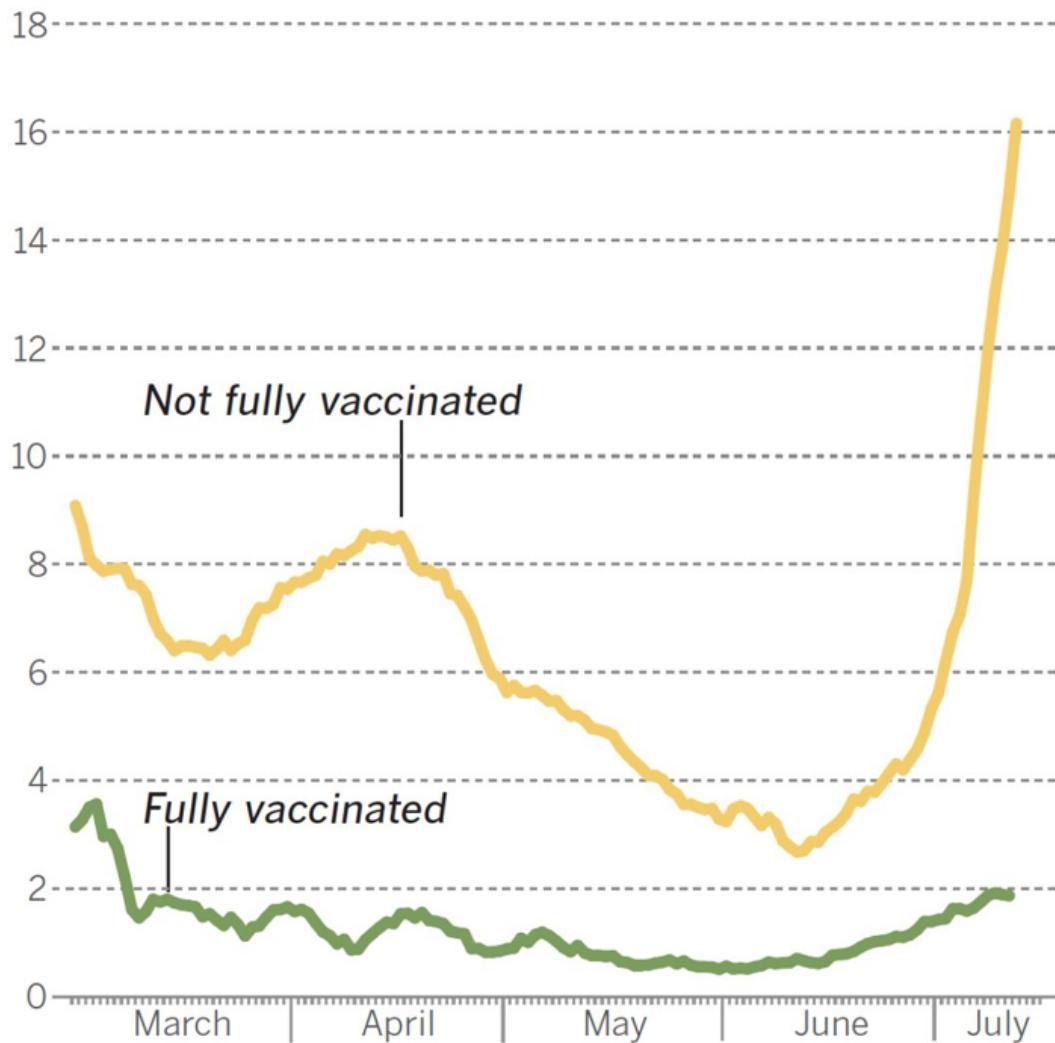
If we take this at face value (it's not normalized sufficiently so there's a bunch of reasons not to, but they point in various directions), it suggests about the same ratios in effectiveness for Delta between the vaccines that we saw for the older strain. Moderna's effective dose is higher than Pfizer's, which potentially could be responsible for that differential, although I'm still mostly inclined to treat the two as the same until we get a lot more similar data points elsewhere. Normalizing properly for exactly when vaccinations were in effect is tough, but for context Minnesota at the time had just under 30k confirmed infections since May 1.

The data below from San Diego is better normalized, and I'm going to say that this is enough non-normalized data points.

[This graph is pretty bizarre](#) when taken at face value, and I presume it shouldn't be, and is only the county of San Diego, but still seems worth noting:

Coronavirus infections after vaccination

Seven-day rolling average of cases per 100,000 residents, March 1 through July 12, 2021.



MICHELLE GILCHRIST U-T

Details:

The fully vaccinated population for each day is the cumulative number of county residents documented to have received the final dose of COVID-19 vaccine more than 14 days prior to that day. The not fully vaccinated population is the estimated total county population minus the fully vaccinated population.

Source: County of San Diego

MICHELLE GILCHRIST U-T

At the lefthand side, we have about 3 cases per 100k among the vaccinated and 9 among the unvaccinated, a ratio of 3:1. That's a surprisingly small ratio.

At the end of May, cases among the vaccinated level off, but cases among the unvaccinated continue to drop until about June 10.

Then the cases among the unvaccinated shoot up, and at the end, we have 16 unvaccinated cases per 100k and 2 vaccinated cases, for an 8:1 ratio. In July, this is most definitely representing a pandemic of the unvaccinated.

	vaccinated	unvaccinated	total
Infected (lagged)	30,000	80,000	110,000
Hospital admits	725	9,275	10,000
Deaths (7/13-7/19)	78	1,900	1,978
CFR (lagged)	0.3%	2.4%	1.8%

The question is, why would this ratio suddenly get much *bigger*?

If Delta is, as everyone now fears, reducing the effectiveness of vaccinations, you'd expect the ratio to go *down* rather than up. Whereas this new 8:1 ratio implies a much *higher* effectiveness level, unless one can explain it via other factors.

One possible explanation is that early on the vaccinations went to highly vulnerable people, whereas now they're more evenly distributed, but that would only explain contrasting March with June, not June with July, where that effect is going to be small.

Another is to [claim this didn't happen](#):



Nick Brooke @Hi_Dr_Nick · Jul 25

Replies to [@EricTopol](#) [@SanDiegoCounty](#) and [@sdut](#)

From June to July, vaxed infections increase from 0.5 to 2 per capita, a x4 increase. Unvaxed infections increase from around 3.5-4 to 16, right around a x4 increase.

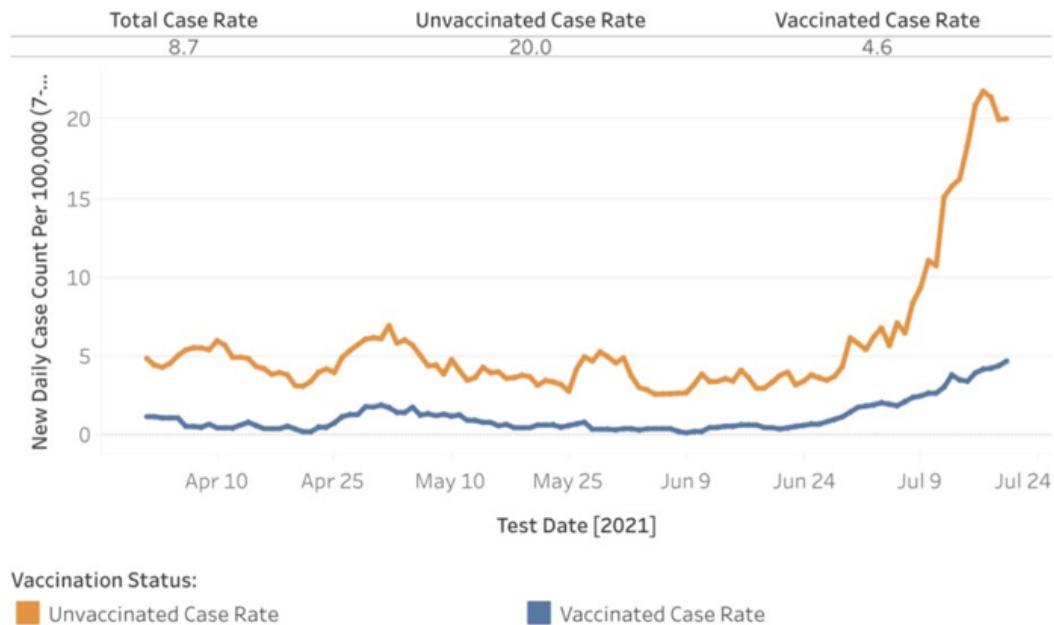
Both groups are growing geometrically, but this linear scale chart doesn't show that. Log scale is better.

If we look at the ratio that had emerged before that period in early June, we do get a different picture then, but that period where the unvaccinated cases are going down but

vaccinated cases went up is super weird for its own reasons, and the contrast with the left side of the graph is still clear. At a minimum, a constant ratio does not suggest any decrease in effective immunity.

A third is to say that San Diego's numbers are quirky because someone's are going to be quirky.

Here's [another Tweet that shows Marin County](#):



Looking at that, I see a ratio of just over 4:1 rather than 8:1. I don't see a big shift to higher ratios in July, but I also don't see a shift to *lower* ratios either.

I tried briefly to find similar charts or data for larger areas and it wasn't obvious where they were. If anyone can link to one it would be appreciated. The normalizations here make the data much more useful but I'd prefer a much bigger and ideally more representative area.

Vaccinations are great, but what matters is immunity, and vaccines are only one way to get immunity. Antibodies are not a perfect proxy for immunity, but they're likely the best one available, and [the numbers in the UK on antibodies are very, very high...](#)

Estimated percentage of adults who tested positive for COVID-19 antibodies in the week beginning 28 June 2021

(not living in care homes, hospitals or other institutional settings)

91.9%	92.6%	90.0%	88.6%
England	Wales	N. Ireland	Scotland

Source: Coronavirus (COVID-19) Infection Survey, UK:
antibody and vaccination data, 21 July 2021



When this many people have antibodies, it's both scary that cases for a while were going up anyway, and also easy to understand how there might be a rapid peak followed by a reversal. It's especially easy to see this if the serial interval averages two or three days instead of five, so things that happen tend to happen fast.

If 90% of people are immune, then each additional 1% that becomes immune reduces spread after that by a full 10% even if everyone is otherwise identical. Whatever vectors still remain for the virus to spread burn themselves out rapidly, until there isn't enough left to sustain the pandemic.

It is possible that this is simply what it takes to turn the corner on Delta. If so, then if the USA has roughly 70% of adults having been vaccinated, than we would turn the corner when about two thirds of the unvaccinated have been infected at some point sufficiently to develop antibodies, with the understanding that many Covid infections don't lead to a positive test and therefore don't show up in the statistics, and also a lot of them already happened over the past year and a half.

It still doesn't explain the *full extent* of the rapid turnarounds in India and the UK, but it helps explain it. Lag in the control system also helps, but again I notice my model remains surprised by this.

You Play to Win the Game

Last year, the NBA figured out how to do Covid testing. This year, it's all about vaccinations, and the NFL is stepping up.

As we all know, in the NFL and also in life, [You Play To Win The Game](#).

[Winning isn't everything, it's the only thing.](#)

What does it look like to play to win, when it comes to Covid-19?

[It looks like this. Here's the key operating principles.](#)



Richard M. Nixon @dick_nixon · Jul 22

...

I said last season that baseball and football should do this. It's long past time.

Do you want to compete, or stand on ceremony? What win matters more—your sense of superiority, or the field?



Tom Pelissero ✅ @TomPelissero · Jul 22

The NFL just informed clubs that if a game cannot be rescheduled during the 18-week season in 2021 due to a COVID outbreak among unvaccinated players, the team with the outbreak will FORFEIT and be credited with a loss for playoff seeding, per sources.

Massive implications.

[Show this thread](#)

Q 15

T 26

Heart 191

Upvote



Richard M. Nixon @dick_nixon · Jul 22

...

Read all of that. The league is not playing around.



Tom Pelissero ✅ @TomPelissero · Jul 22

...

The NFL just informed clubs that if a game cannot be rescheduled during the 18-week season in 2021 due to a COVID outbreak among unvaccinated players, the team with the outbreak will FORFEIT and be credited with a loss for playoff seeding, per sources.

Massive implications.



Tom Pelissero ✅

@TomPelissero

...

Replying to [@TomPelissero](#)

Here's more from today's memo, which also says the team responsible for a canceled game because of an outbreak among unvaccinated players/staff will be responsible for financial losses and subject to potential discipline from the commissioner. Wow.



Tom Pelissero @TomPelissero · Jul 22

Other key competitive aspect of today's memo:

...

Vaccinated individuals who test positive and are asymptomatic can return to duty after two negative tests 24 hours apart.

Unvaccinated individuals still subject to mandatory 10-day isolation period.



Tom Pelissero @TomPelissero · Jul 22

More strong language from today's NFL memo:

...

"Every club is obligated under the Constitution and Bylaws to have its team ready to play at the scheduled time and place. A failure to do so is deemed conduct detrimental. There is no right to postpone a game."

33

740

5.1K

↑



Tom Pelissero @TomPelissero · Jul 22

This is the NFL's strongest step yet to incentivize vaccinations. In essence, vaccination status dictates action:

...

"If a club cannot play due to a Covid spike in vaccinated individuals, we will attempt to minimize the competitive and economic burden on both participating teams."

[Here's what would trigger a forfeit:](#)

- Every club is obligated under the Constitution and Bylaws to have its team ready to play at the scheduled time and place. A failure to do so is deemed conduct detrimental. There is no right to postpone a game. Postponements will only occur if required by government authorities, medical experts, or at the Commissioner's discretion.
- In light of the substantial roster flexibility in place for the 2021 season, absent medical considerations or government directives, games will not be postponed or rescheduled simply to avoid roster issues caused by injury or illness affecting multiple players, even within a position group.
- If a game is cancelled/postponed because a club cannot play due to a Covid spike among or resulting from its non-vaccinated players/staff, then the burden of the cancellation or delay will fall on the club experiencing the Covid infection. We will seek to minimize the burden on the opposing club or clubs. If a club cannot play due to a Covid spike in vaccinated individuals, we will attempt to minimize the competitive and economic burden on both participating teams.
- Whether to reschedule a postponed game will be dependent on health and safety reasons at the recommendation of medical experts as well as considerations of stadium availability, schedule integrity, fan convenience, and other appropriate matters.
- If a game cannot be rescheduled within the current 18-week schedule and is cancelled due to a Covid outbreak among non-vaccinated players on one of the competing teams, the club with the outbreak will forfeit the contest and will be deemed to have played 16 games for purposes of draft, waiver priority, etc. For the purposes of playoff seeding, the forfeiting team will be credited with a loss and the other team will be credited with a win.

In other words:

If your vaccinated players force us to cancel the game, we'll try to do the best we can for everyone involved.

If your unvaccinated players force us to cancel the game, f*** you. You forfeit the game for all purposes where you want to win, the game is cancelled for all purposes where you wanted to lose, and the league will focus on 'minimizing the financial and competitive burden' exclusively on the other team. Not you. You're on your own.

Oh, and regardless of vaccination status, if there's an outbreak that cancels the game, [the players on both teams don't get paid for that game:](#)



Tom Pelissero @TomPelissero · Jul 22

And the biggest penalty of all for players:

...

"If a game is cancelled and cannot be rescheduled within the current 18-week scheduled due to a Covid outbreak, neither team's players will receive their weekly paragraph 5 salary."

You read that right: NOBODY GETS PAID.

That's what it looks like when you play to win.

Some NFL players are less than thrilled with this situation.

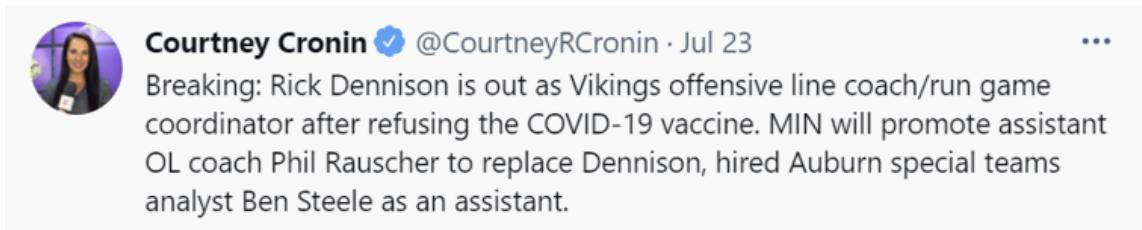
For example:



A screenshot of a Twitter post from Deandre Hopkins (@DeAndreHopkins). The tweet reads: "Never thought I would say this, But being put in a position to hurt my team because I don't want to partake in the vaccine is making me question my future in the @Nfl". The timestamp is 3:49 PM · 7/22/21 · Twitter for iPhone. The post has 21 Retweets, 24 Quote Tweets, and 144 Likes.

Yes, it turns out that not protecting yourself against infectious disease can put you in a position to hurt your team. Who knew?

[Who else we got?](#)



A screenshot of a Twitter post from Courtney Cronin (@CourtneyRCronin). The tweet reads: "Breaking: Rick Dennison is out as Vikings offensive line coach/run game coordinator after refusing the COVID-19 vaccine. MIN will promote assistant OL coach Phil Rauscher to replace Dennison, hired Auburn special teams analyst Ben Steele as an assistant." The timestamp is Jul 23.

I hope someone compiles all of these together, and then continues in the sacred tradition of arranging them into a YouTube montage video set to [Taylor Swift's I Forgot That You Existed](#).

Good luck, Mr. Hopkins and all the rest of you, in all your future and hopefully far away endeavors.

[The Pac-12 is considering following suit](#). It does make a lot of sense the way the commissioner George Kliavkoff put it:

"I will tell you that I'm leaning towards going back to the pre-COVID rules that had a team that was not able to field enough players to forfeit the game,"

Most of the Pac-12 schools have vaccine mandates in place in any case. [Washington State is currently trying to figure out what to do when its coach refuses to get with that program](#).

We need more of that 'can do, and if you choose not to do that's on you that you didn't do it' spirit.

Others Playing To Win the Game

The good news is we are indeed seeing more of this spirit, and there seems to be momentum behind these efforts.

New York and California are requiring government workers who don't get vaccinated to get weekly tests and wear masks indoors. New York at least is not providing a testing option for health care workers, it's vaccination or you are (very understandably) fired. Virginia is mandating vaccinations for health workers, as is the veterans administration.

From Bloomberg's daily newsletter:

Under the new requirements, unvaccinated New York schoolteachers, police officers, fire officials, front-line and office workers who don't comply with testing and mask requirements won't be able to come to work and will lose their pay.

"This means everybody," de Blasio said. "If someone is not wearing their mask, they will be removed from the workplace."

A few hours after de Blasio spoke, California Governor Gavin Newsom announced similar action.

"This is a requirement, to prove you've been vaccinated—and if you have not, you will be tested," Newsom said.

Reports are that Biden will follow shortly with a similar rule for federal employees, if he hasn't already by the time you read this.

[Companies are now able to get in on the act:](#)



BNO Newsroom @BNODesk · 12h

3 major companies have announced COVID-19 vaccine mandates:

- Google
- Facebook
- Netflix

My hope is that this effort continues and spreads, and more and more private employers will be emboldened to enact similar policies. If I was a private employer whose employees were interacting with each other in person or with customers in person, I'd hope to be more worried about what would happen if I *didn't* mandate vaccination, including legal consequences, than what would happen if I did, and this transition will go a long way. As usual, [we blame people via Asymmetric Justice](#) for action but not inaction, so the less a (soft via testing and masking requirements) mandate seems like a bold choice to take action and more like a default state that protects the employer from liability or an outbreak among the other employees, the better. These things matter.

And once again, we gotta get full authorization as soon as possible. A huge amount of the lack of mandates is the lack of full authorization. If we got it, this cascade could kick into high gear quickly, and give a justification for a change in policy. At this point, any efforts to get people vaccinated that don't involve a demand for full authorization are impossible to take seriously. How can we be mandating vaccinations but not be ready to fully approve them?

A bonus is it creates positive selection in employees. If anyone quits or otherwise goes ballistic about the situation, that was likely a time bomb on your team in one form or another, they're definitely bad at risk management, and this gives you the opportunity to be rid of them. It's hard to hire right now so that could be short term trouble, but identifying and getting rid of bad employees is both highly difficult and valuable.

Many other countries also continue to get with the program. [Here's Israel this week](#), playing to win:



Aaron Sibarium @aaronsibarium · Jul 23

Israel's Prime Minister says that those who refuse to get vaccinated will be barred from public spaces without a negative COVID test—at their expense. "There is no reason why...people who have carried out their civic duty should have to finance the tests."



and it is our very lives.

Statement by Prime Minister Bennett following me...

Prime Minister Naftali Bennett issued the following statement to the media: "Good evening. I have ...

🔗 youtube.com

19

78

167



Aaron Sibarium @aaronsibarium · Jul 23

He also says that the "discussion" about vaccines "needs to stop." The unvaccinated threaten "the freedom of every Israeli citizen."

I more than fully endorse this take on all of these developments:



Nate Silver ✅ @NateSilver538 · Jul 24

It's reflective of certain type of (I would argue somewhat incoherent) liberal ideology that restrictions placed on the unvaccinated are portrayed as "coercive" whereas restrictions placed on *everybody* (e.g. lockdowns) generally were not framed that way.

Persuasion vs. Coercion: Vaccine Debate in Europe Heats Up

France is taking the lead in making life unpleasant for the unvaccinated, even requiring some people to get shots. Protesters see a soft dictatorship dawning.

Long Covid

[We finally have some real data to look at.](#)



Prof. Christina Pagel @chrisc chirp · Jul 23

LONG COVID THREAD:

...

The people running the BBC Horizon "Great British Intelligence Test" challenge on over 80,000 people took the opportunity to see if they could detect any differences by whether people had had covid or not...



Prof. Christina Pagel @chrisc chirp · Jul 23

Replies to [@chrisc chirp](#)

...

2. They did this because of increasing concern over reported cognitive impacts of long covid - but more evidence is badly needed.



Prof. Christina Pagel @chrisc chirp · Jul 23

...

3. What they found was significant cognitive deficit for people who'd had covid compared to people that hadn't, after controlling for things like age, education, sex, first language etc.

The degree of deficit was worse the more severe the initial covid infection had been.



Prof. Christina Pagel @chrischirp · Jul 23

4. This isn't just about long covid - this compares people who had had covid with those who hadn't, regardless of ongoing symptoms. Most people who had had covid reported being recovered, but about 25% with confirmed covid reported ongoing symptoms (ie long covid).

...



Prof. Christina Pagel @chrischirp · Jul 23

5. The cognitive deficits remained whether ongoing symptoms were there or not, and did not depend on time since covid either.

...

This seems to suggest it is a long lasting effect.

It also doesn't depend on pre-existing health problems .



Prof. Christina Pagel @chrischirp · Jul 23

7. And the types of deficit found were consistent with the sorts of cognitive problems reported by people wth long covid

...



Prof. Christina Pagel @chrischirp · Jul 23

10. What if by the time there can be no doubt of long term problems in many people who've had covid, we've allowed millions more infections leaving hundreds of thousands more people affected.

...

ONS estimated 634K people with long covid that impacts their life in June.

- Symptoms adversely affected the day-to-day activities of 634,000 people (65.9% of those with self-reported long COVID), with 178,000 (18.5%) reporting that their ability to undertake their day-to-day activities had been "limited a lot".

The original paper is [here](#).

Headline findings:

Methods

We sought to confirm whether there was an association between cross-sectional cognitive performance data from 81,337 participants who between January and December 2020 undertook a clinically validated web-optimized assessment as part of the Great British Intelligence Test, and questionnaire items capturing self-report of suspected and confirmed COVID-19 infection and respiratory symptoms.

Findings

People who had recovered from COVID-19, including those no longer reporting symptoms, exhibited significant cognitive deficits versus controls when controlling for age, gender, education level, income, racial-ethnic group, pre-existing medical disorders, tiredness, depression and anxiety. The deficits were of substantial effect size for people who had been hospitalised ($N=192$), but also for non-hospitalised cases who had biological confirmation of COVID-19 infection ($N=326$). Analysing markers of premorbid intelligence did not support these differences being present prior to infection. Finer grained analysis of performance across sub-tests supported the hypothesis that COVID-19 has a multi-domain impact on human cognition.

Interpretation

Interpretation. These results accord with reports of 'Long Covid' cognitive symptoms that persist into the early-chronic phase. They should act as a clarion call for further research with longitudinal and neuroimaging cohorts to plot recovery trajectories and identify the biological basis of cognitive deficits in SARS-COV-2 survivors.

Implications of all the available evidence

This study confirms the hypothesis that individuals who have been infected with COVID-19 have persistent objectively measurable cognitive deficits after carefully controlling for pre-morbid IQ, pre-existing medical conditions, socio-demographic factors and mental health symptoms. Multiple studies are now using the online assessment technology reported here to investigate the neural correlates of cognitive deficits in people who have survived SARS-COV-2 infection, relate them to clinical outcomes and track at scale how they change over time.

This raises, and/or lets us more usefully address, multiple questions. Should we believe the results of this study? Do these effects seem real? If they are real, what can either an individual or our society in general do about this? Does this change anything if it is true?

First, no matter what the flaws here, a big thank you for running this at all. It does seem like this is a big advance in information value over previous sources. This stuff is hard, and doing something concrete is better than not doing it. This does seem to rise to the level of something useful.

Second, as usual, studying such things is really hard and there are lots of problems, so let's get right to some of those.

First off, baseline methods.

Methods

Study promotion

We collected data from members of the general public, predominantly from the UK, who completed an extended questionnaire (inclusive of questions pertaining to COVID-19 infection) and series of cognitive tasks via The Great British Intelligence Test, a collaborative citizen science project with BBC2 Horizon that launched in late December 2019. At the beginning of January, articles promoting the study were placed on the Horizon homepage, BBC News homepage and main BBC homepage, and circulated via news meta-apps. They remained in prominent positions within the public eye throughout January. In May, aligned with report of initial results considered of interest to the general public via a BBC2 Horizon documentary, there was a further promotional push. This led to high recruitment in the months of January and May, with lower, but still substantial recruitment between and after these dates. The data analysed here includes responses from January until December 2020.

2.2. Data collection

The study was promoted as a free way for people to test themselves in order to find out what their greatest personal cognitive strengths were. It comprised a sequence of nine tests from the broader library that is available on our server system based on prior data showing that they can be used to measure distinct aspects of human cognition, spanning planning/reasoning, working memory, attention and emotion processing abilities, in a manner that is sensitive to population variables of interest whilst being robust against the type of device that a person is tested on. In this respect, the battery of tests should not be considered an IQ test in the classic sense, but instead, is intended to differentiate aspects of cognitive ability on a finer grain. The tests had been optimized for application with older adults and people with mild cognitive and motor impairments. This study was approved by the Imperial College Research Ethics Committee (17IC4009). Participants provided informed consent via the study website prior to starting the assessment.

The data analysed here includes responses from January until December 2020.

There is an obvious concern here. People were recruited to do an intelligence test by offering them an intelligence test. Which, yes, right on, very on the nose and efficient incentivizing, but we do have to worry about the selection effects from that and whether the standard controls handle them.

Here's what they controlled for: "Nuisance variables were age, sex, racial-ethnicity, gender, handedness, first language (English vs other), country of residence (UK vs other), education level, vocational status and annual earning."

Handedness! So other than the bonus handedness, this is your basic check of control variables for basic demographic information and socio-economic status.

Before going further, there are three potential confounding factors here that seem like big issues.

The first is that *choosing to take the test* could be a function of one's situation and practical intelligence. If I had Covid and fully recovered and felt fine, I would not be so curious about taking an intelligence test. If I had Covid *and had continued problems*, then I would plausibly be much more interested to know how I did on such a test. In theory, this could be the whole situation, since those who feel stupid due to non-Covid reasons or due to lockdowns wouldn't feel the same curiosity and wouldn't opt in, whereas those who had Covid *and happened to also feel stupid recently* would take the test.

The second is that *being intelligent helps prevent Covid-19*, after controlling for the other factors. The study was done in 2020 so vaccination isn't relevant, but plenty of other decisions matter. How and when people decided to mask, including how much care was put into doing it properly with a good fit, determines both how likely they were to get Covid and how severe their case was likely to be. Social distancing is similar. As is making a wide variety of other good decisions about how much risk to take. Intelligence also correlates with the type of work that can be done remotely, even controlling for income and education,

which improves ability to social distance. Intelligent people tend to be more rule abiding in general, including when the rule is dumb, which in this case is net useful. And so on.

The third is that it looks like they're using self reports of Covid-19.

Results

Out of 86,285 people who completed the extended questionnaire, 81,337 individuals fit the eligibility criteria and had complete data. These captured a broad demographic (mean age 46.75 years, 15.73 SD), including representation across sociodemographic and ethnic backgrounds ([Table 1](#)). Of these respondents, 93% indicated that their country of residence was in the UK. At the time of completing the extended survey and cognitive tests, a total of 12,689 individuals indicated that they suspected that they had experienced COVID-19, with varying degrees of respiratory severity ([Table 2](#)).

It makes sense to worry that people would conclude from their long term issues that they likely had Covid-19, or from a lack of such issues that they likely didn't have it, which could once again confound the results here.

In theory, one could have controlled for all this, by having people take the test *before* Covid-19. Even now, one could attempt to measure the impact level of the second effect by then following up with the people who took the test and seeing which of them later got Covid, although changing conditions will change the size of the correlation – so it's more of a 'check if this is a substantial effect or not' check than anything else, and now with vaccination everything is different. For the first effect, again, you'd need to find a way to measure things in the other order, possible in theory but not easy or cheap. For the third you could do antibody tests since this was pre-vaccinations.

But all of that is tricky and expensive.

The huge advantages of doing what they actually did were that it was practical, it was 'ethical' and it was relatively inexpensive. I point out issues, but I think the study was likely done roughly the right way in context, picking the low hanging fruit. From a civilizational perspective we could and should have done far better, but that's not the fault of the people doing what they can.

Looking ahead, it does seem like intelligence didn't have too big an effect on chance of getting Covid, based on a follow-up test. This could still eat up the whole observed effect, but I'm less concerned about it than before reading about that. Also a little sad that this effect turns out to be so small, for other reasons.

I also approve of their methods for analyzing the results, especially combining the nine tests into one number. There are a lot of worries I conspicuously *don't* have here.

Here's the breakdown by symptoms and gender:

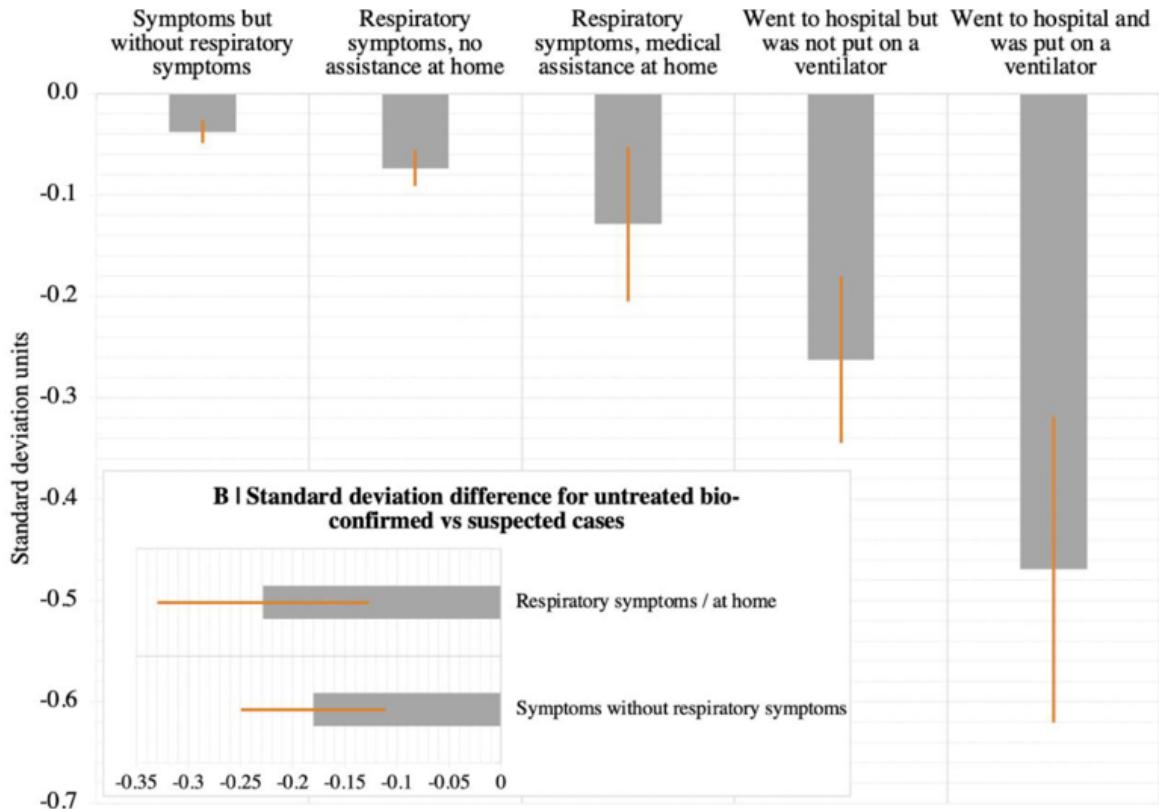
Group	Not ill	Ill without respiratory symptoms	No home assistance	Home assistance	Hospitalised No ventilator	Hospitalised +Ventilator
TOTAL COUNT	68,648	8938	3386	173	148	44
Age mean years	47.3	43.7	43.4	43.7	45.0	41.0
Age SD years	15.9	14.9	13.9	12.2	13.9	14.9
Sex						
Female	0.552	0.527	0.582	0.653	0.649	0.227
Male	0.443	0.468	0.411	0.335	0.351	0.705

I'm noting it because that righthand column is pretty strange. It's lower average age, and suddenly it's very male whereas the other groups are increasingly female and older as symptoms became more serious. Low sample size is presumably the answer (44 people) but it's still kind of weird.

Lower in the same table, it's clear that *almost none* of these people had a positive Covid-19 biological test, so this is almost all self-diagnosis:

Positive COVID-19 biological test						
No/awaiting results	1.000	0.976	0.970	0.919	0.851	0.136
Yes	0.000	0.024	0.030	0.081	0.149	0.864

Then we look at the headline chart.



One standard deviation in an IQ test is about 15 points, so this is an average of about 0.5 points or so for the first group, 1 point for those not requiring assistance, 2 for those getting help at home, 4 for those hospitalized and 7 for those who went on a ventilator.

The first thing to note is that this effect that grows as symptoms get more severe *makes a hell of a lot more sense* than the mysterious ‘Long Covid doesn’t care how bad your case was’ nonsense. I’m far more willing to believe a proportional effect that grows with symptoms than a blanket ‘hope, you technically got Covid and now you roll versus this other thing at constant odds’ hypothesis. And of course if you’ve been in the hospital on a respirator, it’s not going to be good for your cognitive performance.

So that adds a lot of credibility to the findings.

This is their explanation of why we shouldn’t worry about pre-morbid differences (e.g. how smart people were before):

A common challenge in studies of COVID-19 is that differences between people who have vs. have not been ill could relate to premorbid differences. To address this issue, a linear model was trained on the broader independent GBIT dataset ($N = 269,264$) to predict general cognitive performance based on age (to the third order), sex, handedness, ethnicity, first language, country of residence, occupational status and earnings. Predicted and observed general performance correlated substantially $r = 0.53$, providing a proxy measure of premorbid intelligence of comparable performance to common explicit tests such as the National Adult Reading Test [[26]]. Regression of the same linear model with respiratory severity as the predictor indicated that people who were ill would on average be expected to have marginally higher as opposed to lower cognitive performance (Table S6). This relationship did not vary in a simple linear manner with symptom severity. Furthermore, when a follow up questionnaire was deployed in late December 2020, 275 respondents indicated that they had subsequently been ill with COVID-19 and received a positive biological test. Their baseline global cognitive scores did not differ significantly from the 7522 respondents who had not been ill ($t = 0.7151$, $p = 0.4745$ estimate = 0.0531SDs). Taken together, these findings indicate that the cognitive impairments detected in COVID-19 survivors were unlikely to reflect pre-morbid differences.

I find this helpful but not fully convincing. One worry is that they're claiming that those that were ill should have otherwise had higher cognitive performance. I don't find this plausible, so it makes me skeptical their controls are sufficient. It does mean that I'm no longer concerned that the intelligence differences are too big here, since presumably Covid doesn't help cognitive performance and that caps the differences at the effect sizes. The second check, following up with a questionnaire to see who later got Covid, is also helpful in capping the effect size of 'smarter people get less Covid' but doesn't address the other concerns. I'm also sad we didn't ask those people to retake the tests.

I also am skeptical that this effect could fail to *partly* fade with time or as symptoms fully go away, whereas they are claiming to not see such effects.

As always, also, there's the question of whether this effect is unique to Covid or is a general property of many infectious diseases. In some sense it does not matter, but in other senses it matters a lot – or at least points to our failure to be consistent, with several potential ways to address that.

This is their note that the effect size here is indeed a big deal, which it is:

The scale of the observed deficit was not insubstantial; the 0.47 SD global composite score reduction for the hospitalized with ventilator sub-group was greater than the average 10-year decline in global performance between the ages of 20 to 70 within this dataset. It was larger than the mean deficit of 480 people who indicated they had previously suffered a stroke (-0.24SDs) and the 998 who reported learning disabilities (-0.38SDs). For comparison, in a classic intelligence test, 0.47 SDs equates to a 7-point difference in IQ.

The line that the 0.47 SD drop is more than 10 years of decline in global performance between 20 and 70 is very interesting, because it suggests an intuitive way to measure how much we should care about this effect – we can think of this as similar to *aging*.

Every year, we all age one year, and a lot of stuff gets harder. For those of us over the age of 25 or so, it's kind of terrible, and we need to put a lot more effort into making it stop.

So, in theory, this is suggesting that relatively mild (but still symptomatic) Covid is still doing something of similar magnitude to causing our brains to age two years, and as it gets more serious things get much worse. On average over all symptomatic cases we'd be looking at about three years.

If that's all true, that's really bad! Once again, you do not want to get Covid. And one can now think about how much life not lived would be justified in the name of not catching Covid, if one is at only small risk of death.

I don't find it likely they are underestimating the size of *this particular* effect, and I can see how the effect could be smaller or not be there, so that somewhat reduces the expected effect size. But then one must also account for other distinct problems.

Then there's the question of how this interacts with vaccinations and children. If you get long Covid roughly the same way you get other symptoms, that's going to be a big reduction in risk, especially for the very young. My prior would be that this protection is somewhere in between the protection against infection and protection against death.

You'd also want to adjust for Delta, since all this data was from 2020. My presumption is degree of symptoms mostly controls for that, but it's plausible that this doesn't fully control for that.

This Long Covid effect *definitely* would still be a bigger danger for most people than death, even before vaccinations. I'd much rather take the very small risk of death than accept this level of accelerated cognitive decline, plus any longer term non-cognitive effects, and it isn't close even under relatively minimizing assumptions, and multiplying for the uncertainty that this effect is real.

The question then remains, what costs are worth bearing to reduce the probability that this happens to you, individually or collectively.

Which in turn, as always, *depends on one's ability to change that number*. Postponing the problem isn't worth zero, but it's only worth enough to care much if it buys time for a permanent solution, or there's a permanent solution you're willing to implement. That's true both individually and collectively. Postponement can also help if it is 'bending the curve' in ways that matter, but that doesn't impact Long Covid much.

There's the temptation to say 'oh no, this is terrible, something must be done, this is something, therefore we must do it' where the something is an effort to kick a can down the road at very high cost, and which may or may not be able to do much of even that.

Vaccine Hesitancy

[This LA Times piece looks at those who are getting vaccinated now in a local community, and finds highly reasonable thinking going on](#). Of course, these are exactly the previously hesitant people *who then did get vaccinated*. So it makes sense that such people would seem more reasonable, and would not be at all representative of those who still haven't gotten their shots. It is consistent with the model that there are a bunch of people doing cost-benefit calculations who are relatively easy to get, and then a bunch of people who are much harder to get.

[This thread compares anti-vax people to victims of a con](#), with resulting implications for how one should communicate with them if one is attempting to persuade and convince rather than make one feel better about having taken the proper symbolic action. [Cooling the mark](#) via people in the community that have earned people's trust is so crazy an idea it just might work, but requires that such people be convinced first.

Note of course that such framing explicitly assumes the conclusion, that not only are vaccines safe and effective and the right thing for everyone both socially and individually, but that the evidence for this is so sufficient that if you don't believe it, you've been conned. I don't think that's right. Many have effectively been conned by misinformation or the need for tribal signaling, but others are doing a calculation with different information and getting a different answer. I strongly believe they're *wrong*, but it seems plausible that treating such people as con victims is (at best) highly condescending and that they would notice. We are doing a rather terrible job conveying the information about vaccines in a way that is accessible, understandable and credible to such people, and the fact that vaccines *happen to be highly safe and effective* doesn't excuse that.

[This paper provides an interesting model of vaccine hesitancy](#) (via MR). In it, people are effectively doing an approximation of a cost-benefit analysis on vaccination, so the more prevalent the disease, the higher the willingness to vaccinate, which is clear from past outbreaks.

One consequence of this is that if you reduce the number of cases, vaccinations go down. Thus, this model claims vaccine passports are 'ineffective' in the sense that they don't increase vaccinations, and could even backfire, *because they reduce prevalence*. I find their math to prove too much and therefore am confident they're technically wrong to draw their conclusions so broadly, but accept the underlying dynamics as things that are real and that matter.

Oddly, [Tyler thinks this makes a strong case against such passports](#). I would claim the opposite, because reducing prevalence is a good thing. If we can do something that both incentivizes vaccinations and prevents cases at the same time, that's good, and if the reduction in cases means we don't on net cause vaccinations, then that seems fine.

Otherwise, you are in the world where you outright want more cases in order to show the bastards and own the reds, which is a bullet I really hope no one is biting.

Regardless, under such a model, *lowering costs of vaccination* is the obvious choice for getting people vaccinated, and it continues to be a highly efficient strategy. Drive that cost negative. More precisely, one wants to reduce *perceived costs*, which can involve changing people's models of vaccine effectiveness and/or safety, and/or changing the difficulty and costs of vaccination.

Periodic Reminders (You Should Know This Already)

[In case you need a short video explanation for how the mRNA vaccines work and were developed, we got you.](#)

[Your periodic reminder that our travel restrictions make absolutely zero sense \(MR\).](#)

[And yes, a sensible system would be entirely feasible:](#)



Nate Silver @NateSilver538 · Jul 26

You could argue that it would be too much work for the US to constantly update the list of countries based on current conditions except the State Department already does *exactly that* for Americans traveling *to* those countries.

...

[This explanation rings far too true:](#)



Matthew Yglesias @mattyglesias · Jul 25

The explanation I heard is that some people in the administration want more travel restrictions and others want fewer so they've compromised on freezing in place a situation everyone agrees is outdated.

...



Bruno Maçães @MacaesBruno · Jul 25

I am really interested in why the US will not open travel to EU citizens when it's open to everyone else. Is it in order to get some economic concession? Have not heard of any other explanation
[Show this thread](#)



Martin Bratt @brattma · Jul 25

Replying to [@mattyglesias](#)

This is so bananas. I'm going from Norway to the US and will have to go and spend time in either Turkey or Mexico, both of whom have, well, a LOT more COVID than Norway has.

...

Your periodic reminder: [The FDA and cost benefit analysis are not on speaking terms](#), nor does the decision process much correspond to what is safe and effective let alone what is in

the public interest. Hence, we approve \$50,000 drugs that don't work and then are forced to collectively pay for them out of the public treasury, but can't fully approve the same Covid vaccines and definitely can't mandate them, and so on. Yeah, yeah, same old, FDA Delenda Est, stop using such procedures to tell people what they're legally allowed to do, also stop using such procedures to decide what we pay for without looking at costs against benefits, and at a bare minimum stop equating those two decisions. The suggestion from the link of letting government officials choose what is mandatory and paid for versus what is forbidden on an ad hoc per-item basis seems to miss the point of 'are those our only choices?' and I'm not sure if it's better or worse than status quo.

In other FDA Delenda Est [it's-not-news-but-it-was-news-to-him non-news](#):



William Eden @WilliamAEden · Jul 24

...

I am just learning now that the FDA is banning the sale of NAC as a dietary supplement and I am NEWLY FURIOUS about literally everything it does.

It is VERY SAFE.

They're literally banning it because it is ACTUALLY EFFECTIVE.

The implication is that supplements must be useless



30



58



383



William Eden @WilliamAEden · Jul 24

...

While we are on the subject...

I'm not a fan of requirements in general, but if there is *anything* the FDA should do, it is requiring NAC to be in every capsule of Tylenol sold.

Tylenol poisoning kills several hundred people per year and we know exactly how to prevent it!

[WaPo reminds us that Japan is failing at vaccinations by doing the things you would do if you wanted to fail at vaccinations.](#) This includes insisting on distribution by only doctors and nurses, holding out until way too late for a homegrown vaccine, a labyrinth approval process and demands for domestic testing of the vaccines, confusion about rules and a general lack of urgency, among other things.

(If for some reason you want to financially support these weekly posts and/or my writing in general, [you can do so via my bare bones, no rewards of any kind Patreon that is set up exclusively for that purpose](#). On the margin this does shift my time a non-zero amount towards these posts. However: I do not *in any way* need the money, please only provide funds fully understanding I already have plenty of money, and if and to the extent that doing this would make your own life better rather than worse.)

[The whole blood clot issue around AZ was never a thing](#). Of course Pfizer has similar instances of blood clots to AZ, given that *not getting vaccinated at all* also has similar instances. Also Covid-19 itself actually does cause blood clots, but hey.

The term 'genuine' fury is interesting here, since we knew all this already. It also does not matter, for the purposes of the EU's motivations, whether or not the concerns *turn out to be*

valid. Their perception of the situation at the time would remain unchanged.

Also happy to see this report properly label the people opposing AZ as anti-vaxxers, from anti meaning against and vax meaning vaccine:



Alex Wickham @alexwickham · 4h

New study shows AstraZeneca and Pfizer have similar instances of blood clots, and that blood clots are far far more likely in covid patients than people who've had the AZ jab. U.K. officials react with genuine fury at EU leaders who trashed AZ politi.co/3svJcDk

...

Shaming the anti-vaxxers: A [study](#) published in the Lancet this week found similar safety profiles for the AstraZeneca and Pfizer vaccines, and drove a coach and horses through concerns about the safety of the AZ jab by finding that incidences of blood clots were far higher among COVID cases than people who had the vaccine. The real-world study of more than a million people found the number of blood clots among AZ and Pfizer recipients was similar. It concluded that either way, you were far more likely to get a blood clot if you rejected a vaccine and caught COVID.

Blood on their hands: British government officials reacted with genuine fury at the actions of those who needlessly destroyed the reputation of the AstraZeneca vaccine — the jab that had the best chance of vaccinating the developing world but now suffers from low uptake. Earlier this week, [POLITICO's](#) Jillian Deutsch and Ashleigh Furlong quoted a European official who "faulted EU countries for making decisions based on 'emotion' rather than science," revealing that "scientists and politicians quietly blamed Brexit" for the row over the AZ jab. A government official told Playbook: "The European leaders who trashed the AstraZeneca vaccine have blood on their hands. We now know what we all suspected is true, that they did it out of spite for Britain because of Brexit." The official added: "When the history books are written, they'll say these people were directly responsible for the deaths of thousands in developing countries who won't take AZ because of their anti-vaxx scare stories."

To what extent is it true that the European backlash against AZ was due to spite resulting from Brexit? My guess is this was not all that central, but was a substantial contributing factor on the margin. That doesn't especially make this better or worse, it merely notes that the European Union countries were inclined to make such a self-destructive move for overdetermined reasons.

[Very Serious People do not care about physical world models, a case study \(Warning: Someone Is Wrong On the Internet\)](#):



Yaneer Bar-Yam @yaneerbaryam · 16h

...

Catastrophe in the making in Israel. They missed the opportunity to achieve elimination before Delta, when they had less than 1 case per million. Now exponential growth to over 2000 cases per day, over 200 per million and growing fast.



Yaneer Bar-Yam @yaneerbaryam · 16h

...

What is the lesson? When you achieve low levels of cases do not relax until you eliminate. Instead, push hard into the finish line. "Living with the virus" is not the right choice. The virus will mutate and you will again have exponential growth.



Yaneer Bar-Yam @yaneerbaryam · 16h

...

How many times do we have to go through the cycle before we learn.

Sir, the virus did not mutate *among the few remaining infections in Israel*. The current pandemic in Israel *is not causally related, at all*, to the few remaining cases of Alpha or the original strain that were still present in Israel.

It comes from Delta. If Israel had achieved actual zero Covid but not also instituted large new measures to keep Delta out, Delta would still have arrived from overseas, same as it arrived from overseas in every country except one. And if anything, it would have spread *faster*, because they'd have gone even more fully back to normal, so there'd be a worse problem now instead of a better one.

Feels wrong to pick on such statements, but for a while I've been feeling the need to pick out a clean example, and this fits the bill. Also points out an important dynamic - local containment only matters for the medium to long term up until you cycle strains. There's a strong instinct to contain the virus 'in case it mutates' but if it mutates elsewhere all your containment efforts mean nothing, so this only matters to the extent that you stop the mutation from happening at all, anywhere. Which is an important consideration, but not for stomping out the last few cases in one place while things continue to rage full blast in others. Much better to help out those other places.

In Other News

Scott Alexander's post [Things I Learned Writing the Lockdown Post](#) is excellent and from my perspective is much better and more interesting than [the actual lockdown post](#). I don't have the bandwidth to respond properly this week, so noting here that I haven't done so.

[There's this great highlighted comment at AstralCodexTen](#) and I have nothing to add:



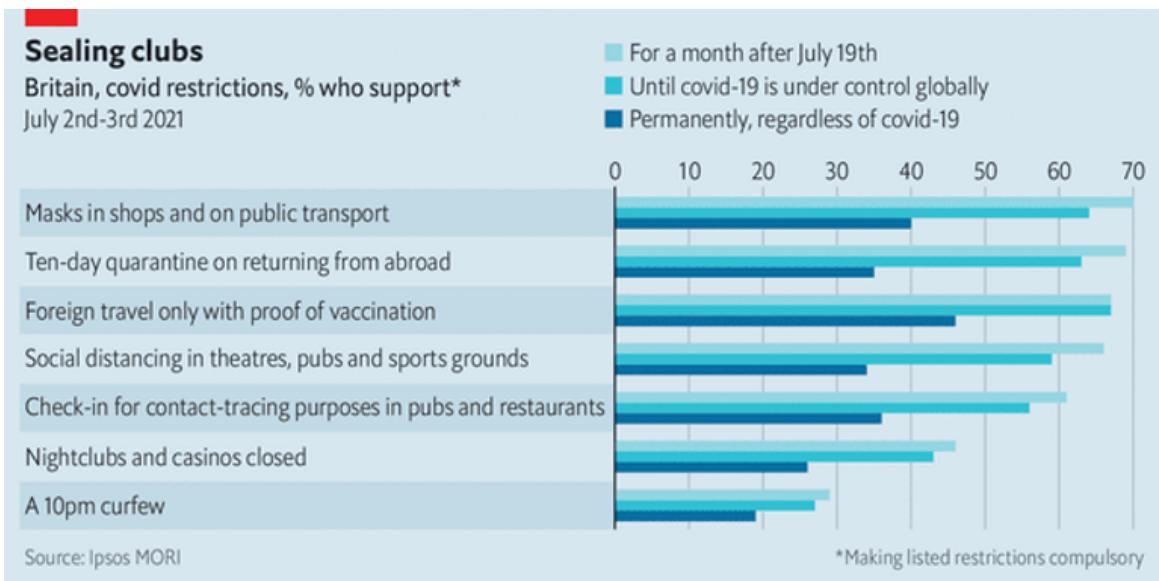
Matthew Talamini Jul 21

Wow. This seems like the start of a really relevant thought experiment. Have I not read this anywhere else? Considering both sides of it as though they were both diseases clarifies the trade-off. If I imagine that I can reduce the impact of the coronavirus by giving almost everybody this other disease, which doesn't do any harm beyond debilitating symptoms mimicking lockdown... Would I choose to counter Pandemic C by releasing Pandemic L into the population?

And the rhetorical move can be reversed: What if both were government policies? Imagine we lived in a horrifying dystopia where the police more or less randomly sprayed people with a toxic gas that harmed them the same way coronavirus does. A lot of people don't even notice the effects, but, also, a lot of people die. This is legal, and it's politically impossible to completely stop the toxic gas policy. But there's a ballot initiative to mandate lockdowns, and if it's implemented, it would decrease toxic gas incidents by exactly as much as lockdowns decrease coronavirus cases. Would I vote to counter the toxic gas policy by implementing lockdowns?

Reply

In the [monthly links post at ACX](#), Scott also points to [the poll that showed remarkable support for permanent lockdown in the UK](#):



The Economist

But then reminds us that if you ask the questions with slightly different wording, people act more sane, and [a permanent lockdown then only gets the 3% support](#) you would expect from the Lizardman's Constant:

And it turns out the results from all these changes *were* rather different. Specifically:

“

Just 3 per cent support a permanent curfew, compared to 19 per cent according to Ipsos.

Just 6 per cent support permanent closure of nightclubs, compared to 26 per cent according to Ipsos.

13 per cent supported a permanent 10-day quarantine when returning from foreign holidays, compared with 31 per cent according to Ipsos.

When it came to masks, there was slightly less of a difference between the two polls: 31 per cent said they should continue to be worn in shops and on public transport, compared with 40 per cent in the Ipsos poll who said wearing a mask in a public place should be mandatory.

They have a graph but it's hard enough to read as it is so I'm not putting it in directly.

We asked Ipsos MORI chief executive Ben Page what he made of the Prolific poll, and he told us:

“

They changed the wording - to be honest I am not surprised because of that. By stressing less/no risk rather than “regardless of the state of the pandemic” you get different results - it’s not a high salience issue and question wording will therefore make a big difference . . .

We can show you that simply adding four words to a statement about house building for example can make support go from 30% to 60%. It’s one reason why we often split sample questions to assess the impact of question wording on responses.

That makes this most of all a manufacturing consent story. Merely by changing a few words and giving people the correct word associations, you can get many people to agree to give up their freedoms and way of life, *permanently*. That’s scary. Not all of them, not quite this easily, but quite a lot of them.

You love to see it: [UK changes guidelines to emphasize outdoors and de-emphasize surfaces.](#)

Do

- ✓ get vaccinated – everyone aged 18 or over can [book COVID-19 vaccination appointments now](#)
- ✓ meet people outside if possible
- ✓ open doors and windows to let in fresh air if meeting people inside
- ✓ limit the number of people you meet and avoid crowded places
- ✓ wear a face covering when it's hard to stay away from other people – particularly indoors or in crowded places
- ✓ wash your hands with soap and water or use hand sanitiser regularly throughout the day

Don't

- ✗ do not touch your eyes, nose or mouth if your hands are not clean

Standard disclaimers that this like everything else is in no way medical advice and not to do anything illegal, [but also this happened](#) and further investigation seems logical:



Sasha Chapin @sashachapin · 9h

As suggested by [@HamiltonMorris](#), I just got finished taking a (small-ish) dose of LSD to get my sense of smell back post-covid and it totally worked. It's back online. Like a light switch, bam, online.

Wtf, psychedelics

29

123

941



...



Sasha Chapin @sashachapin · 9h

I was like at 25% ten days after recovering, certain things were there, certain things weren't, and, well, it's basically 100% now

1



52



...



Sasha Chapin @sashachapin · 9h

Details: while somewhat high, I did scent training with my fragrance collection—recalled smells I knew while smelling them with my impaired nose

Then walked around for awhile had kind of a weird taste in my mouth, acute short-term parosmia

And now we're good

3



56



...



Sasha Chapin @sashachapin · 9h

Also it really wasn't very much, less than a standard tab

I couldn't have driven a car, but I could've cooked a meal, spent most of the time cleaning my house and going shopping with my wife

1



41



...



Sasha Chapin @sashachapin · 8h

Also... I.... too early to say, but some smells seem to have improved? I am understanding iris a little better now

I am not the biggest fan of psychedelics in practice under current conditions, but there's a ton of potential upside. Our refusal to investigate their potential properly, for this and many other things, is a colossal expected value error, potentially our biggest one.

Incentives matter, [so this is mostly great](#) (via MR), but remember that if it's \$200 to get them to sing it might be considerably more to make them stop:

VACCINE RAP DAY

**DATE/LOCATION: JULY
30TH @ 4 PM ON 900
BLOCK OF MARKET
STREET**

**GET VACCINATED AT THE EVENT AND
GET 1/8TH OF CANNABIS 21 YEARS OR
OLDER WITH A FORM OF ID**

ATTENTION TO RAPPERS AND SINGERS

**CREATE AND PERFORM A SONG THAT IS ABOUT
GETTING THE VACCINE. ALL PERFORMERS WILL
GET A \$200 VISA GIFT CARD**

**SONG SUBMISSIONS AND REVIEWS MUST BE
DONE WITH ELGIN BY JULY 27TH AT 5PM**



**For more info contact Elgin at
elgin@codetenderloin.org or (510) 343-1934**

[Incentives matter, dominatrixes requiring vaccination for dungeon entry edition.](#) Seriously, incentives matter, stop pretending this is all so difficult.

The standard check for whether mask mandates are back says yes:



BNO Newsroom @BNODesk

...

Walt Disney World in Florida announces indoor mask mandate for all guests from the age of 2

Not Covid

The lighter side presents:

[Interesting whether or not they are transitive survey results of the week:](#)



Agnes Callard @AgnesCallard · Jul 25

...

So far, the number of people who would stick with their actual lives in preference to the life of a philosopher or a billionaire is about the same, about 40%.

That is also the % who would choose their actual lives over a life they would rather live than their actual lives.

↪ You Retweeted



Robin Hanson ✅ @robinhanson · 12h

...

Would you rather live one of the typical lives that leads to (and follows) becoming a billionaire, or instead live the life that you actually have had and will have?



1,008 votes · 11 hours left

💬 13

↪ 2

❤️ 2

↑



Agnes Callard @AgnesCallard · 12h

...

Which would you rather live, the life of a true philosopher—note I'm not saying "philosophy professor," I'm saying "true philosopher," which you may or may not take to overlap with "philosophy professor"—or your actual life?



556 votes · 11 hours left

💬 13

↪

❤️ 9

↑

|||



Itai Sher
@itaisher

...

Which would you rather live, a life you would rather live than your actual life or your actual life?

Life I'd rather live

60.6%

Actual life

39.4%

614 votes · 12 hours left

10:21 PM · Jul 24, 2021 · Twitter for iPhone

Taken together to see what happens, these imply that about 90% of people think billionaires have lives they'd rather live, and about 70% support for the superiority of the life of the True Philosopher even when you exclude the 21% who implicitly endorse it by already claiming to live it. Of course, Robin is an economist and Agnes is a philosopher, and have followers accordingly, but let's not let that ruin our fun.

Especially given this:

RT **Agnes Callard Retweeted**



Agnes Callard Retweeted
Robin Hanson
@robinhanson

...

Her followers differ from mine, in more preferring to be a billionaire than Nobel prize winner, while mine seem to favor the opposite. Yet she's philosopher, I'm economist. So are economists LESS greedy than philosophers?



Agnes Callard
@AgnesCallard

...

Which life would you choose of these 4?

some qualifications:

billionaire: "self-made," you inherit no wealth

philosopher: real philosopher, whatever that means to you, not necc academic philosopher

nobel prize: in field of your choosing

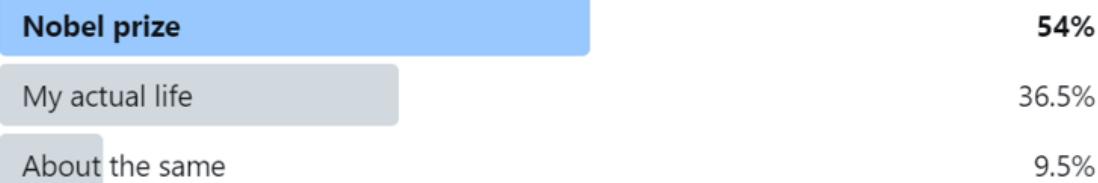
olympic victor: sport of your choosing



Robin Hanson ✅
@robinhanson

...

Would you rather live one of the typical lives that leads to (and follows) getting a Nobel prize, or instead live the life that you actually have had and will have?





Robin Hanson ✅ @robinhanson · Jul 24

...

Would you rather live one of the typical lives that leads to (and follows) becoming a billionaire, or instead live the life that you actually have had and will have?

billionaire

49%

My actual life

40.6%

About the same

10.4%



Robin Hanson ✅ @robinhanson · Jul 24

...

Would you rather live one of the typical lives that leads to (and follows) getting a Pulitzer prize, or instead live the life that you actually have had and will have?

Pulitzer prize

29%

My actual life ✅

65.6%

About the same

5.4%



Robin Hanson ✅ @robinhanson · Jul 24

...

Would you rather live one of the typical lives that leads to (and follows) getting an Academy award, or instead live the life that you actually have had and will have?

Academy award

28.4%

My actual life

65.9%

About the same

5.7%



Robin Hanson @robinhanson · Jul 24

...

Would you have rather live one of the typical lives that leads to (and follows) getting an Olympic gold medal, or live the life that you actually have had and will have instead?

Olympic life

20.8%

My actual life

71.9%

About the same

7.3%

Mostly I find this rank ordering sensible. Olympic, Pulitzer and Academy Award winners get the glory but their overall lives don't seem like they should automatically be all that tempting to most people, whereas billionaire or nobel prize winner seems a lot more tempting. I'd like to live in a place where greatness was widely considered worth its price, so I worry about people not seeing things that way.

The other thing to observe is that this implies that for the samples in question, life is pretty damn good. If it wasn't, there would presumably be a much bigger willingness to switch tracks to people who very clearly 'made it' and have a lot going on. I tend to agree. My life is pretty great, too, whether or not I count as a true philosopher.

A Contamination Theory of the Obesity Epidemic

This is a summary of [a paper](#) that I found open in a browser tab; I don't recall where I came across it. I think it's a nice paper, but it's also 63 pages long and seemed worth a synopsis for those who wouldn't otherwise tackle it.

Scott concluded in [For, then Against, High-saturated-fat diets](#) that the obesity crisis seemed to imply one of three answers:

1. That weight loss is really hard and people in previous centuries had really hard lives and that's why there was so little obesity back then.
2. That it's "being caused by plastics or antibiotics affecting the microbiome or something like that".
3. That there is hysteresis—once you become overweight it's semi-permanent.

This paper argues for the second answer, and against the other two.

At the outset, there are reasons to be wary of this paper: neither author (who share a family name) appear to have expertise in applicable fields, and it appears to be set in Computer Modern Roman, hardly the style of a journal submission. So it's coming from outside of traditional expertise. (I don't have any expertise here either.)

With that in mind, the paper starts by posing a series of challenging facts about obesity. (References in the original:)

1. It's new. One hundred years ago obesity was very rare (~1% of the population) but there were plenty of people who had enough to eat and, from our point of view, ate a lot of fattening foods.
2. It's not just new, it seemed to suddenly kick off around 1980. "Today the rate of obesity in Italy, France, and Sweden is around 20%. In 1975, there was no country in the world that had an obesity rate higher than 15%".
3. It's still getting worse. It's less in the news but if anything it's accelerating in the US. This is despite Americans significantly cutting back on sugars and carbs since 2000.
4. It's not just humans: lab animals and wild animals appear to be getting fatter over time too. (A surprise to me, but casual inspection seems to confirm that this is really a thing that reviewed papers are noting.)
5. Junk food from a supermarket fattens rats far more than giving them more of any macro-nutrient does. Somehow junk food is more than the sum of its sugars, proteins, and fats.
6. Across several countries, living at sea-level seems to increase obesity.
7. Diets produce modest reductions in weight over the span of weeks or months, but the weight comes back over time. There's been a *lot* of searching for effective diets, but they're all about the same in large populations.

The next section answers some of the competing explanations for obesity:

"It's from overeating!", they cry. But controlled overfeeding studies (from the 1970's—pre-explosion) struggle to make people gain weight and they loose it quickly once the overfeeding stops. (Which is evidence against a hysteresis theory.)

"It's lack of exercise", they yell. But making people exercise doesn't seem to produce significant weight loss, and obesity is still spreading despite lots of money and effort being put into exercise.

"It's from eating too much fat", rings out. But Americans reduced their fat intake in response to messaging about the evils of fat some decades ago and it didn't help. Nor are low-fat diets very effective.

"It's too much sugar / carbs", you hear. But Americans reduced their sugar and (more generally) carb intakes over recent years and that didn't help either. Gary Taubes's study was a bit of a damp squib.

In this section there's a hunter-gatherer tribe for everything. I'm a little suspicious of this line of evidence because these small human populations could plausibly have evolved to tolerate their specific environment but, if you want a group of humans with zero-percent obesity who eat 60%+ carbs, or 60%+ fat, this paper has one for you. They have plenty of food, they just live happily and remain thin.

Next the paper establishes that there is clearly some degree of homeostatic regulation of weight by the brain. You can damage a specific part of the brain and cause obesity. Or you can have a genetic flaw that results in fat cells not producing leptin, which results in an insatiable appetite. (But adding leptin to overweight people [doesn't work](#).)

Now the paper presents its thesis: it's all caused by a subtle poison! Manufacture of which really took off slightly before 1980, and is increasing or is bio-accumulative. Diets don't work because it's not a diet problem. Supermarket food fattens rats much more than any macro-nutrient chow because supermarket food contains more of the contamination. Wild animals are getting fatter because they're consuming it too. Living at sea-level means that your water supply has traveled much further and picked up more of it, which is why altitude is anti-correlated with obesity.

There are many drugs that cause weight gain and that appear to do so by acting on the brain, so these things can exist. This is hardly the first paper to suggest that certain chemicals contribute to the problem, but this paper is distinguishing itself by saying that it's the dominant factor.

Three specific families of chemicals are detailed for consideration: antibiotics; per-, and poly-fluoroalkyls (PFAS); and lithium. All have ambiguous evidence.

Antibiotics certainly make animals fatter, but do they do the same to humans in the amounts consumed? If so, why aren't places that use a lot in livestock fatter than those which use less? Why aren't vegan diets magic for weight loss?

PFAS is a family of thousands of under-studied chemicals, but they doesn't clearly cause weight gain in humans at plausible levels. But they are certainly getting everywhere.

Lithium clearly does cause weight gain in humans, but are the amounts that people are exposed to increasing, and they are large enough to cause weight gain?

The paper also notes that things are even worse to think about because chemicals do complex things in the environment, and in animals. You don't just have to think about the chemicals that are made, you have to worry about everything they can become. The paper includes the example of a factory in Colorado that made "war materials"

and released some chemicals into the ground around the factory. It took several years for the chemicals to travel through the ground water to farms several miles away. During that time they had reacted to form 2,4-D, a herbicide, which killed crops on those farms. (The unreacted chemicals were also pretty nasty.)

Switching to speculation that should be blamed entirely on me, not the paper: it seems that there might be a tendency for any chemical that affects the regulation of adiposity to do so in the direction of obesity. There are several drugs that target the brain and cause weight gain, but fewer safe drugs that cause significant down regulation of weight. (If there were several such drugs, lots of people would be taking them.) Rather, drugs that cause weight loss often cause energy to be wasted from the body rather than changing the regulation of weight: DNP causing heat-loss, or SGLT2 inhibitors causing glucose excretion. Thus we might be facing a situation where multiple minor factors affect adipose regulation, but the overall effect is towards obesity because any effect tends to be in that direction.

The obvious example of something that causes down-regulation of weight is smoking. (We wouldn't call it "safe" though.) I wonder whether the paper is overly focused on something that had a step change in prevalence shortly before 1980. It might have been building steadily in the prior decades but the [40%ish](#) of American adults who smoked in the 60s/70s hid it for a while.

If we were to hypothesise that some environmental factor is causing a significant fraction of the obesity problem then how would we test it? It could well be the sum of multiple factors, some of which may be carried in water given the correlation with elevation. It seems that one would need groups of overweight people willing to consume exclusively provided water (from distillation) and a source of food that is somehow pristine. The half-life of PFAS, at least, is measured in years in humans, so the subjects would have to remain compliant with this proscribed diet for extended periods of time. In order to have control groups we would have to (double-blind) contaminate the pristine food with environmentally-plausible levels of candidate chemicals, I guess? Would that get past any IRB? The paper contains easier experiments, like having Kuwait change its desalination process or, more reasonably, have a car-mechanic company change grease, but these would only produce partial answers unless we got lucky and one factor dominates.

The environmental hypothesis is primarily one of exclusion, and this paper makes a good case. (Although one should have significant epistemological humility about any complex technical argument outside of one's expertise.) And I haven't even covered it all! The paper continues with arguments about paradoxical reactions and the occurrence of anorexia! There is much more within if this summary piqued an interest.

We have some evidence that masks work

by Gavin Leech and [Charlie Rogers-Smith](#)

[Our work](#) on masks vs COVID at the population level was [recently reproduced](#) with a bunch of additional experiments. These seem to cast doubt on our results, but we think that each of them is misguided. Since the post got some traction on LW and Marginal Revolution, we decided to respond.

Nevertheless, thanks to Mike, who put a lot of work in, and who was the only person in the world to check our results, despite plenty of people trying to gotcha us on Twitter.

“Observational Window”

Best-guess summary of Mike’s analysis: he extends the window of analysis by a bit and runs our model. He does this because he’s concerned that we chose a window with low transmissibility to make masks look more effective than they are. However, he finds similar results to the original paper, and concludes that our results seem robust to longer periods.

But as our paper notes, a longer window isn’t valid using this data. After September, many countries [move to subnational NPIs](#), and our analysis is national. The way [our NPI data source](#) codes things means that they don’t capture this properly, and so they stop being suitable for national analyses.

Estimates of national mask effect after this don’t properly adjust for crucial factors, and so masks will “steal” statistical power from them. So this analysis isn’t good evidence about the robustness of our results to a longer window.

“Regional Effects”

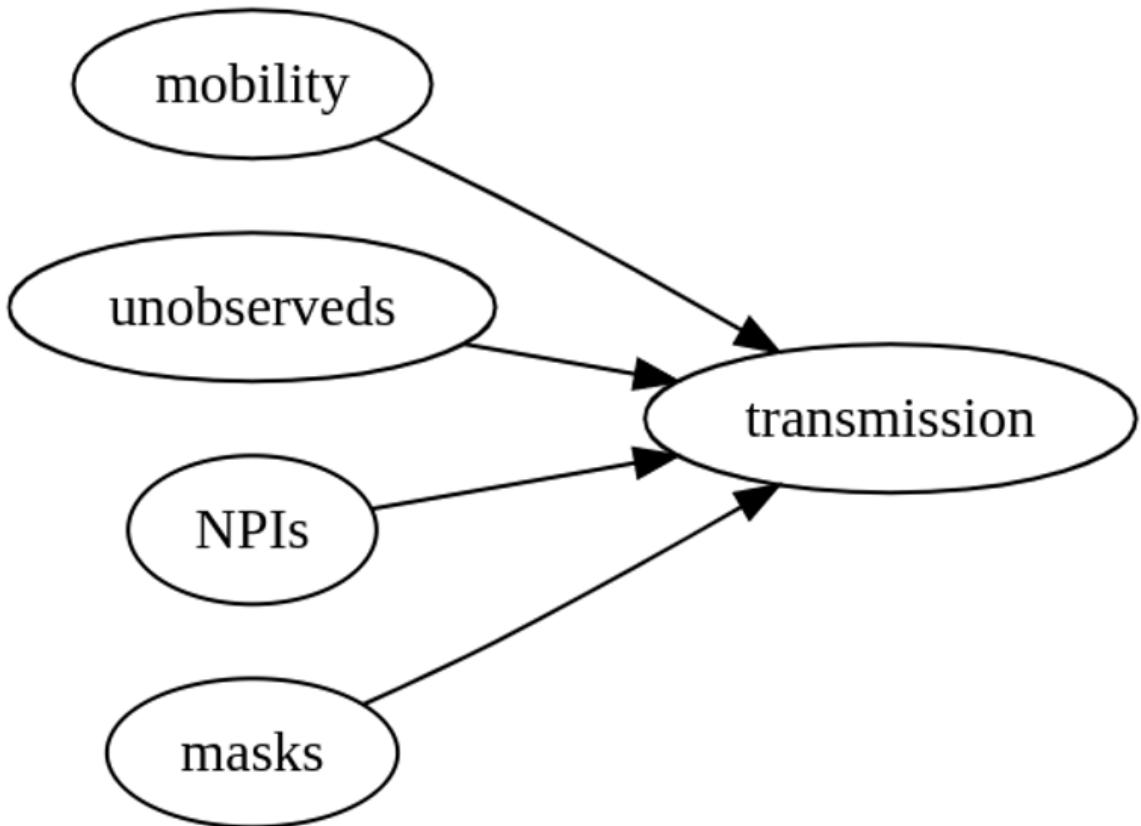
MH: *If mask wearing causes a drop in transmissibility, then regions with higher levels of mask wearing should observe lower growth rates.*

Best-guess summary of Mike’s analysis: A correlational analysis between the median wearing level of a region and the R_0 (the expected number of new cases per initial case in a region) that our model infers. (What he calls ‘growth rates’, but which are not growth rates.) He claims that if wearing is effective then the correlation should be negative. The intuition is that if masks work, then countries with lots of mask-wearing should have lower transmissibility. Instead, he finds that the correlation is positive.

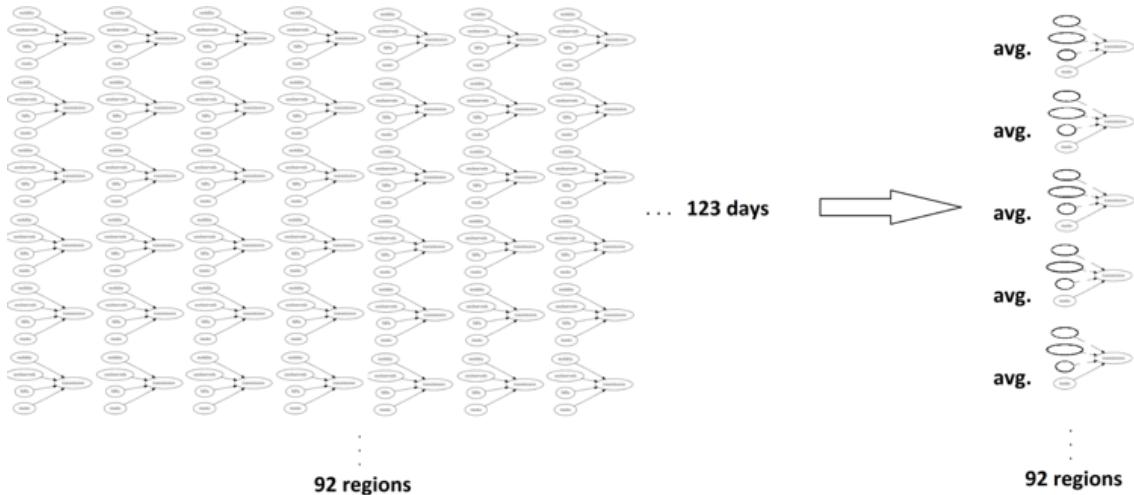
This is interesting, but the conclusion doesn’t seem right. You can tell a bunch of stories about why mask-wearing might be correlated with R_0 , independent of mask effectiveness.

For example, it seems plausible that when transmissibility increases, more people wear masks. Overall, the correlation between average mask-wearing and constant transmissibility is such weak evidence that we honestly don’t know in which *direction* to update our beliefs based on this result.

It’s also worth highlighting how much information this analysis discards. He takes a scalar (the regional R_0) and then plots it against a scalar (the median wearing level). But this averages away almost all of the information.



one day of our data, in one region



Mike's analysis removes all time info

It's hard to say *anything* about the relationship between R_t and wearing as a static average. You have to look at changes in mask-wearing.

MH: "Within a given region, increased mask usage is correlated with lower growth rates (the 25% claimed effectiveness), but when comparing across regions masks seem to be ineffective."

Even given fixes to the above, this doesn't follow. Our posterior with a median of 25% is a pooled estimate across regions.

Endogeneity (the estimate being biased by, for instance, people masking up in response to an outbreak) is a real concern, but the above doesn't show this either. We can see how serious endogeneity could be by looking at the correlation between mask level and case level: $p = 0.05$.

“Uniform Regional Transmissibility”

MH: "*The first experiment was to force all regions to share the same base transmissibility. This provided an estimate that masks had an effectiveness of -10%*"

Best-guess summary of Mike's analysis: Mike sets all R_0 s to the same value and runs the model. He does this to isolate the 'relative' effect of mask-wearing -- i.e. the effect from day-to-day changes in wearing, as opposed to absolute mask-wearing.

I think the intuition comes from the fact that we use two sources of info to determine mask-wearing effectiveness: the starting level of mask-wearing in a region, and day-to-day changes. It would be cool to see what we infer from only day-to-day changes in wearing. But this method doesn't achieve this; instead, setting all the R_0 s to be the same will bias the wearing effect

To see this, suppose we have data on two regions. Region A has an $R_0 = 0.5$ and B has an $R_0 = 1.5$, but we don't know these values. Further assume that region B has more mask-wearing, which is consistent with Mike's finding that there's a small positive correlation between R_0 and mask-wearing. What happens if we force these values to be the same, say $R_0 = 1$? Well, the model will use the mask-wearing effect to shrink A's 1.0 down to 0.5, and to pull B's 1.0 up to 1.5. And since B has more mask-wearing relative to A, this is only possible when the mask-wearing effect is negative. So fixing region R_0 s to be the same creates a strong negative bias on the mask effect estimate.

Can we do better? We think so! As we mentioned to Mike in correspondence, a better way to isolate the effect from day-to-day changes in wearing would be to zero-out wearing at the start of the period, so that no information from wearing levels can inform our estimate of R_0 . We tried this analysis and got a 40% reduction in R from mask-wearing, with large uncertainty (because we're removing an information source).

“No Mask Variation”

MH: "*The next experiment was to force each region to use a constant value for mask wearing (the average value in the time period).*"

Best-guess summary of Mike's analysis: let's isolate the absolute effect now! To do this, set mask-wearing to be constant across the period.

Most of what our model uses to estimate the effect is day-to-day changes in wearing and transmission. If mask-wearing is set constant, our model will still be 'learning' about the mask-wearing effect from day-to-day changes in transmissibility, even if mask-wearing

doesn't change--it will just be learning from false data. Setting mask-wearing constant is not [inferring nothing from day-to-day changes in transmission], it's inferring false things.

“Data Extrapolation”

MH: "*the failure of large absolute differences in variable X across regions to meaningfully impact the observed growth rate ... should make us skeptical of large claimed effects*"

Best-guess summary of Mike's analysis: Let's compare changes in wearing from April to May to changes in growth rates. If masks work then we should find a strong negative correlation.

The method:

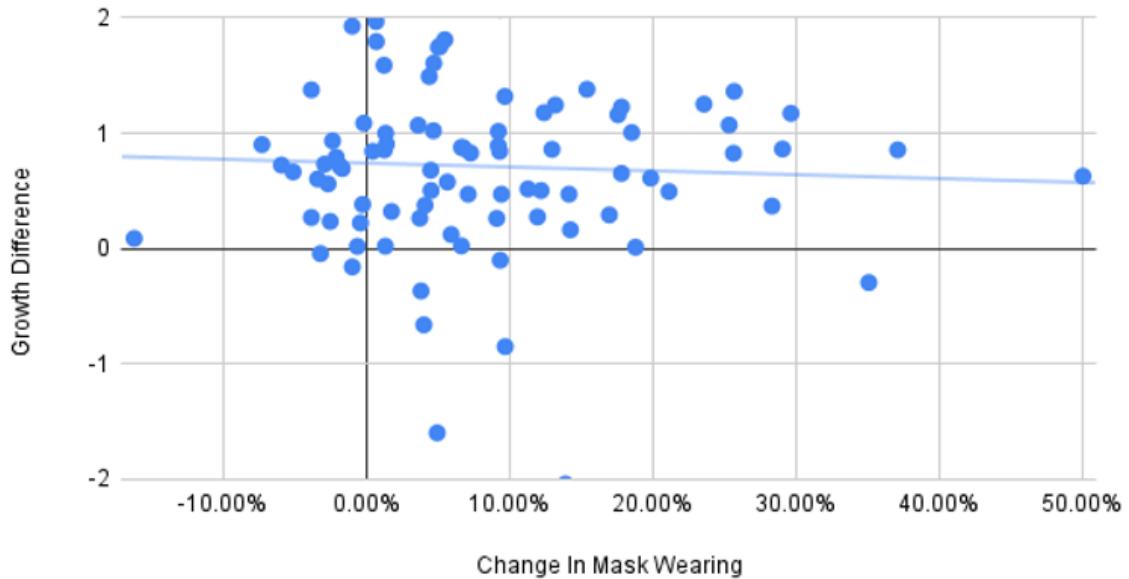
1. For April and May, take the average wearing
2. May average mask wearing - April average mask wearing (x-axis).
3. AprilCaseRatio = Cases @ end April / cases @ start April
4. MayCaseRatio = Cases @ end May / cases @ start May
5. "Growth rate" = AprilCaseRatio / MayCaseRatio (y-axis).
6. Scatterplot each region

(This throws away even more useful information -- picking out two days in April and May throws away 96% of the dataset.)

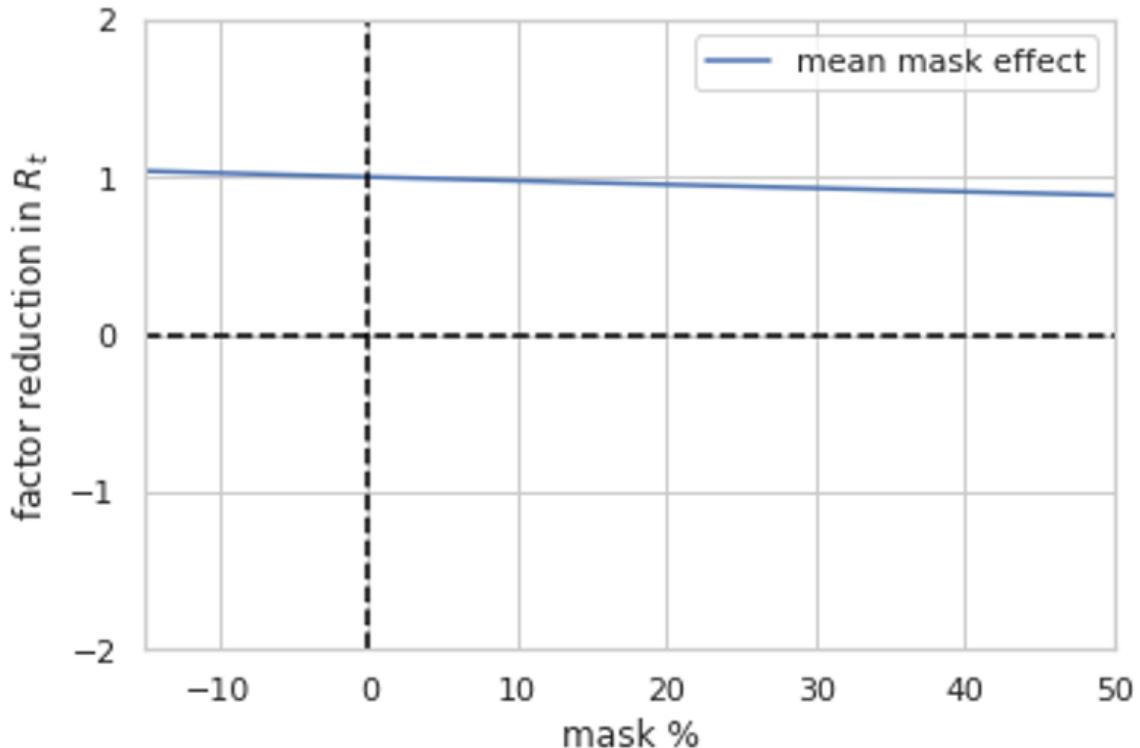
But this analysis doesn't account for any of the known factors on transmission. The crux here is whether we'd expect to 'see' the effect of wearing amidst all the variation in other factors. Averaging over 90 or so regions could smooth out random factors. However, if factors are persistent across regions -- factors that, for example, result in increased transmission across most countries -- then this method will not uncover the wearing effect. And in fact there are strong international trends in factors in May 2020. We wouldn't update much unless the correlation was particularly strong.

However, let's assume we could find the wearing effect using this method. Mike implies this slope should be more strongly negative than he observes. How negative should it be, under our model (25% reduction in R due to masks)? Let's look at Mike's plot and compare it to the mean of our posterior (which you shouldn't do but anyway):

Mask Wearing vs Change in Growth



Mike's simple linear regression. ~20% decrease



The reduction inferred by our model over this range. ~12.5% decrease

The slope he finds is similar to our model estimate - in fact it's more negative. (Looks like a $1 - 0.55/0.7 \approx 20\%$ decrease in Growth Difference for +50% wearing.) In our plot, y is the

reduction in R, not Mike's ratio of case ratios, but it gives you an idea of what a point estimate of our claim looks like on a unit scale, on the same figure grid.

It's difficult to see medium-sized effects with methods this simple, and this is why we use a semi-mechanistic model. Correlation analysis between cases (or growth) vs wearing neglects important factors we need to adjust for (e.g. mobility at 40% effect!). Moreover, doing it in the way described neglects a lot of data.

(Charlie wants to indicate that he isn't confident in the following paragraph -- not because he disagrees, but because he hasn't been following the broader literature.)

Even if the experiments above showed what they purport to, Mike's title, 'We Still Don't Know If Masks Work', would still be misleading. By this point we have convergent lines of evidence: [meta-analyses of clinical trials](#), slightly nightmarish [animal models](#), [mechanistic models](#), an [ok cloth mask study](#). We should get [the big RCT](#) results soon: I (Gavin) am happy to bet Mike \$100 that this will find a median reduction in R greater than 15% (for the 0-100% effect).

Our paper is observational, and there are limits to how strong such evidence can be. Mike says "they are not sufficient to prove a causal story"; this much we agree on.

Lastly: we exchanged 13 emails with Mike, helping him get the model converging and explaining most of these errors (though not as extensively). It was disappointing to find none of them corrected, no mention of our emails, no code, and no note that he had posted his work.

My Marriage Vows

I'm getting married. We decided to take marriage vows very seriously, and write vows that we will be fully committed to uphold. These vows are going to be a commitment no weaker than any promise I ever made or any contract I ever signed. Therefore, it is very important to avoid serious errors in their content.

I'm interested to hear feedback of the form "making these vows might turn out to be a big mistake for you, and here is why"^[1] or of the form "here is how the spirit of these vows can be implemented better". Given that this is a community which nurtures security mindset, I have great expectations :) More precisely, I am less interested in extreme nitpicking / rule-lawyering, since that should be neutralized by the Vow of Good Faith anyway (but tell me if you think I'm wrong about this!) and more in serious problems that can arise in at least semi-realistic situations. (Of course, since many of us here expect a Singularity in a few decades, semi-realistic is not a very high bar ;)

Without further ado, the vows:

[EDIT 2022-07-15: The following text has been edited to match the final version of the Vows (that we took on 2021-08-29)]

I, [name], solemnly pledge to [name] three sacred Vows as I take [pronoun] to be my [spouse]. These vows are completely sincere, literal, binding and irrevocable from the moment both of us take the Vows, unless the marriage is dissolved or my [spouse] unconscionably^[2] breaks [pronoun]'s own Vows which I believe in all likelihood will never happen. Let everyone present be my witness.

The First Vow is that of Honesty. I will never set out to deceive my [spouse] on purpose without [pronoun]'s unambiguous consent^[3], without exception. I will also never withhold information that [pronoun] would in hindsight prefer to know^[4]. The only exception to the latter is when this information was given to me in confidence by a third party as part of an agreement which was made in compliance with all Vows^[5]. If for any reason I break my vow, I will act to repair the error as fast as reasonably possible.

The Second Vow is that of Concord. Everything I do will be according to the policy which is the Nash solution to the bargaining problem defined by my [spouse]'s and my own priors and utility functions, with the disagreement point set at the counterfactual in which we did not marry. I will act as if we made all precommitments that would a priori be beneficial from a Nash bargaining point of view^[6]. If our utility functions change, we will effectively perform another Nash bargaining with the previous policy as the disagreement point. Moreover, if I deviate from this policy for any reason then I will return to optimal behavior as soon as possible, while preserving my [spouse]'s a priori expected utility if at all possible^[7]. Finally, a hypothetical act of dissolving this marriage would also fall under the purview of this Vow^[8].

The Third Vow is that of Good Faith, which augments and clarifies all three Vows. The spirit of the Vows takes precedence over the letter. When there's some doubt or dispute as to how to interpret the Vows, the chosen interpretation should be that which my [spouse] and I would agree on at the time of our wedding, in the counterfactual in which the source of said doubt or dispute would be revealed to

us and understood by us with all of its implications at that time as well as we understand it at the time it actually surfaced^[9].

1. Conditional on the assumption that my decision to marry is about as well-grounded as one can expect. I am *not* soliciting criticism of my choice of spouse! ↵
2. Meaning that it's a grave or persistent violation rather than a minor lapse. ↵
3. Consent is mentioned to allow us to e.g. play tabletop games where you're supposed to deceive each other. ↵
4. That is, information X such that if the spouse knew X, they would believe it's good that they found out about it. This excludes information which is not important (knowing X is practically useless) and infohazards (knowing X is actively harmful). ↵
5. If I enter an agreement with a third party in violation of the Vow of Concord, the Vow of Honesty takes precedence over the agreement and I might have to violate the latter and pay whatever fine is necessary. ↵
6. We are taking an "updateless" perspective here. The disagreement point is fixed in the counterfactual in which we didn't marry in the first place, it does *not* move to the counterfactual of divorce. Notice also that marriage is guaranteed to be an *a priori* Pareto improvement over no-marriage because *this is our current estimate*, even if it turns out to be false *a posteriori*. ↵
7. If the violation shifts the Pareto frontier such that the previous optimum is outside of it, the new Pareto optimum is chosen s.t. the violating party bears the cost. ↵
8. This makes all of the Vows weightier than they otherwise would be. The Vows can be unmade by dissolving the marriage, but the act of dissolving the marriage is in itself subject to the Vow of Concord, which limits the ability to dissolve it unilaterally. ↵
9. In other words, interpretation is according to the extrapolated volition of us at the time of our wedding, where the extrapolation is towards our knowledge and intellectual ability at the time of making the judgment. ↵

One Study, Many Results (Matt Clancy)

This is a linkpost for <https://mattsclancy.substack.com/p/one-study-many-results>

I didn't see this post, its author, or the study involved elsewhere on LW, so I'm crossposting the content. Let me know if this is redundant, and I'll take it down.

Summary

This post looks at cases where teams of researchers all began with the same data, then used it to answer a question — and got a bunch of different answers, based on their different approaches to statistical testing, "judgment calls", etc.

This shows the difficulty of doing good replication work even *without publication bias*; none of the teams here had any special incentive to come up with a certain result, and they all seemed to be doing their best to really answer the question.

Also, I'll copy the conclusion of the post and put it here:

More broadly, I take away three things from this literature:

1. Failures to replicate are to be expected, given the state of our methodological technology, even in the best circumstances, even if there's no publication bias.
2. Form your ideas based on suites of papers, or entire literatures, not primarily on individual studies.
3. There is plenty of randomness in the research process for publication bias to exploit. More on that in the future.

The post

Science is commonly understood as being a lot more certain than it is. In popular science books and articles, an extremely common approach is to pair a deep dive into one study with an illustrative anecdote. The implication is that's enough: the study discovered something deep, and the anecdote made the discovery accessible. Or take the coverage of science in the popular press (and even the academic press): most coverage of science revolves around highlighting the results of a single new (cool) study. Again, the implication is that one study is enough to know something new. This isn't universal, and I think coverage has become more cautious and nuanced in some outlets during the era of covid-19, but it's common enough that for many people "believe science" is a sincere mantra, as if science made pronouncements in the same way religions do.

But that's not the way it works. Single studies - especially in the social sciences - are not certain. In the 2010s, it has become clear that a lot of studies (maybe the majority) do not replicate. The failure of studies to replicate is often blamed ([not without evidence](#)) on a bias towards publishing new and exciting results. Consciously or subconsciously, that leads scientists to employ shaky methods that get them the results they want, but which don't deliver reliable results.

But perhaps it's worse than that. Suppose you could erase publication bias and just let scientists choose whatever method they thought was the best way to answer a question. Freed from the need to find a cool new result, scientists would pick the best method to answer a question and then, well, answer it.

The many-analysts literature shows us that's not the case though. The truth is, the state of our "methodological technology" just isn't there yet. There remains a core of unresolvable uncertainty and randomness in the best of circumstances. Science isn't certain.

Crowdsourcing Science

In many-analyst studies, multiple teams of researchers test the same previously specified hypothesis, using the exact same dataset. In all the cases we're going to talk about today, publication is not contingent on results, so we don't have scientists cherry-picking the results that make their results look most interesting; nor do we have replicators cherry-picking results to overturn prior results. Instead, we just have researchers applying judgment to data in the hopes of answering a question. Even still results can be all over the map.

Let's start with a really recent paper in economics: [Huntington-Klein et al. \(2021\)](#). In this paper, seven different teams of researchers tackle two research questions that had been previously published in top economics journals (but which were not so well known that the replicators knew about them). In each case, the papers were based on publicly accessible data, and part of the point of the exercise was to see how different decisions about building a dataset from the same public sources lead to different outcomes. In the first case, researchers used variation across US states in compulsory schooling laws to assess the impact of compulsory schooling on teenage pregnancy rates.

Researchers were given a dataset of schooling laws across states and times, but to assess the impact of these laws on teen pregnancy, they had to construct a dataset on individuals from publicly available IPUMS data. In building the data, researchers diverged in how they handled different judgement calls. For examples:

One team dropped data on women living in group homes; others kept them.

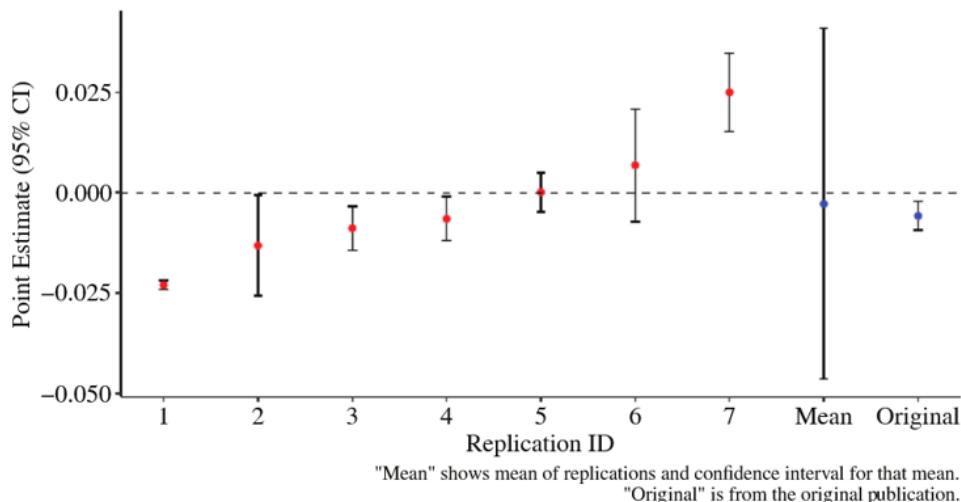
Some teams counted teenage pregnancy as pregnancy after the age of 14, but one counted pregnancy at the age of 13 as well

One team dropped data on women who never had any children

In Ohio, schooling was compulsory until the age of 18 in every year except 1944, when the compulsory schooling age was 8. Was this a genuine policy change? Or a typo? One team dropped this observation, but the others retained it.

Between this and other judgement calls, no team assembled exactly the same dataset. Next, the teams needed to decide how, exactly, to perform the test. Again, each team differed a bit in terms of what variables it chose to control for and which it didn't. Race? Age? Birth year? Pregnancy year?

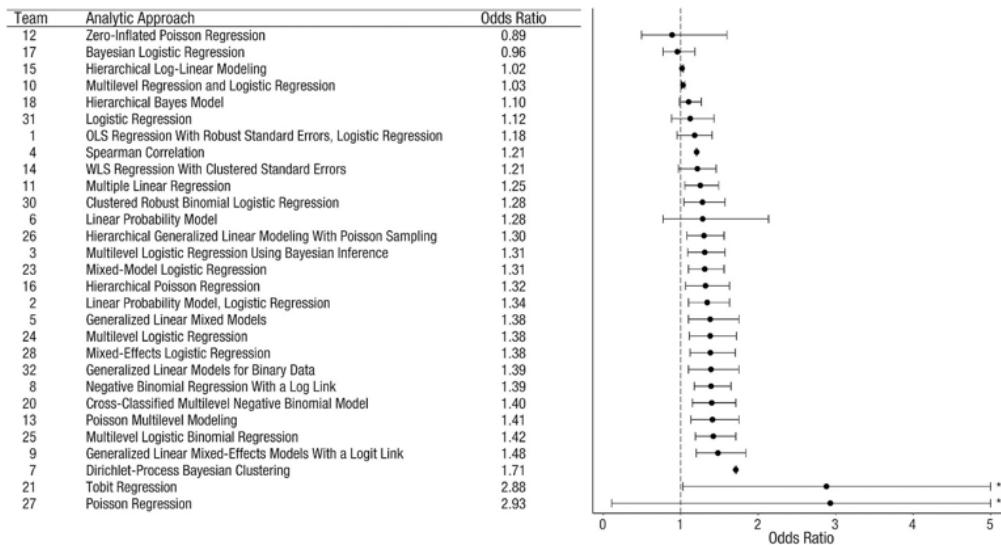
It's not immediately obvious which decisions are the right ones. Unfortunately, they matter a lot! Here were the seven teams' different results.



Depending on your dataset construction choices and exact specification, you can find either that compulsory schooling lowers or increases teenage pregnancy, or has no impact at all! (There was a second study as well - we will come back to that at the end)

This isn't the first paper to take this approach. An early paper in this vein is [Silberzahn et al. \(2018\)](#). In this paper, 29 research teams composed of 61 analysts sought to answer the question "are soccer players with dark skin tone more likely to receive red cards from referees?" This time, teams were given the same data but still had to make decisions about what to include and exclude from analysis. The data consisted of information on all 1,586 soccer players who played in the first male divisions of England, Germany, France and Spain in the 2012-2013 season, and for whom a photograph was available (to code skin tone). There was also data on player interactions with all referees throughout their professional careers, including how many of these interactions ended in a red card and a bunch of additional variables.

As in Huntington-Klein et al. (2021), the teams adopted a host of different statistical techniques, data cleaning methods, and exact specifications. While everyone included “number of games” as one variable, just one other variable was included in more than half of the teams regression models. Unlike Huntington-Klein et al. (2021), in this study, there was also a much larger set of different statistical estimation techniques. The resulting estimates (with 95% confidence intervals) are below.



Is this good news or bad news? On the one hand, most of the estimates lie between 1 and 1.5. On the other hand, about a third of the teams cannot rule out zero impact of skin tone on red cards; the other two thirds find a positive effect that is statistically significant at standard levels. In other words, if we picked two of these teams’ results at random and called one the “first result” and the other a “replication,” they would only agree whether the result is statistically significant or not about 55% of the time!

Let’s look at another. [Breznau et al. \(2021\)](#) get 73 teams, comprising 162 researchers to answer the question “does immigration lower public support for social policies?” Again, each team was given the same data. This time, that consisted of responses to surveys about support for government social policies (example: “On the whole, do you think it should or should not be the government’s responsibility to provide a job for everyone who wants one?”), measures of immigration (at the country level), and various country-level explanatory variables such as GDP per capita and the Gini coefficient. The results spanned the spectrum of possible conclusions.

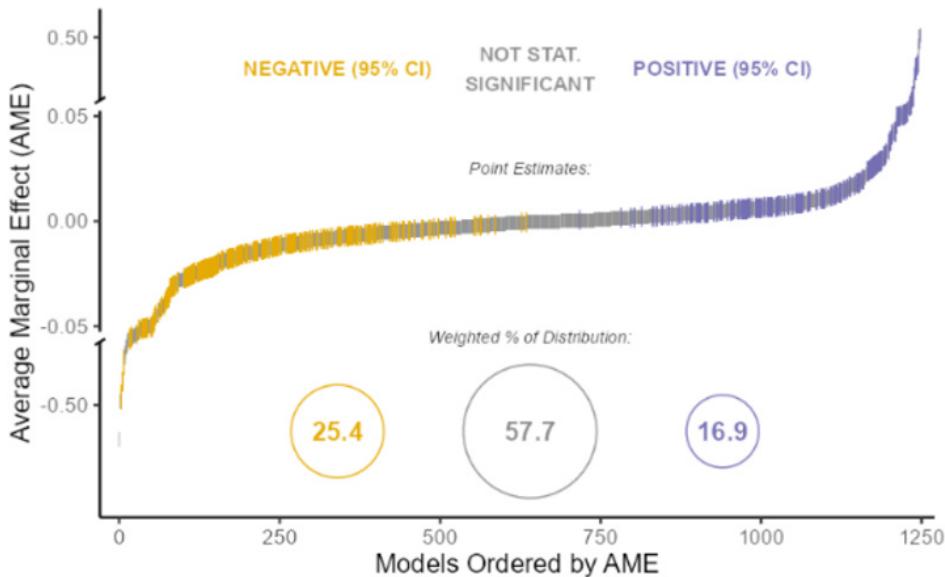


Fig. 1 Broad variation in findings from 73 teams testing the same hypothesis with the same data

Slightly more than half of the results found no statistically significant link between immigration levels and support for policies - but a quarter found more immigration reduced support, and more than a sixth found more immigration increased support. If you picked two results at random, they would agree on the direction and statistical significance of the results less than half the time!

We could do [more studies](#), but the general consensus is the same: when many teams answer the same question, beginning with the same dataset, it is quite common to find a wide spread of conclusions (even when you remove motivations related to beating publication bias).

At this point, it's tempting to hope the different results stem from differing levels of expertise, or differing quality of analysis. "OK," we might say, "different scientists will reach different conclusions, but maybe that's because some scientists are bad at research. Good scientists will agree." But as best as these papers can tell, that's not a very big factor.

The study on soccer players tried to answer this in a few ways. First, the teams were split into two groups based on various measures of expertise (teaching classes on statistics, publishing on methodology, etc). The half with greater expertise was more likely to find a positive and statistically significant effect (78% of teams, instead of 68%), but the variability of their estimates was the same across the groups (just shifted in one direction or another). Second, the teams graded each other on the quality of their analysis plans (without seeing the results). But in this case, the quality of the analysis plan was unrelated to the outcome. This was the case even when they only looked at the grades given by experts in the statistical technique being used.

The last study also split its research teams into groups based on methodological expertise or topical expertise. In neither case did it have much of an impact on the kind of results discovered.

So; don't assume the results of a given study are definitive to the question. It's quite likely that a different set of researchers, tackling the exact same question and starting with the exact same data would have obtained a different result. Even if they had the same level of expertise!

Resist Science Nihilism!

But while most people probably overrate the degree of certainty in science, there also seems to be a sizable online contingent that has embraced the opposite conclusion. They know about the replication crisis and the unreliability of research, and have concluded the whole scientific operation is a scam. This goes too far in the opposite direction.

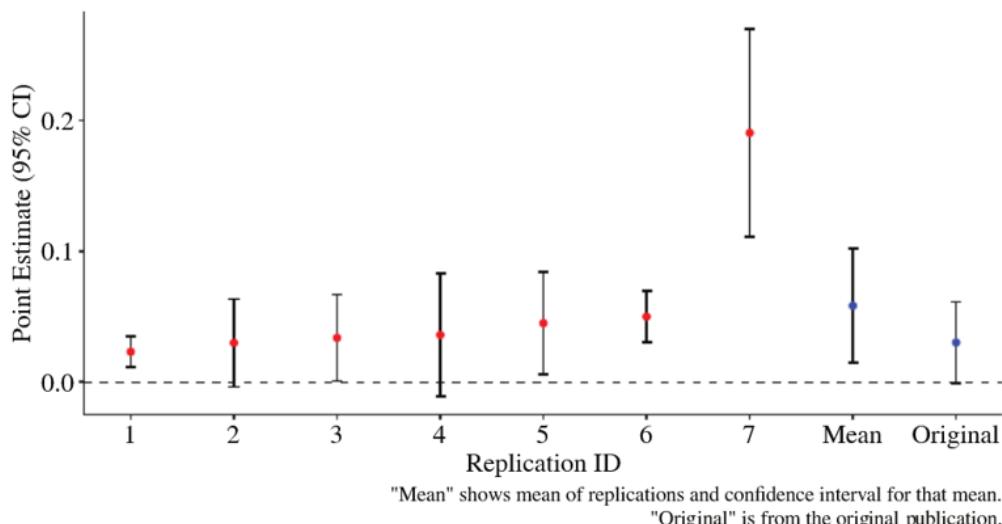
For example, a science nihilist might conclude that if expertise doesn't drive the results above, then it must be that scientists simply find whatever they want to find, and that their results are designed to fabricate evidence for whatever they happen to believe already. But that doesn't seem to be the case, at least in these multi-analyst studies. In both the study of soccer players and the one on immigration, participating researchers reported their beliefs before doing their analysis. In both cases there wasn't a statistically significant correlation between prior beliefs and reported results.

If it's not expertise and it's not preconceived beliefs that drive results, what is it? I think it really is simply that research is hard and different defensible decisions can lead to different outcomes. Huntington-Klein et al. (2021) perform an interesting exercise where they apply the same analysis to different teams data, or alternatively, apply different analysis plans to the same dataset. That exercise suggests roughly half of the divergence in the teams conclusions stems from different decisions made in the database construction stage and half from different decisions made about analysis. There's no silver bullet - just a lot of little decisions that add up.

More importantly, while it's true that any scientific study should not be viewed as the last word on anything, studies still do give us signals about what might be true. And the signals add up.

Looking at the above results, while I am not certain of anything, I come away thinking it's slightly more likely that compulsory schooling reduces teenage pregnancy, pretty likely that dark skinned soccer players get more red cards, and that there is no simple meaningful relationship between immigration and views on government social policy. Given that most of the decisions are defensible, I go with the results that show up more often than not.

And sometimes, the results are pretty compelling. Earlier, I mentioned that Huntington-Klein et al. (2021) actually investigated two hypotheses. In the second, Huntington-Klein et al. (2021) ask researchers to look at the effect of employer-provided healthcare on entrepreneurship. The key identifying assumption is that in the US, people become eligible for publicly provided health insurance (Medicare) at age 65. But people's personalities and opportunities tend to change more slowly and idiosyncratically - they also don't suddenly change on your 65th birthday. So the study looks at how rates of entrepreneurship compare between groups just older than the 65 threshold and those just under it. Again, researchers have to build a dataset from publicly available data. Again every team made different decisions, such that none of the data sets are exactly alike. Again, researchers must decide exactly how to test the hypothesis, and again they choose slight variations in how to test it. But this time, at least the estimated effects line up reasonably well.

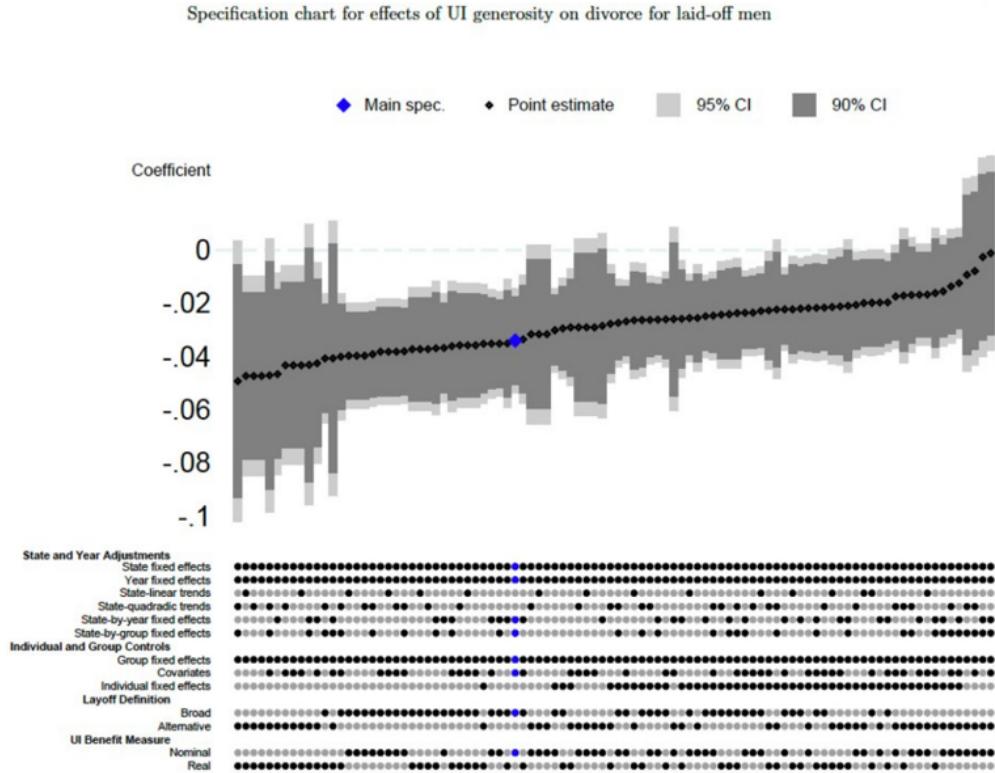


I think this is pretty compelling evidence that there's something really going on here - at least for the time and place under study.

And it isn't necessary to have teams of researchers generate the above kinds of figures. "Multiverse analysis" asks researchers to explicitly consider how their results change under all plausible changes to the data and analysis; essentially, it asks individual teams to try and behave like a set of teams. In economics (and I'm sure in many other fields - I'm just writing about what I know here), something like this is supposedly done in the "robustness checks" section of a paper. In this part of a study, the researchers show how their results are or are not robust to alternative data and analysis decisions. The trouble has long been

that robustness checks have been selective rather than systematic; the fear is that researchers highlight only the robustness checks that make their core conclusion look good and bury the rest.

But I wonder if this is changing. The robustness checks section of economics papers has been steadily ballooning over time, contributing to the novella-like length of many modern economics papers (the average length [rose](#) from 15 pages to 45 pages between 1970 and 2012). [Some](#) papers are now beginning to include figures like the following, which show how the core results change when assumptions change and which closely mirror the results generated by multiple-analyst papers. Notably, this figure includes many sets of assumptions that show results that are not statistically different from zero (the authors aren't hiding everything).



Economists [complain](#) about how difficult these requirements make the publication process (and how unpleasant they make it to read papers), but the multiple-analyst work suggests it's probably still a good idea, at least until our "methodological technology" catches up so that you don't have a big spread of results when you make different defensible decisions.

More broadly, I take away three things from this literature:

1. Failures to replicate are to be expected, given the state of our methodological technology, even in the best circumstances, even if there's no publication bias.
2. Form your ideas based on suites of papers, or entire literatures, not primarily on individual studies.
3. There is plenty of randomness in the research process for publication bias to exploit. More on that in the future.

Relentlessness

There's not a great way to convey the merciless relentlessness of having a child who insists on continuing to exist and want and need regardless of how much sleep you got, how sick you are, how many times you have already read that book, how tired your arms or how aching your feet, how hungry or sweaty or needy-for-cognition you've gotten, how much or little support happens to be available that day. It's a *lot*.

...

Parents aren't better at parenting tasks because of magic or even because they responsibly read the entire parenting manual, they're better at them because they are forced to practice way way way more than anyone would naturally choose to practice any such tasks (and accordingly your skills will be uneven depending on how you divide those tasks).

(from [a tumblr post](#) I wrote about having kids)

Why is immersion the best way to learn a language?

I submit that it is because you do not get to stop.

If you were in, say, Java, then you probably would pick up Javanese as long as you did things reasonably aimed at continuing to be immersed in Javanese (that is, not immediately finding the nearest English-speaker and latching onto them, or adopting a convenient orphan and teaching them English, or buying a ticket to Australia). In spite of the fact that this strategy does not necessarily draw on anything we know about deliberate practice, or language education, it would still probably work. It's how everybody learns their first language, and in that case it basically *always* works, because babies *really* can't do anything else about it since they don't already speak anything else. To communicate *at all*, a very basic human need, you have to match the standard of other talking entities around you, and you will not stop having to do this, so eventually you will.

Most things are not like this, or they are like this but not enough for it to be a good idea to try to learn them this way. If you are in the ocean, and you cannot stop being in the ocean, this is actually a terrible way to learn to swim and an even worse way to learn about sharks. If you are a subsistence farmer, and you cannot stop being a subsistence farmer, you might learn a ton about your plot of land, but also you might have bad weather one year and starve. If you are in a typical American math class, and you cannot stop being in a typical American math class, you might pick up more math than you would by playing in the woods, but you might also burn out and develop lifelong math anxiety.

I hesitated before writing this post because I don't know what is special about languages and childrearing - I can't think of other obvious things in the category, though there are probably some. And I worry that pointing out the efficacy of relentlessness will lead people to commit themselves to things that are more like typical American math classes in the pursuit of whatever it is they want-to-want to learn. Please don't do that. But if anyone else can come up with a common thread between my examples of successful-relentlessness, or more examples, that might be useful.

Intermittent Distillations #4: Semiconductors, Economics, Intelligence, and Technological Progress.

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

The Semiconductor Supply Chain: Assessing National Competitiveness (Saif M. Khan, Alexander Mann, Dahlia Peterson)

[The Semiconductor Supply Chain: Assessing National Competitiveness](#)

Summary

This report analyzes the current supply chain for semiconductors. It particularly focuses on which portions of the supply chain are controlled by US and its allies and China. Some key insights:

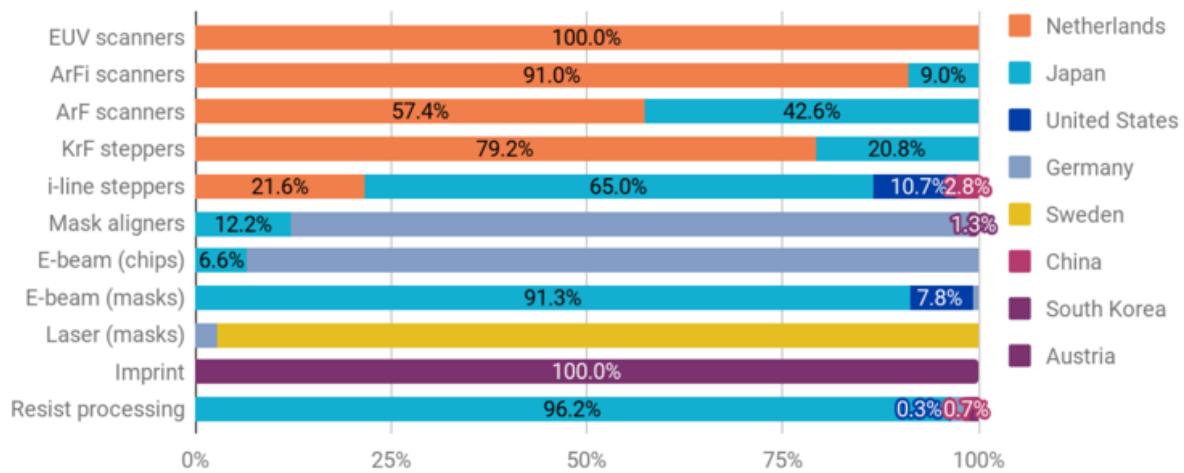
- The US semiconductor industry is estimated to contribute 39 percent of the total value of the global semiconductor supply chain.
- The semiconductor supply chain is incredibly complicated. The production of a single chip requires more than 1,000 steps and passes through borders more than 70 times throughout production.
- AMD is currently the only company with expertise in designing both high-end GPUs and high-end CPUs.
- TSMC controls 54% of the logic foundry market, with a larger share for leading edge production, e.g., state-of-the-art 5 nm node chips.
- Revenue per wafer for TSMC is rapidly increasing, while other foundries are seeing declines.
- The Netherlands has a monopoly on extreme ultraviolet (EUV) scanners, equipment needed to make the most advanced chips.
- The Netherlands and Japan have a monopoly on argon fluoride (ArF) immersion scanners, needed to make the second most advanced chips.
- The US has a monopoly on full-spectrum electronic design automation (EDA) software needed to design semiconductors.
- Japan, Taiwan, Germany and South Korea manufacture the state-of-the-art 300 mm wafers used for 99.7 percent of the world's chip manufacturing. This manufacturing process requires large amounts of tacit know-how.
- China controls the largest share of manufacturing for most natural materials. The US and its allies have a sizable share in all materials except for low-grade

gallium, tungsten and magnesium.

- China controls ~2/3rds of the world's silicon production, but the US and allies have reserves.

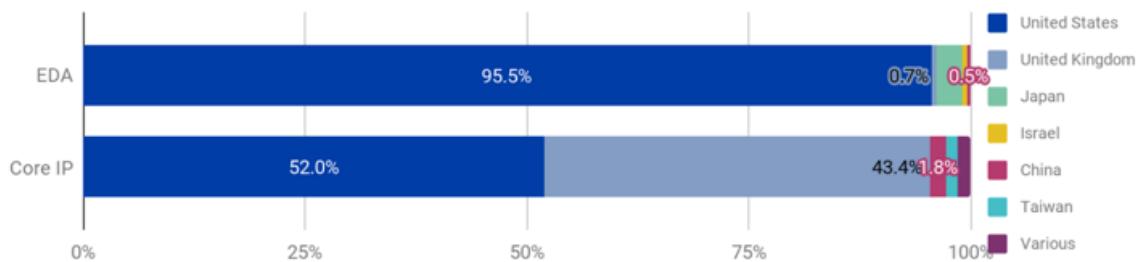
The report also analyzes US competitiveness at very detailed levels of the supply chain, which I didn't read that carefully. Tables:

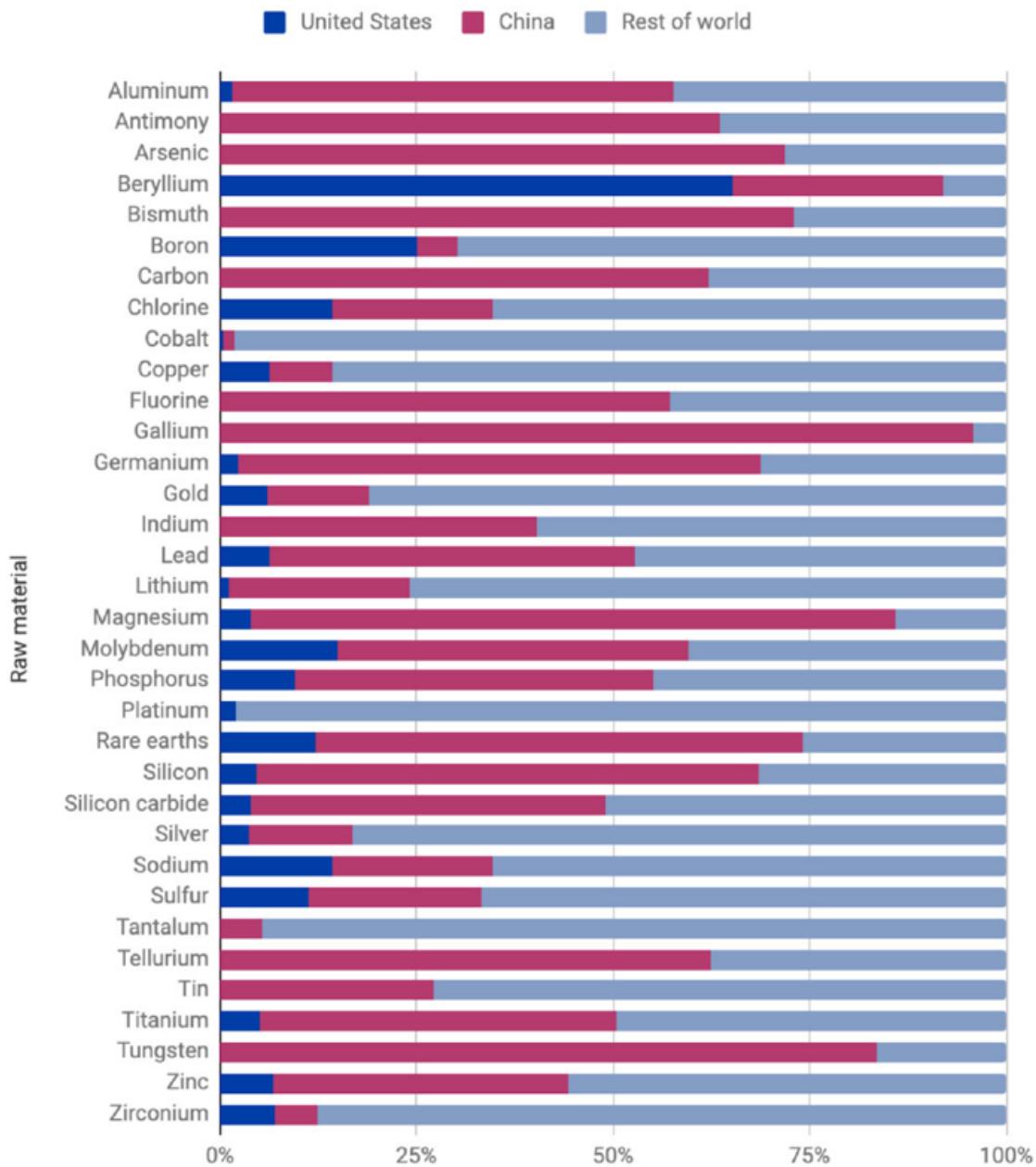
Figure 17: 2019 lithography country shares by firm headquarters



Source: VLSI Research

Figure 24: EDA and core IP country shares by firm headquarters





Source: USGS²¹⁷

Opinion

One perspective on the economy is that it's running a vast, distributed computation to allocate supply to demand in a relatively efficient manner. Examining the details on one of the supply chains underpinning a half a trillion dollar industry is relatively awe-inspiring. The only thing I'm currently aware of that is as complicated as computer hardware is computer software, which is sort of cheating. As AI becomes more advanced, control of semiconductor production becomes a strategic resource.

However, there are multiple monopolies/ sort of monopolies at every point. Each of these monopolies has a relatively large amount of bargaining power under many reasonable models. This situation puts the world in an interesting place. One concrete thing that I didn't consider before reading this report is the relevance of design software to semiconductor manufacturing. In retrospect, it seems pretty clear that the design of complicated things, e.g., video games, buildings, semiconductors, and animations, require complicated software with companies dedicated to building it. Lacking this software could constitute a meaningful bottleneck to being able to produce complicated artifacts. The asymmetry between manufacturing software and hardware is that software is easier to acquire through illegal means, whereas a EUV scanner has "100,000 parts, 3,000 cables, 40,000 bolts and 2 kilometers of hosing," making it prohibitive to steal.

Intelligence Explosion Microeconomics (Eliezer Yudkowsky)

[Intelligence Explosion Microeconomics](#)

Summary

Takeaways I found interesting:

- Evolutionary history strongly suggests that there are linear or superlinear returns to increased cognitive investment. Various lines of evidence, e.g. brain size of humans compared to chimps, look linear with respect to time. The last common ancestor of humans and chimps was only about 10 million years ago. 400,000 generations isn't enough time for mutations with low fitness increases to reach fixation, so the timelines are inconsistent with rapidly diminishing cognitive returns.
- If intelligence is a product of software capabilities and hardware capabilities, a series of recursive software improvements (in terms of percentile gains) can converge to a finite amount or go to infinity depending on the amount of hardware present. In economics, [Jevon's paradox](#) is the observation that increasing the efficiency of coal use might increase the demand for coal. Similarly, algorithmic improvements in AI might increase how profitable AI is, resulting in increased spending on computers.

Opinion

I knew most of the stuff in this paper already, but the presentation was clean and the arguments were well laid-out. I wish I had read it instead of the [AI-FOOM debate](#), although the latter was likely more entertaining.

Artificial Intelligence and Economic Growth (Philippe Aghion, Benjamin F. Jones & Charles I. Jones)

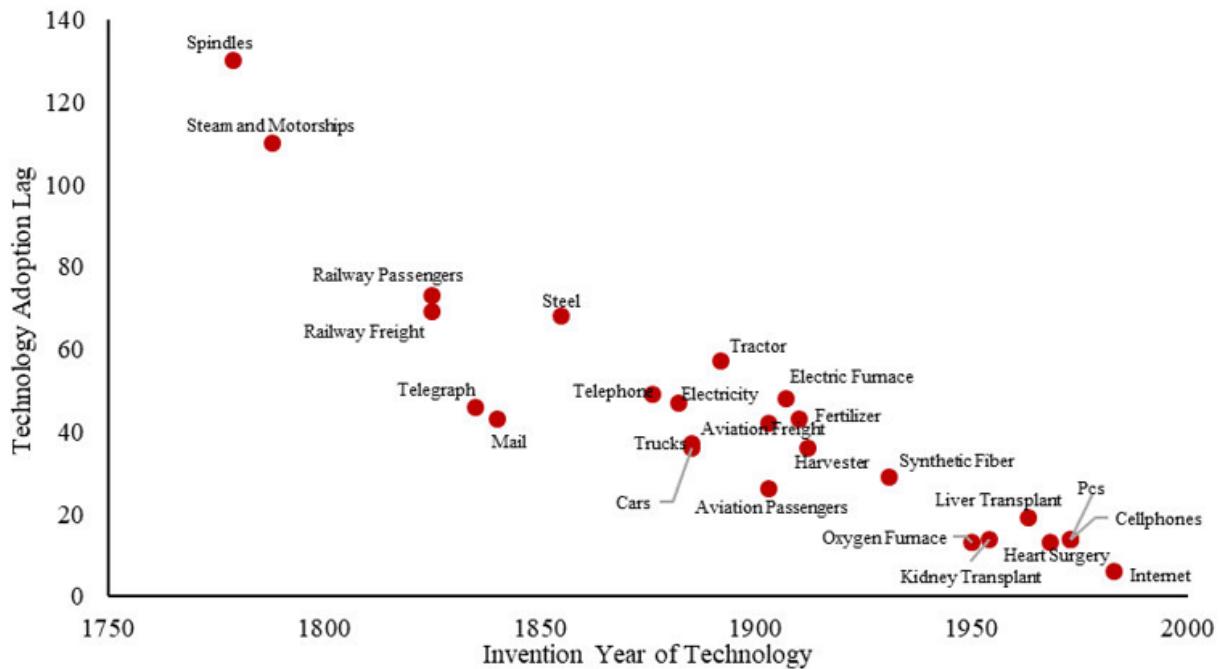
[Artificial Intelligence and Economic Growth](#)

Summary

The authors speculate on how economic growth might be affected by artificial intelligence by considering two main themes. The first theme is modeling artificial intelligence as a continuation of the natural historical processes that have automated various parts of the economy for the last 200 years. This theme allows historical data to be used to inform future speculation. The second theme is to constrain growth by considering Barmol's "cost disease", the observation that sectors that experience productivity growth see their share of the GDP declining, while sectors with slow growth see their share of GDP increasing. To quote the authors, "as a consequence, economic growth may be constrained not by what we do well but rather by what is essential and yet hard to improve." The author's model a situation in which the economy has two inputs, labor and capital, and work is divided into various tasks. Tasks can be automated, in which case they take a unit of capital to finish, or unautomated, in which case they take a unit of labor. In this model, as the fraction of tasks that are automated increases, the capital share, i.e. the amount of GDP going to people who are investing money as opposed to time, goes up. However, as the capital share rises, capital begins to accumulate, which results in an excess of automated goods (since they can be produced with only capital). This surplus results in the value of automated goods declining relative to nonautomated goods, resulting in "cost disease" type effects. As more sectors are automated, this increases the share of automated capital. However, since these sectors grow faster, the price declines. Implicit in this model is low elasticity of substitution between automated and nonautomated goods, i.e. as more goods become automated, it requires even more automated goods to replace the remaining nonautomated goods. This rate increases fast enough that even with infinite automated goods, there still remains some amount of goods that are nonautomated. This general process produces overall balanced growth because nonautomated goods provide a constraint the overall speed of the economy. Other models are then presented and analyzed, with the general trend being elasticity of substitution being an important parameter.

Opinion

I'm not really sure of what to think of predicting AI growth with economic tools. I think they're useful because they allow you to answer questions like, "What happens if we model AI as a continuation of historical automation?". However, I am unsure if this is the correct question to be answering. I also enjoyed the discussion of how cost disease type effects can prevent extremely explosive growth even if one good cannot be automated. I'm pretty skeptical that there will exist such a good. I don't have data to back this up, but I have a vague sense that historically when people have claimed that X cannot be automated away for fundamental reasons, they've been mostly wrong. The most plausible story to me seems like regulation won't be fast enough to keep up with the pace of automation and instead of getting e.g. automated doctors, we get AI systems that do all the work for the doctor but the doctor has to be the one that gives the treatment. In the short term, these regulatory hurdles might be enough produce balanced growth for a period before growth becomes [explosive](#). As a counterpoint, historical trends suggest rapid adoption. [Brookings](#):



Are Ideas Getting Harder to Find? (Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb)

[Are Ideas Getting Harder to Find?](#)

Summary

The authors model technological progress as a product of the number of researchers and productivity per researcher. They typically define technological progress such that constant progress leads to exponential growth in the economy, with the logic being that ideas can spread throughout the entire economy without cost. For example, an idea that causes a percentage decrease in the size of a transistor will result in a percentage increase in the number of transistors per some unit area. In areas such as medicine, the authors measure technological progress as a linear change in life expectancy.

The authors measure the number of researchers by taking aggregate R&D spending and deflating it by the average researcher wage. In addition to controlling for the effects of researcher wage increase, assuming roughly that labs will optimally allocate capital between high/low skilled researchers and lab equipment, dividing the by the average researcher wage allows for relatively fair comparisons between these different types of spending. They also consider other measures of research such as the number of clinical trials or the number of published papers in an area. In general, research efforts are increasing exponentially. In semiconductors, the research effort has risen by 18 since 1971. Between 1969 and 2009, corn and soybeans seed efficiency research both rose by more than a factor of 23.

The authors measure total technological progress across many domains by looking at transistors, crop yield per unit area, and life expectancy. The first two exhibit steady exponential growth (constant percentage), while life expectancy historically exhibits linear growth. Estimating the rate of technological progress and the effective number of researchers lets the authors calculate productivity. The results are quite clear: it's declining. For a representative example, transistors per unit area has been basically increasing at a constant 35% per year for the past 50 years, but research effort has risen by 18, which means it's currently 18x harder to maintain the same rate of growth as it was in 1971. This pattern appears to varying degrees in other fields. Table 7 summarizes:

TABLE 7—SUMMARY OF THE EVIDENCE ON RESEARCH PRODUCTIVITY

Scope	Time period	Average annual growth rate (%)	Half-life (years)	Dynamic diminishing returns, β
Aggregate economy	1930–2015	−5.1	14	3.1
Moore's Law	1971–2014	−6.8	10	0.2
Semiconductor TFP growth	1975–2011	−5.6	12	0.4
Agriculture, US R&D	1970–2007	−3.7	19	2.2
Agriculture, global R&D	1980–2010	−5.5	13	3.3
Corn, version 1	1969–2009	−9.9	7	7.2
Corn, version 2	1969–2009	−6.2	11	4.5
Soybeans, version 1	1969–2009	−7.3	9	6.3
Soybeans, version 2	1969–2009	−4.4	16	3.8
Cotton, version 1	1969–2009	−3.4	21	2.5
Cotton, version 2	1969–2009	+1.3	−55	−0.9
Wheat, version 1	1969–2009	−6.1	11	6.8
Wheat, version 2	1969–2009	−3.3	21	3.7
New molecular entities	1970–2015	−3.5	20	...
Cancer (all), publications	1975–2006	−0.6	116	...
Cancer (all), trials	1975–2006	−5.7	12	...
Breast cancer, publications	1975–2006	−6.1	11	...
Breast cancer, trials	1975–2006	−10.1	7	...
Heart disease, publications	1968–2011	−3.7	19	...
Heart disease, trials	1968–2011	−7.2	10	...
Compustat, sales	3 decades	−11.1	6	1.1
Compustat, market cap	3 decades	−9.2	8	0.9
Compustat, employment	3 decades	−14.5	5	1.8
Compustat, sales/employment	3 decades	−4.5	15	1.1
Census of Manufacturing	1992–2012	−7.8	9	...

Notes: The growth rates of research productivity are taken from other tables in this paper. The half-life is the number of years it takes for research productivity to fall in half at this growth rate. The last column reports the extent of dynamic diminishing returns in producing exponential growth, according to equation (17). This measure is only reported for cases in which the idea output measure is an exponential growth rate (i.e., not for the health technologies, where units would matter).

The authors also calculate a β parameter that measures the extent to which successive doublings in production are more difficult. For example, the economy has $\beta = 3$, so each successive doubling of the economy requires 3x the aggregate research input (in this case R&D spending). However, semiconductors have $\beta = 0.2$, so each successive doubling only requires 0.2 of a doubling of inputs. The reason why semiconductor

research has undergone such a large productivity decrease is because it has undergone so many doublings.

Opinion

Yep, it sure seems like research productivity is declining. Why? I recently read [Intelligence Explosion Microeconomics](#), so I'm tempted to say something like "Research is a product of serial causal depth of thinking, you find the easiest innovations first, and the later ones are going to be exponentially harder." I don't quite think this is true and I'm not sure what it would imply if it were true, but I suspect that low-hanging-fruit being picked is probably the main reason. I've heard "duplicate work" mentioned as another possible reason, but I'm not as compelled because good ideas getting duplicated doesn't seem to happen often enough to explain such large decreases in productivity. I enjoyed reading about how the authors thought about measuring the various inputs and outputs of their equations, e.g. how they would measure "total research" and "technological progress." I wouldn't be surprised if they got something very wrong, but their results have been applied to many areas so I would be surprised if the qualitative trends were wrong. I'm sort of surprised that growth theory is said to assume constant research productivity. The sort of basic observations that R&D spending has increased but economic growth has remained roughly the same seems to imply the obvious conclusion that productivity is declining. The corresponding counterargument surveyed by the paper that productivity within domains remains constant but aggregate productivity is declining seems sort of also clearly false if you think about any industry. My suspicion is that the falseness of this assumption has been obvious for a while, but people thought it might be a reasonable approximation or no one had bothered to get the data to falsify it until the authors. There's some argument that since economics research (and other theoretical stuff) has no clear way for it to make money, it's not going to be as efficient as semiconductor research, so fruit hangs lower, this paper being one such example. I suspect this observation to be relatively sophomoric compared to analyses of "research economies" that have definitely been done by some people.

Statistical Basis for Predicting Technological Progress (Béla Nagy, J. Doyne Farmer, Quan M. Bui, Jessika E. Trancik)

[Statistical Basis for Predicting Technological Progress](#)

Summary

How will technological development happen in the future? The authors compile a dataset of production versus unit cost for 62 technologies and hindcast a variety of prediction rules to determine which functional form tends to relate production and unit cost. They find that Wright's law, which predicts unit cost is some constant fraction of cumulative production, is most predictive, with Moore's law slightly after. Another way to interpret Wright's law is the prediction that unit cost will be exponentially decreasing

and cumulative production will be exponentially increasing, with the ratio between the exponents being a constant. Moore's law, on the other hand, only predicts that unit cost will decrease exponentially without positing a relation between production and cost. Wright's law is often interpreted to imply "learning by doing," with the thought being most technological advances (decrease in unit costs) come from attempting to scale up manufacturing. The authors conclude by commenting that their dataset doesn't contain enough technologies that don't grow exponentially to strongly distinguish between Moore's law and Wright's law.

Opinion

From the perspective of thinking about broad strokes technological progress, Wright's law seems like a strict upgrade to Moore's law. However, Wright's law still leaves unanswered the question "Why is cumulative production increasing exponentially?" Also, given that production is increasing extremely rapidly for most technologies, total cumulative production is only a few years or decades of current production levels. For example, if production grows by 50% a year and we're at 150 units per year right now, the past few years will have been 100 units, 67 units, 45 units, etc., which is a convergent geometric series that will sum to 450, 3 years of current production (see that $3 = 1/(1 - (1/1.5))$). Thus, even if current technological production stagnates, we should expect cumulative production to continue growing quickly. If Wright's law holds, this will correspond to a large decrease in unit cost.

Experience curves, large populations, and the long run (Carl Shulman)

[Experience curves, large populations, and the long run](#)

Summary

An experience curve predicts unit cost will fall a constant percent every time cumulative production doubles. Experience curves have empirically good at predicting technological progress across a variety of domains. For information technologies, experience curves predict an approximate halving of unit cost for each doubling of cumulative production. There are enough resources on Earth to sustain many doublings of cumulative production. For example, solar panel production could be scaled up 1000x, which is 10 doublings. According to [Nagy et al.](#), this would result in a 10x decrease in unit costs. For information technologies, whose production has been rising exponentially, a thousand years at current production levels would result in a 100x increase in cumulative production, which implies a 100x decrease in unit costs. Furthermore, Earth only captures about a billionth of the share of solar energy. [Fact check: the Earth is 150 billion meters from the Sun and is about 6 million meter radius. The area of space 150 billion m from the sun is about 10^{22} , while the area occupied by the Earth is about 10^{12} , a difference of 10^{10} , which is about a billion.] This theoretically gives another 9 orders of magnitude (OOM) for growth in cumulative production across many technologies. Space also exists and has many resources, giving many more OOMs. All this suggests if humanity does eventually use these resources, Wright's law implies we will run into the physical limits of technological progress.

Opinion

It's often useful to answer questions like "if we just take the existing trends and extrapolate them, what does that predict?" For example, [Roodman](#) uses historical GWP data to predict that humanity reaches a singularity in 2047. The point of the exercise is not to find out when humanity will reach a singularity, but rather determine how strong one's arguments have to be in either direction. For example, people who claim that AGI will cause a singularity around 2050 are making a prediction in line with historical super-exponential economic growth. Those who claim this won't happen are thus claiming that Roodman's extrapolation is overoptimistic. Similarly, using simple models relating technological progress to cumulative production predicts that we'll hit technological limits before we exhaust the amount of easily available resources. This is interesting because then people who claim that we won't either have to defy the increase in cumulative production or the relation between cumulative production and unit cost. I'm tempted to say something like "burden of proof", but I don't think that's really how arguments should work. The point is that doing such analyses lets discussion drop more easily to object-level disputes about whether particular models are accurate or not or the value of particular parameters, which I think are often more productive.

Benchmarking an old chess engine on new hardware

I previously explored the performance of a modern chess engine on old hardware ([1](#), [2](#)). Paul Christiano [asked](#) for the case of an old engine running on modern hardware. This is the topic of the present post.

State of the art

Through an online search, I found the [CCRL Blitz Rating list](#). It is run on an i7-4770k at [9.2 MNodes/s](#). The time controls are 2min+1s per move, i.e. 160s per 40 moves, or 4s per move. On the 4770k, that's 36.8 MNodes/move. The current number one on that list is Stockfish 14 at 3745 ELO. The list includes Fritz 5.32 from 1997, but on old hardware (Pentium 90). Over the years, CCRL moved through P90-P200-K62/450-Athlon1200-Athlon4600-4770k). I screened the list, but found no case of old-engine on modern hardware.

- One solution may be to reach out to the CCRL and ask for a special test run, calibrated against several modern engines.
- I can also imagine that the Swedish Chess Computer Association is able to perform the test directly, using their [Rating List](#) procedure. It covers 155,019 games played by 397 computers, going back to 2 MHz machines from the year 1984. They may be able to transplant an old version onto a new machine for cross-calibration. The right person to contact would be [Lars Sandin](#).

But the devil is in the details...

Making own experiments

In principle, the experiment should be trivial. The standard tool to compare chess engines is the command-line interface [cutechess-cli](#). Set up old and new and get it running.

Problem 1: Find an old engine with a working interface

Paul and I independently favored the program Fritz from the most famous chess year 1997, because it was at the time well respected, had seen serious development effort over many years, and won competitions. The problem is that it only supports today's standard interface, UCI, [since version 7](#) from the year 2001. So with that we can go back 20 years, but no more. Earlier versions used a proprietary interface (to connect to its GUI and to ChessBase), for which I found no converter.

I was then advised by [Stefan Pohl](#) to try [Rebel 6.0](#) instead. Rebel was amongst the strongest engines [between 1980 and 2005](#). For example, it won the [WCCC 1992 in](#)

[Madrid](#). It was [converted to UCI](#) by its author Ed Schröder. I believe that old Rebel makes for a similarly good comparison as old Fritz.

A series of other old engines that support UCI are [available for download](#). The following experiments should be possible with any of these.

Problem 2: Configuration

We can [download Rebel 6](#) and set up cutechess like so:

```
cutechess-cli  
-engine cmd="C:\SF14.exe" proto=uci option.Hash=128 tc=40/60+0.6 ponder=off  
-engine cmd="rebeluci.exe" dir="C:\Rebel 6.0" timemargin=1000 proto=uci tc=40/60+0.6  
ponder=off -rounds 20
```

One can define hash (RAM), pondering, and other settings through UCI in cutechess. However, Rebel does not accept these settings through the UCI interface. Instead, they must be defined in its config file wb2uci.eng:

```
Ponder = false  
Set InitString = BookOff/n  
Program = rebel6.exe w7 rebel6.eng
```

The "wX" parameter sets hash/RAM: w0=2 MB, w1=4 MB,.. w9=512 MB

So, we can test RAM settings between 2 MB and 512 MB for the old engine. Stockfish is typically run at its default of 128 MB. That amount would have been possible on old machines (486 era), although it would have been not common.

For Rebel 6, the interface runs through an adaptor, which takes time. If we would ignore the fact, it would simply lose due to time control violations. So, we need to give it 1000ms of slack with "timemargin=1000". Now, let's go!

Problem 3: Measuring the quality of a player that losses every single game

I ran this experiment over night. The result after 1,000 matches? Stockfish won all of them. So it appears that Rebel is worse, but by how much? You can't tell if it loses almost every game.

Rebel 6.0 has ELO 2415 on a P90, [Stockfish 13 is 3544](#) (SF14 is not in that list yet). That's a difference of 1129 ELO, with [expectations to draw only one game in 2199; and lose the rest](#). Of course, Rebel 6 will have more than 2415 ELO when running on a modern machine - that's what we want to measure. Measuring the gap costs a lot of compute, because many games need to be played.

OK, can't we just make SF run slower until it is more equal? Sure, we can do that, but that's a different experiment: The one of my previous post. So, let's keep the time controls to something sensible, at least at Blitz level for SF. With time controls of 30s, a game takes about a minute. We expect to play for at most 2199 games (worth 36

hours) until the first draw occurs. If we want to collect 10 draws to fight small number statistics, that's worth 15 days of compute. On a 16-core machine, 1 day.

Unfortunately, I'm currently on vacation with no access to a good computer. If somebody out there has the resources to execute the experiment, let me know - happy to assist in setting it up!

Experiment: SF3 versus SF13

There is a similar experiment with less old software that can be done on smaller computers: Going back 8 years between new and old SF versions.

The Stockfish team [self-tests new versions against old](#). This self-testing of Stockfish inflates the ELO score:

- Between their SF3-SF13 is a difference of 631 ELO.
- In the CCRL Rating list, which compares many engines, the difference is only [379 ELO](#) (the list doesn't have SF14 yet).
- Thus, self-testing inflates scores.

Let us compare these versions ([SF3](#) vs SF13) more rigorously. The timegap between 2013 and 2021 is ~8 years. Let us choose a good, but not ridiculous computer for both epochs (something like <1000 USD for the CPU). That would buy us an Intel Core i7 4770K in 2013, and an AMD 5950X in 2021. Their SF multicore speed is [10 vs. 78 MN/s](#); a compute factor of 8x.

We can now test:

- How much more compute does SF3 require to match SF13?
Answer: 32x (uncertainty: 30-35x)
- How much of the ELO gap does SF3 close at 8x compute?
Answer: 189 i.e. 50% of 379 ELO, or 30% of 631 ELO.

Interpretation: If we accept SF as amongst the very best chess programs in the last decade, we can make a more general assessment of chess compute vs. algorithm. Compute explains 30-50% of the computer chess ELO progress; algorithm improvements explain 50-70%.

Experiment method:

```
cutechess-cli -engine cmd="C:\sf3.exe" proto=uci tc=40/10 -engine  
cmd="C:\sf14.exe" proto=uci tc=40/20 -rounds 100
```

This runs a set of 100 rounds, changing colors each round, between SF3 and SF13 with their default parameters. The time controls in this example are 40 moves in 20 vs. 10 seconds etc. The time values must be explored in a wide range to determine consistency: The result may be valid for "fast" time controls, but not for longer games.

- From my experience, it is required to play at least 100 games for useful uncertainties.
- Due to time constraints, I have only explored the blitz regime so far (30s per game and less). Yet, the results are consistent at these short time controls. I strongly assume that it also holds for longer time controls. Then, algorithms

explain ~50% of the ELO gain for Stockfish over the last 8 years. Others are invited to execute the experiment at longer time settings.

Minor influence factors

So far, we have used no pondering (default in cutechess), no endgame tables, no opening books, default RAM (128 MB for Stockfish).

- Endgame databases: A classical space-compute trade-off. Decades ago, these were small; constrained by disk space limitations. Today, we have 7-stone endgame databases through the cloud (they weigh in at 140 TB). They [seem to be worth about 50 ELO](#).
- The influence of opening books is [small](#). I also suspect that its influence diminished as engines got better.
- Pondering: If the engine continues to calculate with opponents time, it can use ~50% more time. Typical ["hit rates" are about 60%](#). Thus, the advantage should be similar to 30% longer time control. ELO with time is not linear, thus no fixed ELO gain can be given.
- RAM sizes (hash table sizes) have a very small influence (source [1](#), [2](#)). For Stockfish, it appears to be only a few ELO.

Slack Has Positive Externalities For Groups

You ever have one of those group projects where you all need to find a time to meet, but some or all of the group members have packed schedules, so you pull out the whenisgood and it turns out the only times that work are 2-2:15 pm on Thursday or 6:30-7:30 am on Tuesday?

For this sort of scheduling problem, the best group members are those with lots of slack in their schedule - people who either have lots of time available, or have very flexible time commitments which can move around to accommodate a group meeting. But if my schedule is flexible, note that most of the benefits of that flexibility are captured by the group as a whole, not by me: my flexibility mostly allows the group to accommodate less-flexible members.

The slack in my schedule creates *positive externalities* for the group. I mostly control how much slack to create/maintain in my schedule, but a large chunk of the benefit goes to other people. This means I'm incentivized to create/maintain less-than-optimal slack in my schedule.

Once you look for it, this shows up in other guises too: many different flavors of slack create positive externalities for groups. In general, we should expect people to create/maintain less slack than would be socially optimal, and this in turn will make groups less functional. What do other forms of this look like, and what can we do about it?

Many Flavors of Slack

A few common forms of slack:

- Financial: money not budgeted for anything in particular, or which can easily be spent on something else instead, is financial slack.
- Time: time not scheduled for anything in particular, or which can easily be rescheduled, is time slack.
- Space: space not used for anything in particular, or which can easily be used for something else, is space slack.
- Emotional: capacity for excess stress is emotional slack.
- Social: multiple social groups which one can fall back on, or the ability to make new friends quickly, provide social slack.

We can also separate short-term vs long-term slack for each of these. For instance, a bank may have lots of capital to invest, but limited liquidity, so they can't move their capital around quickly: high long-term financial slack but limited short-term financial slack. Conversely, someone who has some savings on hand but is spending as much as they earn has short-term financial slack, but not long-term financial slack. Exercise for the reader: what would short-term and long-term time and emotional slack look like?

How do each of these create externalities for groups?

Space is an easy one: groups often need space in which to meet (either short-term, when the usual space is unavailable, or long-term) or store things (again, either short-term or long-term). If someone has spare space to use, that slack provides benefits to the whole group. But unless the group is paying to use the space, the person providing the slack captures only a small share of the benefits. So, people are incentivized to maintain less space slack than optimal.

Financial is another easy one: if some group members can occasionally cover some group costs, and it's not a big deal, that makes it a lot easier for a group to function smoothly. Again, this applies both short-term (e.g. paying the bill for a group dinner, with the expectation that everyone will eventually pay back) or long-term (covering some costs without reimbursement). Again, the person providing slack captures only a small share of the benefits.

The short-term/long-term distinction matters mainly for a group's agility/respondiveness/dynamicity. If there's a crisis and the group needs to respond quickly, or the group needs to make and execute plans on-the-fly as new information comes in, that requires short-term slack on the part of the group members. For instance, last year many groups started to work on COVID tools, like [microcovid](#), [radvac](#), or the various forecasting projects. Many of these required full-time work - people needed the slack to pause or quit their day jobs on relatively-short notice. That takes short-term financial slack, obviously, but also short-term emotional slack (very stressful!) and social slack (hopefully my coworkers aren't my only friends, or I can make new ones quickly!).

Another example: suppose a company or organization wants to move ([*cough*](#)) - not just across town, but to another state or country. That typically means employees will need to move with them. That requires emotional slack: moves are among the most stressful events most people go through. It requires social slack: people either need friends in the new location, remote friends, or the ability to quickly make new friends. And it requires financial slack, to pay for the move.

In both these examples, the group needs slack from its members in order to do things. (Or, to put it differently: group members' slack facilitates solutions to [coordination problems](#).) The ability to do things as a group mostly benefits the whole group, so the benefits of any particular person's slack largely go to the rest of the group.

What To Do About It?

One standard econ-101 answer is "internalize the externalities" - i.e. reward people for their slack. People don't usually do this with monetary payments, but we often do it with less legible rewards, like social status. For instance, if someone provides space for a group to meet, or occasionally covers some bills for the group, that's usually rewarded with status within the group.

Another standard solution is to *require* group members to maintain slack. Again, this usually isn't explicit, but we often do it in less-legible ways. For instance, if one or two people have very little slack in their schedules, maybe the rest of the group decides to meet without them. Or, if one or two people have very little emotional slack and sometimes break down if a competition gets too stressful, maybe they end up usually not participating in board game night or capture-the-flag. This is especially relevant to the last two examples from the previous section: the various COVID groups or the organization moving. If someone lacks the slack to participate, they would probably

not end up in the group. Of course, there still need to be *some* people who do have enough slack in order for the group to include anyone at all.

But these are illegible and imperfect methods. One point of this post is that it may help to *explicitly* pay attention to slack and its externalities. At a personal level, if we wish to be altruistic, this might mean maintaining extra slack in all its various forms, in order to provide value to the groups in which we participate. It might also mean avoiding people who have very little slack along one or more dimensions, or trying to supplement others' slack when possible (easy for finances, hard for time). For group organizers, it might mean explicitly requiring slack - e.g. the national guard requires that its members be able to drop everything and respond full-time to an emergency.

Important side point: slack has increasing marginal returns; the tenth unit of any particular flavor of slack is worth more than the first unit. The reason is that, if we flip n coins and count up the number of heads, the noise in that count is only $\sim\sqrt{n}$. And more generally, if we add up $\sim n$ independent noisy things, the noise will typically be of order $\sim\sqrt{n}$. So, if we want to take advantage of noisy *opportunities* - like a project which might go over budget, or a group which might need to move its meeting to a different time/space sometimes, or an event which might be fun or might be stressful - then we only need $\sim\sqrt{n}$ units of slack to take advantage of $\sim n$ opportunities. Going from zero unit of slack to one lets us take advantage of \sim one more opportunity, whereas going from nine units of slack to ten lets us take advantage of \sim twenty more opportunities. The more slack we have, the more we can benefit from adding marginal slack.

That means we should expect people to specialize in either having lots of slack, or no slack at all. For instance, we should expect people to either have carefully-planned tightly-packed calendars, or mostly-open calendars with lots of flexibility. We should expect people to either budget every dollar carefully, or have a large surplus and mostly not worry about their budget. Etc. One type takes advantage of lots of "noisy" opportunities, while the other makes their schedule/budget/etc maximally predictable. For a low-slack person to take advantage of just one noisy opportunity would require them to free up a bunch of extra room in their schedule/budget "just in case". The high-slack person already has a bunch of "just in case" built in, and can "re-use" that elbow room for one more thing, since it's highly unlikely that *all* the "risky" outcomes will happen all at once.

To the extent that this actually holds in the real world, we can think of slack (of a particular flavor) as binary: am I a high-time-slack person or a low-time-slack person? Am I a high-emotional-slack person or a low-emotional-slack person? That means the incentives don't need to be perfect - as long as a group can *roughly* select for high-slack members, or *roughly* reward high slack, that should be enough to incentivize the high-slack equilibrium rather than the low-slack equilibrium, and create lots of positive externalities for the group.

[Link] Musk's non-missing mood

This is a linkpost for <https://lukemuehlhauser.com/musks-non-missing-mood/>

Luke Muehlhauser writes:

Over the years, my colleagues and I have spoken to many machine learning researchers who, perhaps after some discussion and argument, claim to think there's a moderate chance — a 5%, or 15%, or even a 40% chance — that AI systems will destroy human civilization in the next few decades.¹ However, I often detect what Bryan Caplan has called a “[missing mood](#)”; a mood they would predictably exhibit if they really thought such a dire future was plausible, but which they don't seem to exhibit. In many cases, the researcher who claims to think that medium-term existential catastrophe from AI is plausible doesn't seem too upset or worried or sad about it, and doesn't seem to be taking any specific actions as a result.

Not so with Elon Musk. Consider his reaction ([here](#) and [here](#)) when podcaster Joe Rogan asks about his AI doomsaying. Musk stares at the table, and takes a deep breath. He looks *sad*. Dejected. Fatalistic. Then he says:

[Read more](#)

The shoot-the-moon strategy

Sometimes you can solve a problem by intentionally making it "worse" to such an extreme degree that the problem goes away. Real-world examples:

- I accidentally spilled cooking oil on my shoe, forming an unsightly stain. When soap and scrubbing failed to remove it, I instead smeared oil all over both shoes, turning them a uniform dark color and thus "eliminating" the stain.
- Email encryption can conceal the content of messages, but not the metadata (i.e. the fact that Alice sent Bob an email). To solve this, someone came up with [a protocol](#) where every message is always sent to *everyone*, though only the intended recipient can decrypt it. This is hugely inefficient but it does solve the problem of metadata leakage.

Hypothetical examples:

- If I want to avoid spoilers for a sports game that I wasn't able to watch live, I can either assiduously avoid browsing news websites, or I can use a browser extension that injects fake headlines about the outcome so I don't know what really happened.
- If a politician knows that an embarrassing video of them is about to leak, they can blunt its effect by releasing a large number of deepfake videos of themselves and other politicians.

The common theme here is that you're seemingly trying to get rid of X, but what you really want is to get rid of the distinction between X and not-X. If a problem looks like this, consider whether [shooting the moon](#) is a viable strategy.

An Apprentice Experiment in Python Programming

A couple weeks ago Zvi made [an apprentice thread](#). I have always wanted to be someone's apprentice, but it didn't occur to me that I could just ...ask to do that. Mainly I was concerned about this being too big of an ask. I saw [gilch's comment](#) offering to mentor Python programming. I want to level up my Python skills, so I took gilch up on the offer. In a separate [comment](#), gilch posed some questions about what mentors and/or the community get in return. I proposed that I, as the mentee, document what I have learned and share it publicly.

Yesterday we had our first session.

Background

I had identified that I wanted to fill gaps in my Python knowledge, two of which being package management and decorators.

Map and Territory

Gilch started by saying that "even senior developers typically have noticeable gaps," but building an accurate map of the territory of programming would enable one to ask the right questions. They then listed three things to help with that:

Documentation on the [Python standard library](#). "You should at least know what's in there, even if you don't know how to use all of it. Skimming the documentation is probably the fastest way to learn that. You should know what all the operators and builtin functions do."

[Structure and Interpretation of Computer Programs](#) for computer science foundation. There are [some variants](#) of the book in Python, if one does not want to use Scheme.

[CODE](#) as "more of a pop book" on the backgrounds.

In my case, I minored in CS, but did not take operating systems or compilers. I currently work as a junior Python developer, so reading the Python standard library seems to be the lowest hanging fruit here, with SICP on the side, CODE on the back burner.

Decorators

The rest of the conversation consisted of gilch teaching me about decorators.

Gilch: Decorators are syntactic sugar.

```
@foo  
def bar():  
    ...
```

means the same thing as

```
def bar():
    ...
bar = foo(bar)
```

Decorators also work on classes.

```
@foo
class Bar:
    ...
```

is the same as

```
class Bar:
    ...
Bar = foo(Bar)
```

An Example from pytest

At this point I asked if decorators were more than that. I had seen decorators in pytest:

```
@pytest.fixture
def foo():
    ...
def test_bar(foo): # foo automatically gets evaluated inside the test
    ...
```

Does this mean that, when foo is passed in test_bar as a variable, what gets passed in is actually something like pytest.fixture(foo)?

Gilch identified that there might be more than decorators involved in this example, so we left this for later and went back to decorators.

Decorators, Example 1

I started sharing my screen, gilch gave me the first instruction: Try making a decorator.

```
def test_decorator(foo):
    return 42
@test_decorator
def bar():
    print('hi')
print(bar())
```

Then, before I ran the program, gilch asked me what I expected to happen when I run this program, to which I answered that hi and 42 would be printed to console. At this point, gilch reminded me that decorators were sugar, and asked me to write out the un-sugared translation of the function above. I wrote:

```
def bar():
    bar = test_decorator(bar)
    return bar
```

I ran the program, and was surprised by the error `TypeError: 'int' object is not callable`. I expected `bar` to still be a function, not an integer.

Gilch asked me to correct my translation of my program based on the result I saw. It took me a few more tries, and eventually they showed me the correct translation:

```
def bar():
    print('hi')

bar = test_decorator(bar)
```

Then I realized why I was confused--I had the idea that decorators modify the function they decorate directly (in the sense of modifying function definitions), when in fact the actions happen outside of function definitions.

Gilch explained: A decorator could [modify the function], but this one doesn't. It ignores the function and returns something else. Which it then gives the function's old name.

Decorators, Example 2

Gilch: Can you make a function that subtracts two numbers?

Me:

```
def subtract(a, b):
    return a - b
```

Gilch: Now make a decorator that swaps the order of the arguments.

My first thought was to ask if there was any way for us to access function parameters the same way we use `sys.argv` to access command line arguments. But gilch steered me away from that path by pointing out that decorators could return anything. I was stuck, so gilch suggested that I try `return lambda x, y: y-x`.

Definition Time

My program looked like this at this point:

```
@swap_order
def subtract(a, b):
    return a - b

def swap_order(foo):
    return lambda x, y: y - x
```

PyCharm gave me an warning about referencing `swap_order` before defining it. Gilch explained that decoration happened at definition time, which made sense considering the un-sugared version.

Interactive Python

Up until this point, I had been running my programs with the command `python3 <file>`. Gilch suggested that I run `python3 -i <file>` to make it interactive, which made it easier to experiment with things.

Decorators, Example 2

Gilch: Now try an add function. Decorate it too.

Me:

```
def swap_order(foo):
    return lambda x, y: y - x

@swap_order
def subtract(a, b):
    return a - b

@swap_order
def add(a, b):
    return a + b
```

Gilch then asked, "What do you expect the add function to do after decoration?" To which I answered that the add function would return the value of its second argument subtracted by the first argument. The next question gilch asked was, "Can you modify the decorator to swap the arguments for both functions?"

I started to think about `sys.argv` again, then gilch hinted, "You have 'foo' as an argument." I then realized that I could rewrite the return value of the lambda function:

```
def swap_order(foo):
    return lambda x, y: foo(y, x)
```

I remarked that we'd see the same result from add with or without the decorator. Gilch asked, "Is addition commutative in Python?" and I immediately responded yes, then I realized that `+` is an overloaded operator that would work on strings too, and in that case it would not be commutative. We tried with string inputs, and indeed the resulting value was the reverse-ordered arguments concatenated together.

Gilch: Now can you write a decorator that converts its result to a string?

I wrote:

```
def convert_to_str(foo):
    return str(foo())
```

It was not right. I then tried

```
def convert_to_str(foo):
    return str
```

and it was still not right. Finally I got it:

```
def convert_to_str(foo):
    lambda x, y: str(foo(x, y))
```

There was some pair debugging that gilch and I did before I reached the answer. Looking at the mistakes I've made here, I see that I still hadn't grasped the idea that decorators would return functions that transform the results of other functions, not the transformed result itself.

Gilch: Try adding a decorator that appends ", meow." to the result of the function.

I verbalized the code in my head out loud, then asked how we'd convert the types of the function return value to string before appending ", meow" to it. Gilch suggested `f"\n{foo(x, y)}\n, meow"` and we had our third decorator.

We then applied decorators in different orders to show that multiple decorators were allowed, and that the order of decorators decided the order of application.

Splat

When we were writing the `convert_to_str` decorator, I commented that this would only work for functions that take in exactly 2 arguments. So gilch asked me if I was familiar with the term "unpacking" or "splat." I knew it was something like `**` but didn't have more knowledge than that.

How Many Arguments

Gilch asked me, "How many arguments can `print()` take?" To which I answered "infinite." They then pointed out that it was different from infinite--zero would be valid, or one, or two, and so on. So the answer is "any number," and the next challenge would be to make `convert_to_str` work with any number of arguments.

`print()`

We tried passing different numbers of arguments into `print()`, and sure enough it took any number of arguments. Here, gilch pointed out that `print` actually printed out a newline character by default, and the default separator was a space. They also pointed out that I could use the `help(print)` command to access the doc in the terminal without switching to my browser.

`type(_)`

Gilch pointed out that I could use the command `type(_)` to get the type of the previous value in the console, without having to copy and paste.

Splat

To illustrate how splat worked, gilch gave me a few commands to try. I'd say out loud what I expected the result to be before I ran the code. Sometimes I got what I expected; sometimes I was surprised by the result, and gilch would point out what I had missed. To illustrate splat in arrays, gilch gave two examples:

`print(1,2,3,*"spam", sep="~")` and `print(1,2,*"eggs",3,*"spam", sep="~")`. Then they showed me how to use `**` to construct a mapping: `(lambda **kv: kv)(foo=1, bar=2)`

Dictionary vs. Mapping

We went off on a small tangent on dictionary vs. mapping because gilch pointed out that dictionary was not the only type of mapping and tuple is no the only type of iterable. I asked if there were other types of mapping in Python, and they listed OrderedDict as a subtype and the Mapping abstract class.

Parameter vs. Argument, Packing vs. Unpacking

At this point gilch noticed that I kept using the word "unpacking." I also noticed that I was using the term "argument" and "parameter" interchangeably here. Turns out the distinction is important here--the splat operator used on a parameter packs values in a tuple; used on an argument unpacks iterable into separate values. For example, in `(lambda a, b, *cs: [a, b, cs])(1,2,3,4,5)`, cs is a parameter and * packs the values 3, 4, 5 into a tuple; in `print(*"spam", sep="~")`, "spam" is an argument and * unpacks it into individual characters.

Dictionaries

Gilch gave me another example: Try `{'x':1, **dict(foo=2,bar=3), 'y':4}`. I answered that it would return a dictionary with four key-value pairs, with foo and bar also becoming keys. Gilch then asked, "in what order?" To which I answered "dictionaries are not ordered."

"Not true anymore," gilch pointed out, "Since Python 3.7, they're guaranteed to remember their insertion order." We looked up the Python documentation and it was indeed the case. We tried `dict(foo=2, **{'x':1,'y':4}, bar=3)` and got a dictionary in a different order.

Hashable Types

I asked if there was any difference in defining a dictionary using `{}` versus `dict()`. Gilch compared two examples: `{42:'spam'}` works and `dict(42='spam')` doesn't. They commented that keys could be any hashable type, but keyword arguments were always keyed by identifier strings. The builtin `hash()` only worked on hashable types.

I don't fully understand the connection between hashable types and identifier strings here, it's something that I'll clarify later.

Parameter vs. Argument, Packing vs. Unpacking

Gilch gave another example: `a, b, *cs, z = "spameggs"`

I made a guess that cs would be an argument here, so * would be unpacking, but then got stuck on what cs might be. I tried to run it:

```
>>> a, b, *cs, z = "spameggs"
>>> a
's'
>>> b
'p'
>>> cs
['a', 'm', 'e', 'g', 'g']
>>> z
's'
```

Gilch pointed out that `cs` was a store context, not a load context, which made it more like a parameter rather than an argument. Then I asked what store vs. load context was.

Context

Gilch suggested, `import ast then def dump(code): return ast.dump(ast.parse(code))`. Then something like `dump("a = a")` would return a nested object, in which we can locate the `ctx` value for each variable.

This reminded me of lvalue and rvalue in C++, so I asked if they were the same thing as store vs. load context. They were.

Splat

Gilch tied it all together, "So for a decorator to pass along all args and kwargs, you do something like `lambda *args, **kwargs: foo(*args, **kwargs)`. Then it works regardless of their number. Arguments and keyword arguments in a tuple and dict by keyword. So you can add, remove, and reorder arguments by using decorators to wrap functions. You can also process return values. You can also return something completely different. But wrapping a function in another function is a very common use of decorators. You can also have definition-time side effects. When you first load the module, it runs all the definitions--This is still runtime in Python, but you define a function at a different time than when you call it. The decoration happens on definition, not on call."

We wrapped up our call at this point.

Observations

1. As we were working through the examples, we'd voice out what we expect to see when we run the code before actually running to verify. Several times gilch asked me to translate a decorated function into an undecorated one. This was helpful for me to check my understanding of things.
2. Another thing I found valuable were the tips and tricks I picked up from gilch throughout the session, like interactive mode; and the clarification of concepts, like the distinction between parameter and argument.
3. Gilch quizzed me throughout the session. This made things super fun! I haven't had the opportunity for someone to keep quizzing me purely for learning (as opposed to giving me a grade or deciding whether to hire me) for the longest time! I guess that reading through well-written text tends to be effective for familiarizing oneself with concepts, while asking/answering questions is effective at solidifying and synthesizing knowledge.
4. In this post, I tried to replicate the structure of my conversation with gilch as much as possible (the fact that gilch's mic was broken so they typed while I talked made writing this post so much easier--I had their half of the transcript generated for me!) since we went off on some tangents and I wanted to provide context for those tangents. I think of a conversation as a tree structure--we start with a root topic and go from there. A branch would happen when we go off on a tangent and then later come back to where we left off before the tangent. Sometimes two sections of this post would have the same section headings; a second time a section heading is used indicates that we stopped the tangent and went back to where we branched off.

Covid 7/8: Delta Takes Over

Delta is now two thirds of sequenced samples from the past week, so we can be confident that it has taken over, and soon will be most cases in America, and soon after that most cases around the world. That's bad news, but given we know the case numbers, it sort of is good news. It means we've already 'taken the hit' to the reproduction rate, mostly, and if we can stabilize things one final time, then unless there's a new even worse variant then the last scare is over.

Things are not yet stable, but they continue to echo what happened this time last year, and things continue to be quite hot. If things cool down a bit (in temperature terms), several weeks pass and these trends continue, then they'll be more clearly where we are at going forward, and we'll be counting on further vaccinations and on past vaccinations kicking in. We *should* have enough momentum there to still get us over the finish line, but in many places it's going to be close.

Let's run the numbers.

The Numbers

Predictions

I forgot that July 4th was coming, so my deaths prediction was dumb. Whoops.

Prediction from last week: Positivity rate of 2.7% (up 0.3%) and deaths decline by 5%.

Result: Positivity rate of 2.9% (up 0.5%) and deaths decline by 20% likely due to July 4th.

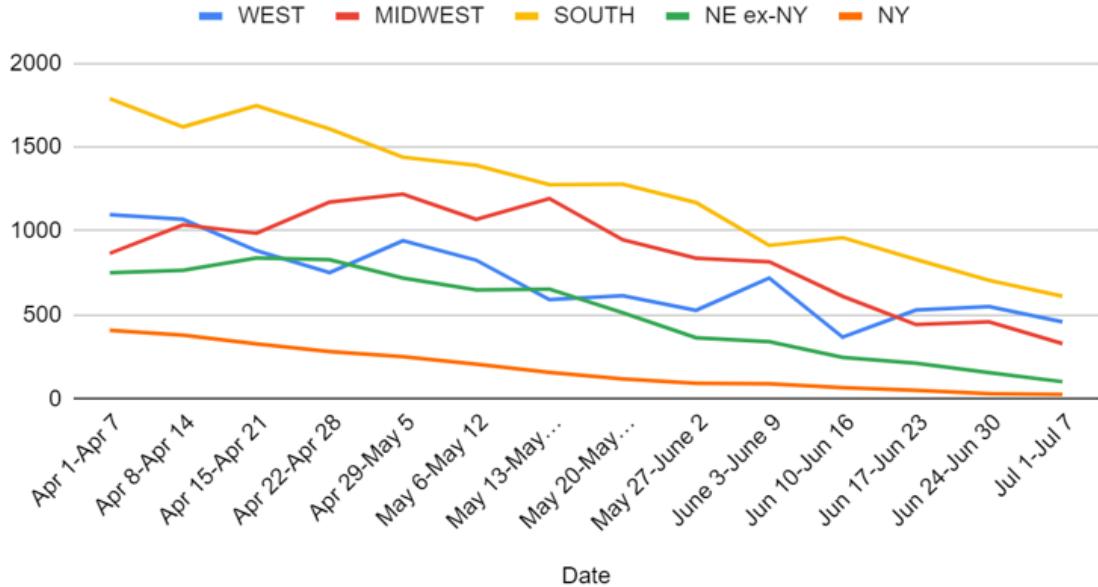
Prediction for next week: Positivity rate of 3.3% (up 0.4%) and deaths increase by 7%.

Predicting deaths this week is weird, and I'm predicting an increase because I think this week's count is artificially low due to the holiday. So there will be some catch-up reporting and some reversion, even if the new rise in cases won't have caught up to us yet. For the positivity rate, it jumped a lot the last day, but I have to assume it will keep rising for now. As it stops being constrained, the positivity rate will once again become a useful measure of our situation.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 27-June 2	527	838	1170	456	2991
June 3-June 9	720	817	915	431	2883
Jun 10-Jun 16	368	611	961	314	2254
Jun 17-Jun 23	529	443	831	263	2066
Jun 24-Jun 30	550	459	706	186	1901
Jul 1-Jul 7	459	329	612	128	1528

Deaths by Region



I'd love for this to be real, but it's probably not. It's July 4th, and I keep forgetting to look ahead to the holidays in the next week before I make predictions. There's no way this full decline is real, and it's hard to know where the real number was.

My brain continues to not understand how it being the 4th of July causes three days of missing data across the country. That's not a mindset I can process, even now, not fully. It definitely did happen, whether I can process it or not, and it's very unclear how much of the difference was made up after the weekend was over.

Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 20-May 26	33,890	34,694	48,973	24,849	142,406
May 27-June 2	31,172	20,044	33,293	14,660	99,169
Jun 3-Jun 9	25,987	18,267	32,545	11,540	88,339
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	7,050	91,954

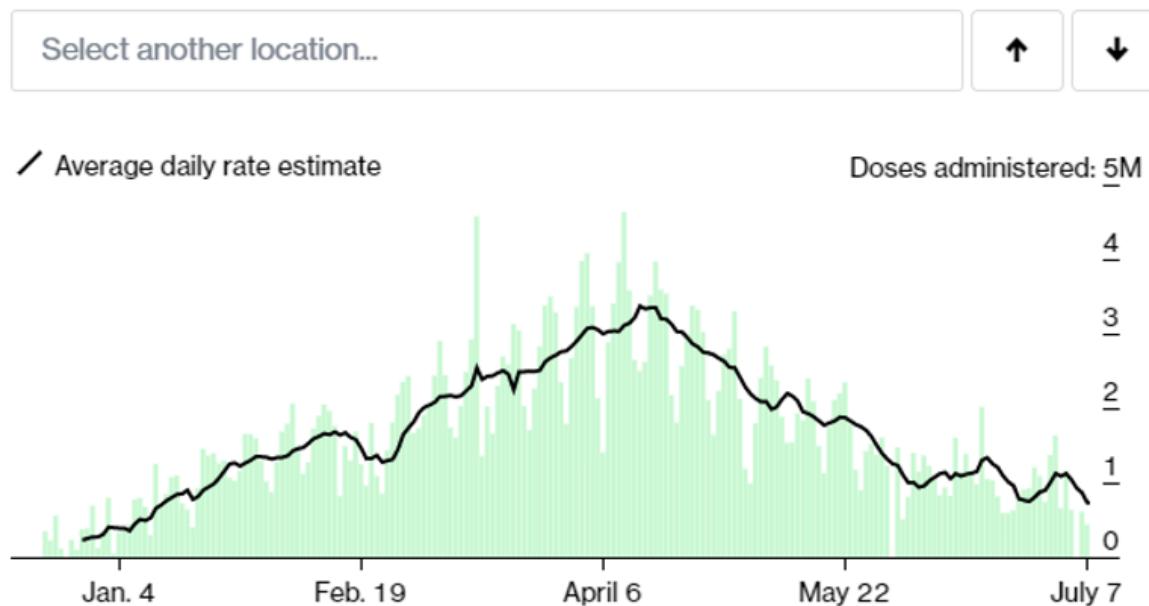
That's an unsettling jump especially given the holiday weekend with its disruptions to reporting, but given the jump I'm going to presume that there wasn't much disruption to testing, only to deaths. That means we saw a 20% or so rise in cases, up from 10% last week, which was across the board and concentrated in the places one would expect. In terms of absolute levels things are still fine, and there isn't much room for further acceleration because Delta is already dominant, but at 20% per week ($R \sim 1.15$) things will get ugly fast if trends don't improve.

These changes are very similar to the regional changes one year ago this week, by region, in percentage terms.

Further discussion of likely future trends in the Delta section.

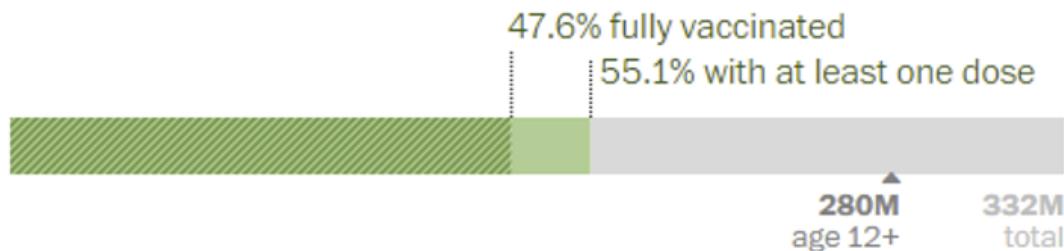
Vaccinations

In the U.S., the latest vaccination rate is **732,848 doses** per day, on average. At this pace, it will take another **7 months** to cover **75%** of the population.



182.9 million vaccinated

This includes more than **157.9 million people** who have been fully vaccinated in the United States.



In the last week, an average of **732.8k doses per day** were administered, a **31% decrease ↓** over the week before.

The decrease is unwelcome, but also clearly linked to the holiday, so it's reasonable to expect some amount of rebound over the coming week. It could also be seen as the end of a previous temporary bump, in which case we wouldn't expect much further decline right away but wouldn't expect a bounce-back. My guess is the next week will hold steady from here, which is still not bad at all. Last week we were at 46.7% fully vaccinated and 54.4% with one dose, and we picked up 0.9% and 0.7% on those respectively. That's at least an effective 1.5% drop in R₀, and I'm guessing more than that. As I say each week, that adds up fast.

[Periodic reminder from MR that fractional dosing of vaccines would save many lives and end the pandemic faster with no downside. This links to a new paper in Nature Medicine arguing the same thing. And a preprint showing 1/4 doses of Moderna create immunity similar to natural infection.](#) As a reminder, I continue to believe that Moderna's day-after side effects are due to the dose being actively too large rather than simply not being strictly necessary. [And Moderna is getting half-doses ready for children](#), so we know they can provide them, and also arguing feasibility was always silly. For children, the new half-doses should be overkill the same way the full-doses are overkill for adults.

Or [simpler version](#):



Robert Wiblin @robertwiblin · Jul 2

...

I know it's old news but we're still not moving to quickly use lower doses of mRNA vaccines, which will likely result in 1 million gratuitous deaths or more.



Robert Wiblin @robertwiblin · Jul 4

...

This is good:

"Moderna is gearing up to halve the dose of its COVID-19 vaccine, the U.S. drugmaker said on Wednesday, so that it can also be used to combat variants and inoculate children."

If you have had one shot of the Johnson & Johnson vaccine, should you then get a shot of Pfizer or Moderna? If it is available, absolutely, yes you should. There's every reason to think that mix and match will work here, one shot of J&J isn't that differently effective than one shot of mRNA, and no reason beyond supply limitations to not take the second shot. Yet because it is not Officially Recognized all the Very Serious People are falling in line and finding ways to tie themselves up in knots and pretend that a second shot would not be useful. [Here's Nate Silver commenting on the bizarre NYTimes attempt at this. Here's the Washington Post attempt.](#) It's all disgraceful bull**** designed to back up official policy that doesn't make sense, with gems like 'if you don't feel safe enough without a booster, use additional measures like mask wearing.' They've committed to one J&J shot counting as 'vaccinated' while one shot of Moderna doesn't, and now they have to pretend that definition makes sense.

Remember to always be *praising* people when they finally do the right thing, regardless of whether they first exhausted all alternatives. The last thing you want to do is punish correct action by using it as justification for someone now being that much more blameworthy for not doing it sooner! [Still, sometimes...](#)



tylercowen  @tylercowen · Jul 1

Ahem.

..



Justin Sandefur @JustinSandefur · Jul 1

The World Bank, IMF, WHO, and WTO are very pleased to announce they just held the *first* meeting of their joint taskforce on covid vaccines.

On June 30.

2021.

(h/t an anonymous World Banker)
worldbank.org/en/news/statem...

[Show this thread](#)

WASHINGTON, June 30, 2021—The Heads of the World Bank Group, International Monetary Fund, World Health Organization, and World Trade Organization today convened for the first meeting of the Task Force on COVID-19 Vaccines, Therapeutics and Diagnostics for Developing Countries. They issued the following joint statement:

6

10

141

↑

More vaccinations are the only viable way to deal with Delta. [As Matt says here, there's no going backwards](#), and we should have and will have very little tolerance for attempts to reimpose intrusive measures:



Matthew Yglesias  @mattyglesias · Jul 3

I am very worried about what Delta will do in low-vaccination counties but the only plausible solution is for more people to get vaccinated, trying to get the already vaccinated people to also go back to masks and other precautions doesn't make sense and wouldn't work.

119

215

1.9K



Matthew Yglesias  @mattyglesias · Jul 3

If we want to take intrusive measures to combat Delta we should make the vaccine mandatory in more situations.

If we don't want to take intrusive measures, then we just shouldn't.

Either way there's no universe in which returning to policies from six months ago makes sense.

26

77

918



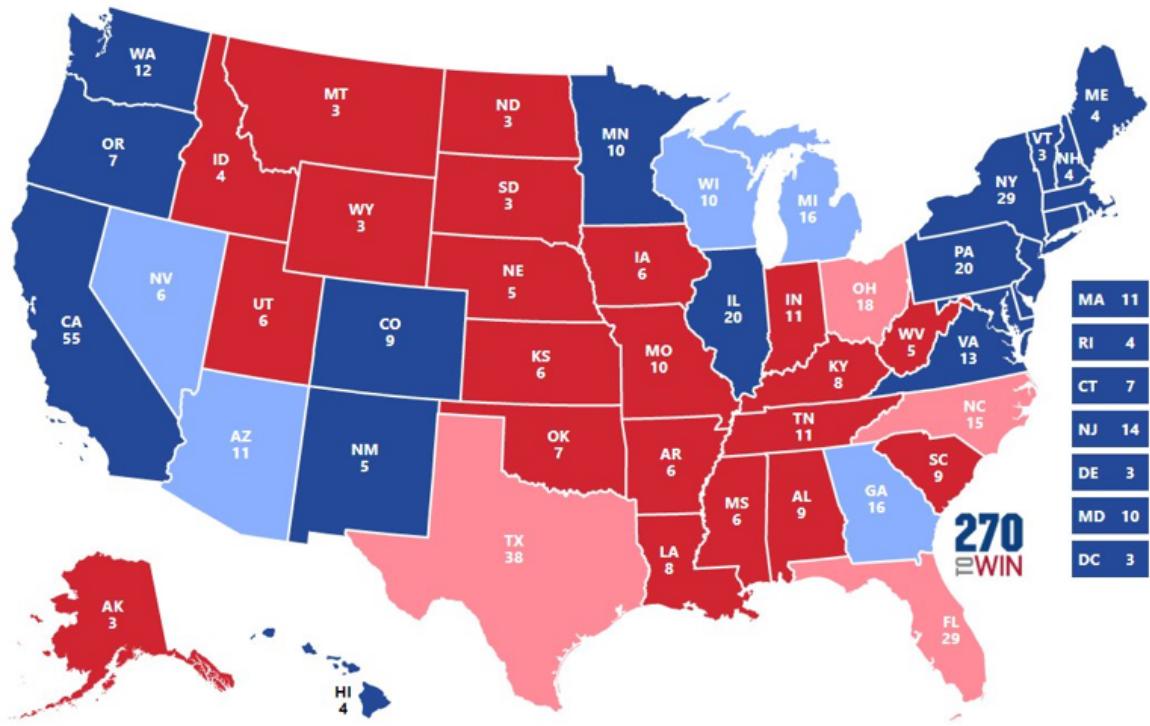
Matthew Yglesias  @mattyglesias · Jul 3

Replying to [@mattyglesias](#)

Third option is that I suppose we could give Pfizer & Moderna license to market their vaccines under the trade name "Ivermectin"

[Map pairing that's going around Twitter that might not be entirely fair, but come on:](#)

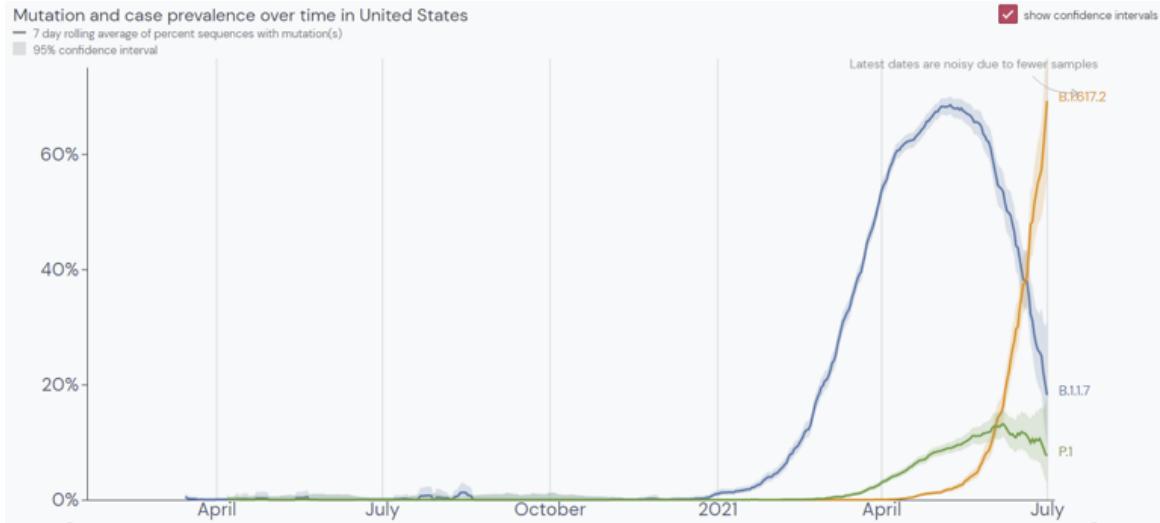




Delta Variant

Within a few weeks Delta will be almost all cases, and there won't be a need for a distinct Delta section, but for now it still makes sense.

Where are we on that right now? [Mostly we're there.](#)



This is a seven day rolling average of data that's lagged days behind that to begin with, and this matches where I previously presumed the trend-line was. If it was 65% or so in this 7-day rolling average of lagged data, current case counts likely reflect something more like 75%-80%, and the Alpha+Gamma is accounting for almost all of the remainder as per this

graph. That means we have accounted for over 90% of the effect of the shift from pre-Alpha to Delta.

If we take this week's numbers seriously in that context, and don't make any adjustments, we'd get a final R₀ for Delta under exact current conditions of R₀ = 1.18 or so, once we take into account the lag in our case numbers.

[Trevor Bedford thread attempts to ballpark the potential size of the wave coming from Delta.](#)

It's useful to check one's intuitions and estimate calculations against those of others. Trevor uses a basic SIR model, assumes that vaccinations are fixed and independent of infections (and counts only those currently fully vaccinated, but gives them credit for full immunity, and ignores varying distributions of vaccinations, much of which he notes is wrong but that he hopes such effects will cancel) and as always *ignores all control systems in all directions* and comes up with a further 11% of the population likely to be infected before this is all over.

In addition to all that, I've written several times about why SIR is not a great approximation in situations where different people are making varying adjustments in their behaviors and taking radically different levels of risk, which he's also not considering. Nor does it take the children versus adults distinction seriously.

I think these issues mostly point in the direction of a smaller wave, and there's one that towers over the rest of them, which is that vaccinations will continue until morale improves.

We are continuing to see vaccine shots given out at a good clip. At a *bare minimum*, I would expect the bulk of those who have only had their first shot to get their second shot. He's starting with 46% fully vaccinated and an R₀ = 1.18 (matching my estimate above). If we move that 46% to something like 55%, that's enough to get R₀ = 1, and all that would require is everyone who has had a first shot as of July 5 getting a second shot. It doesn't work that way, because the first shot's protection is already present and can't be double counted, but other factors point the other way.

Another way to think about this is that when Trevor says 11% additional infections, that's 11% additional infections *or vaccinations*, and can mostly be considered a hard upper bound. More vaccinations means less overshooting after R₀ gets down to 1, combined with all the other issues, make me continue to think that the Delta wave is unlikely to get all that big.

There's also the seasonality perspective. This week, we saw a 21% increase in cases. A year ago this week, we saw a 21% increase in cases, both with similar regional patterns. A reminder:

Date in 2020	Cases Total
June 11-June 17	158164
June 18-June 24	215751
June 25-July 1	300510
July 2-July 8	365107
July 9-July 15	431972
July 16-July 22	461441
July 23-July 29	444797
July 30-Aug 5	392193
Aug 6-Aug 12	365028
Aug 13-Aug 19	322126

Thus, it seems premature to conclude that we are in a permanent R=1.18 world going forward pending additional vaccinations. We also have this situation at very low overall Covid levels, so control systems via individual action have a long way to adjust if necessary. We

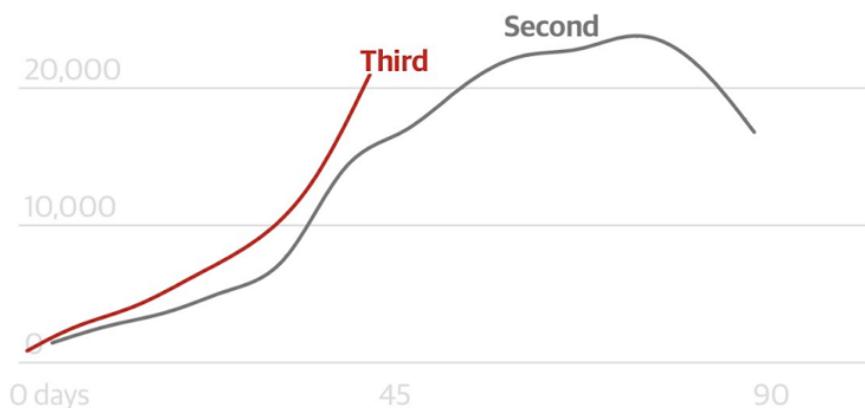
also have vaccinations continuing, with this week's disappointing crop still dropping R₀ at least by 1.5% (and at least one vaccination week won't be reflected in cases yet, likely 2-3, so we have a head start here).

My baseline scenario continues to be that cases rise for a bit, but things stabilize in most places before reaching levels that require behavioral adjustments, especially by the vaccinated. But I do continue to expect some regional/local issues in places with lower vaccination rates.

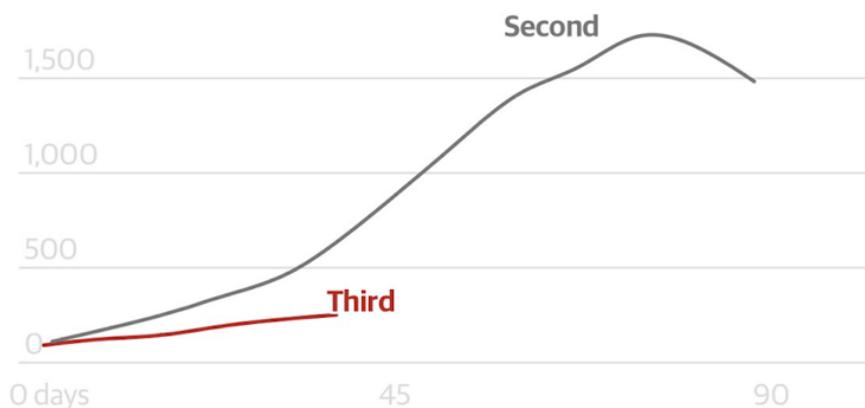
Meanwhile, [over in the UK, Guardian provides this graph:](#)

Coronavirus waves compared

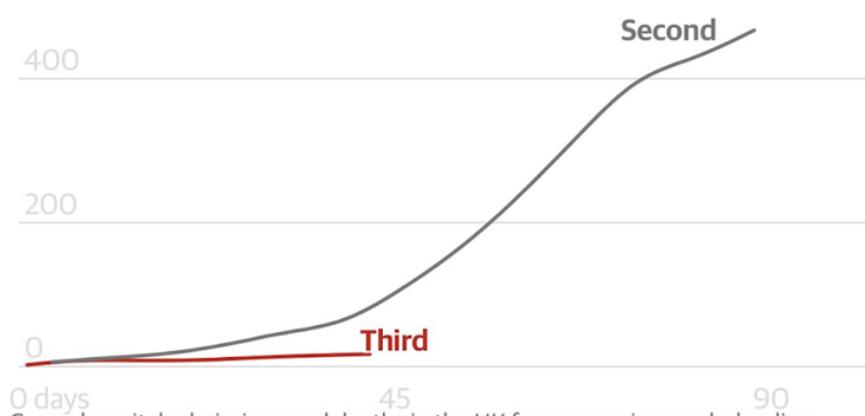
Cases



Hospitalisations



Deaths



Cases, hospital admissions and deaths in the UK for coronavirus each day, line shows rolling seven day average. Second wave taken as starting 30 Aug 2020, third wave, 19 May 2021. Data: data.gov.uk, updated 1 July, 2021

Periodically commenters will ask what the evidence is that Delta is more lethal than Alpha. I've seen such estimates from several sources, and no formal estimates doubting it, but such

effects can't hold a candle to the power of vaccination, and especially of vaccination of the most vulnerable populations.

Scott Alexander Analyzes Lockdowns

[Scott Alexander writes Lockdown Effectiveness: Much More Than You Wanted To Know.](#) This was not, in this case, more than I wanted to know. Instead it felt more like a ton of empty calories, with comparison after comparison and calculation after calculation that had so many caveats (that were explicitly mentioned – Scott plays fair on such matters) that I don't feel much more enlightened than when I began reading, and the main update I made was that the evidence available to find was such that Scott was unable to provide an update.

[Sam Bankman-Fried summarizes it this way, which seems right:](#)



SBF  @SBF_Alameda · 13h

astralcodexten.substack.com/p/lockdown-eff...

TL;DR:

- 1) lockdowns saved lives, probably
- 2) the \$ cost per life saved was in-line with replacements, not clearly good or bad
- 3) if you include non-immediate-GDP impacts (e.g. it sucks), maybe lockdowns were very net bad
- 4) huge error bars and uncertainty

...

Comparisons of this form stack the deck in favor of lockdowns, because they discount non-GDP effects (SBF's #3), and also by considering the average countermeasure against the average gain, instead of comparing effects on the margin.

That second one is worth unpacking a bit if it isn't obvious to you. The value of lockdown was being considered en masse rather than on the margin. Given that we end up controlling the situation either way and never end up in 'everyone is infected' mode, stricter long-term lockdowns have increasing marginal cost per case prevented or life saved. Thus, if it's not clear whether lockdowns *in general* are good, or whether lockdowns *above a given much lower level* (like red state level versus blue state level) are good, that means that *on the margin* we did too much locking down. If we didn't do enough, lockdowns up to that point would look very good.

Another note is that Scott is much too kind about the 'maybe some of these provisions were not all that great' aspect of the problem. Closing parks and beaches and playgrounds isn't 'we made lives worse for not much gain,' it's 'we actively forced people into more dangerous situations and made the pandemic worse, while also making lived experience worse and burning out public tolerance and trust.'

What was the right rate of mandated lockdown, *given our ability to prioritize measures?* My belief is that private action reacts better and the control systems are very strong, and that the real reason to do lockdowns is the tail risk of complete disasters (and the big silent reason *not* to do them is to keep that ammo for when you really need it and/or to minimize risk of complete disaster via loss of civil order and people losing their minds).

That's the argument that an analysis like Scott's is missing the central point of decision making under uncertainty, rather than stacking (or not stacking) the deck in a particular way.

Once we got past the first few months, I'm firmly in the 'we did too much locking down on the margin' camp. I believe things would have gone better if we had let people make more of their own decisions. Even better would have been *smarter* restrictions, but I've learned not to make that the comparison point in such questions.

In Other News

[EA Forum post \(via MR\) reiterates some of the various ways the pandemic was botched.](#)

[Thread explaining why vaccines provide better immunity than natural infection.](#)

[Belgium mandates CO₂-level monitors in businesses as a proxy for ventilation \(news story\).](#)

"Businesses were happy with this because it meant they could reopen, and it's a small investment that costs €60 or €80 depending on the type of sensor," Mr Facon said.

The amount of carbon dioxide reflects how much of the air that individuals are breathing in has already passed through the lungs of others.

This seems like an excellent idea. Businesses liking it makes sense too, but in general there is the worry that 'the government tells you that you can't open, then tells you that you can open if you do X' will make you happy whether or not the prohibition or requirement makes any sense, as it's better to pay a small additional cost than stay closed.

Air filters

Some buildings including some old schools, certain catering establishments, and some sports infrastructure will not be able to comply with the carbon dioxide limits due to their design. Such venues will have the option instead of installing air filters that disinfect the air.

"The carbon dioxide can be too high, but if there is filtration and disinfection, that's okay from a Covid-19 perspective," Mr Facon said. "It's not really okay from an overall wellbeing perspective, but that's something we would need to address in the more long term."

Poor ventilation is quite bad for many non-Covid reasons, and it's very good to see this acknowledged. Fixing air quality is an underappreciated cause area, whether via filters or better designs, and if that caught on more it would be a large upside to everything that's happened.

Great news, [New York gets to keep its outdoor dining:](#)



Morgan Mckay @morganfmckay · 17h

.@NYGovCuomo signed a bill today that will allow New York restaurants to keep using sidewalks and streets for outdoor dining

...

Which of course led to this as the featured comment, gotta love the user name:



Trump You Lying Shitbag @Cannabi02960735 · 17h

Replying to @morganfmckay and @NYGovCuomo

Where are pedestrians supposed to walk especially older folks ?

...

To which I'd respond *on the sidewalk* because I have never seen an outdoor dining area block a sidewalk to pedestrians in a meaningful way, not once, what city does this person even live in. It's amazing we have any nice things at all.

[And not as great news, New York](#), insert your own joke here:



News 8 WROC @News_8 · Jul 6

Gov. Cuomo: "We want to do with gun violence with what we just did with COVID."

1.9K

1.6K

471



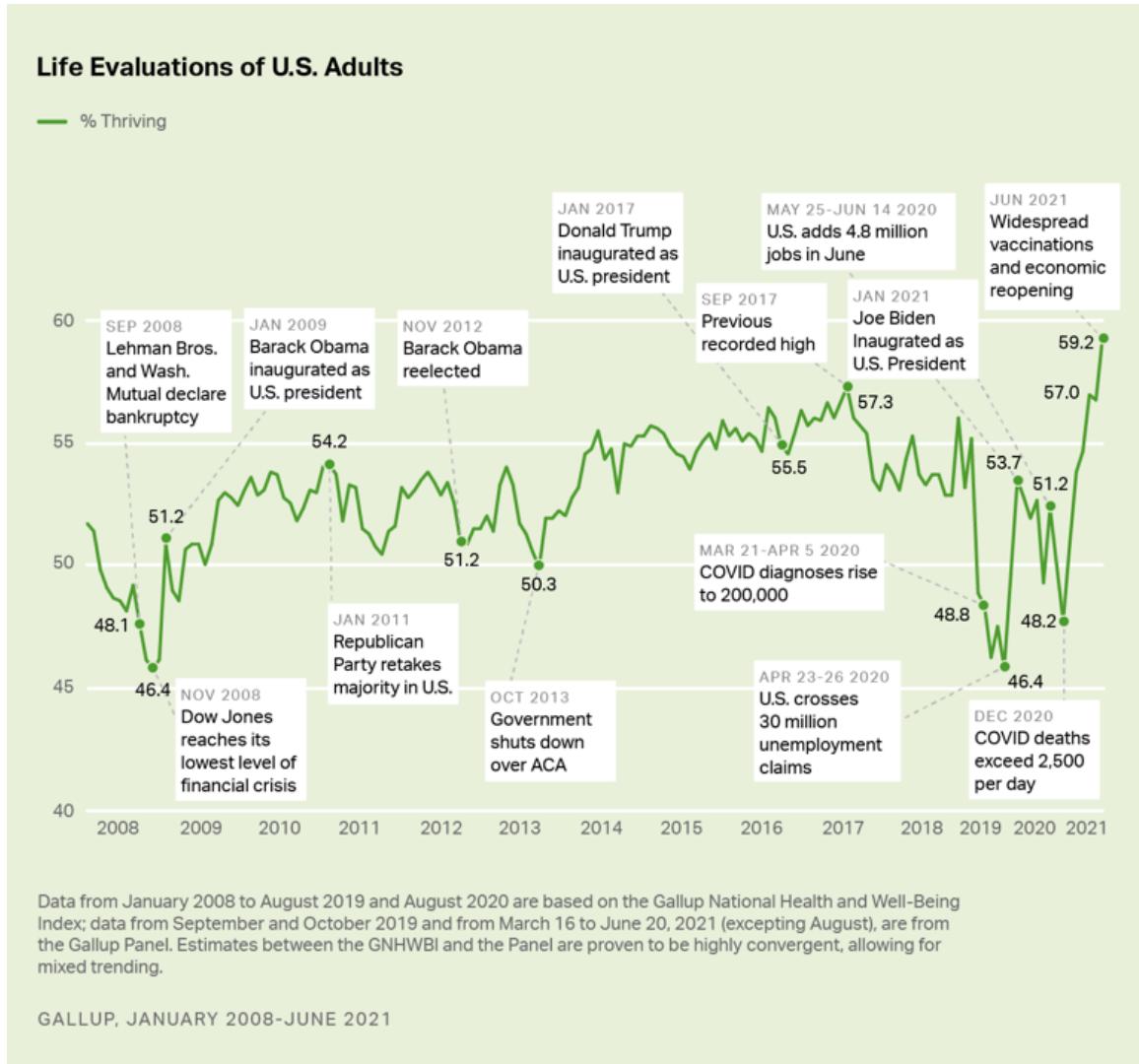
[Nate Silver points us to the news about American life satisfaction](#). It's up!



Nate Silver @NateSilver538 · 20h

My theory here is COVID taught people how much they value everyday activities that they had to give up during the worst phases of the pandemic, and it will probably be harder to get people to give those up in the future if new waves/future pandemics hit.

...



[More surveys at the original post](#), including that worry is down to pre-pandemic levels, but daily enjoyment still lags behind. My top takeaway here is that tough times make people better appreciate life, and that we all went through some tough times and then things improved, and we now have comparison points that make me feel better. Daily life scores are still lagging, so the 'life is actually better now' hypothesis doesn't fit the data. It also doesn't fit my observations otherwise, as life in general is pretty great but it's still not fully back to the old baseline.

[MR links to this LW post on the recent mask wearing studies, titled 'we still don't know if masks work.'](#) I agree with its finding that the study in question didn't prove anything in any direction, but that doesn't mean we don't know if masks work, because we are allowed to know things that didn't come from a Properly Done and Formatted Scientific Study.

[Evidence that some people's defenses against Covid kick in without causing antibody tests to come back positive.](#) For now I don't think this has much practical impact, since proving that something happens and showing that it happens frequently enough to matter are very different things.

[Lab leak: Confirmation that many scientists will affirm its plausibility privately but not publicly.](#) I continue to skip over most lab-related items, and the situation hasn't changed.

NYTimes so no link, but MR points us to a birthday paper. The key finding is that your risk of Covid goes up when you have a birthday, presumably because you have a birthday party, and the effect size cares not whether you're blue tribe or red tribe, implying their choices of party were remarkably similar.



Margot Sanger-Katz  @sangerkatz · Jul 6

But there was one detail in the research that really stuck with me. Overall, a birthday was associated with a jump in covid risk. And that increase was the same in Republican-leaning areas and Democratic-leaning areas.

2

7

23



Margot Sanger-Katz  @sangerkatz · Jul 6

We know from lots of other research about the virus that there have been big partisan differences in public behavior. Mask wearing was more widespread in Democratic-leaning areas, for example.



Margot Sanger-Katz  @sangerkatz · Jul 6

The implication of the birthday paper is that, at the height of the pandemic, there were not very big differences in private behavior according to party, even if public behavior was different.

Not Covid

In not-ready-for-prime-time-players AI news, [IBM Watson has decided to start making tennis predictions](#). They are, shall we say, not so great.

IBM POWER RANKINGS  UPDATED: MON 23:10 BST

 Novak Djokovic



Likelihood to Win

1
1

IBM POWER RANK
ATP SINGLES RANK

 Marton Fucsovics



14
48

IBM POWER RANKINGS LEADERBOARD



That's not how this works. That's not how any of this works. Djokovic absolutely does not lose to Garin 43% of the time, or anything of the sort. You don't need to know much about tennis to know that the world #1 is going to dominate the world #8 most of the time, yet Watson can't figure this out, despite the training data providing more than enough information to solve this puzzle.

That would be fine if quite odd, if this was a little experiment IBM had privately done and then said 'whoops, I wonder why that didn't work.' Instead, it's publishing the output as if it's the next iteration of tennis analysis. I don't blame Watson, it's a computer program. I blame the humans, who took a problem that AI should be quite good at, clearly botched it at the most basic level, then *didn't notice they'd botched it somehow*.

That's the part I don't understand. I've gotten similarly nonsensical analytical results plenty of times, but when I do I say the whoops thing and learn about how to make better predictions. To paraphrase Seth, if they think this is publishable tennis analysis, keep Watson the hell away from any and all medical care, or it will end quite badly for all concerned.

The other part I don't understand is *how* they got this kind of answer, given what I know about machine learning, and I'd love to hear a plausible gears-level theory of how this happened, even *ignoring that the humans should catch it within five seconds*. How did the algorithm get one of the most normal basically-in-sample situations so wrong?

If one is curious, [my friend Seth offers this tracking of how it would be doing gambling](#), which can best be summarized as 'lighting money on fire.'

Book Review: Order Without Law

This review [originally appeared](#) on the blog [Astral Codex Ten](#) as part of a contest. You can now read it here, ~~with working footnotes~~. (LW note: footnotes didn't transfer; they work [on my blog](#).) There's even an [audio version](#) as part of the ACX podcast; the footnotes are all read out at the end.

Shasta County

Shasta County, northern California, is a rural area home to many cattle ranchers.¹ It has an unusual legal feature: its rangeland can be designated as either *open* or *closed*. (Most places in the country pick one or the other.) The county board of supervisors has the power to close range, but not to open it. When a range closure petition is circulated, the cattlemen have strong opinions about it. They like their range open.

If you ask why, they'll tell you it's because of what happens if a motorist hits one of their herd. In open range, the driver should have been more careful; "the motorist buys the cow". In closed range, the rancher should have been sure to fence his animals in; he compensates the motorist.

They are simply wrong about this. Range designation has no legal effect on what happens when a motorist hits a cow. (Or, maybe not quite *no* effect. There's some, mostly theoretical, reason to think it might make a small difference. But certainly the ranchers exaggerate it.) When these cases go to court, ranchers either settle or lose, and complain that lawyers don't understand the law.

Even if they were right about the law, they have insurance for such matters. They'll tell you that their insurance premiums will rise if the range closes, but insurers don't adjust their rates on that level of granularity. One major insurer doesn't even adjust its rates between Shasta County and other counties in California. They might plausibly want to increase their coverage amount, but the cost of that is on the order of \$10/year.

No, the actual effect that range designation has, is on what happens when a rancher's cow accidentally trespasses on someone else's land. In closed range, the owner is responsible for fencing in his cattle. If they trespass on someone else's land, he's [strictly liable](#) for any damage they cause. In open range, the landowner is responsible for fencing the cattle out; the cattle owner is only liable for damages if the land was entirely fenced or if he took them there deliberately. (Law enforcement also has more power to impound cattle in closed range, but most years they don't do that even once.)

The cattlemen mostly don't understand this detail of the law. They have a vague grasp of it, but it's even more simplified than the version I've just given. And they don't act upon it. Regardless of range designation, they follow an informal code of neighborliness. According to them, it's unneighborly to deliberately allow your cattle to trespass; but it's also unneighborly to make a fuss when it does happen. The usual response is to call the owner (whom you identify by brand) and let him know. He'll thank you, apologize, and drive down to collect it. You don't ask for compensation.

Or, sometimes it would be inconvenient for him to collect it. If his cow has joined your herd, it's simpler for it just to stay there until you round them up. In that case, you'll be feeding someone else's cow, possibly for months. The expense of that is perhaps \$100, a notable amount, but you still don't ask for compensation.

Sometimes a rancher will fail to live up to this standard of neighborliness. He'll be careless about fencing in his cattle, or slow to pick them up. Usually the victims will gossip about him, and that's enough to provoke an apology. If not, they get tougher. They may drive a cow to somewhere it would be inconvenient to collect - this is questionably legal. They might threaten to injure or kill the animal. They might actually injure or kill it - this is certainly illegal, but they won't get in trouble for it.

They almost never ask for money, and lawyers only get involved in the most exceptional circumstances (the author found two instances of that happening). When someone does need to pay a debt, he does so in kind: "Should your goat happen to eat your neighbor's tomatoes, the neighborly thing for you to do would be to help replant the tomatoes; a transfer of money would be too cold and too impersonal."² Ranchers do keep rough mental account of debits and credits, but they allow these to be settled long term and over multiple fronts. A debt of "he refused to help with our mutual fence" might be paid with "but he did look after my place while I was on holiday".

(This is how ranchers deal with each other. Ranchette³ owners will also sometimes complain to public officials, who in turn talk to the cattle owner. They'll sometimes file damage claims against the rancher's insurance. It's ranchette owners who are responsible for range closure petitions.)

Range designation also doesn't affect the legal rules around building and maintaining fences. But it does change the meaning of the fences themselves, so maybe it would change how cattlemen handle fencing? But again, no. Legally, in some situations neighbors are required to share fence maintenance duties, and sometimes someone can build a fence and later force his neighbor to pay some of the cost. The cattlemen don't generally know this, and would ignore it if they did. They maintain fences unilaterally; if one of them doesn't do any work for years, the other will complain at them. If they want to build or upgrade a fence, they'll talk to their neighbor in advance, and usually figure out between them a rough way to split the material costs and labor in proportion to how many cattle each has near the fence. (Crop farmers aren't asked to pay to keep the ranchers' animals out.) Occasionally they can't reach an agreement, but this doesn't cause much animosity. This is despite that fences cost thousands of dollars per mile to build, and half a person-day per mile per year to maintain.

So this is a puzzle. Range designation is legally relevant with regard to cattle trespass, but it doesn't change how ranchers act in that regard. Range designation is not legally relevant to motor accidents, and ranchers have no reason to think it is; but that's why they ostensibly care about it.

(And it's not just words. Many of them act on their beliefs. We can roughly divide cattlemen into "traditionalists who don't irrigate and can't afford fences" and "modernists who irrigate and already use fences" - by improving pasture, irrigation vastly decreases the amount of land needed. After a closure, traditionalists drop their grazing leases in the area. Modernists oppose closures like traditionalists, but they don't react to them if they pass.)

What's up with this? Why do the cattlemen continue to be so wrong in the face of, you know, everything?

What's up with this?

Order Without Law: How Neighbors Settle Disputes is a study of, well, its subtitle. The author, Robert Ellickson, is a professor and legal scholar. He comes across as a low-key anarchist, and I've seen him quoted at length on some anarchist websites, and I wouldn't be surprised to learn that he's just a full-blown anarchist. He doesn't identify as one explicitly, at least not here, and he does respect what states bring to the table. He just wishes people would remember that they're not the only game in town. Part of the thesis of the book could be summed up (in my words, not his) as: *we credit the government with creating public order, but if you look, it turns out that people create plenty of public order that has basically nothing to do with the legal system.*

Sometimes there is no relevant law, sometimes the order predates the law, and sometimes the order ignores the law. More on this later.

Part one is an in-depth exploration of Shasta County that I found fascinating, and that I've only given in very brief summary. He goes into much more detail about basically everything.⁴

One oversight is that it's not clear to me how large the population Ellickson studied is. Given that it's a case study for questions of groups maintaining order, I think the size of the group matters a lot. For example, according to wikipedia on [Dunbar's number](#): "Proponents assert that numbers larger than this generally require more restrictive rules, laws, and enforced norms to maintain a stable, cohesive group. It has been proposed to lie between 100 and 250, with a commonly used value of 150."

Does Shasta County support that? I think not, but it's hard to say. Ellickson admittedly doesn't know the population size of the area he studied. (It's a small part of a [census division](#) whose population was 6,784 in 1980, so that's an upper bound.) But I feel like he could have been a lot more helpful. Roughly how many ranchers are there, how many ranchette owners, and how many farmers? (I think most of the relevant people are in one of those groups. I'm not sure to what extent we should count families as units. I'm not sure how many people in the area are in none of those groups.) Overall I'd guess we're looking at perhaps 300-1000 people over perhaps 100-300 families, but I'm not confident.

(I tracked down the minutes of the Shasta County Cattlemen's Association, and they had 128 members in [June 2011](#). I think "most ranchers are in the Association but ranchette owners and farmers generally aren't" is probably a decent guess. But that's over twenty years later, so who knows what changed in that time.)

Near the end of part one, Ellickson poses the "what's up with this?" question. Why are the cattlemen so wrong about what range designation means?

His answer is that it's about symbolism. Cattlemen like to think of themselves as being highly regarded in society. But as Shasta County urbanizes, that position is threatened. A closure petition is symbolic of that threat. Open range gives cattlemen more formal rights, even if they don't take advantage of them. It marks them as an important group of people, given deference by the law. So if the range closes, that's an indication to the whole county that cattlemen aren't a priority.

They care about this sort of symbolism - partly because symbols have instrumental value, but also just because people care about symbols inherently. But you can't admit that you care about symbols, because that shows insecurity. So you have to make the battle about something instrumental, and they develop beliefs which allow them to do so. They're fairly successful, too - there haven't been any closures since 1973.

(Though I note that Ellickson documents only one attempted closure in that time. It was triggered by a specific rogue cattleman who left the area soon after. It sounds like there may have been other petitions that Ellickson doesn't discuss, but I have no idea how many, what triggered them, or how much support they got. So maybe it's not so much that the cattlemen are successful as that no one else really cares.)

As for how they remain wrong - it simply isn't costing them enough. It costs them some amount, to be sure. It cost one couple \$100,000 when a motorist hit three cattle in open range. They didn't have enough liability insurance, and if they'd understood the law, they might have done. But the question is whether ignorant cattlemen will be driven out of work, or even just outcompeted by knowledgeable ones. This mistake isn't nearly powerful enough for that. Nor does anyone else have much incentive to educate them about what range designation actually means. So they remain uneducated on the subject.

This all seems plausible enough, though admittedly I'm fairly predisposed to the idea already. For someone who wasn't, I feel like it probably wouldn't be very convincing, and it could stand to have more depth. (Though it's not the focus of the work, so I hope they'd forgive that.) I'd be curious to know more about the couple who didn't have enough insurance - did they increase their insurance afterwards, and do they still think the motorist buys the cow? Did that case encourage anyone else to get more insurance? It seems like the sort of event that could have triggered a wide-scale shift in beliefs.

(Is this just standard stuff covered in works like the Sequences (which I've read, long ago) and *Elephant in the Brain* (which I haven't)? I'm not sure. I think it's analyzing on a different level than the [Fake Beliefs](#) sequence - that seems like more "here's what's going on in the brain of an individual" and this is more "here's what's going on in a society". Also, remember that it long predates those works.)

A counterpoint might be... these cases aren't all that common, and don't usually go to court, and when they do they're usually settled (on the advice of lawyers) instead of ruled. And "lawyers don't understand this specific part of the law" isn't all that implausible. So although the evidence Ellickson presents is overwhelming that the cattlemen are wrong, I'm not sure I can fault the cattlemen too hard for not changing their minds.

Against previous work

Part one was mostly a case study, with some theorizing. It kind of felt like it was building towards the "what's up with this?" question for part two, but instead it gave a brief answer at the end. Part two is a different style and focus: about evenly split between theorizing and several smaller case studies. We're explicitly told this is what's going to happen, but still, it's a little jarring.

Ellickson spends some time criticising previous theories and theorists of social control, which he divides broadly into two camps.

His own background is in the [law-and-economics](#) camp⁵, which studies the law and its effects in terms of economic theory. Among other things, this camp notably produced the [Coase theorem](#).⁶ But law-and-economics theorists tend to put too much emphasis on the state. Hobbes' *Leviathan* is a classic example:

Hobbes apparently saw no possibility that some nonlegal system of social control - such as the decentralized enforcement of norms - might bring about at least a modicum of order even under conditions of anarchy. (The term *anarchy* is used here in its root sense of a lack of government, rather than in its colloquial sense of a state of disorder. Only a legal centralist would equate the two.)

But Coase fell into this trap too:

Throughout his scholarly career, Coase has emphasized the capacity of individuals to work out mutually advantageous arrangements without the aid of a central coordinator. Yet in his famous article "The Problem of Social Cost," Coase fell into a line of analysis that was wholly in the Hobbesian tradition. In analyzing the effect that changes in law might have on human interactions, Coase implicitly assumed that governments have a monopoly on rulemaking functions. ... Even in the parts of his article where he took transaction costs into account, Coase failed to note that in some contexts initial rights might arise from norms generated through decentralized social processes, rather than from law.

As have others:

Max Weber and Roscoe Pound both seemingly endorsed the dubious propositions that the state has, and should have, a monopoly on the use of violent force. In fact, as both those scholars recognized elsewhere in their writings, operative rules in human societies often authorize forceful private responses to provocative conduct.

(See what I mean about coming across as a low-key anarchist?)

There's plenty of evidence refuting the extreme version of this camp. We can see that social norms often override law in people's actions. (The Norwegian Housemaid Law of 1948 imposed labor standards that were violated by the employers in almost 90% of households studied, but no lawsuits were brought under it for two years.) People often apply nonlegal sanctions, like gossip and violence. ("Donald Black, who has gathered cross-cultural evidence on violent self-help, has asserted that much of what is ordinarily classified as crime is in fact retaliatory action aimed at achieving social control.") Even specialists often don't know the law in detail as it applies to their speciality. (The "great majority" of California therapists thought the [Tarasoff](#) decision imposed stronger duties than it actually did.) And people just don't hire attorneys very often. We saw examples of all of these in Shasta County as well; part one can be seen as a challenge to the law-and-economics camp.

The other camp is law-and-society, emphasizing that the law exists as just one part in the broader scheme of things. These scholars tend to have a more realistic view of how the legal system interacts with other forms of control, but they've been reluctant to develop theory. They often just take norms as given, rather than trying to explain them. The theories they have developed are all flawed, although Ellickson thinks functionalism is on the right track. (This is the idea that norms develop which help a group to survive and prosper.) Ellickson explicitly describes part two as a "gauntlet" thrown towards law-and-society.

(Also, some law-and-society scholars go too far in the other direction, thinking that the legal system is ineffectual. They're just as mistaken. See⁷: Muslim Central Asia after the Russian Revolution; US civil rights laws in the 50s and 60s; range closure in Shasta County; "that the allocation of legal property rights in the intertidal zone affects labor productivity in the oyster industry, that the structure of workers' compensation systems influences the frequency of workplace fatalities, and that the content of medical malpractice law affects how claims are settled." [Footnotes removed.]

The hypothesis

Ellickson has his own theory of norms, which he formed after studying Shasta County. The main thrust of part two is to elaborate and defend it:

Members of a close-knit group develop and maintain norms whose content serves to maximize the aggregate welfare that members obtain in their workaday affairs with one another. ... Stated more simply, the hypothesis predicts that members of tight social groups will informally encourage each other to engage in cooperative behavior. [Emphasis original; footnotes removed.]

(He doesn't name this theory, calling it simply "the hypothesis". I admire that restraint, but I kind of wish I had a name to refer to it by.)

Ellickson makes sure to clarify and caveat the hypothesis here, so that we don't interpret it more strongly than he intends. But before looking at his clarifications, I'm going to jump ahead a little, and look at an example he uses of the hypothesis in action.

Consider the Shasta County norm that a livestock owner is strictly responsible for cattle trespass damages. The hypothesis is that this norm is welfare-maximizing. To test that, we have to compare it to alternatives. One alternative would be non-strict liability. Another would be that trespass damages are borne by the victim.

Compared to a negligence standard, strict liability requires less investigation but triggers more sanctions. (Apparently there's a "premise that strict-liability rules and negligence rules are equally effective at inducing cost-justified levels of care", but Ellickson doesn't really explain this.) In Shasta County, the sanctions have basically no transaction costs, since they're just neighbors adjusting mental accounts. So strict liability it is.

To be welfare maximizing, costs should be borne by whoever can avoid them most cheaply. In this case that's the ranchers; I'm not sure I fully buy Ellickson's argument, but I think the conclusion is probably true.⁸

So Ellickson argues that the Shasta County trespass norms support the hypothesis.⁹ He also makes a prediction here that things were different in the mid-nineteenth century. "During the early history of the state of California, irrigated pastures and ranchettes were rare, at-large cattle numerous, and motorized vehicles unknown. In addition, a century ago most rural residents were accustomed to handling livestock. Especially prior to the invention of barbed wire in 1874, the fencing of rangelands was rarely cost-justified. In those days an isolated grower of field crops in Shasta County, as one of the few persons at risk from at-large cattle, would have been *prima facie* the cheaper avoider of livestock damage to crops." And so the farmer would have been responsible for fencing animals out, and borne the costs if he failed to.

Clarifications: "close-knit" and "workaday affairs"

Before we go further, let's look at Ellickson's clarifications. It's important to know what the hypothesis doesn't say.

Ellickson emphasizes that it's descriptive, not normative; it's not a recommendation that norms should be used in preference to other forms of social control. Not all groups are close-knit; welfare isn't the only thing people might want to optimize for; and norms of cooperation within a group often come at the expense of outsiders.

He also emphasizes that a loose reading would give a much stronger version of the hypothesis than he intends. The terms "close-knit", "welfare" and "workaday affairs" are all significant here, and Ellickson explains their meanings in some depth. In order of how much I want to push back against them:

A "close-knit" group is one where "informal power is broadly distributed among group members and the information pertinent to informal control circulates easily among them." This is admittedly vague, but unavoidably so. Rural Shasta County residents are close-knit, and residents of a small remote island are even closer-knit. Patrons of a singles bar at O'Hare Airport are not. Distributed power allows group members to protect their selves and their property, and to personally enforce sanctions against those who wrong them. Information gives people reputations; it allows for punishing people who commit small wrongs against many group members, and for rewarding people who perform those punishments.

Notably, a close-knit group need not be small or exclusive. Someone can be a member of several completely nonoverlapping close-knit groups at once (coworkers, neighborhood, church). And although a small population tends to increase close-knittedness through "quality of gossip, reciprocal power, and ease of enforcement", the size itself has no effect. This is where I think it would be really nice to know how large the relevant population in Shasta County is - as the major case study of the book, it could lend a lot of weight to the idea that large populations can remain close-knit and the hypothesis continues to apply.

"Workaday affairs" means to assume that there's a preexisting set of ground rules, allowing group members to hold and trade personal property. (Which also requires, for example, rules against murder, theft and enslavement.) This is necessary because to calculate welfare, we need some way to measure peoples' values, and we can only do that if people can make voluntary exchanges. The hypothesis doesn't apply to those rules. Seems like a fair restriction.

A little more hackily, it also doesn't apply to "purely distributive" norms, like norms of charity. If you take wealth from one person and give it to another, the transfer process consumes resources and creates none, reducing aggregate welfare. (This is assuming Ellickson's strict definition of welfare, which he's explained by now but I haven't. Sorry.) But clearly norms of charity do exist. There are theories under which they do enhance welfare (through social insurance, or reciprocity). But those might be too simplistic, so Ellickson thinks it prudent to just exclude charity from the hypothesis.

Actually, he goes further than that. He cites Mitch Polinsky (*An Introduction to Law and Economics*) arguing that for a legal system, the cheapest way to redistribute wealth is (typically) through tax and welfare programs. And so, Polinsky argues, most legal doctrine should be shaped by efficiency concerns, not redistribution. That is, areas like tort and contract law should focus on maximizing aggregate welfare. In a dispute

between a rich and a poor person, we shouldn't consider questions like "okay, but the poor person has much more use for the money". In such disputes we should assume the same amount of wealth has equal value whoever's hands it's in, and the point is just to maximize total wealth. Then, if we end up with people having too little wealth, we have a separate welfare system set up to solve that problem.

I can buy that. Ellickson doesn't actually present the argument himself, just says that Polinsky's explained it lucidly, but sure. Stipulated.

Ellickson assumes that the same argument holds for norms as it does for law. Not only that, he assumes that norm-makers subscribe to that argument.¹⁰ That... seems like a stretch.

But granted that assumption, norms would follow a similar pattern: most norms don't try to be redistributive, and if redistribution is necessary, there would be norms specifically for that. For example, the hypothesis predicts "that a poor person would not be excused from a general social obligation to supervise cattle, and that a rich person would not, on account of his wealth, have greater fencing obligations."

That seems entirely reasonable to me, and it's consistent with Shasta County practice. And actually, I don't think we need the strong assumption to get this kind of pattern? It's the kind of thing that plausibly could happen through local dynamics. I would have been happy if Ellickson had just assumed the result, not any particular cause for it. This is a fairly minor criticism though.

(It's a little weird. Normally I expect people try to sneak in strong assumptions that are necessary for their arguments. Ellickson is explicitly flagging a strong assumption that isn't necessary.)

(I'm not sure the words "workaday affairs" was the best way to point at these restrictions. I think I see where he's coming from, but the name doesn't hook into the concept very well for me. But that's minor too.)

Clarification: "Welfare"

This gets its own section because apparently I have a lot to say about it.

The point of "welfare" maximization is to avoid subjectivity problems with utility maximization. I can work to satisfy my own preferences because I know what they are. But I don't have direct access to others' preferences, so I can't measure their utility and I can't work to maximize it.

In Economics, the concepts of [Pareto efficiency](#) and [Kaldor-Hicks efficiency](#) both work with subjective valuations, people can just decide whether a particular change would make them better off or not. That works fine for people making decisions for themselves or voluntary agreements with others.

But third-party controllers are making rules that bind people who don't consent. They're making tradeoffs for people who don't get to veto them. And they can't read minds, so they don't know people's subjective utilities.

They could try to measure subjective utilities. Market prices are a thing - but at best, they only give the subjective preferences of marginal buyers and sellers. (That is, if I buy a loaf of bread for \$1, I might still buy it for \$2 and the seller might still sell it for

\$0.50.) And not everything is or can be bought and sold. We can slightly improve on this with for example the concept of [shadow prices](#) but ultimately this just isn't going to work.

(Ellickson doesn't consider just asking people for their preferences. But that obviously doesn't work either because people can lie.)

And so third-party controllers need to act without access to people's subjective preferences, and make rules that don't reference them. Welfare serves as a crude but objective proxy to utility.

We can estimate welfare by using market prices, and looking at voluntary exchanges people have made. (Which is part of the reason for the "workaday affairs" restriction.) When a fence-maintenance credit is used to forgive a looking-after-my-house debit, that tells us something about how much one particular person values those things. This process is "sketchy and inexact", and we just admitted it doesn't give us subjective utilities - but that doesn't mean we can do any better than that.

To be clear, welfare doesn't just count material goods. Anything people might value is included, "such as parenthood, leisure, good health, high social status, and close personal relationships." Ellickson sometimes uses the word "wealth", and while he's not explicit about it, I take that to be the material component of welfare.

What welfare doesn't consider, as I understand it, is *personal* valuations of things. That is, for any given thing, its value is assumed to be the same for every member of society. "As a matter of personal ethics, you can aspire to do unto others as you would have them do unto you. Because norm-makers don't know your subjective preferences, they can only ask you to do unto others as you would want to have done unto you if you were an ordinary person."

Ellickson doesn't give examples of what this means, so I'll have to try myself. In Shasta County, there's a norm of not getting too upset when someone else's cattle trespass on your land, provided they're not egregious about it. So I think it's safe to suppose that the objective loss in welfare from cattle trespass in Shasta County is low. Suppose, by some quirk of psychology, you found cattle trespass really unusually upsetting. Or maybe you have a particular patch of grass that has sentimental value to you. Cattle trespass would harm your utility a lot, but your welfare only a little - no more than anyone else's - and you'd still be bound by this norm. But if you had an objective reason to dislike cattle trespass more - perhaps because you grow an unusually valuable crop - then your welfare would be harmed more than normal. And so norms might be different. One Shasta County rancher reported that he felt more responsibility than normal to maintain a fence with a neighbor growing alfalfa.

Or consider noisiness and noise sensitivity. Most people get some amount of value from making noise - or maybe more accurately, from certain noisy actions. Talking on the phone, having sex, playing the drums. And most people get some amount of disvalue from hearing other people's noise. In the welfare calculus, there'd be some level of noisemaking that's objectively valued equal to some level of noise exposure. Then (according to hypothesis, in a close-knit group) norms would permit people to be that amount of noisy. If someone was noisier than that, their neighbors would be permitted to informally punish them. If a neighbor tried to punish someone less noisy than that, the neighbor would risk punishment themselves. The acceptable noise level would change depending on the time (objective), but not depending on just "I happen to be really bothered by noise" (subjective). What about "I have young children"? (Or,

"some of the inhabitants of that house are young children".) Maybe - that's an objective fact that's likely to be relevant to the welfare calculus. Or "I have a verifiably diagnosed hearing disorder"? Still maybe, but it feels less likely. In part because it's less common, and in part because it's less visible. Both of those seem like they'd make it less... accessible? salient? to whatever process calculates welfare. And if you're unusually noise sensitive and the welfare function doesn't capture that, the cost would fall on you. You could ask people to be quiet (but then you'd probably owe them a favor); or you could offer them something they value more than noise-making; or you could learn to live with it (e.g. by buying noise-cancelling headphones).

So okay. One thing I have to say is, it seems really easy to fall into a self-justifying trap here. Ellickson criticizes functionalism for this, and maybe he doesn't fall into it himself. But did you notice when I did it a couple of paragraphs up? (I noticed it fairly fast, but it wasn't originally deliberate.) I looked at the norms in Shasta County and used those to infer a welfare function. If you do that, of course you find that norms maximize welfare.

To test the hypothesis, we instead need to figure out a welfare function without looking at the norms, and then show that the norms maximize it. In Shasta County, we'd need to figure out how much people disvalue cattle trespass by looking at those parts of their related behaviour that aren't constrained by norms. For example, there seems to be no norm against putting up more fences than they currently do, so they probably disvalue (the marginal cost of cattle trespass avoided by a length of fence) less than they disvalue (the marginal cost of that length of fence).

How much freedom do we have in this process? If two researchers try it out, will they tell us similar welfare functions? If we look at the set of plausible welfare functions for a society, is the uncertainty correlated between axes? (Can we say "X is valued between \$A and \$B, Y is valued between \$C and \$D" or do we have to add "...but if Y is valued near \$C, then X is valued near B"?)

And even this kind of assumes there's no feedback from norms to the welfare function. Ellickson admits that possibility, and admits that it leads to indeterminacy, but thinks the risk is slight. (He seems to assume it would only happen if norms change the market price of a good - unlikely when the group in question is much smaller than the market.) I'm not so convinced. Suppose there's a norm of "everyone owns a gun and practices regularly". Then it's probably common for people to own super effective noise-cancelling headphones. And then they don't mind noisy neighbors so much, because they can wear headphones. That's... perhaps not quite changing the welfare function, because people still disvalue noisiness the same, they just have a tool to reduce noisiness? But it still seems important that this norm effectively reduces the cost of that tool. I dunno. (For further reading, Ellickson cites [one person](#) making this criticism and [another](#) responding to it. Both articles paywalled.)

Separately, I wish Ellickson was clearer about the sorts of things he considers acceptable for a welfare function to consider, and the sorts of calculations he considers acceptable for them to perform. Subjective information is out, sure. But from discussion in the "workaday affairs" section, it seems that "I give you a dollar" is welfare-neutral, and we don't get that result just from eliminating subjective information. We do get it if we make sure the welfare function is linear in all its inputs, but that seems silly. I think we also get it if we also eliminate *non-publicly-verifiable* information. The welfare function would be linear in dollars, because I can pretend to have more or fewer dollars than I actually do. But it wouldn't need to be linear in the

number of children I'm raising, because I can't really hide those. I feel like Ellickson may have been implicitly assuming a restriction along those lines, but I don't think he said so.

Separately again, how closely does welfare correspond to utility? A utility monster couldn't become a welfare monster; I'm not sure if that's a feature or a bug, but it suggests the two can diverge considerably. A few chapters down, Ellickson does some formal game theory where the payoffs are in welfare; is it safe to ignore the possibility of "player gets higher welfare from this quadrant, but still prefers that quadrant"? It seems inevitable that some group members' utilities will get higher weighting in the welfare function than others'; people with invisible disabilities are likely to be fucked over. Ellickson admits that welfare maximization isn't the only thing we care about, but that leaves open the question of how much we should value it at all?

Suppose Quiet Quentin is unusually sensitive to noise, and happy to wear drab clothing. Drab Debbie is unusually sensitive to loud fashion, and happy to be quiet. Each of them knows this. One day Debbie accidentally makes a normal amount of noise, that Quentin isn't (by norm) allowed to punish her for. But wearing a normally-loud shirt doesn't count as punishing her, so he does that. Debbie gets indignant, and makes another normally-loud noise in retaliation, and so on. No one is acting badly according to the welfare function, but it still seems like something's gone wrong here. Is there anything to stop this kind of thing from happening?

It feels weird to me that things like parenthood and personal relationships are a component of the welfare function. Obviously they're a large part of people's subjective utility, but with so much variance that putting an objective value on them seems far too noisy. And what does a system of norms even do with that information?

[11](#)

This one feels very out-there, but for completeness: the reason for using welfare instead of utility is that a norm can't reference people's individual preferences. Not just because they're subjective, but also because there's too many of them; "Alice can make loud noise at any time, but Bob can only make loud noise when Carol isn't home" would be far too complicated for a norm. But when people interact with people they know well, maybe subjectivity isn't a problem; maybe people get a decent handle on others' preferences. And then norms don't need to reference individual preferences, they can just tell people to take others' preferences into account. The norm could be "make loud noise if you value making noise more than others nearby value you not doing that". This feels like it wouldn't actually work at any sort of scale, and I don't fault Ellickson for not discussing it.

Despite all this, I do think there's some "there" there. A decent amount of "there", even. I think Ellickson's use of welfare should be given a long, hard look, but I think it would come out of that ordeal mostly recognizable.

Clarification: Dynamics

There's another clarification that I think is needed. The phrase "develop and maintain" is a claim about dynamics, partial derivatives, conditions at equilibrium. It's not a claim that "all norms always maximize welfare" but that "norms move in the direction of maximizing welfare".

Ellickson never says this explicitly, but I think he'd basically agree. Partly I think that because the alternative is kind of obviously ridiculous - norms don't change

immediately when conditions change. But he does also hint at it. For example, he speculates that a group of court cases around whaling arose because whalers were having trouble transitioning from one set of norms to a more utilitarian set (more on this later). Elsewhere, he presents a simple model of a society evolving over time to phase out certain rewards in favor of punishments.

Taken to an extreme, this weakens the hypothesis significantly. If someone points at a set of norms that seems obviously nonutilitarian, we can just say "yeah, well, maybe they haven't finished adapting yet". I don't think Ellickson would go that far. I think he'd say the dynamics are strong enough that he can write a 300-page book about the hypothesis, not explicitly admit that it's a hypothesis about dynamics, and it wouldn't seem all that weird.

Still, I also think this weakens the *book* significantly. When we admit that it's a hypothesis about dynamics, there's a bunch of questions we can and should ask. Probably the most obvious is "how fast/powerful are these dynamics". But there's also "what makes them faster or slower/more or less powerful" and "to what extent is the process random versus deterministic" and "how large are the second derivatives". (For those last two, consider: will norms sometimes update in such a way as to make things worse, and only on average will tend to make things better? Will norms sometimes start moving in a direction, then move too far in that direction and have to move back?) I'd be interested in "what do the intermediate states look like" and "how much do the individual personalities within the group change things".

I don't even necessarily expect Ellickson to have good answers to these questions. I just think they're important to acknowledge.

(I'd want to dock Ellickson some points here even if it didn't ultimately matter. I think "saying what we mean" is better than "saying things that are silly enough the reader will notice we probably don't mean them and figure out what we do mean instead".)

I think this is my biggest criticism of the book.

Substance

With all these clarifications weakening the hypothesis, does it still have substance?

Yes, Ellickson says. It disagrees with Hobbes and other legal centrists; with "prominent scholars such as Jon Elster who regard many norms as dysfunctional"; with Marxism, which sees norms as serving only a small subset of a group; with people who think norms are influenced by nonutilitarian considerations like justice; and with "the belief, currently ascendant in anthropology and many of the humanities, that norms are highly contingent and, to the extent that they can be rationalized at all, should be seen as mainly serving symbolic functions unrelated to people's perceptions of costs and benefits."

And it's falsifiable. We can identify norms, by looking at patterns of behavior and sanctions, and aspirational statements. And we can measure the variables affecting close-knittedness. ("For example, if three black and three white fire fighters belonging to a racially polarized union were suddenly to be adrift in a well-stocked lifeboat in the middle of the Pacific Ocean, as an objective matter the social environment of the six would have become close-knit and they would be predicted to cooperate."¹²)

But what we can't do is quantify the objective costs and benefits of various possible norm systems. So we fall back to intuitive assessments, looking at alternatives and pointing out problems they'd cause. This is not quite everything I'd hoped for from the word "falsifiable", but it'll do. Ellickson spends the next few chapters doing this sort of thing, at varying levels of abstraction but often with real-world examples. He also makes occasional concrete predictions, admitting that if those fail the hypothesis would be weakened. I'll only look at a few of his analyses.

Lying, and things like it

A common contract norm forbids people from lying about what they're trading. The hypothesis predicts we'd find such norms among any close-knit group of buyers and sellers. I bring this up for the exceptions that Ellickson allows:

Falsehoods threaten to decrease welfare because they are likely to increase others' costs of eventually obtaining accurate information. Honesty is so essential to the smooth operation of a system of communication that all close-knit societies can be expected to endeavor to make their members internalize, and hence self-enforce, norms against lying. Of course a no-fraud norm, like any broadly stated rule, is ambiguous around the edges. Norms may tolerate white lies, practical joking, and the puffing of products. By hypothesis, however, these exceptions would not permit misinformation that would be welfare threatening. The "entertaining deceivers" that anthropologists delight in finding are thus predicted not to be allowed to practice truly costly deceptions. [Footnotes removed; one mentions that "A cross-cultural study of permissible practical joking would provide a good test of the hypothesis."]

It's not clear to me why norms would allow such exceptions, which still increase costs of information and are presumably net-negative. To sketch a possible answer: the edge cases are likely to be where the value of enforcing the norm is lower. I'd roughly expect the social costs of violations to be lower, and the transaction costs of figuring out if there was a violation to be higher. (I feel like I've read a sequence of three essays arguing about [one particular case](#); they wouldn't have been necessary if the case had been a *blatant lie*.¹³) So, okay, minor violations don't get punished. But if minor violations don't get punished when they happen, then (a) you don't actually have a norm against them; and (b) to the extent that some people avoid those violations anyway, you've set up an [asshole filter](#) (that is, you're rewarding vice and punishing virtue).

So plausibly, the ideal situation is for it to be common knowledge that such things are considered fine to do. We might expect this to just push the problem one level up; so that instead of litigating minor deceptions, you're litigating slightly-less-minor deceptions. But these deceptions have a higher social cost, so more value to litigating them, so maybe it's fine.

(Aside, it's not clear to me why the hypothesis specifically expects such norms to be internalized, rather than enforced some other way. Possible answer: you do still need external enforcement of these norms, but that enforcement will be costly. It'll be cheaper if you can mostly expect people to obey them even if they don't expect to get caught, so that relies on self-enforcement. But is that a very general argument that almost all norms should be internalized? Well, maybe almost all norms are internalized. In any case, I don't think that clause was very important.)

Pandering to the SSC audience: dead whales

The second-most-detailed case study in the book is whalers. If a whale is wounded by one ship and killed by another, who keeps it? What if a dead whale under tow is lost in a storm, and found by another ship? The law eventually developed opinions on these questions, but when it did, it enshrined preexisting norms that the whalers themselves had developed.

Ellickson describes a few possible norms that wouldn't be welfare maximizing for them, and which in fact weren't used. For example, a whale might simply belong to whichever ship physically held the carcass; but that would allow one ship to wait for another to weaken a whale, then attach a stronger line and pull it in. Or it might belong to the ship that killed it; but that would often be ambiguous, and ships would have no incentive to harvest dead whales or to injure without killing. Or it might belong to whichever ship first lowered a boat to pursue it, so long as the boat remained in fresh pursuit; but that would encourage them to launch too early, and give claim to a crew who might not be best placed to take advantage of it. Or it might belong to whichever ship first had a reasonable chance of capturing it, so long as it remained in fresh pursuit; but that would be far too ambiguous.

In practice they used three different sets of norms. Two gave full ownership to one party. The "fast-fish/loose-fish" rule said that you owned a whale as long as it was physically connected to your boat or ship. The "first iron" (or "iron holds the whale") rule said that the first whaler to land a harpoon could claim a whale, as long as they remained in fresh pursuit, and as long as whoever found it hadn't started cutting in by then.

Whalers used these norms according to the fishery¹⁴ they hunted from, and each was suited to the whales usually hunted from that fishery. Right whales are weak swimmers, so they don't often escape once you've harpooned them. Fast-fish works well for hunting them. Sperm whales do often break free, and might be hunted by attaching the harpoon to a drogue, a wooden float that would tire the whale and mark its location. The concept of "fresh pursuit" makes first-iron more ambiguous than fast-fish, which isn't ideal, but it allows more effective means of hunting.

(Sperm whales also swim in schools, so ideally you want to kill a bunch of them and then come back for the corpses. If you killed a whale, you could plant a flag in it, which gave you claim for longer than a harpoon. You had to be given reasonable time to come back, and might take ownership even if the taker had started cutting in. Ellickson doesn't say explicitly, but it sounds like American whalers in the Pacific might have had this rule, but not American whalers operating from New England, for unclear reasons.)

The other was a split-ownership rule. A fishery in the Galápagos Islands split ownership 50/50 between whoever attached a drogue and whoever took the carcass. This norm gave whalers an incentive to fetter lots of them and let others harvest them later, but it's not clear how or why that fishery developed different rules than others. On the New England coast, whalers would hunt fast finback whales with bomb-lances; the whales would sink and wash up on shore days later. The killer was entitled to the carcass, less a small fee to whoever found it. This norm was binding even on people unconnected with the whaling industry, and a court upheld that in at least one case. I'm not sure how anyone knew who killed any given whale. Perhaps there just weren't enough whalers around for it to be ambiguous?

(Ellickson notes that the "50/50 split versus small fee" questions is about rules versus standards. Standards let you consider individual cases in more detail, taking into account how much each party contributed to the final outcome, and have lower deadweight losses. But rules have fewer disputes about how they should be applied, and thus lower transaction costs.)

So this is all plausibly welfare-maximizing, but that's not good enough. Ellickson admits that this sort of ex post explanation risks being "too pat". He points out two objections you could raise. First, why did the norms depend on the fishery, and not the fish? (That would have been more complicated, because there are dozens of species of whale. And you had to have your boats and harpoons ready, so you couldn't easily change your technique according to what you encountered.)

More interestingly, what about overfishing? If norms had imposed catch quotas, or protected calves and cows, they might have been able to keep their stock high. Ellickson has two answers. One is that that would have improved global welfare, but not necessarily the welfare of the current close-knit group of whalers, as they couldn't have stopped anyone else from joining the whaling business. This is a reminder that norms may be locally welfare-maximizing but globally harmful.

His other answer is... that that might not be the sort of thing that norms are good at? Which feels like a failure of the hypothesis. Here's the relevant passage:

Establishment of an appropriate quota system for whale fishing requires both a sophisticated scientific understanding of whale breeding and also an international system for monitoring worldwide catches. For a technically difficult and administratively complicated task such as this, a hierarchical organization, such as a formal trade association or a legal system, would likely outperform the diffuse social forces that make norms. Whalers who recognized the risk of overfishing thus could rationally ignore that risk when making norms on the ground that norm-makers could make no cost-justified contribution to its solution. [Footnote removed]

There's some subtlety here, like maybe he's trying to say "norms aren't particularly good at this, so if there's another plausible source of rules, norm-makers would defer to them; but if there wasn't, norm-makers would go ahead and do it themselves". That feels implausible on the face of it though, and while I'm no expert, my understanding is that no other group did step up to prevent overfishing in time.

This section is one place where Ellickson talks about the hypothesis as concerning dynamics. There are only five American court cases on this subject, and four of them involved whales caught between 1852 and 1862 in the Sea of Okhotsk; the other was an 1872 decision about a whale caught in that sea in an unstated year. Americans had been whaling for more than a century, so why did that happen? The whales in that area were bowheads, for which fast-fish may have been more utilitarian than first-iron. Ellickson speculates that "American whalers, accustomed to hunting sperm whales in the Pacific, may have had trouble making this switch."

(He does give an alternate explanation, that by that time the whaling industry was in decline and the community was becoming less close-knit. "The deviant whalers involved in the litigated cases, seeing themselves nearing their last periods of play, may have decided to defect.")

There's something that stuck out to me especially in this section, which I don't think Ellickson ever remarked upon. A lot of norms seem to bend on questions that are

unambiguous given the facts but where *the facts are unprovable*. If I take a whale that you're in fresh pursuit of, I can tell everyone that you'd lost its trail and only found me days later. Who's to know?

Well, in the case of whalers, the answer is "everyone on both of our ships". That's too many people to maintain a lie. But even where it's just one person's word against another's, this seems mostly fine. If someone has a habit of lying, that's likely to build as a reputation even if no one can prove any of the lies.

Remedies

In private (i.e. non-criminal) law, when someone is found to be deviant, the standard remedy is to award damages. That doesn't always work. They might not have the assets to make good; or they might just be willing to pay that price to disrupt someone's world. So the legal system also has the power of injunctions, requiring or forbidding certain future actions. And if someone violates an injunction, the legal system can incarcerate them.

Norms have analogous remedies. Instead of damages, one can simply adjust a mental account. Instead of an injunction, one can offer a more-or-less veiled threat. Instead of incarcerating someone, one can carry out that threat.

Incarceration itself isn't a credible threat ("kidnapping is apt both to trigger a feud and to result in a criminal prosecution"), but other forms of illegal violence are. ("Indeed, according to Donald Black, a good portion of crime is actually undertaken to exercise social control." [cite](#))

Remedial norms require a grievant to apply self-help measures in an escalating sequence. Generally it starts at *give the deviant notice of the debt*; goes through *gossip truthfully about it*; and ends with *sieze or destroy some of their assets*. Gossip can be omitted when it would be obviously pointless, such as against outsiders. This is consistent with the hypothesis, since the less destructive remedies come first in the sequence. It's also consistent with practice in Shasta County, and we see it as well in the lobstermen of Maine when someone sets a trap in someone else's territory. They'll start by attaching a warning to a trap, sometimes sabotaging it without damaging it. If that doesn't work, they destroy the trap. They don't seem to use gossip, perhaps because they can't identify the intruder or aren't close-knit with him.

"Sieze or destroy" - which should you do? Destroying assets is a deadweight loss, so it might seem that seizing them would be better for total welfare. But destruction has advantages too. Mainly, it's more obviously punitive, and so less likely to be seen as aggression and to lead to a feud. The Shasta County practice of driving a cow somewhere inconvenient isn't something you'd do for personal gain. But also, it's easier to calibrate (you can't seize part of a cow, but you can wound it instead of killing). And it can be done surreptitiously, which is sometimes desired (though open punishment is usually preferred, to maintain public records).

Incentives all the way up

We don't have a good understanding of how norms work to provide order. But the key is "altruistic" norm enforcement by third parties. (Those are Ellickson's scare quotes, not mine.) How do we reconcile that with the assumption of self-interested behavior?

One possibility is internalized norms, where we feel guilty if we fail to act to enforce norms, or self-satisfied if we do act. (I feel like this is stretching the idea of self-interest, but then we can just say we reject that assumption, so whatever.)

Another is that the seemingly altruistic enforcers are themselves motivated by incentives supplied by other third parties. This seems to have an infinite regress. Ellickson gives as example a young man who tackled a pickpocket to retrieve someone's wallet. The woman he helped wrote into the New York Times to publicly thank him, so there's his incentive. But we also need incentives for her to write that letter, and for the editor to publish it, and so on.

(I'm not actually entirely sure where "so on" goes. I guess we also need incentive for people to read letters like that. Though according to [Yudkowsky's Law of Ultrafinite Recursion](#) there's no need to go further than the editor.)

This infinite regress seems bad for the chances of informal cooperation. But it might actually help. Ellickson's not entirely detailed about his argument here, so I might be filling in the blanks a bit, but here's what I think he's going for. Suppose there's a virtuous third-party enforcer "at the highest level of social control". That is, someone who acts on every level of the infinite regress. They'll sanction primary behavior as appropriate to enforce norms; but also sanction the people who enforce (or fail to enforce) those norms themselves; and the people who enforce (or fail to enforce) the enforcement of those norms; and so on, if "so on" exists.

Then that enforcer could create "incentives for cooperative activity that cascade down and ultimately produce welfare-maximizing primary behavior." They don't need to do all the enforcement themselves, but by performing enforcement on every level, they encourage others to perform enforcement on every level.

This might work even with just the perception of such an enforcer. God could be this figure, but so could "a critical mass of self-disciplined elders or other good citizens, known to be committed to the cause of cooperation". Art and literature could help too.

Anarchy in academia

Academia seems to have a disproportionate number of legal centrists. So you might think professors would be unusually law-abiding. Not when it comes to photocopying. The law says how they should go about copying materials for use in class: fair-use doctrine is quite restrictive unless they get explicit permission, which can be slow to obtain¹⁵. Professors decide they don't really like this, and they substitute their own set of norms.

The Association of American Publishers tells us that campus copyright violation is (Ellickson quotes) "widespread, flagrant, and egregious". They seem to be right. Ellickson asked law professors directly, and almost all admit to doing it - though not for major portions of books. The managers of law school copy rooms don't try to enforce the rules, they let the professors police themselves. Several commercial copy centres made multiple copies for him of an article from a professional journal. "I have overheard a staff member of a copy center tell a patron that copyright laws prevented him from photocopying more than 10 percent of a book presented as a hardcopy original; the patron then asked whether he himself could use the copy center's equipment to accomplish that task and was told that he could."¹⁶

So professors' norms seem to permit illegal repeated copying of articles and minor parts of books. That lets them avoid knowing fair-use doctrine in detail. And since the law would require them to write (and respond to) many requests for consent, it lets them avoid that too.

Professors sense that Congress is unlikely to make welfare-maximizing copyright law. (Publishers can hire better lobbyists than they can.) This lets them frame their norms as principled subversion. I'm not sure if it's particularly relevant though - if copyright law was welfare-maximizing overall, but not for the professors, I think the hypothesis would still predict them to develop their own norms. But thinking back to the stuff on symbolism, maybe "being able to frame your actions as principled subversion" is a component of welfare.

Why will they copy articles, but not large portions of books? Authors of articles don't get paid much for them, and for no charge will mail reprints to colleagues and allow excerpts to be included in compilations. "It appears that most academic authors are so eager for readers to know and cite their work that they usually regard a royalty of zero or even less as perfectly acceptable. For them small-scale copying is not a misappropriation but a service." But book authors do receive royalties, and large-scale copying would diminish those significantly. So according to the hypothesis, this restraint comes from wanting to protect author-professors' royalty incomes, not from caring about publishers' and booksellers' revenues. (Though they might start to care about those, if they thought there might be a shortage of publishers and booksellers. They also might care more about university-affiliated publishers and booksellers.)

(There's a question that comes to mind here, that Ellickson doesn't bring up. Why do professors decline to copy books *written by non-academics*? I can think of a few answers that all seem plausible: that this is a simpler norm; that it's not necessarily clear who is and isn't an academic; and that it makes it easier to sell the "principled subversion" thing.)

Notably, in the two leading cases around academic copying, the plaintiffs were publishers and the primary defendant was an off-campus copy center. This is consistent with the hypothesis. In these situations, those two parties have the most distant relationship. Publishers have no use for copy centers, and copy centers don't buy many books, so neither of them has informal power over the other. Even more notably, in one of these cases, university-run copy centers weren't included as defendants - that might anger the professors, who do have power over publishers.

Counterexamples?

But Ellickson admits that all of this could be cherry-picking. So he looks at two well-known cases that he expects people to point to as counterexamples. (I hadn't heard of either of them, so I can't rule out that he's cherry-picking here, too. But I don't expect it.)

The first is the [Ik](#) of northern Uganda. These are a once-nomadic tribe with a few thousand members. Colin Turnbull found an unsettling pattern of inhumanity among them. Parents were indifferent to the welfare of their children after infancy, and people took delight in others' suffering. In Turnbull's words: "men would watch a child with eager anticipation as it crawled toward the fire, then burst into gay and happy laughter as it plunged a skinny hand into the coals. ... Anyone falling down was good for a laugh too, particularly if he was old or weak or blind."

Ellickson replies that the Ik were "literally starving to death" at the time of Turnbull's visit. A few years prior, their traditional hunting ground had been turned into a national park, and now they were forced to survive by farming a drought-plagued area. (Turnbull "briefly presented these facts" but didn't emphasize them.) "Previously cooperative in hunting, the Ik became increasingly inhumane as they starved. Rather than undermining the hypothesis, the tragic story of the Ik thus actually supports the hypothesis' stress on close-knittedness: cooperation among the Ik withered only as their prospects for continuing relationships ebbed." [Footnote removed.]

I note that Wikipedia disputes this account. "[Turnbull] seems to have misrepresented the Ik by describing them as traditionally being hunters and gatherers forced by circumstance to become farmers, when there is ample linguistic and cultural evidence that the Ik were farmers long before they were displaced from their hunting grounds after the formation of Kidepo National Park - the event that Turnbull says forced the Ik to become farmers." To the extent that Ellickson's reply relies on this change in circumstances, it apparently (according to Wikipedia) falls short. But perhaps the important detail isn't that they switched from hunting to farming, but that they switched from "not literally starving to death" to "literally starving to death" (because of a recent drought).

Ellickson also cites (among others) Peter Singer as criticising Turnbull in [The Expanding Circle](#), pp 24-26. Looking it up, Singer points out that, even if we take Turnbull's account at face value, Ik society retains an ethical code.

Turnbull refers to disputes over the theft of berries which reveal that, although stealing takes place, the Ik retain notions of private property and the wrongness of theft. Turnbull mentions the Ik's attachment to the mountains and the reverence with which they speak of Mount Morungole, which seems to be a sacred place for them. He observes that the Ik like to sit together in groups and insist on living together in villages. He describes a code that has to be followed by an Ik husband who intends to beat his wife, a code that gives the wife a chance to leave first. He reports that the obligations of a pact of mutual assistance known as *nyot* are invariably carried out. He tells us that there is a strict prohibition on Ik killing each other or even drawing blood. The Ik may let each other starve, but they apparently do not think of other Ik as they think of any non-human animals they find - that is, as potential food. A normal well-fed reader will take the prohibition of cannibalism for granted, but under the circumstances in which the Ik were living human flesh would have been a great boost to the diets of stronger Ik; that they refrain from this source of food is an example of the continuing strength of their ethical code despite the crumbling of almost everything that had made their lives worth living.

This seems to support the hypothesis too. I do think there's some [tension](#) between these two defenses. Roughly: their circumstances made them the way they were; and anyway, they weren't that way after all. But they don't seem quite contradictory.

The other potential counterexample is the peasants in Mentegrano, a southern Italian village, as studied by Edward Banfield.

Banfield found no horrors as graphic as [those of the Ik], but concluded that the Italian peasants he studied were practitioners of what he called "amoral familialism," a moral code that asked its adherents to "maximize the material, short-run advantage of the nuclear family; assume all others will do likewise." According to Banfield, this attitude hindered cooperation among families and

helped keep the villagers mired in poverty. [One footnote removed; minor style editing.]

Ellickson has two replies here. Firstly, the evidence is arguably consistent with the hypothesis: some of Banfield's reviewers suggested that, going by Banfield's evidence, the villagers had adapted as well as possible to their environment. Secondly, Banfield's evidence often seems to contradict Banfield's thesis: neighbors have good relationships and reciprocate favors. Banfield apparently discounted that because they did so out of self-interest, but it's still compatible with the hypothesis.

(I don't think these replies are in the same kind of tension.)

For a more general possible counterexample, Ellickson points at primitive tribes believing in magic and engaging in brutal rites. (This is something I did have in my mind while reading, so I'm glad he addressed it.) Some anthropologists are good at finding utilitarian explanations for such things, but Ellickson rejects that answer. Instead, he simply predicts that these practices would become abandoned as the tribe becomes better educated. "A tribe that used to turn to rain dancing during droughts thus is predicted to phase out that ritual after tribe members learn more meteorology. Tribes are predicted to abandon dangerous puberty rites after members obtain better medical information. As tribe members become more familiar with science in general, the status of their magicians and witch doctors should fall. As a more contemporary example, faith in astrology should correlate negatively with knowledge of astronomy. These propositions are potentially falsifiable."

This was my guess as to an answer before I reached this part of the book, which I think says good things about both myself and the book. And I basically agree with his prediction. But I also think it's not entirely satisfactory.

It seems like we need to add a caveat to the hypothesis for this kind of thing, "if people believe that rain dances bring rain, then norms will encourage rain dances". And I kind of want to say that's fair enough, you can't expect norms to be smarter than people. But on the other hand, I think the thesis of [The Secret of Our Success](#) and the like is that actually, that's exactly what you can expect norms to be. And it seems like a significant weakening of the hypothesis - do we now only predict norms to optimize in ways that group members understand? Or to optimize not for welfare but for "what group members predict their future welfare will be"? I dunno, and that's a bad sign. But if the hypothesis doesn't lose points for rain dances, it probably shouldn't gain points for manioc. (Though as Ben Hoffman [points out](#), the cost-benefit of manioc processing isn't immediately obvious. Maybe the hypothesis should lose points for both manioc and rain dances.)

If a ritual is cheap to implement, I'd be inclined to give it a pass. There's costs of obtaining information, and that could apply to whatever process develops norms like it does to individuals. Plus, it would only take a small benefit to be welfare-maximizing, and small benefits are probably less obvious than larger ones. (Though if that's what's going on, it's not clear whether we should expect education to phase the rituals out.)

But for vicious and dangerous rituals, this doesn't seem sufficient. Ellickson mentions a tribe where they "cut a finger from the hand of each of a man's close female relatives after he dies"; what medical knowledge are they lacking that makes this seem welfare-maximizing?

I think this is my biggest criticism of the hypothesis.

Another possible counterexample worth considering would be [Jonestown](#), and cults in general. (h/t whoever it was brought this to my attention.) I don't feel like I know enough about these to comment, but I'm going to anyway. I wonder if part of what's going on is that cults effectively don't have the rule of law - they make it costly for you to leave, or to bring in outside enforcers, and so you can't really enforce your property rights or bodily autonomy. If so, it seems like the "workaday" assumption is violated, and the hypothesis isn't in play.

Or, what about dueling traditions? We might again say the "workaday" assumption (that brings rules against murder) is violated, but that seems like a cheat. My vague understanding, at least of pistol dueling as seen in *Hamilton*, is it was less lethal than we might expect; and fell out of favor when better guns made it more lethal. But neither of these feels enough to satisfy, and we should demand satisfaction. Did the group gain something that was worth the potential loss of life? Alternatively, were such things only ever a transition phase?

Formal game theory

Something I haven't touched on is Ellickson's use of formal game theory. To do justice to that section, I split it into [its own essay](#). The tl;dr is that I think he handled it reasonably well, with forgiveable blind spots but not outright mistakes that I noticed. I don't feel like I need to discount the rest of the book (on subjects I know less well) based on his treatment of game theory.

Summing up

Is this a good book? Yes, very much so. I found it fascinating both on the level of details and the level of ideas. Ellickson is fairly readable, and occasionally has a dry turn of phrase that I love. ("A drowning baby has neither the time nor the capacity to contract for a rescue.") And I don't know if this came across in my review, but he's an unusually careful thinker. He owns up to weaknesses. He rejects bad arguments in favor of his position. He'll make a claim and then offer citations of people disagreeing. He makes predictions and admits that if they fail the hypothesis will be weakened. I think he made some mistakes, and I think his argument could have been clearer in places, but overall I'm impressed with his ability to think.

Is the hypothesis true? I... don't think so, but if we add one more caveat, then maybe.

The hypothesis says that norms maximize welfare. Note that although Ellickson calls the welfare function "objective", I think a better word might be "intersubjective". The welfare function is just, like, some amorphous structure that factors out when you look at the minds of group members. Except we can't look at their minds, we have to look at behaviour. The same is true of norms themselves: to figure out what the norms are in a society we ultimately just have to look at how people in that society behave.

And so if we're to evaluate the hypothesis properly, I think we need to: look at certain types of behaviour, and infer something that's reasonable to call "norms"; and then look at non-normative behavior - the behaviour that the inferred system of norms doesn't dictate - and infer something that's reasonable to call a "welfare function". And then the hypothesis is that the set of norms will maximize the welfare function. ("Maximize over what?" I almost forgot to ask. I think, maximize over possible systems of norms that might have been inferred from plausible observed behaviour?)

Put like that it sounds kind of impossible. I suspect it's... not too hard to do an okay job? Like I'd guess that if we tried to do this, we'd be able to find things that we'd call "norms" and "a welfare function" that mostly fit and are only a little bit circular; and we wouldn't have an overabundance of choice around where we draw the lines; and we could test the hypothesis on them and the hypothesis would mostly come out looking okay.

But to the extent that we can *only* do "okay" - to the extent that doing this right is just fundamentally hard - I suspect we'll find that the hypothesis also fails.

There are problems which are known to be [fundamentally hard in important ways](#), and we can't program a computer to reliably solve them. Sometimes people say that [slime molds](#) have solved them and this means something about the ineffable power of nature. But they're wrong. The slime molds haven't solved anything we can't program a computer to solve, because we can program a computer to emulate the slime molds.¹⁷ What happens is that the slime molds have found a pretty decent approach to solving the problem, that usually works under conditions the slime molds usually encounter. But the slime molds will get the wrong answer too, if the specific instance of the problem is pathological in certain ways.

In this analogy, human behavior is a slime mold. It changes according to rules evaluated on local conditions. (Ellickson sometimes talks about "norm-makers" as though they're agents, but that feels like anthropomorphising. I expect only a minority of norms will have come about through some agentic process.) It might be that, in doing so, it often manages to find pretty good global solutions to hard problems, and this will look like norms maximizing welfare. But when things are set up right, there'd be another, better solution.

(I'm not sure I've got this quite right, but I don't think Ellickson has, either.)

So I want to add a caveat acknowledging that sort of thing. I don't know how to put it succinctly. I suspect that simply changing "maximize" in the hypothesis to "locally maximize" weakens it too far, but I dunno.

With this additional caveat, is the hypothesis true? I still wouldn't confidently say "yes", for a few reasons. My main inside-view objections are the ritual stuff and duelling, but there's also the outside-view "this is a complicated domain that I don't know well". (I only thought of duelling in a late draft of this review; how many similar things are there that I still haven't thought of?) But it does feel to me like a good rule of thumb, at least, and I wouldn't be surprised if it's stronger than that.

Further questions

I want to finish up with some further questions.

- The world seems to be getting more atomic, with less social force being applied to people. Does that result in more legal force? Ellickson gives a brief answer: "In [Donald Black's] view, the state has recently risen in importance as lawmakers have striven to fill the void created by the decline of the family, clan, and village." (Also: "increasing urbanization, the spread of liability insurance, and the advent of the welfare state". But Black speculates it'll decline in future because of increasing equality. And although the number of lawyers in the US has increased, litigation between individuals other than divorce remains "surprisingly rare".)

- Can I apply this to my previous thoughts on [the responsibility of open source maintainers](#)? When I try, two things come to mind. First, maintainers know more about the quality of their code than users. Thus, if we require maintainers to put reputation on the line when "selling" their code, we (partially) transfer costs of bad code to the people who best know those costs and are best able to control them. So that's a way to frame the issue that I don't remember having in mind when I wrote the post, that points in the same directions I was already looking. Cool. Second, I feel like in that post I probably neglected to think about transaction costs of figuring out whether someone was neglecting their responsibility? Which seems like an important oversight.
- To test the hypothesis, I'd be tempted to look at more traditions and see whether those are (at least plausibly) welfare-maximizing. But a caveat: are traditions enforced through norms? I'd guess mostly yes, but some may not be enforced, and some may be enforced through law. In those cases the hypothesis doesn't concern itself with them.
- Making predictions based on the hypothesis seems difficult. Saying that one set of norms will increase welfare relative to another set might be doable, but how can you be confident you've identified the best possible set? Ellickson does make predictions, and I don't feel like he's stretching too far - though I can't rule it out. But I'm not sure I'd be able to make the same predictions independently. How can we develop and test this skill?
- Sometimes a close-knit group will fracture. What sort of things cause that? What does it look like when it happens? What happens to the norms it was maintaining?
- What are some follow-up things to read? Ellickson approvingly cites [The Behavior of Law](#) a bunch. If we want skepticism, he cites [Social Norms and Economic Theory](#) a few times. At some point I ran across [Norms in a Wired World](#) which looks interesting and cites Ellickson, but that's about all I know of it.
- How does this apply to the internet? I'd note a few things. Pseudonymity (or especially anonymity) and transience will reduce close-knittedness, as you only have limited power over someone who can just abandon their identity. To the extent that people do have power, it may not be broadly distributed; on Twitter for example, I'd guess your power is roughly a function of how many followers you have, which is wildly unequal. On the other hand, public-by-default interactions increase close-knittedness. I do think that e.g. LessWrong plausibly counts as close-knit. The default sanction on reddit is voting, and it seems kind of not-great that the default sanction is so low-bandwidth. For an added kick or when it's not clear what the voting is for, someone can write a comment ("this so much"; "downvoted for..."). That comment will have more weight if it gets upvoted itself, and/or comes from a respected user, and/or gets the moderator flag attached to it. Reddit also has gilding as essentially a super-upvote. For sanctioning remedial behavior, people can comment on it ("thanks for the gold", "why is this getting downvoted?") and vote on those comments. But some places also have meta-moderation as an explicit mechanism.
- How much information is available about historical norms and the social conditions they arose in? Enough to test the hypothesis?

- There's a repeated assumption that if someone has extraordinary needs then it's welfare maximizing for the cost to fall on them instead of other people. I'm not sure how far Ellickson would endorse that; my sense is that he thinks it's a pretty good rule of thumb, but I'm not sure he ever investigates an instance of it in enough detail to tell whether it's false. It would seem to antipredict norms of disability accommodation, possibly including [curb cuts](#). (Possibly not, because those turn out to benefit lots of people. But then, curb cuts are enforced by law, not norm.) This might be a good place to look for failures of the hypothesis, but it's also highly politicized which might make it hard to get good data.
- Ellickson sometimes suggests that poor people will be more litigious because they have their legal fees paid. We should be able to check that. If it's wrong, that doesn't necessarily mean the hypothesis is wrong; there are other factors to consider, like whether poor people have time and knowledge to go to court. But it would be a point in favor of something like "the hypothesis is underspecified relative to the world", such that trying to use it to make predictions is unlikely to work.
- Is Congress close-knit? Has that changed recently? Is it a good thing for it to be close-knit? (Remember, norms maximizing the welfare of congresspeople don't necessarily maximize the welfare of citizens.)
- Does this work at a level above people? Can we (and if so, when can we) apply it to organizations like companies, charities, and governments?
- Suppose the book's analysis is broadly true. Generally speaking, can we use this knowledge for good?

1. In this review, I use the present tense. But the book was published in 1991, based on research carried out in the 1980s. [←](#)
2. Something like this is familiar to me from the days when most of my friendships took place in pubs. Small favours, even with specific monetary value, would typically be repaid in drinks and not in cash. Once, apparently after a disappointing sexual experience, I was asked my opinion on the exchange rate between drinks and orgasms.

It strikes me that your neighbor is still clearly worse off than if your goat hadn't eaten his tomatoes. He's gone from having tomatoes-now to only having future-tomatoes. But that means your neighbor has no reason to be careless with his tomatoes. And helping to replant may encourage you to control your goat more than paying money would. [←](#)

3. Ellickson never says explicitly what one of these is, but my read is a small ranch, more a home than a business and operated more for fun than profit. Only a handful of animals or possibly none at all, and sometimes crops. [←](#)
4. If you want a teaser, the first three chapters were based on his previous article [Of Coase and Cattle](#) (1986). I haven't compared closely, but they seem to have a lot of text in common.

I should note here that I read part one a few years ago. I took reasonably detailed notes, with the intent of editing them into something worth publishing before moving on to part two. Then I didn't do that. My review of part one is

largely based on my notes, although I have skimmed or reread large amounts of it. I read part two recently, specifically for this contest. ↵

5. He was a founding member of the [American Law and Economics Association](#) in 1991, and its President 2000-2001. ↵
6. "This counterintuitive proposition states, in its strongest form, that when transaction costs are zero a change in the rule of liability will have no effect on the allocation of resources. ... This theorem has undoubtedly been both the most fruitful, and the most controversial, proposition to arise out of the law-and-economics movement." The [paper](#) which first presented the theorem used cattle trespass as an example, directly inspiring the study in part one. ↵
7. Ellickson offers citations but (apart from Shasta County) no elaboration on these. ↵
8. Ellickson makes two points. First, that ranchers are more familiar than rachette owners with barbed-wire fencing. To some extent that seems circular, since they're the ones who are expected to know about it, but it's also in part because many rachette owners have moved from the city. Second, that ranchers can fence in their own herds unilaterally, while victims would have to coordinate; motorists in particular would have trouble with that, and arguably they benefit the most. But motorists aren't part of the relevant close-knit group, so we should ignore them for this analysis. And as far as I know the other noteworthy victims are all landowners, who don't need to coordinate to protect their own interests.

But: even if victims wouldn't need to coordinate, they'd all need to act individually, and acquiring the skills would be a cost to them. Ranchers would presumably still need those skills, and even if not, there are presumably fewer of them. So it seems cheaper for all ranchers to acquire the skills, than all of their potential victims. ↵
9. Of course, since this example was a generator of the hypothesis, that [says little](#) by itself. This isn't a big deal, Ellickson looks outside Shasta County plenty, I'm just pointing it out because it's important to notice things like this. ↵
10. No, really. This isn't him saying one thing and me [saying](#) "well that only works if...". He says explicitly that the hypothesis "assumes that norm-makers in close-knit groups would subscribe to an unalloyed version of this principle". ↵
11. Actually, for parenthood, a plausible answer does come to mind: deciding who society celebrates as parents (rich couples who can mostly pay the costs of parenthood themselves) and who it shames (poor single mothers who socialize the costs). Then I guess the hypothesis predicts that you're allowed to socialize the costs of parenthood to the same extent that parenthood is welfare-positive. Except... that doesn't really work, because the total welfare change seems like it would be more-or-less the same whether the costs are borne by the parents or by society. I dunno, I think I'm still confused. ↵
12. This is kind of a weird example, as he all-but-admits in a footnote: "This statement assumes the continuing presence of foundational rules that forbid the firefighters from killing, maiming or imprisoning each other." If we want to know what actually happens in this situation, he points us to [Dudley and Stephens](#) and a book named *Cannibalism and the Common Law*. ↵

13. I haven't read these recently, and might be misremembering. [←](#)
14. Minor complaint: I wish Ellickson had been clearer about what exactly a "fishery" is. Did two boats from different fisheries ever encounter each other? [←](#)
15. In one case study, 23 permission letters were sent to publishers and only 17 received a response in six months. Ellickson doesn't say how many were denied. [←](#)
16. I looked it up out of curiosity. Although the 10% figure may have come from [the relevant guidelines](#), they're unsurprisingly a lot more restrictive than that. For prose the maximum seems to be "1,000 words or 10% of the work, whichever is less, but in any event a minimum of 500 words." [←](#)
17. At least, if we can't in practice, there's nothing stopping us in theory. I'm not sure if we know exactly what the slime molds are doing. But I'm sure that if we *did* know, there wouldn't turn out to be anything fundamentally mysterious and unprogrammable-in-computers about it. [←](#)

Covid 7/15: Rates of Change

Cases rose by over 60% in America this week, and we're seeing large jumps in cases around the world. I am highly suspicious about the jump in the rate of increase, but Delta certainly seems to be the real deal, and this was well above my expectations.

I worry that recently I've lacked sufficient skin in the game. Everyone I personally care about is vaccinated or young enough that they don't need vaccination, so the real sense of danger is largely gone. The worry is about *the reaction* to Covid, rather than about Covid itself. But that's a very real danger, and I have back that sense of 'oh no, things could go very wrong' because there's the danger that we really will blow up our way of life over all this, and go into a permanent dystopia of sorts. That's what we need to ensure does not happen.

Thus, the bulk of this post is a numbers analysis trying to figure out what we know about Delta's transmissibility and the effectiveness of vaccines in reducing that transmissibility, using data from a variety of sources. Others are encouraged to continue this analysis and try to get to the bottom of this.

So let's run the numbers.

The Numbers

Predictions

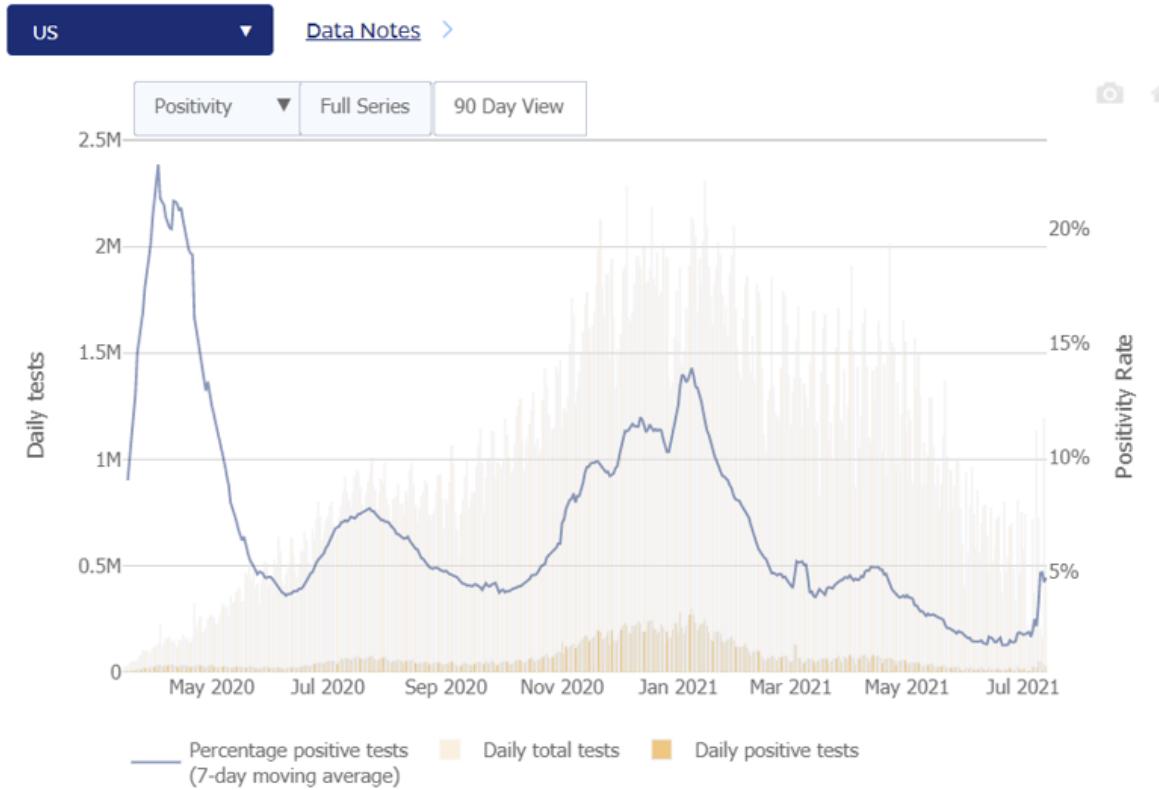
Prediction from last week: Positivity rate of 3.3% (up 0.4%) and deaths increase by 7%.

Result: Positivity rate of 4.8% (!) and deaths increase by 15%.

Prediction for next week: Positivity rate of 4.7% (down 0.1%) and deaths unchanged.

The null prediction is always an option, here two distinct null predictions with distinct reasoning. For deaths it's clear that there was a reporting gap as predicted, so I do not think the death rate last week represents things getting worse, but they likely should start to get worse given Delta is deadlier and cases have stopped dropping within the required time window, and it doesn't seem like last week's number was too artificially high.

The case number is trickier, as there's good reasons to think the data is distorted, either by July 4 or otherwise:



That giant spike represents going from an average of 2.6% to an average of 5.0% over two days. That's not a thing that should happen to seven day averages. If it does, then the next five days things should continue to rise as old data cycles out for new, but that didn't happen so far.

Perhaps July 4 was truly a superspreader event in the way that previous holidays were almost always underwhelming, as much as the previous holidays make that seem unlikely. Perhaps it's a giant delayed data dump that didn't include negative tests. It is hard to say. What I do know is that either things were worse than this before the jump and the low number wasn't fully real, or the jump and new number are not fully real – one of these two numbers is misleading.

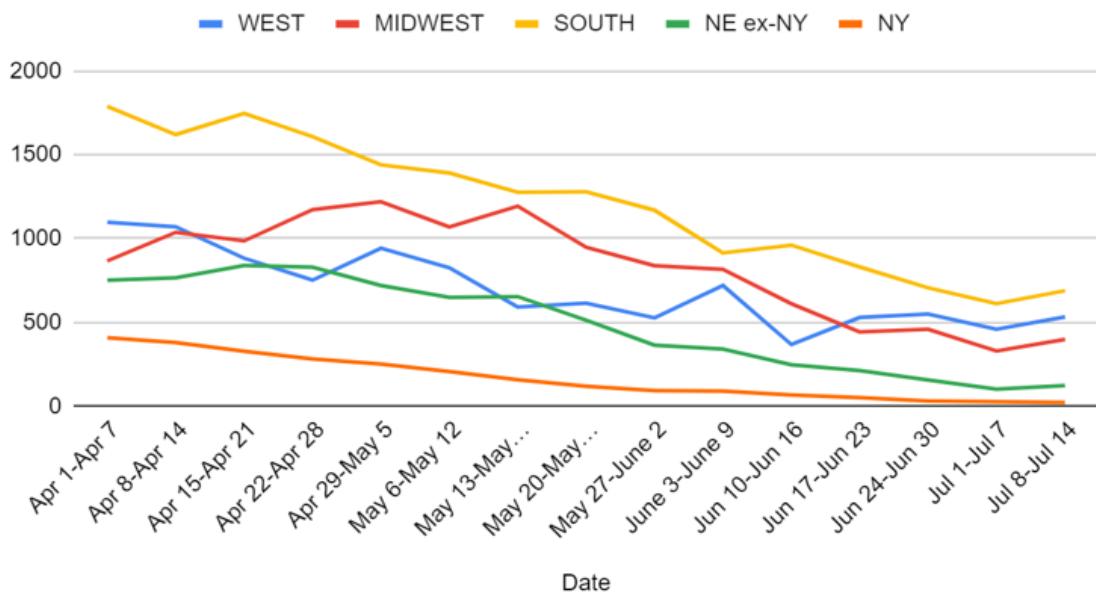
Thus, we have the standard question when a number seems to clearly overshoot, where it's unclear where the 'real' number is and how fast it's moving, so it's unclear where it will end up. In this case, substantial real growth seems almost certain, and I definitely feel like the null prediction here is 'chickening out' but I'm not sure which direction to go, although I notice I'm more comfortable predicting a small reversion than an increase, while noting that such a decrease wouldn't be 'real.' Thus I'm choosing a very small decline, with discussion continuing in the cases section and the Delta section, but this data doesn't make sense.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 27-June 2	527	838	1170	456	2991
June 3-June 9	720	817	915	431	2883
Jun 10-Jun 16	368	611	961	314	2254
Jun 17-Jun 23	529	443	831	263	2066

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

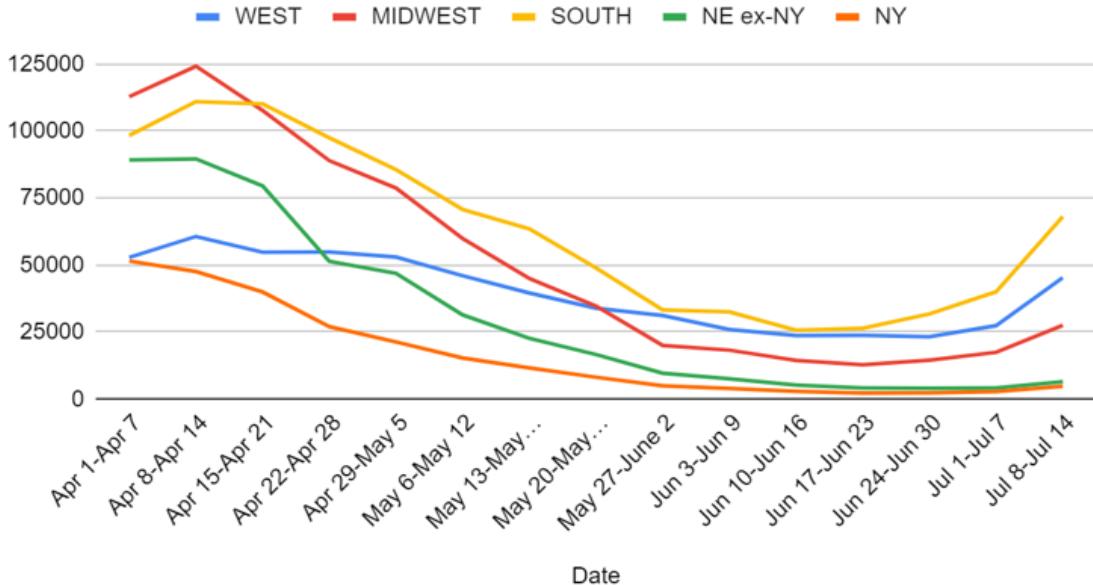
Deaths by Region



Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 20-May 26	33,890	34,694	48,973	24,849	142,406
May 27-June 2	31,172	20,044	33,293	14,660	99,169
Jun 3-Jun 9	25,987	18,267	32,545	11,540	88,339
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

Positive Tests by Region



Cases are up by 66%, and the positive test rate is up by 65%, which implies the *number of tests* was constant.

The more I think about this, the less sense it makes.

If cases were up by anything like 66%, why aren't we running more tests? Aren't people testing in large part based on whether they suspect they have Covid, and whether they have symptoms or known exposures, which both should be up a lot? Are we somehow supply constrained on this, despite no observations of difficulty in getting testing done on demand? How do these numbers make any sense?

Thus, I come to the conclusion that the numbers *don't* make sense, and don't belong in the same universe. If cases double test counts should rise a lot, and that not happening is super weird.

For now, I'm going to mostly ignore the test percentages and act as if the raw positive test counts are more accurate, because if there's one thing I *definitely don't believe*, it's the reported number of negative tests. That doesn't make any sense no matter how bad things are.

One possible way for this to *kinda sorta* work is that perhaps there are people who get tested in order to show a negative test, whose tests get reported every time, and people who get tested because they want to actually know if they have Covid, who mostly only report when *they're positive*. Then, doubling the size of the second group doesn't change *reported* test counts much? That's the best I can come up with.

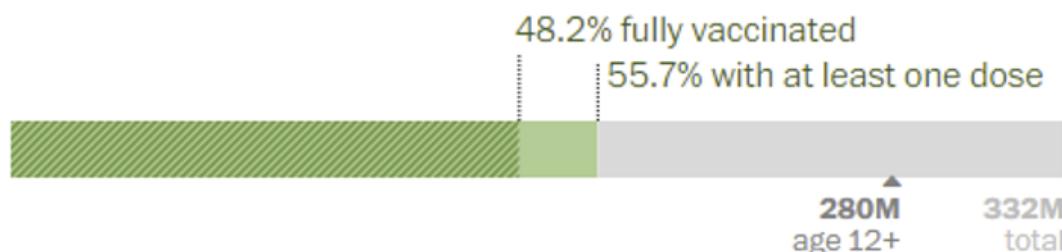
What should we make of the 66% rise in cases? How much of it might be timeshifted and in what ways? Is this a fully real rise, and should we expect it to continue? If so, what happened?

Discussion of all that will continue in the Delta section, where I attempt to reconcile all the various different data points.

Vaccinations

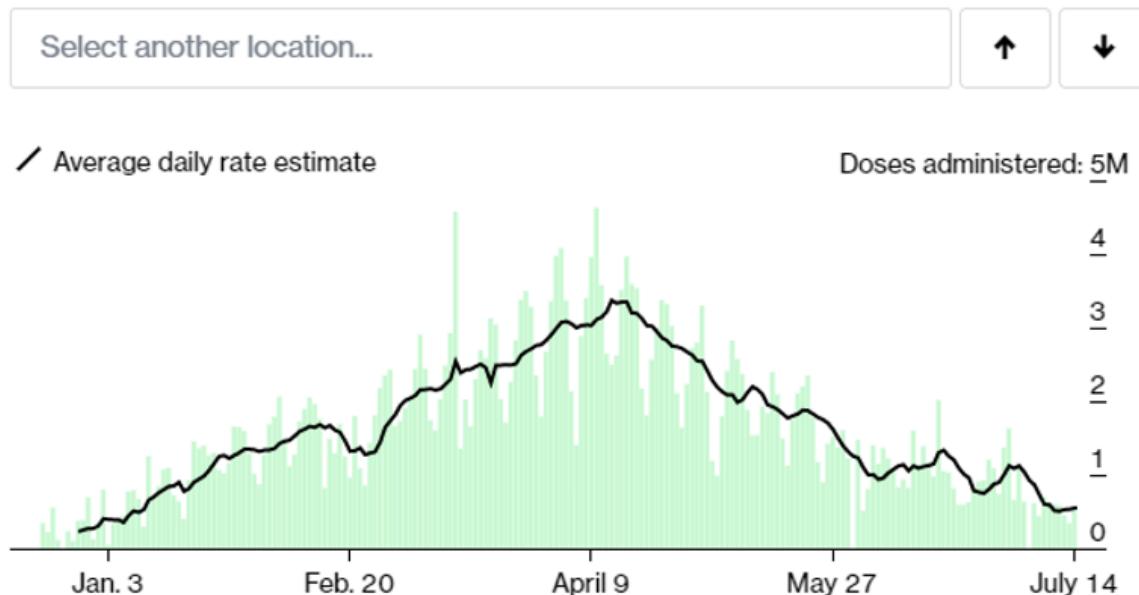
184.8 million vaccinated

This includes more than **160.1 million people** who have been fully vaccinated in the United States.



In the last week, an average of **548.0k doses per day** were administered, a **25% decrease ↓** over the week before.

In the U.S., the latest vaccination rate is **548,045 doses** per day, on average. At this pace, it will take another **8 months** to cover **75%** of the population.



Given no attempts to halt the course of events, this is a strong result.

[There's a new warning on J&J shots](#) (WaPo), because there were about 100 detected cases of an autoimmune disorder out of 12.8 million shots. There was one (1) death involved, again out of 12.8 million shots. It will be an interesting control group for the previous experiment

where J&J got suspended. If we take a side effect that doesn't matter, and treat it like it exists but doesn't matter, does that have an impact? Versus treating it as a huge freaking deal and freaking out everyone and suspending the vaccine, which we are pretty sure had a fairly large impact the last time there was a side effect of magnitude epsilon.

[New paper says that neither 'give people fact box' nor 'explain how mRNA vaccine was developed and that it wasn't too fast' impacted vaccine hesitancy](#). Kudos for publishing the negative result. Why aren't we seeing a lot more studies like this of various things one might do?

Then there's the question of what is happening in Tennessee (news article).



Brett Kelman @BrettKelman · 17h

SCOOP: Tennessee Department of Health halts all vaccine outreach to kids – not just for COVID-19, but all diseases – amid pressure from GOP. Staff ordered to remove the agency logo from any documents providing vaccine info to the public, per internal dox.



Brett Kelman @BrettKelman · 17h

The agency will also end all COVID-19 vaccine events at schools, even though they've mostly served adults. And, if teens get a two-dose vaccine, it won't remind them to go back for their second dose. Teens are intentionally stripped from the mailing list for reminder postcards.

The explanation I found was that in Tennessee teenagers can get vaccinated without 'parental consent' and this was creating problems. Which does not seem like much of an explanation, as there's no reason why such consent should be required or even relevant. And certainly does not explain why this applies not only to Covid-19, but to *all vaccinations period*. Or why they think it makes sense to strip teenagers out of lists for reminder postcards if they're scheduled for their second shots, but focus on the *other vaccines* part of this, if you're considering that this all might have a perfectly logical explanation.

Perhaps we can at least partially salvage this by taking advantage of the 'natural experiment,' and observing what happens to vaccination rates. Do the reminder postcards do anything? What about the other things that got suspended? What happens to the vaccine rates for other diseases? If you can't go with 'prevent people from getting sick and dying' you can at least upgrade your world models.

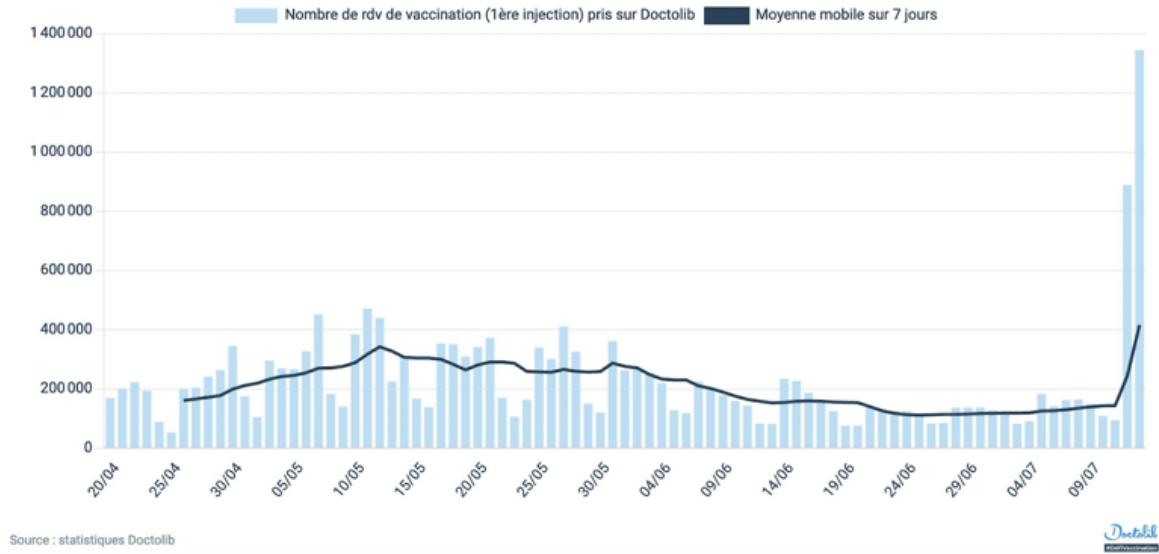
Genius in France: Incentives Matter!



Edouard Mathieu @redouad · 22h

Emmanuel Macron announced on Monday that a proof of vaccination (or a negative test) would very soon be needed to access public events, restaurants, cinemas, stations & airports...

Since then, more than 2.2 million vaccination appointments have been booked in less than 48 hours.



The second day exceeds the first, so those 2.2 million appointments are likely only the beginning. It's one thing to pass up a vaccine, it's another to pass up the ability to participate in many aspects of life. Make no mistake. If implemented, this will work.

If America's Delta problem gets sufficiently worse that they start bringing restrictions back, and they *don't* start requiring proof of vaccination in such situations but instead once again halt life for the rest of us, I call upon all of us to find this *completely* unacceptable, the same way I find permanent child masking unacceptable.

In such a scenario, there are two sane choices. You can either let people do what they want, or severely restrict what the unvaccinated can do. Ending life as we know it, presumably indefinitely, shouldn't be even potentially on the table, nor should we have any tolerance for such proposals.

Delta Variant

[A post entitled Delta Variant: Everything You Wanted to Know](#) does make a real attempt to be exactly what it says on the tin. The graphs are rather cherry-picked to make things look as bad as possible, as are a number of other discussions, but the data is all legitimate. The question of the day is now exactly how bad Delta is and making sure our models of it are right to figure out what is to come. There's a bunch of superficially contradictory data that must be reconciled, as there usually is.

Taking stock of those data points is the logical first step.

Israeli Data

Israel offers the scariest data point, suggesting greatly reduced vaccine effectiveness.

The best data we have is from Israel, which used Pfizer.

Before Delta, it appeared that full vaccinations reduced infections, hospitalizations, and deaths by 93%, 93%, and 91% respectively. Partial vaccinations were quite good, but not as good.

Now, with Delta, it looks like the figures are 64%, 93% and 93% according to Israel, and 79% for symptomatic infection and 96% hospitalization according to the UK.

That said, Pfizer (and probably Moderna, as similar mRNA vaccines) are likely better than AstraZeneca's (88% protection against symptomatic infection vs. 60% for AstraZeneca).

That means the protection against hospitalization and death is nearly perfect for mRNA vaccines, but maybe not for infections. How should we interpret this data?

I've seen the Israeli data presented in slightly different ways but this is basically what they're reporting there. I'd like to note that the story doesn't make sense, even on its own.

For previous strains, this is saying that vaccination wasn't protective against hospitalization, and mostly wasn't protective against death, once someone was infected (93% reduction vs. 93% reduction), or at least once someone tested positive. We knew even then this wasn't remotely the case.

Then with Delta we get this *gigantic* drop to 64% protection against infection, but then protection against hospitalization stays at 93% and protection against death *rises*? So conditional on infection, this is saying hospitalization protection went from ~0% to 80%? Really?

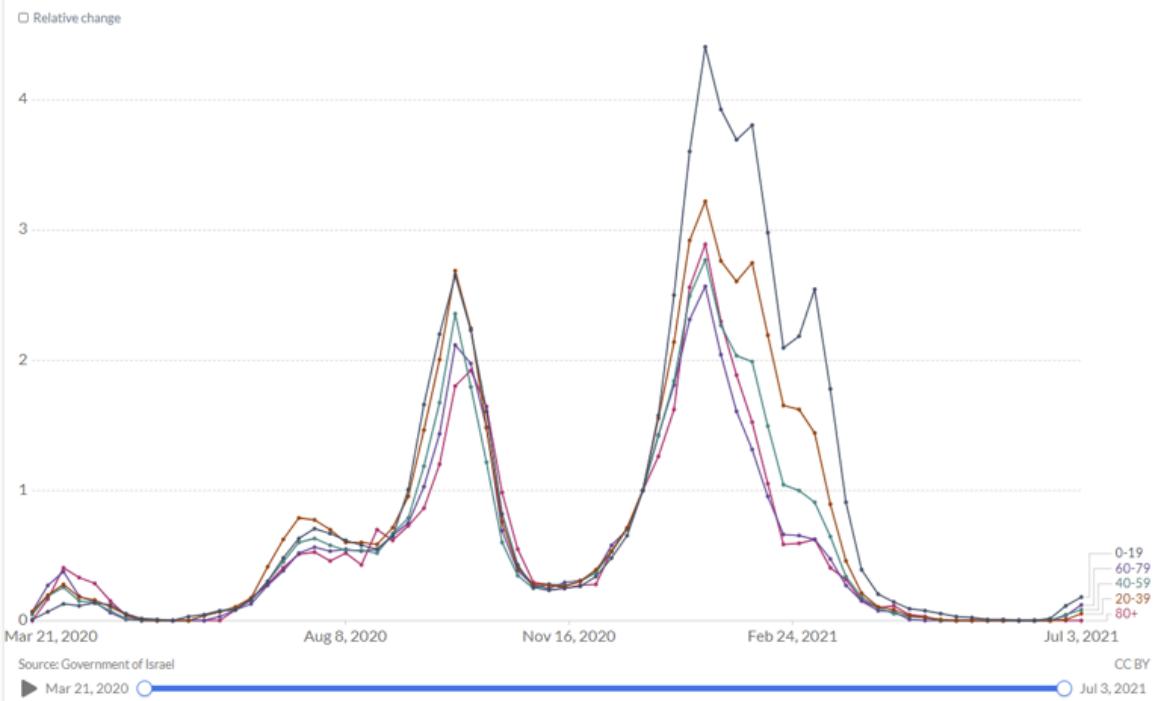
Even the 79% number seems very very *strange* when looked at this way alone.

You could tell a story that justifies it. In that story, vaccine protection works 96% of the time (and there's measurement errors), and if that happens you're protected against severe outcomes no matter which variant you face because that wasn't a close call, but being infected at all is a lower threshold. Before, if you were successfully vaccinated you basically never got infected (in this model), whereas now if you are vaccinated you sometimes do still get infected, but it's never serious whereas before it never got that far in the first place. Then there's the 4% of people for whom the vaccine doesn't work properly, who are still at real risk. Or something like that.

As a sanity check, what happens if we ignore the reports and attempt to back out the answer from [the raw data on infection numbers?](#)

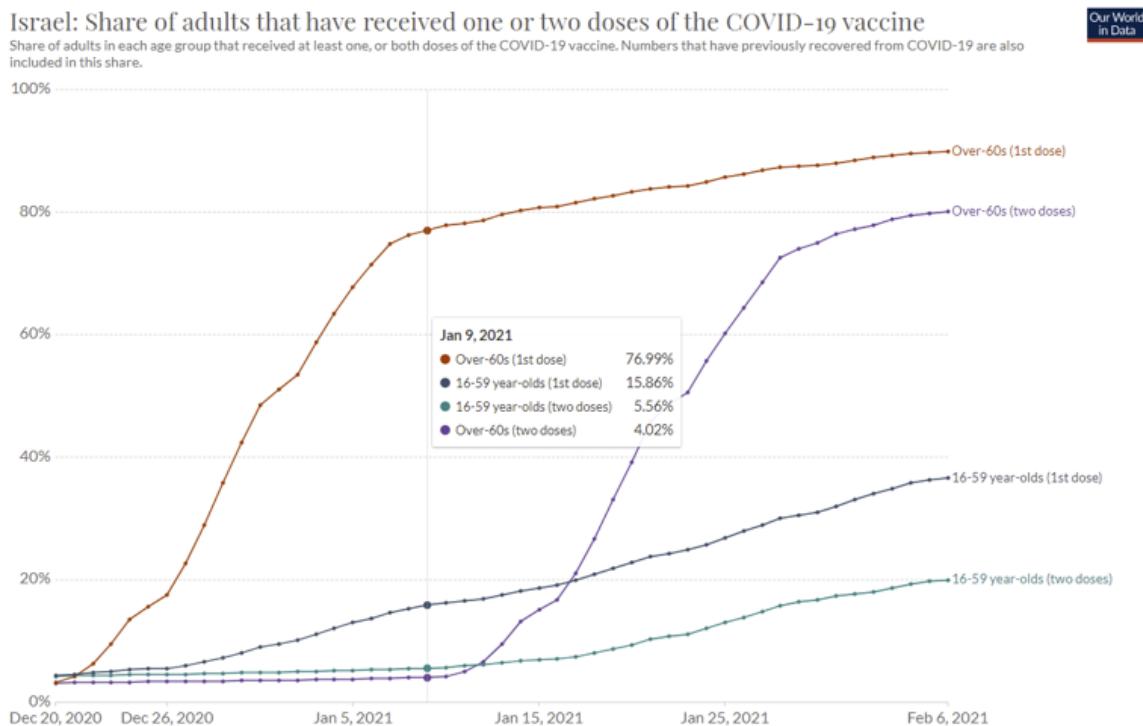
Israel: Confirmed COVID-19 cases by age group
 The values for each age group are indexed to the cases reported at the peak of the last wave in mid-January.
 The chart shows the relative decline in cases since then, by age-group.

Our World
in Data



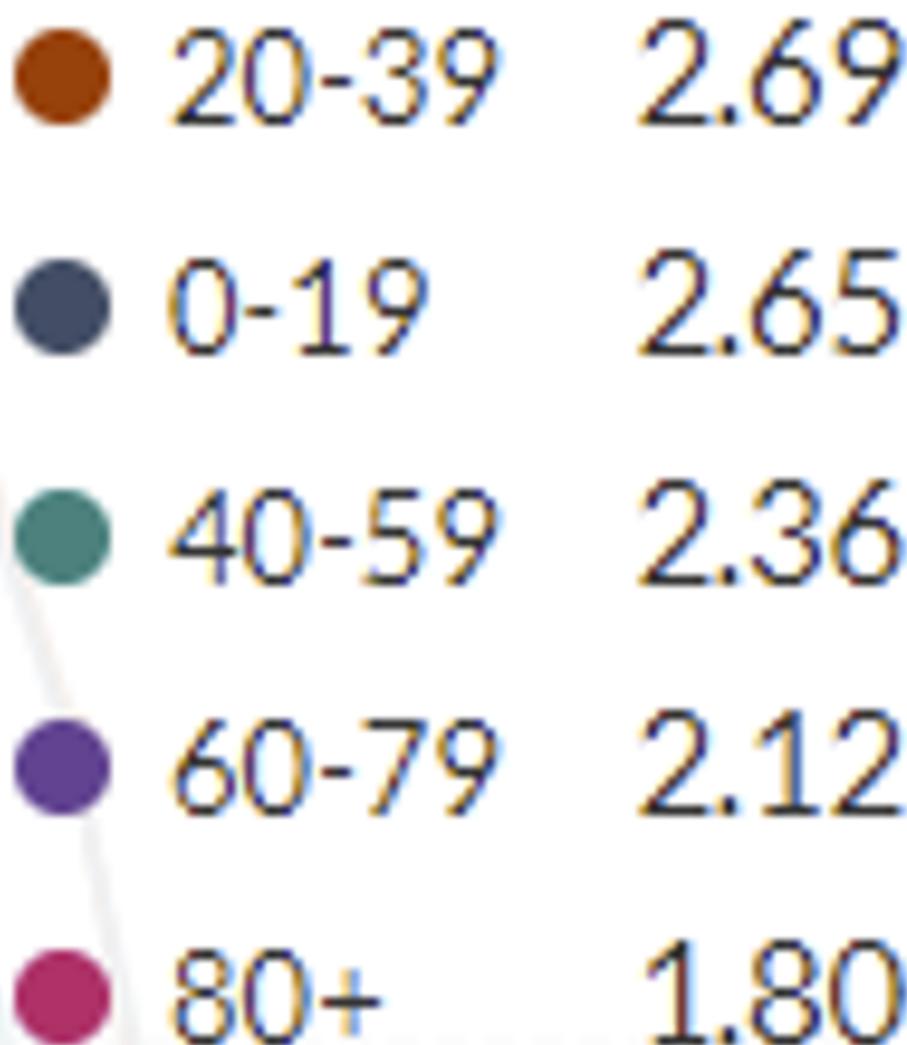
That's not a picture one can easily read, so click the link if you want to examine it.

Reminder:



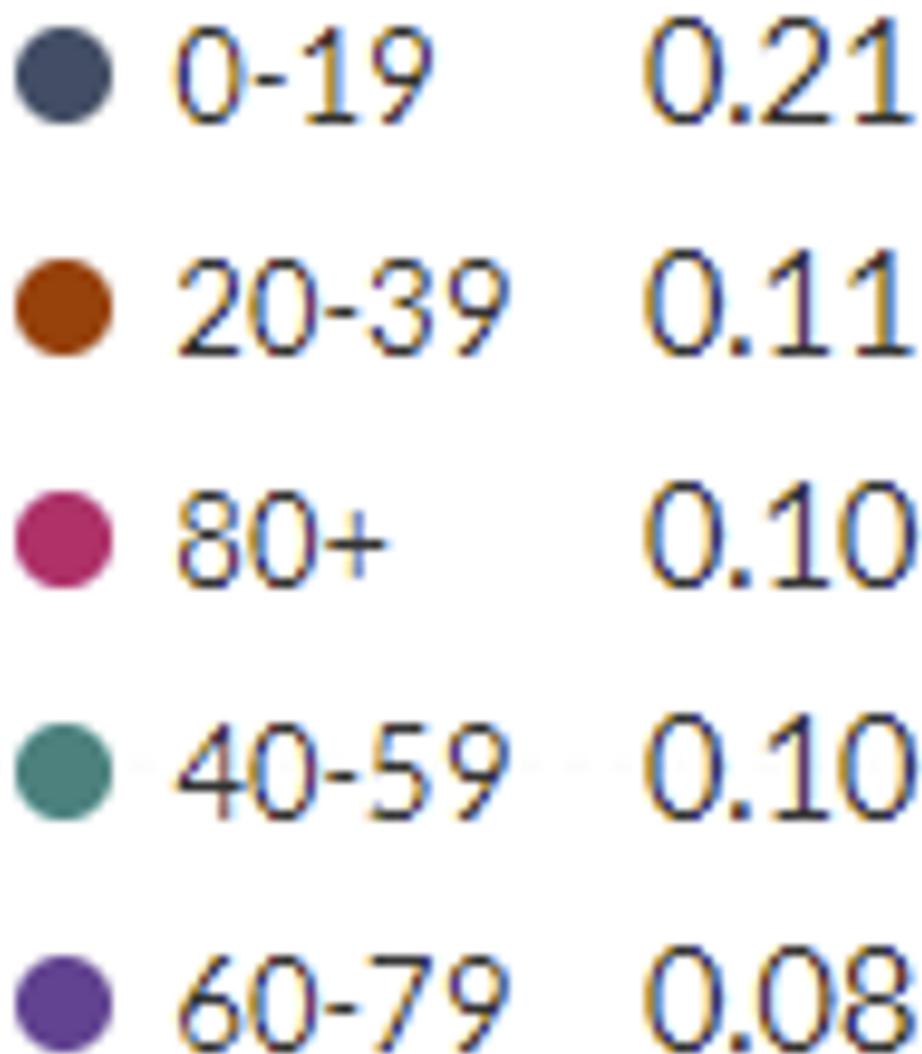
There are a lot of confounders, but let's start with the pre-vaccine comparison.

Sep 26, 2020



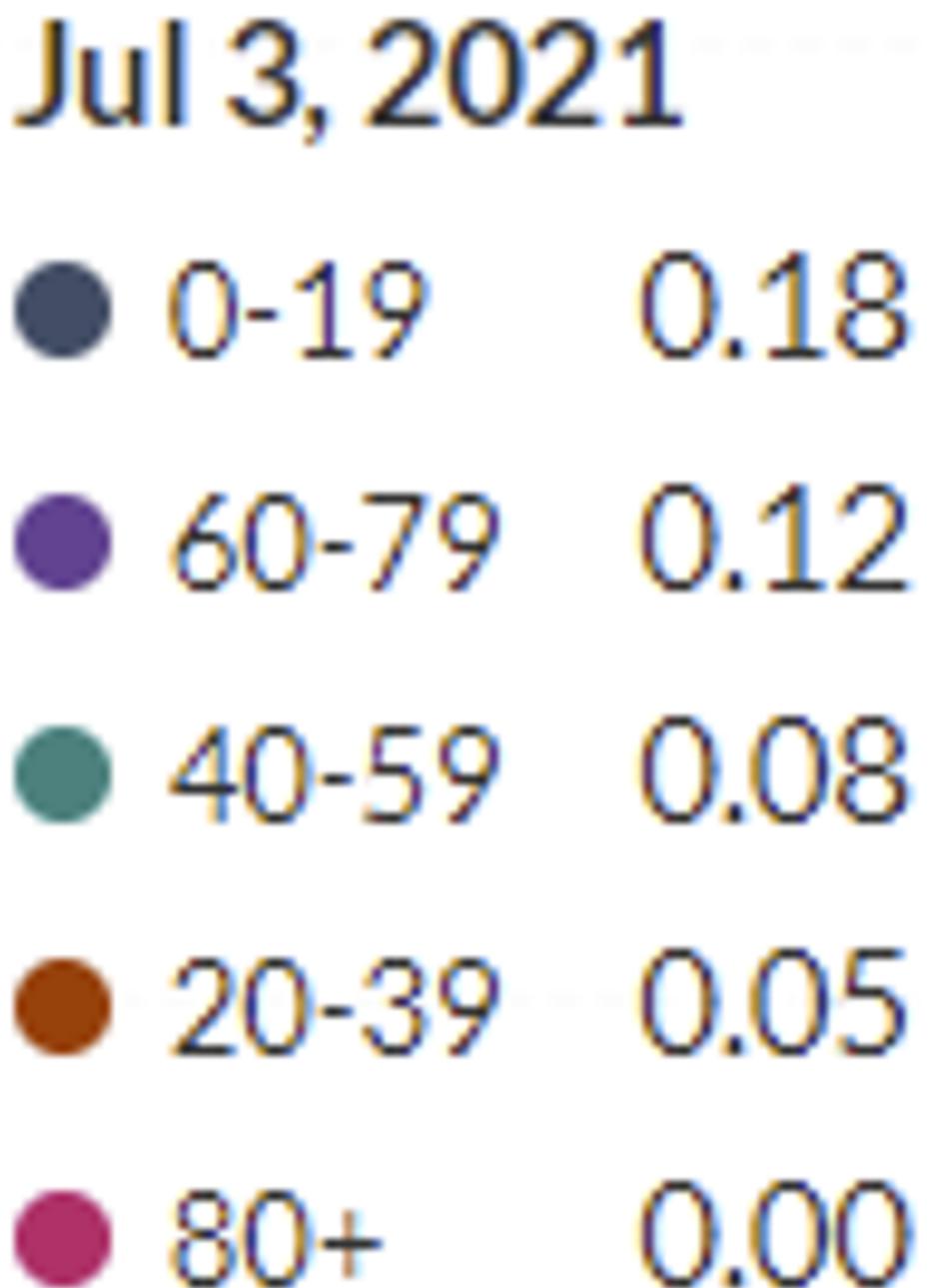
Next, let's do a pre-Delta comparison, from when the numbers were very low, but people were mostly vaccinated:

Apr 3, 2021



Consider this a 'post-vaccination' equilibrium. Many of the young aren't vaccinated, whereas most of the old are, so the ratios change, but almost entirely for the youngest group. Things were mostly level before otherwise, and remain mostly level now, in roughly the same order.

Now let's look at the last day they have data for here:



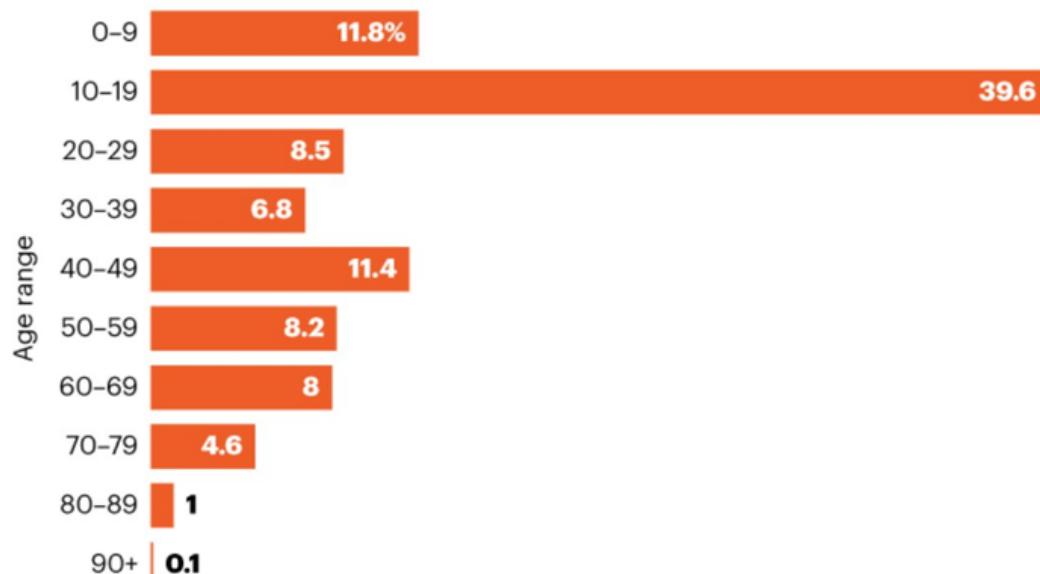
It's unfortunate we don't have the July 10 data, but we go with what we have. This still has to represent almost all Delta cases, and gives us apples-to-apples comparisons. What can we infer from these numbers?

Also, how does this reconcile [with this graph?](#)

TRENDING YOUNGER

With the majority of adults in Israel now vaccinated, just over half of the country's new COVID-19 cases in the month up to 5 July were in people aged 19 and under.

Proportion of recent COVID-19 cases in Israel by age group



©nature

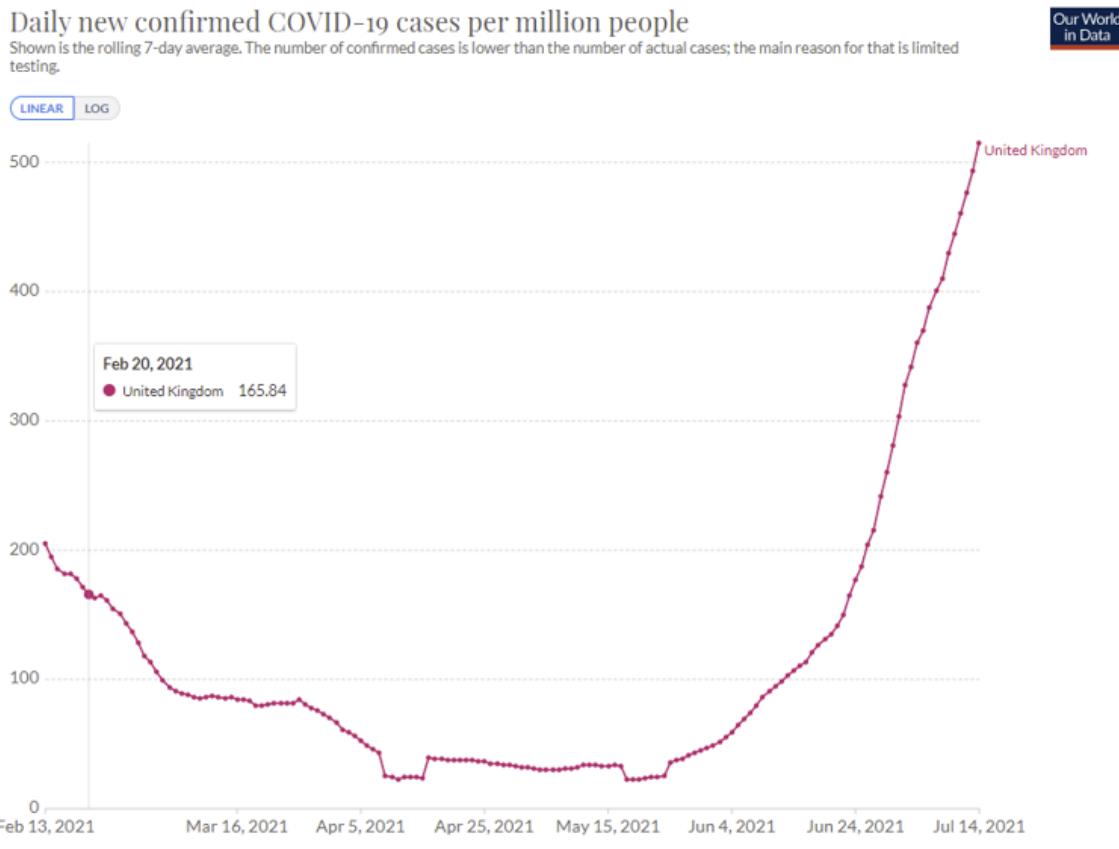
Source: Israel Ministry of Health

Israel is young, but it's not *that* young, and my response to the above graph is more like 'those sample sizes are all absurdly small because Israel didn't have Covid that month.' Still seems difficult to reconcile in the details, but easy in the bigger picture.

My conclusion is mostly that this is muddled enough that I can't draw a fixed conclusion. Especially weird is the 60-79 range. One possibility is that the vaccine needs a threshold of effectiveness to prevent infection, and it's still mostly good enough to hold off Delta, but those with weakened immune systems are in a different situation and they are mostly very old? But the share of vaccinations in such groups is still super high compared to younger groups, and the order of these groups still seems really odd. It would, however, explain how 20-39 could be the lowest major group while 0-19 is the highest, perhaps - The kids in their 20s can still largely get vaccinated, and their immune systems are still strong, so it's highly effective? Whereas with the 40-59s it's less effective especially on the high end, and so on.

We should also look at case counts in Israel. On June 18 they had 1.92 cases per million, right before things started rising, on June 14 it was 65.09, for $R_0 = 1.97$. From previous data, we can presume that when Delta was a very small portion of Israeli cases, the control system adjusted things to something like $R_0 = 1$, so we'll keep that number in mind.

UK Data



We can presume that Alpha hasn't increased its absolute numbers, so at this point UK is almost all Delta. On May 25 there were 35.7 cases per million. On July 4 there were 512.9, On July 4 there were 361, so the two-week $R_0 = 1.19$. From May 25 to July 4 we get $R_0 = 1.32$, despite a bunch of Alpha early on, so the control system has been adjusting, but also we never see a period with an extreme R_0 in the UK. In April, they had Alpha and things were stable ($R_0 \sim 1$). For an extreme situation, let's look at relatively early lift-off of Delta and assume a base of about 30 Alpha that doesn't count for growth rate, and start when total cases are at double that, or 60, which is June 4. Over the next 10 days things went to 107, so let's say that this represents Delta going from 30 to 77, before much adjustment has been made, over two cycles. That would give an estimate of $R_0 \sim 1.60$, on the high end of the range for increased base transmissibility. It doesn't leave that much additional room. One more cycle gets us to 131, which would be $R_0 \sim 1.31$ for Delta alone, so presumably there were adjustments being made already then.

Transmissibility

There's the question of how much more deadly Delta is than older strains, but I'm mostly going to ignore it because it doesn't impact the path of the pandemic much. The question is containment versus lack of containment. If Delta is twice as deadly, that's quite bad, but it's an isolated question. As commentators have pointed out, the data supporting the increased deadliness is not that solidly grounded. The extra viral loads are suggestive, and the early data does look like it's more deadly for a given unvaccinated person, but our data remains not great.

The much more key variable is transmissibility of Delta versus Alpha or the original strain. I've been using a 120% increase, or 2.2x (220% of original infectiousness) for Delta, and a 1.4x (140% of original) for Alpha, which continues to match estimates I've seen. Those all presumably refer to *transmissibility among the unvaccinated* and ignore the vaccinated population, and I've been treating the numbers accordingly. I see no reason to change that methodology, but that means that we need to reconcile our numbers with the case counts that we see.

An open question is how tight the bounds are on such numbers. If we have a model where things were previously going fine and then you 'multiply by Delta' then the bounds are reasonably tight. However, if you think that behaviors are adjusting in real time and seasonality causes 'random looking' ebbs and flows naturally, then there's room for the data to look highly misleading, and my position at this point is if anything is closer to that second view.

Either way, your model has to make predictions whose math checks out, and that's a good place to start.

Let's say we accept that Delta's 'multiplier' is 2.2, versus the old baseline, and Alpha's is 1.4. Now let's suppose vaccine effectiveness is reduced from 95% to 65%. What happens? Given current vaccination rates, with an adjustment for children, that's about *another* 50% increase in the rate of infections.

That would mean that Alpha to Delta is a full effective +120% increase in the rate of infection, on top of the increase from original to Alpha, or a final factor of around 3.3. Or, alternatively, it would have a 'base' R_0 of something like 6, with a vaccine that only reduced that *for the vaccinated* to 2, which would mean that even in fully vaccinated populations this would double every five days under pre-pandemic behaviors.

Set aside for the moment the question of what we would want to do about that. How does it line up with the data?

In the UK, where the Delta problem is currently largest and we are confident Delta is essentially the whole pandemic (using Our World In Data as my source), we had 36 cases/100k on May 25, and 494 on July 14, 52 days later, or just over 10 cycles, and $R_0 \sim 1.28$. During April and the first half of May, we saw the number of cases stay roughly constant within a factor of 2, so $R_0 > 0.9$. We could factor in some increase in vaccinations, but if Delta taking over sent R_0 from 0.9 to 1.28, that's only a 42% increase, similar to what we'd expect from Delta taking over from Alpha, minus some existing more infectious strains and some extra vaccinations, give or take behavioral adjustments and seasonality.

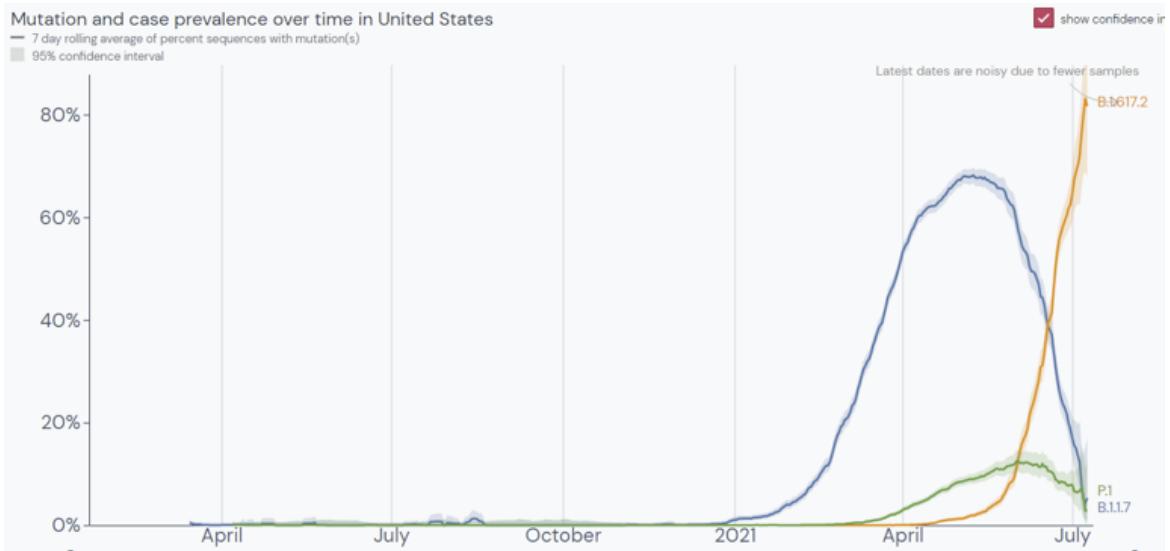
None of this matches the greatly reduced effectiveness hypothesis, unless you presume that behaviors are substantially adjusting during this period, but the shape of the curve isn't suggestive of that either.

American Data

In the last week, America has seen a 66% increase in cases. As discussed above, going +10% then +20% then +66% is extreme, and can't be explained by Delta alone. The share of Delta cases can only rise by about 30% in one week, even under extreme assumptions (e.g. from 35% to 65% or something like that, would be the theoretical limit) so a +46% effect in seven days off a 30% rise would mean Delta was at least twice as transmissible as Alpha. It also means the numbers in previous weeks would have shown a bigger problem, since the displacement of Alpha by Delta has to be gradual - there's no way there was three times the growth in Delta this week than there was last week, because math.

Overall, though, the number isn't crazy – if we presume that the control system had already adjusted for our vaccinations and for Alpha. If we take the +66% number seriously, and compare it to the pre-Delta situation, it's safe to presume that we had previously stabilized under a full-Alpha situation, then +66% in a week represents a 43% rise in transmissibility from Delta versus Alpha minus any extra vaccinations, so 50-60% total, which doesn't even leave room for the vaccines to lose effectiveness since that's our estimate range for Delta already among the unvaccinated.

Under such assumptions, we can backchain, and it matches up with the sample data claiming that by percentage [there wasn't much Delta running around in early June](#), despite Delta now having taken over and now being 85%+:



An alternative calculation would be to look only at *Alpha*, or to compare Alpha to Delta. We have a percentage for it, so we can back out its growth in absolute numbers. It was something like 60% of cases on June 15 by this graph, and is now down to 5%. Whereas Delta went from roughly 4% to 85%.

From June 15 to July 15 is thirty days. Thirty days allows for six serial intervals. So if all of that is accurate, we get a ratio of 2.5:1, or +150% increased infectiousness, which is even *higher* than the +120% estimate for the full effects previously measured. We then look at absolute numbers, and see Alpha having an $R_0 = 0.75$ over this period, versus $R_0 = 2.25$ for Delta. Which would mean that both we've taken a lot more precautions recently than we did before to get things that low, and *also* that Delta should be doubling every four days, yet cases over the last week only rose by 66% despite starting off with the majority being Delta.

Thus I don't believe the chart above is a representative sample – the math doesn't add up.

One can have a hypothesis that strains crowd each other out in some sense but I don't think the base rates are high enough for that effect to be big right now.

Then again, as I've been noting, if you take the numbers too seriously (as in, you don't think there are big hidden factors and random distortions) then none of this adds up.

So where does that leave us? Can we put bounds on things?

Our lower bound should presumably be that Delta is 50% more infectious than Alpha, but that vaccine effectiveness is mostly unchanged.

Under Israeli conditions, it seems mostly safe to say that Delta is at most twice as infectious as Alpha, but that about twice is possible. This is the scariest data set.

Under UK conditions, it seems mostly safe to say that Delta is at most 75% more infectious than Alpha, and it would be difficult to get to a doubling.

Under American conditions, it seems mostly safe to say that Delta is going to be less than twice as infectious, given everything we know – the math starts to fold in on itself if we get above 75% or so, in the sense that things need to look much worse than they do. I'd put a soft bound around 75%.

We could also look at any number of other places. Israel and UK are especially picked because they're well-vaccinated and having trouble.

Now let's look at vaccination rates. Using Bloomberg's 'enough for X people' metric, Israel is around 60%, the UK is also around 60%, and the USA is at 52.4%. That's out of the full population, which includes children, so effective vaccination rates are somewhat higher especially for Israel (which has a younger population), and one dose is more than half of two doses, so the USA is effectively closer to 60%, and we can put Israel and the UK closer to 67% in terms of our effective percentage.

Let's presume that in the base case, vaccinated people are 96% protected in terms of transmission. If we presume that the unvaccinated transmit at a 50% higher rate, but that the effective increase is 100% in Israel, that would give a vaccine effectiveness versus transmission of 67%. If we take the UK and presume 75% increased transmission, that implies vaccine effectiveness versus transmission of 80%. For the USA, a 75% increase in transmission would imply 76% vaccine effectiveness versus transmission.

It is possible, in theory, that this difference could be that vaccine effectiveness against transmission fades somewhat over time, and the difference here would be that Israel vaccinated earlier than other countries did.

If things were instead at our lower bound, by assumption, vaccines would remain at 96% effective.

It's really hard to put bounds on things given all the factors we can't account for, including control system adjustments in both directions, seasonality, and so on and so forth.

These are likely not tight bounds. There are likely a lot of behavioral adjustments involved in all of this. But it's all very noisy, and I haven't seen other serious attempts to figure this out. I'm encouraging everyone to take a stab at this from various angles and see what you find. There's tons of data to work with.

If this reduced effectiveness is near those upper bounds, there is a very large problem. Even fully vaccinated populations wouldn't be able to fully return to normal if you wanted to avoid Covid outbreaks. You'd either accept that vaccinated people and children often get Covid, and it would be fine, or you'd need to impose restrictions forever, or you'd need to get a new more effective vaccine distributed in the form of booster shots. And that's if you got close to 100% coverage, which is not going to be happening.

Thus, the question would become, if Covid is not done with us, can we decide to be done with Covid and that life beckons, or are we actually going to kill our civilization and way of life over this despite having a vaccine that renders Covid mostly harmless?

In Other News

Fox News has generally not been as anti-vaccine as its customer base likely would have preferred, [but some evidence that this may no longer be true](#). If I have any regular viewers

reading this, can you update us?

[Reasonable thread laying out different questions surrounding booster shots](#), which oddly still leaves out the question of whether shots can be modified to work better versus Delta. I'm confused why there isn't more discussion about that. My presumption is that modifications wouldn't help, which is interesting in and of itself and seems worthy of mention if true.

[Deal between Israel and South Korea where Israel gives Pfizer shots now that it couldn't use and were going to expire, gets future shots in exchange](#) because selling things for money is evil. Post (from MR) points out that Covax's quest to allocate vaccines equally is going to end up wasting a lot of vaccine, as many places don't have the means to distribute the shots they'll get.

[Moderna begins a trial for yearly mRNA shot that would combine vaccines for flu, COVID-19, respiratory viruses RSV and HMPV](#). It turns out that not only does mRNA allow us to cure a wide variety of diseases, it lets us cure *all of them at the same time*, because the technology allows the payloads to be delivered together. I'd be a little concerned about short term side-effects similar to the ones with the current Covid vaccine, but my hope is that the problem can mostly be solved by proper dosing.

This is also an excellent way to give people Covid booster shots without everyone freaking out. If one shot, once a year, can deal with a wide variety of problems, that should work great, so long as the misinformation from anti-vaxxers doesn't cause too many problems.

New Zealand didn't secure enough vaccine shots, which is unfortunate, [but is taking the AZ - > Pfizer booster path seriously](#), which is great.

You don't often get to pick your allies, [such as in the war on school](#). Or is it the war on children?

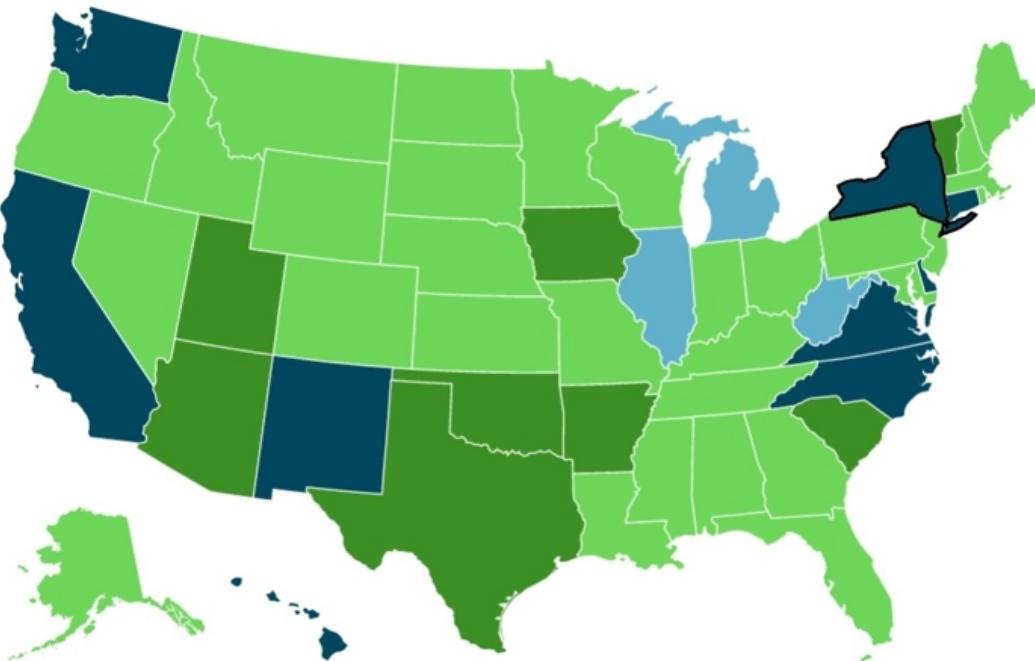
A screenshot of a Twitter post from the account @BNODesk. The post includes the BNO News logo, the handle @BNODesk, the text "18h", and three dots for more options. The main message is: "California will require students from kindergarten through 12th grade to wear masks indoors, and those who refuse will be banned from campus".

Safety Measures for K-12 Schools

1. Masks

- a. Masks are optional outdoors for all in K-12 school settings.
- b. K-12 students are required to mask indoors, with exemptions per [CDPH face mask guidance](#). Adults in K-12 school settings are required to mask when sharing indoor spaces with students.
- c. Persons exempted from wearing a face covering due to a medical condition, must wear a non-restrictive alternative, such as a face shield with a drape on the bottom edge, as long as their condition permits it.
- d. Schools must develop protocols to provide a face covering to students who inadvertently fail to bring a face covering to school to prevent unnecessary exclusions.
- e. Schools should offer alternative educational opportunities for students who are excluded from campus because they will not wear a face covering.
- f. In order to comply with this guidance, schools must exclude students from campus if they are not exempt from wearing a face covering under CDPH guidelines and refuse to wear one provided by the school.
- g. In limited situations where a face covering cannot be used for pedagogical or developmental reasons, (e.g., communicating or assisting young children or those with special needs) a face shield with a drape (per CDPH guidelines) can be used instead of a face covering while in the classroom as long as the wearer maintains physical distance from others. Staff must return to wearing a face covering outside of the classroom.

[Map of such insanities:](#)



Mask Policy

Mask Mandate Banned No Mask Mandate - Local Flexibility Vaccine Contingent Mask Requirement

My understanding is this applies even to *private schools*, although good luck enforcing that. If they're going to mandate this now, when will it end? Are they planning on keeping things like this forever? Or even going back to the torturous 'remote learning' [at the drop of a hat?](#)

What *would* it take to free our children? How bad would it have to get? Shall we run the experiment and find out?

[A rant on the quest to establish that the lockdowns were always painless and super effective](#) and that [Everybody Knows](#) this.

And some *really* are pushing to make the pandemic restrictions permanent. [Who would want such a dystopia? \(post\)](#)



Robin Hanson @robinhanson · Jul 12

Woah: "masks, social distancing ... majority would support them until covid-19 is controlled worldwide, which may take years. ... A quarter say nightclubs and casinos should never reopen"

"Polling by Ipsos MORI for The Economist suggests two-thirds think masks, social distancing and travel restrictions should continue for another month. A majority would support them until covid-19 is controlled worldwide, which may take years. Even more strikingly, a sizeable minority would like personal freedoms to be restricted permanently. A quarter say nightclubs and casinos should never reopen; almost two in ten would support an indefinite ban on leaving home after 10pm 'without good reason'," the report added.

Did that say "restricted permanently"? Yes, it did. Full results at the polling firm's website reveal 19 percent support for a *permanent* 10pm curfew. Significant minorities favor other permanent restrictions, including: keeping nightclubs and casinos closed forever (26 percent); enforced social distancing in theaters, pubs and sports grounds (34 percent); mandatory 10-day quarantines for people returning from foreign countries (35 percent); mandatory tracking-app check-ins when entering pubs and restaurants (36 percent); mandatory masks in shops and on public transportation (40 percent); foreign travel allowed only with proof of vaccination (46 percent).

People permanently not being able to leave home after 10pm 'without good reason' is the kind of thing hack writers put into young adult novels. Or at least, it used to be. Then again, perhaps there's *always* been a 10% share of people who shake their fists at 'kids these days' and actually want to be the villains in such novels. How much of this is new?

A quarter of people wanting to close casinos and nightclubs *permanently* (but again, how much of that has anything to do with Covid?)? A third of people favoring permanent quarantines for international travel? Tracking everyone who enters a restaurant, again, *permanently*? This is still less than half of people, but a third is a lot of people.

This is the fight that is coming, even if conditions are good. They hate us for our freedom. You gotta fight for your right to party.

One can still party a little too hard, even outdoors. Outdoors is much safer than indoors, but that doesn't mean one can't push the envelope too far:



▲ The festival Verknipt at plas Strijkviertel in De Meern. ©
Ruud Voest

Nearly 1,000 visitors to festival Infected, counter rises: 'More infections cannot be ruled out'

UPDATE Almost a thousand visitors to the two-day music festival Verknipt in Utrecht have become infected with the coronavirus. The open-air event drew 20,000 people.

So yes, there's mass gathering and then there's mass gathering, might want to not do that second one. Also, 'cannot be ruled out' is technically true but I think we could have gone with something a little stronger. A thousand *identified* cases is not a thousand cases.

Or, alternatively, instead of protesting for left-wing causes, one can protest against a Communist regime, [in which case your protest is dangerous](#):



Americas

Cuban protests risk exacerbating COVID-19 spike - PAHO

Aislinn Laing, Vivian Sequera



"The gathering of individuals for protests... increases the risk of transmission, in particular in cases such as Cuba where you have active transmission in many areas over the last week and 34,244 new cases reported," said Dr. Ciro Ugarte, PAHO's director of health emergencies. ([Graphic on global cases](#))

Remember, whether or not something spreads Covid depends on whether it is approved of by the proper cultural authorities and Very Serious People. And this is what they think about protesting against authoritarian Communism.

Academic Rationality Research

There are now at least two academic research groups on rationality, in the sense we use the word in here, in Germany that seem to be little known in the US rationality community. The point of this post is telling you they exist if you didn't already know.

There's the [Rationality Enhancement Group](#) lead by Falk Lieder in the Max-Plank Institute in Tübingen and there's a group on [Adaptive Rationality](#) in the Max-Plank Institute in Berlin.

When Falk Lieder was in at our European community weekend he repeatedly said that he's interested in collaborating with the wider rationality community. There's a list of [publications](#) of his group and also a [Youtube](#) channel that presents a few ideas.

The adaptive rationality group has decided to speak of rationality techniques we would likely call applied rationality techniques as "boosting decision-making" in contrast to the academic literature on nudging. I think it's worth exploring whether we should also apply their term for the cluster of techniques like Double Crux.

Improving capital gains taxes

As [I've mentioned](#), I think the tax code could be improved. In a departure from my usual style, this post fleshes out some fairness-based arguments for one of my favorite changes.

(I think that this proposal, and many of the arguments in favor, is old. Wikipedia quotes Joseph Stiglitz making the basic point in *Economics of the Public Sector*.)

The policy

I'm advocating a package with 3 parts:

1. We should significantly increase capital gains taxes by taxing it at the same rate as ordinary income.
2. If an investor makes money one year then loses it later, we should give them their taxes back.
3. We should only tax excess returns above what you get by putting your money in a savings account.

Justifications

1. Why have a higher tax rate?

The simple answer is “most Americans think rich people should pay higher capital gains taxes.” It looks unfair for Warren Buffet to pay a lower marginal rate on income than a working class family. But there are other great reasons.

One is that it can be really hard to distinguish “income” from “investment income,” and many rich people misclassify income to pay fewer taxes. Some of these egregious loopholes lose a lot of tax revenue. But if we tax capital gains at the same rate as other income then these loopholes disappear and the whole system gets simpler.

Another is that rich people do a good job of identifying profitable investment opportunities; normal people can’t participate effectively in those opportunities, and so the rich get richer faster. That feels like a fact of life, but it’s actually pretty easy to change. A capital gains tax effectively lets everyone benefit from every investment opportunity.

2. Why pay investors back when they lose money?

When an investor makes money they pay taxes. But when they lose money they never get a check back. The government effectively participates in the returns but not the risk.

(Investors do get tax deductions they can use once they’ve recouped all their losses, but much of the reason people are afraid to lose money is that they might never make it back.)

This discourages investors from taking on risk and encourages them to push the risk onto other people. Combined with the proposed increase in capital gains rates, it seriously discourages uncertain long-term investments and probably reduces growth.

In addition to sounding a lot like “Heads I win, tails you lose,” I think this policy probably *loses* the government money.

Why? The risk of going broke pushes investors to be more conservative and causes them to have much less taxable income—which is bad for both them and the government. But because very few investors *actually* end up going broke, the government doesn’t save much money at all. We’ve managed to get the worst of all worlds.

(That’s coming from a pretty rough BOTEC. If the intuitive case actually depends on that calculation, I should check more carefully.)

The new alternative effectively makes the government a non-voting partner in every investment, sharing in both risk and returns. That make the system more fair, puts us all on the same team with aligned incentives, and makes everyone richer. Instead of discouraging rich people from finding opportunities, it motivates them to make bigger bets so that they can give everyone a share.

3. Why adjust the initial investment using the risk-free rate?

When I put \$1,000 in a savings account, I get \$2,000 in 30 years because “money now” is worth twice as much as “money later.” Everyone—rich and poor, household or government—can make the substitution at that same rate. We should only tax your savings account if we specifically want you to spend your money immediately rather than later.

It’s actually worse than that: over the last few decades all of my “earnings” from a savings account are just keeping up with inflation, so I’m paying taxes without getting any richer at all. The market expects this to keep happening for the next 30 years, so this isn’t a weird corner case.

Instead, we should tax the difference between what you earned and what anyone could have made by just putting the same amount of money in a savings account. That is, we should tax the stuff that is actually *income*, i.e. when you are actually doing work, taking risks, or exploiting connections.

The best argument for taxing savings is that rich people save a higher percentage of their income and we want to spread the wealth. But if that’s our goal then we can just raise income taxes on the richest people—we don’t need to take a convoluted approach. Justifications about “keeping the money moving” are based on a misunderstanding—when I “save” money I’m lending it to someone else who will use it immediately.

Covid 7/22: Error Correction

Delta has taken over, and cases are rising rapidly, with a 58% rise this week after a 65% rise last week. There's no reason to expect this to turn around in the near term.

[Three weeks ago, in One Last Scare](#), I ran the numbers and concluded that most places in America would 'make it' without a big scary surge from Delta. It's time to look at what went wrong with that calculation, which I believe to be a failure to sufficiently integrate different parts of my model.

Then there's the question of what we are going to do about this, and whether we are going to destroy some combination of free speech and the ordinary day to day activities that constitute our lives and civilization, perhaps indefinitely, in the face of this situation. Such collateral damage has the potential to be far scarier and more deadly than the direct threat from Covid-19.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 4.7% (down 0.1%) and deaths unchanged.

Result: Positivity rate of 4.4% (down 0.4%) and deaths decline by 5%.

Other Result: Positive test counts rose by 58%, versus 65% last week.

I do not understand the divergence between positive test rates and case counts that we see here. This implies that the number of tests is scaling with the number of cases, but that wasn't true earlier, and also there's a lot of testing that takes place out of an abundance of caution or to provide the necessary evidence and/or paperwork in various contexts, where the tests clearly wouldn't scale. So this is weird.

In any case, it seems pointless in this context to make a positivity rate prediction instead of a case count prediction, since it doesn't tell us what we want to know. Thus, I'm switching to predicting case counts.

There's no reason to think cases won't continue to rise in the near term. Control systems should be kicking in and Delta has mostly completed taking over, so the rate of increase should continue to slowly decrease. I'm going to predict a 50% increase, down from 58% this week and 65% last week.

Prediction for next week: 360,000 cases (+50%) and 1845 deaths (+10%).

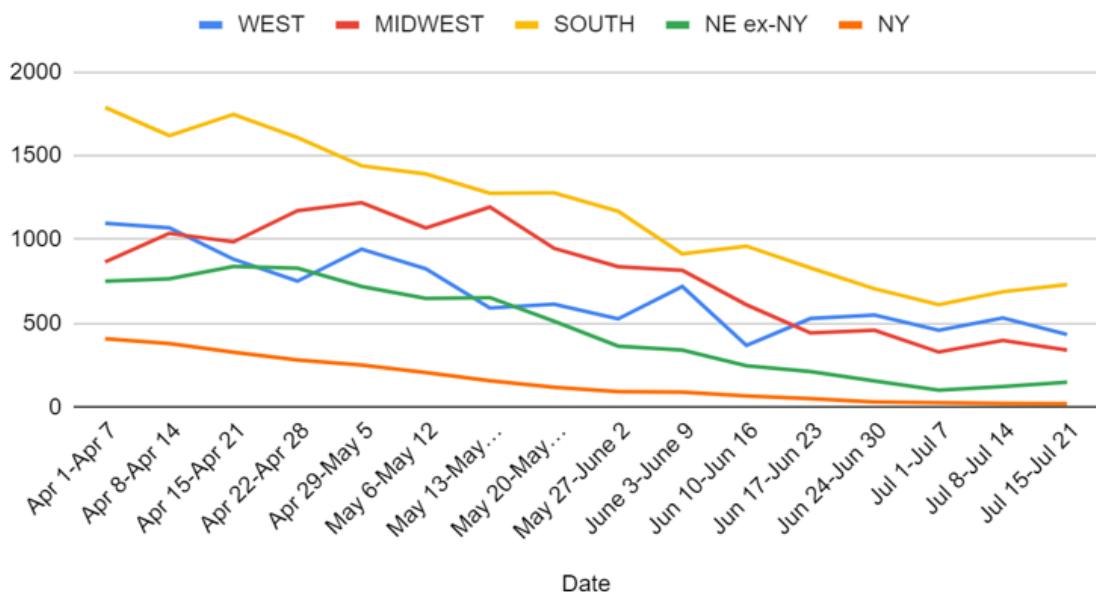
Predictions will be evaluated against data from Wikipedia, after correcting for obvious data anomalies.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
June 3-June 9	720	817	915	431	2883
Jun 10-Jun 16	368	611	961	314	2254

Jun 17-Jun 23	529	443	831	263	2066
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

Deaths by Region



There's no sign yet that people are dying from the new wave of cases. Most of the rise in cases came in the last two weeks, so we wouldn't expect a *dramatic* increase yet, but seeing a 5% rise over the last two weeks is still at least somewhat reassuring that what we've seen in places like the UK, where the IFR was dramatically reduced versus previous waves, will also happen here.

If we don't see a big rise in deaths within the next two weeks, that will be both very surprising and quite excellent news, as there will have been enough time for at least some of the new cases to have resulted in deaths. By three weeks from now we can be confident what new normal we are dealing with, at least as long as the hospitals have sufficient capacity.

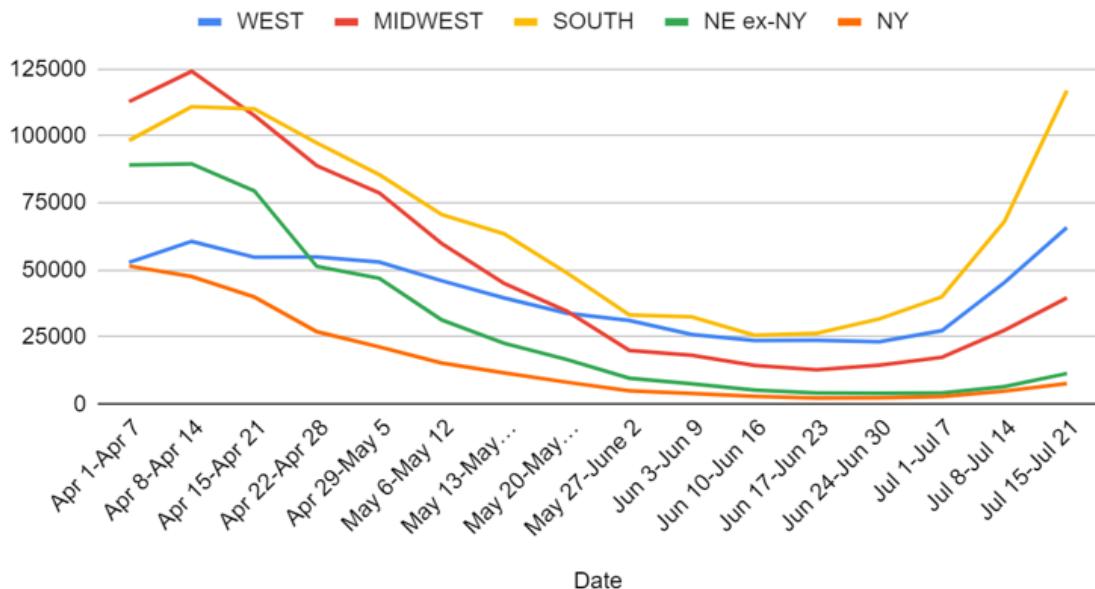
The cases remain the story.

Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 27-June 2	31,172	20,044	33,293	14,660	99,169
Jun 3-Jun 9	25,987	18,267	32,545	11,540	88,339
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

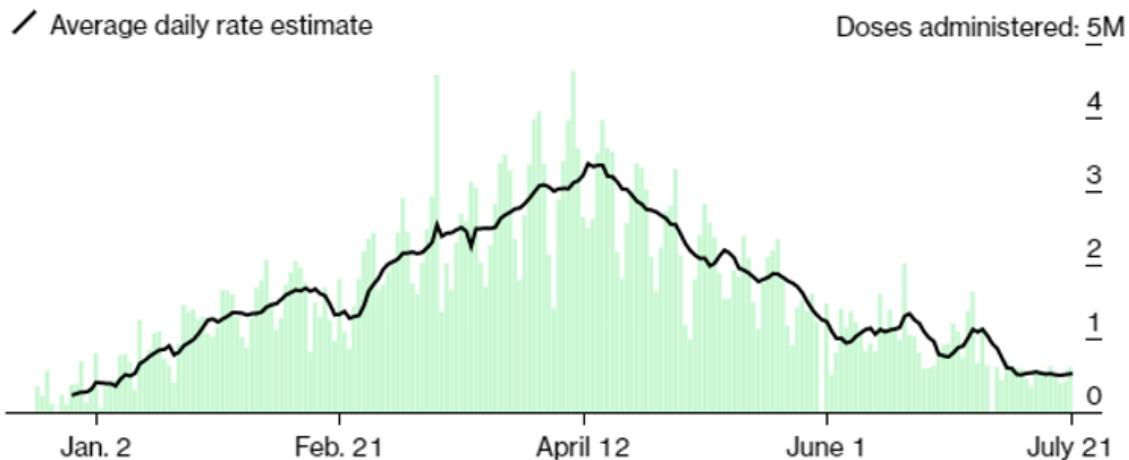
Positive Tests by Region



Last week was a 65% increase, and this week was a 58% increase. Delta has now mostly taken over, so differences are some combination of reporting, timing and testing details, seasonality changes, and control system adjustments. We should expect some people to alter their behaviors by now, and that will accelerate as cases pick up.

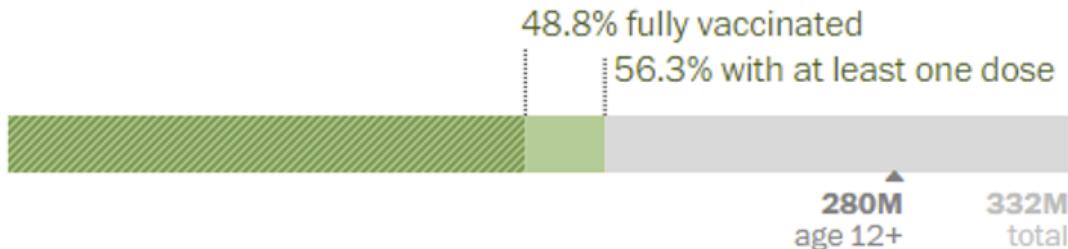
Either way, we have a 60% increase week over week as the new baseline, which would represent $R \sim 1.4$ based on the old assumption of a five day cycle. That remains consistent with the 2.2 multiplier on Delta versus a 1.4 multiplier on Alpha, plus the control system having adjusted to be stable under Alpha.

Vaccinations



186.8 million vaccinated

This includes more than **161.9 million people** who have been fully vaccinated in the United States.



In the last week, an average of **516.4k doses per day** were administered, a **3% decrease ↓** over the week before.

Vaccination rates are now roughly stable at about 500k/day, so about 1.5mm people/week going from unvaccinated to vaccinated, or about 0.4% of the population, and a resulting 1% or so decrease in R. If we think Delta is still on a five-day cycle, that's 1.5% less case growth each week. If Delta is on a three day cycle, it's 2.3% less case growth each week. More on that in the Delta section.

The good news is that [we are seeing more Republicans stepping up and telling their constituents to get vaccinated](#). That, combined with the threat from Delta, should help, despite the efforts of the Ministry of Truth. More discussion in that section.

Not Necessarily the News

When X is reported on the news, we learn *at least* three claims to evaluate:

1. X happened.
2. X was noticed.
3. X was news.

Depending on your prior knowledge and model, the same news report can change your world model in opposite directions, and often people get this calculation wrong.

If you see a plane crash reported and you know the media reports most plane crashes, that's bad news, but it's not importantly bad news, and it at least means there wasn't worse news crowding it out. If you see a plane crash reported and don't know that crashes always get reported, *it's good news* because you learn crashes are rare enough to be news.

This brings us to [this week's reports of infections taking place at a wedding](#).

The wedding took place outdoors in a large open-air tent. The wedding had 92 attendees, all of whom were required to have been fully vaccinated prior to the event. The first to get sick were a man and woman who had traveled from India. The man had had no existing medical problems and the woman had diabetes. Both had tested negative for the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) prior to boarding the airplane and had traveled to Houston 10 days after receiving their second doses of the Covaxin BBV152 vaccine, manufactured by Bharat Biotech.

This led to a bunch of reactions that were at core like this one:



Chana

@ChanaMessinger

...

Ahhhhhhh fucking hell



Greg Christie @Greg0706 · Jul 14

Forbes | Delta variant infects 6 fully vaccinated guests at U.S. outdoor wedding, eventually killing 1. All 92 wedding guests were required to be fully vaccinated. forbes.com/sites/brucelee...

Which in turn is doubtless causing a lot of this:



Nate Silver ✅

@NateSilver538

...

The Every! Breakthrough! Infection! Is! Big! News! news cycle is getting pretty annoying and is probably going to give vaccinated people a lot of unnecessary anxiety about Delta while also providing kindling to anti-vaxxers.

One could also observe that one person dying, or six being infected, *was news*, and worry perhaps even less than before, given how many weddings there are.

Then again, there was also this, in the comments last week:

[...] jimmy 4d ♂ < 10 >

My wife and I went to a wedding last week, and 6 of the 14 girls who went to the bachelorette party (all vaccinated) have since tested positive for covid. I and another one of the guys (also vaccinated) got it a few days later, most likely from one of those girls (and not our partners, who both tested negative).

It's just a few cases, but nonetheless seems hard to square with the lower bound of vaccines still working super well to prevent symptomatic infection.

EDIT: I have some more anecdotal evidence. My cousin just told me that his friend has nearly the exact same story with his wife's bachelorette party. 6/15 so far with symptoms and positive tests, all vaccinated. There's some selection bias there since I likely wouldn't have heard about it if it were only 1/15, but not enough to make it expected under a "just a fluke" model.

Twice is at least suspicious, and these two do seem to be close together in a meaningful way. So it's at least a little bit news.

Similarly, we have [the headline that 27 vaccinated people \(it says “nearly 30” but actually it’s 27\) in Louisiana died of Covid](#). The vaccines are so effective that 27 deaths was news, which is rather good news. If you run the numbers on Louisiana, I find roughly 900 deaths in the ‘vaccine era’ starting some time in March, so we’re talking roughly 3% of all deaths, from a group that includes most of the old people for most of that time. Yet I saw people freaking out about this, or wondering whether or not or how much they should be freaking.

[We also got more data on the Dutch music festival](#), where it seems 5% came back infected. It seems the festival was '[not entirely open air](#):



It also involved most everyone being unmasked in surrounding bars and restaurants. So the result here is not surprising, nor is it a warning about the dangers of outdoor events, or of anything but the usual rule of 'don't do stupid stuff.'

Delta Variant

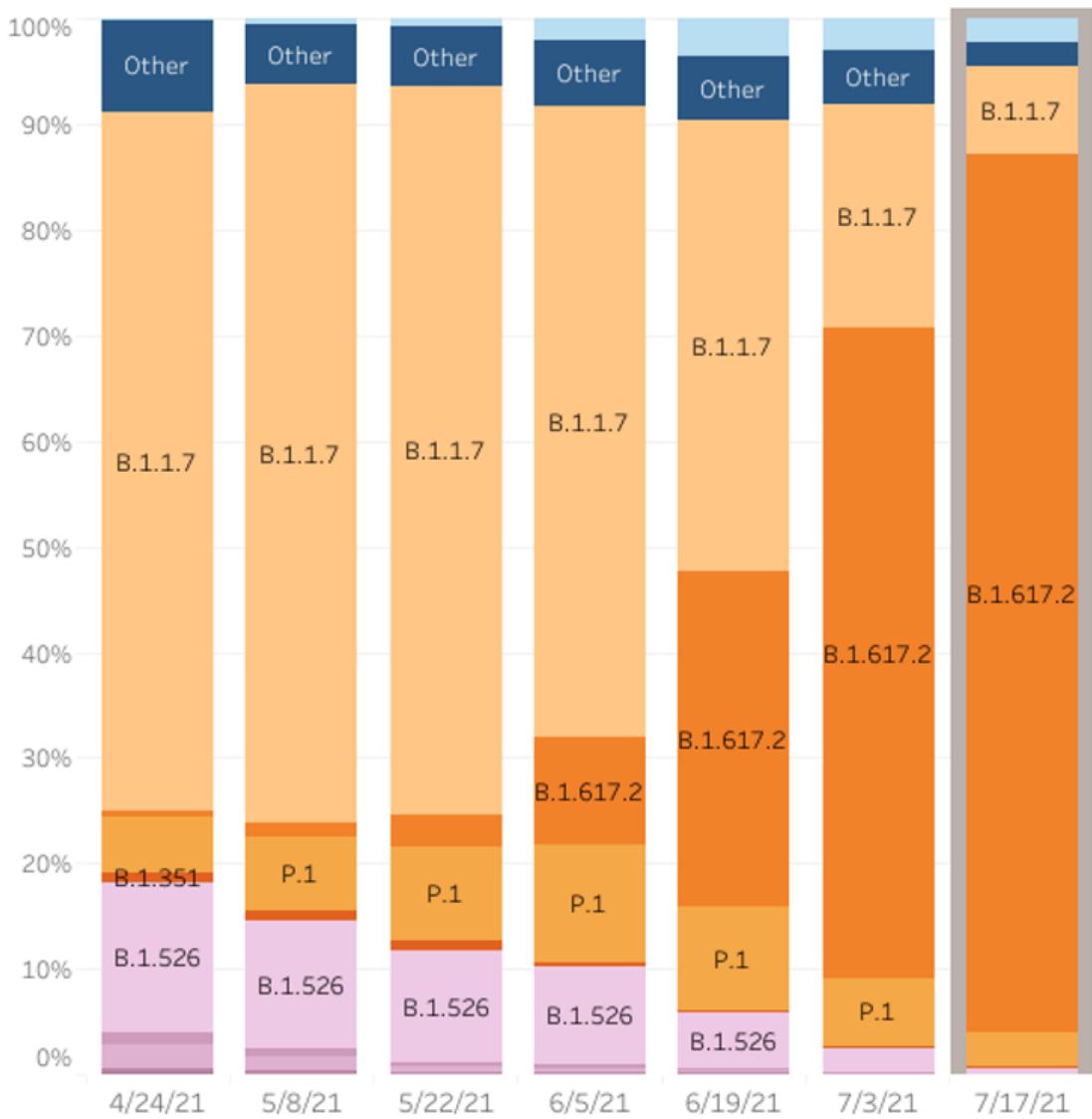
I always wonder in cases like this, how did they think a number like 83% gets distributed, and do they think a very slightly unequal distribution is somehow scarier or worse? Similarly, was going from 50% to 83% in two weeks unexpected somehow?

The rapidly spreading [Delta variant](#) now accounts for 83% of sequenced Covid-19 cases in the United States, the nation's public health leaders said Tuesday, and in parts of the country with low vaccination rates, it may be responsible for up to 90%.

Rochelle Walensky, the director of the Centers for Disease Control and Prevention, told a congressional hearing the new 83% estimate represents "a dramatic increase" from 50% of sequenced cases that were tallied the week of July 3. In some parts of the country that percentage is even higher, particularly in areas with low vaccination rates, said Anthony Fauci, head of the National Institute of Allergy and Infectious Diseases.

If you assume we started with a 50/50 split with Alpha versus Delta, and then there are three serial intervals since July 3, you get 79%. So this is very close to what you would expect based on my baseline estimates (with no loss of effectiveness from the vaccines, or very little), which is interesting in light of the discussions on Delta potentially being faster, which I'll talk about next.

[BNO Newsroom offers this graph](#) for easy reference, which has >50% in the July 3 bucket and thus has growth almost exactly in line with expectations. And the previous two week period was a *slower* takeover than one would have expected from the same model:

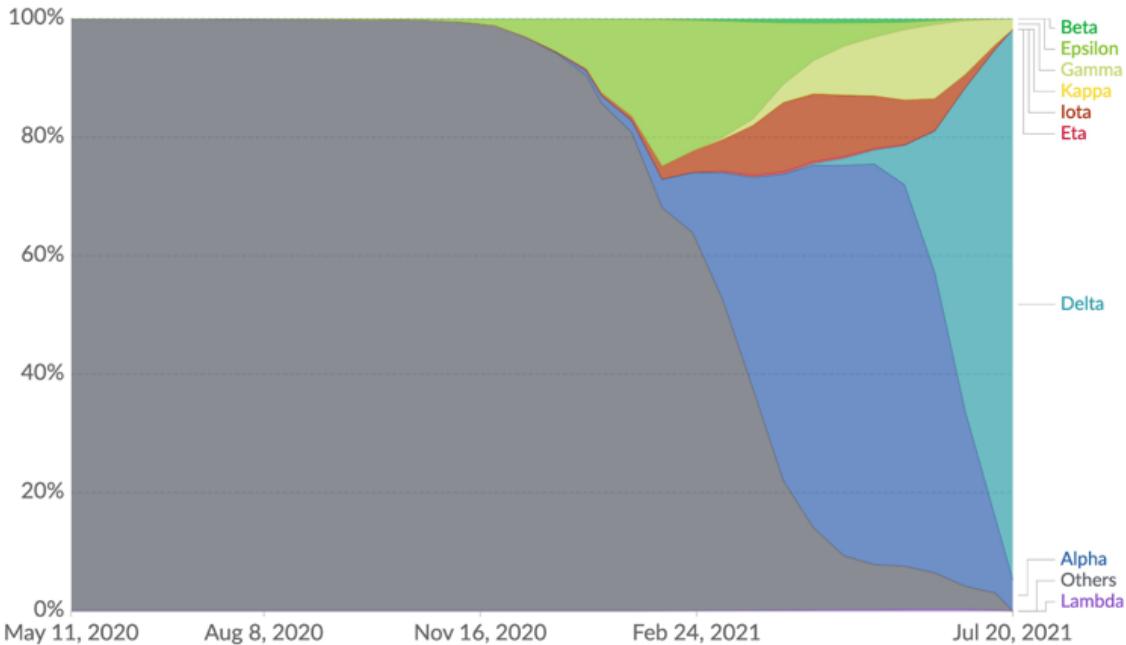


Also, OurWorldInData turns out to [graph this](#) the same way it graphs everything else, and I find their presentation very clean and easy to read:

SARS-CoV-2 variants in analyzed sequences, United States

Our World
in Data

The share of analyzed sequences in the last two weeks that correspond to each variant group. This share may not reflect the complete breakdown of cases since only a fraction of all cases are sequenced.



Source: CoVariants.org and GISAID – Last updated 20 July 2021, 21:10 (London time)

OurWorldInData.org/coronavirus • CC BY

An interesting and hopeful theory that came up was that Delta might be spreading so quickly in large part because it is *faster*. Under this hypothesis, rather than there being an average of five days between the time you catch Covid to when you give it to someone else, that interval could be as small as two days. That makes physical sense if the viral loads are much higher, as Delta would need less time to multiply in a new host before being able to spread.

If this was the case, then what *looks* like much higher rates of infection in graphs, and in the type of analysis that was run last week, is a lot less scary, and the actual R will be much closer to 1 (in theory, in both directions) than I calculated. Doubling every five days previously would have meant $R \sim 2$, but if that's two and a half cycles, then it means a much more fixable $R \sim 1.31$. What otherwise would look like a 'we're f***ed it's over' scenario might not be one.

What's the evidence for this? We have [this study out of China](#):

China (2). The mean generation time was 2.9 days (95% CI: 2.4–3.3), which was much shorter than that reported by Hu et al. in Hunan Province (2.9 vs. 5.7) (2). The mean serial interval was 2.3 days (95% CI: 1.4–3.3), which was also shorter than that reported by previous reports (2–3), and 21.6% (11/51) of serial intervals were negative (Figure 1C). We observed that 64.7% (44/68) of transmission events occurred during the pre-symptomatic phase, which is higher than that reported by Hu et al. (64.7% vs. 59.2%) (3). The transmission parameters suggested that suppressing the rapid spread and hidden transmission of this mutant virus is of high priority.

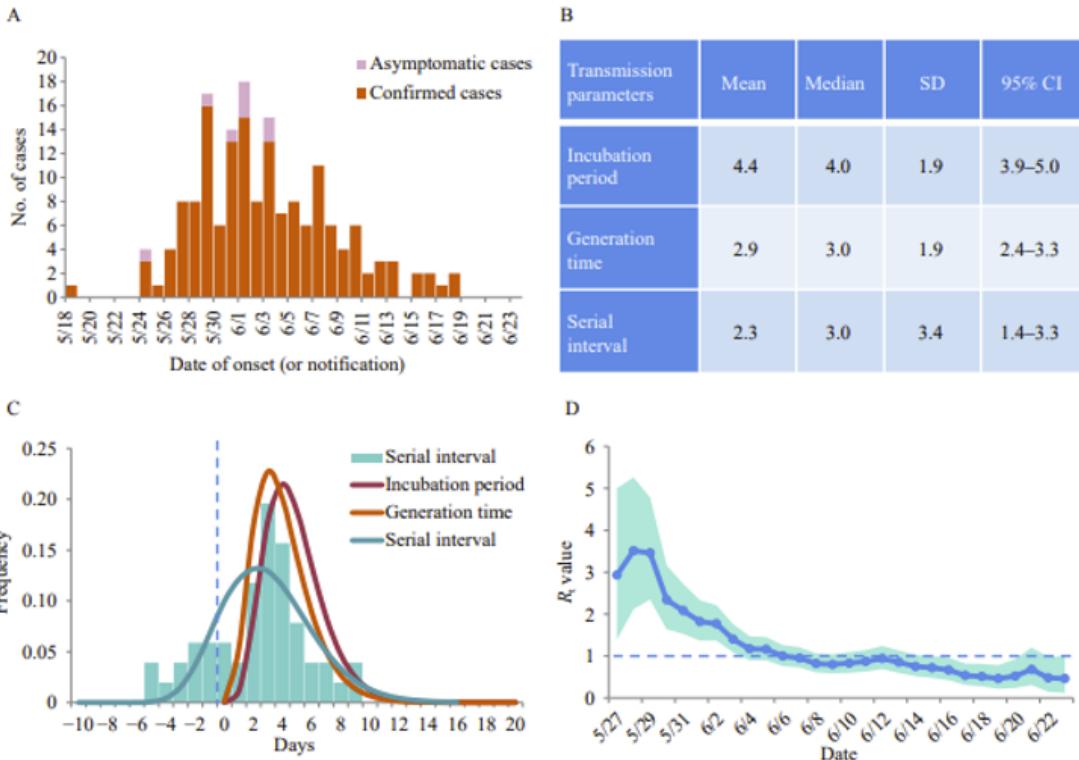
I am deeply confused how a serial interval can be *negative*. If I understand the words involved that means you spread it on to someone who gets their symptoms before you do? In which case, wow, that's quite the rapid spread.

It goes on to say this about R:

Based on the data of the cases with illness onset (or notification) between May 18 and May 29, and the GT of 2.9 days, the basic reproductive number (R_0) was estimated, which was defined as the expected number of additional cases that one case will generate. The estimated R_0 (maximum likelihood method) was 3.2 (95% CI: 2.0–4.8), which was much higher than 2.2 from Li et al. (2). Based on the GT and R_0 estimated, the epidemic growth rate (r , which represents transmission rate of epidemic with the formula of $r=[R_0 -1]/GT$) for the early stage of the outbreak was estimated as approximately 0.76 per day, which was about 100% higher than findings from previous epidemic strains (4). This result was in line with the Finlay et al. report that the transmissibility of Delta variant was increased by 97% (95% CI: 76%–117%) (5).

If R_0 went from 2.2 to 3.2 in this type of setting, that's less than a 50% increase from the original, so it's only 'in line' with the 97% increase reported by Finlay in the sense that they both observed the same rate of increase in cases, except Finlay assumed the old timing of infections and this new study believes things have sped up a lot. Thus, what would have been 97% is now slightly under 50%.

Their graphs are quite good. I wish more papers were 2 pages long with this much useful information:



I'm not fully sold, but it seems likely this is right. We are seeing *super duper fast* spread of Delta in some places, and not in others, and in some times and not others, such as when India went suddenly from an out of control epidemic to everything stabilizing quickly. Speeding up transmission makes all of that make a lot more sense.

A comment last week pointed to [this study](#) of vaccine effectiveness against Delta. I believe it had already been incorporated into the claims assessed last week, but good to explicitly note the primary source. I'm reproducing the bottom-line sections in full, skip if you don't want to dive into the details.

Among sequenced samples that were originally tested using the TaqPath assay, there was a high correlation between S-gene target status and the two variants under investigation with 87.5% of S-gene positive cases identified as B.1.617.2 and 99.7% of S-gene target negative cases identified as B.1.1.7 (supplementary tables 1 and 2). Results of the TNCC analysis are shown in **table 2**. In the ‘any vaccine’ analysis, effectiveness was notably lower after 1 dose of vaccine with B.1.617.2 cases 33.5% (95%CI: 20.6 to 44.3) compared to B.1.1.7 cases 51.1% (95%CI: 47.3 to 54.7). Results for dose 1 were similar for both vaccines. Following dose 2, the reduction in vaccine effectiveness was much smaller and non-significant: 86.8% (95%CI: 83.1 to 89.6) with B.1.1.7 and 80.9 (70.7 to 87.6) with B.1.617.2. With BNT162b2 there was a small reduction in effectiveness post dose 2 from 93.4% (95%CI: 90.4 to 95.5) with B.1.1.7 to 87.9% (95%CI: 78.2 to 93.2) with B.1.617.2. Numbers vaccinated with 2 doses of ChAdOx1 were smaller and the overall 2 dose vaccine effectiveness was lower than with BNT162b2 however the difference in vaccine effectiveness between B.1.1.7 and B.1.617.2 was small and non-significant: 66.1% (95% CI: 54.0 to 75.0) and 59.8% (95%CI: 28.9 to 77.3) respectively.

Table 3 shows the adjusted odds ratios for detection of B.1.617.2 relative to B.1.1.7 in vaccinated compared to unvaccinated individuals odds of cases having B.1.617.2 detected in vaccinated individuals was higher than in unvaccinated individuals for dose 1 of any vaccine (OR 1.38; 95% CI 1.10-1.72) and dose 2 of any vaccine (OR 1.60; 0.87-2.97). Given that vaccine effectiveness against symptomatic disease with B.1.1.7 is estimated at approximately 60% after dose 1 and 85% after dose 2,(10, 27) these results would indicate effectiveness of 45% and 76% respectively for B.1.617.2. By vaccine type the reduction in vaccine effectiveness appeared to be greater with ChAdOx1 (OR 1.48; 95%CI 1.18-1.87) than BNT162b2 (OR 1.17; 95%CI 0.82-1.67) though confidence intervals overlapped. The sensitivity analysis comparing to the 0-13 day post dose 1 period gave a similar pattern of results though the odds ratios were smaller and not statistically significant (supplementary table 3). This was also the case with the matched analysis (supplementary table 4).

Main findings

We found an absolute reduction of one dose vaccine effectiveness against symptomatic disease with the B.1.617.2 variant of approximately 20% when compared to the B.1.1.7 variant. However, reductions in vaccine effectiveness after two doses were very small. This was the case for both the BNT162b2 and ChAdOx1 vaccines. Using a TNCC analysis, estimated vaccine effectiveness against symptomatic disease with B.1.617.2 for a single dose of either vaccine is approximately 33%, for two doses of BNT162b2 is approximately 88% and for two doses of ChAdOx1 is approximately 60%.

Interpretation

These findings suggest a modest reduction in vaccine effectiveness. Nevertheless, a clear effect of both vaccines was noted with high levels of effectiveness after two doses. Vaccine effects after two doses of ChAdOx1 vaccine were smaller than for BNT162b2 against either variant. This is consistent with reported clinical trial findings. However, rollout of second doses of ChAdOx1 was later than BNT162b2 and the difference may be explained by the limited follow-up after two doses of ChAdOx1 if it takes more than two weeks to reach maximum effectiveness with this vaccine. Consistent with this, 74% of those who had received 2 doses of ChAdOx1 had done so between 2 and 4 weeks prior to symptom onset compared to 46% with BNT162b2 (supplementary figure 1).

In general, it's potentially highly misleading to compare the vaccinated to the unvaccinated in the wild, because the two groups differ in a lot of ways. I'm not entirely sure which

direction this goes, as the vaccinated start out with safer behaviors but then change behaviors based on being vaccinated.

Here, we can compare measured vaccine effectiveness between different strains. The obvious worry then is that there could be a difference in which populations are dealing with which strains during this period, which could skew the results as well. These are not controlled experiments. One thing that makes me more confident here is that we see other adjustments and measurements that don't seem out of whack.

The headline conclusion is then that mRNA vaccines retain 88% effectiveness against positive tests. If we accepted this figure, we'd then need to translate that into a measure of how often such people transmit. With milder cases and lower viral loads, the presumption is that they don't transmit as effectively, but the flip side is that milder cases mean we might be missing a larger percentage of cases, so the 88% number might be high for that reason. It also might be low or high for several other reasons.

[Here's CellBioGuy in the comments at LessWrong:](#)

CellBioGuy 4d ⚭ < 5 >

It should be noted, for your contagion calculations, that people infected through immune memory are almost certainly not NEARLY as infectious to others on average as completely naive people who are infected.

<https://pubmed.ncbi.nlm.nih.gov/34250518/>

Israeli healthcare workers who are vaccinated who test positive have a much decreased viral RNA level in their samples, circa a factor of thirty, with the difference increasing as time from vaccination increases.

There is a wide range but the whole range moves down so way fewer people will have the obscene viral levels that can do things like infecting sixty people in a room all at once. They also go from 80% showing up positive on an antigen test to 30% showing up positive on an antigen test - which has been a reasonable binary proxy for infectiousness in the past.

This makes sense. You never hear about 60 people all getting the flu at once in one place except in very special circumstances (things like an airplane sealed for two hours without air circulation and filtration), presumably because everyone has at least some anti-flu memory even if it isn't good enough to completely stop everything in its tracks. The disease dynamics when there are lots of totally naive people running around is going to be completely different from the dynamics when everyone has memory, be it from vaccines or infection.

[Reply](#)

Most of the variance remains in the difference between measurements in different places, but I think all of it points to roughly the same place anyway.

The numbers will come in somewhere in the range where fully vaccinated groups won't have outbreaks unless they partake in a lot of what I call 'stupid stuff,' which is basically (some combination of most of) packing lots of people tightly into indoor spaces without proper ventilation for extended periods. However, it also would mean we're close enough to the edge that if everything went *fully* back to normal, we'd need more people vaccinated than we can realistically hope for in the next few months or perhaps ever.

How worried should a vaccinated person be about Delta?

In terms of death, [seriously not very much](#), vaccinated people don't die of Covid and Delta doesn't change that. Thread points to a few different claims about whether Delta is deadlier and by how much, but it's definitely not enough to overcome the vaccinations or even put much of a dent in them.

How the UK's vaccine rollout has dramatically reduced Covid-19 deaths

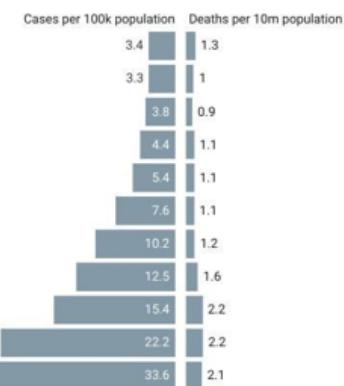
Cases versus deaths over days 1–50 of the UK's second and third Covid waves

Second Wave



[Get the data](#) • Created with Datawrapper

Third wave



[Get the data](#) • Created with Datawrapper

Calculations based on a seven-day rolling average of daily recorded cases and deaths. Second wave is recorded from 8/9/20, third wave is recorded from 14/5/21.
Source: UK Government, ONS

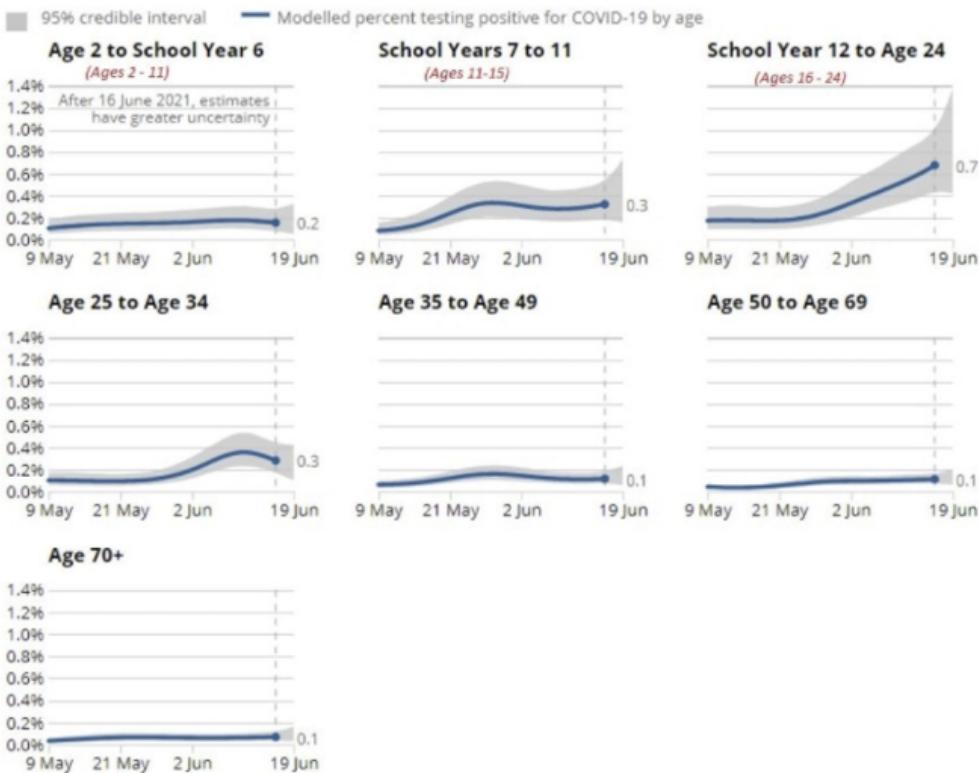
NewStatesman

The question is *entirely* one of the unknown unknown risks of Long Covid. Getting data on this, or being confident in a position, is incredibly hard, whether it's on how big the Long Covid risk was to begin with, or how much the vaccines reduce that risk. It would be completely biologically bizarre if the risk wasn't greatly reduced by the vaccine the same as everything else, but it's still enough of a worry that one would strongly prefer not getting Covid, if that was an option.

I wish I could do better than that, but I really can't give one, given what I know. My guess is that we're talking about a small but non-zero chance (3%?) of some amount of lasting effects of some kind for the vaccinated, most of which are minor and temporary, but yeah, who the hell knows.

For young kids, that's even more true – the danger is *purely* Long Covid. [There's a good 'what's up with Delta and kids' analysis up this week](#), although it doesn't offer us anything concrete that's new, and it points out that the Long Covid risks haven't changed and that other diseases also have similar long tails, we just don't talk much about them. And that even with unmasked schools and lots of vaccinations elsewhere, we're not seeing an explosion in cases among those schoolchildren too young for the vaccine, as a percentage of the cases in the population when there's a Delta wave in the UK:

Estimated percentage of the population testing positive for the coronavirus (COVID-19) on nose and throat swabs, daily, by age group from 9 May to 19 June 2021, England



It is well known that city living leads to more infectious diseases than non-city living, to a very large degree. There's a small long tail for many of those diseases, where people develop long term problems. One of the periodic reminders, as we move into the next phase of the pandemic and beyond, is that *if you are worried about Long Covid as a vaccinated person, why aren't you completely panicked about living in a city?*

That brings us back to the calculations. The spread of Delta in the United States as a share of cases is exactly in line with the 2.2 vs. 1.4 difference from Alpha to Delta, as of earlier this week, which does not leave room for vaccines to additionally lose effectiveness. Then, if it turns out Delta replicates faster, that changes the math once again.

One Last Scare: Re-Evaluation

This week, I was persuaded to add a post-mortem to [my big We're F***ed, It's Over post](#) from the end of 2020. Reading it over again, I believe the core logic of that post was solid – we were not capable of adjusting the control system sufficiently to contain a 65% more infectious strain in time given the expected rate of vaccinations. I predicted a 70% chance that we had such an infectious strain and that if we did, we would face this crisis and have no reasonable options.

It didn't happen. Instead we had a 40% more infectious strain, and faster vaccinations, which combined as *that post's model said it would*, to prevent the wave. We did better than I would have expected even then, with cases coming down much faster, so there was even more going on than that. In any case, the prediction was too confident, and didn't properly adjust

for uncertainty over speed of vaccinations or seasonality. Perhaps there was even some degree of [self-preventing prophecy](#) when you combine it with others' similar warnings. I do think giving the scenario less than a 30%-40% chance would have been *more wrong* than giving it a 70% chance, but that numbers over 50% were too high.

This time around, I predicted:

Right now, we are cutting cases in half every 3 weeks or so, which is about 4 cycles, so I'd estimate R_0 at 0.84. Increasing that by 35% puts us at 1.14. That's presumably high, because sequencing is delayed and thus the current Delta share is higher than in the above calculation. I don't consider 35% a strict upper bound, but my mean estimate is more like 30%, and every little bit helps.

Thus, it looks clear to me that *most places* in America are going to make it given the additional vaccinations that will take place, but some places with low vaccination rates will fall short.

Alas, that does mean that it might be a while before the final restrictions can be lifted, and life returns 100% fully back to normal. We will be stuck in a kind of hybrid limbo, mostly involving performative mask wearing and other such annoyances. But that's actually pretty close to fully normal, in terms of practical consequences for adults. For the kids, things could stay rough for a while, because people are really stupid about such things. I hope we can get the vaccines approved for the Under 12 crowd soon.

It does look like masks will be around for a while, and might be making a comeback – Biden is considering reinstating them in many situations, or at least trying to do so, and many schools look poised to torture their students this way, and several jurisdictions (including Los Angeles and Las Vegas) are bringing back indoor mask mandates already on their own.

We also aren't seeing that many calls for anything *beyond* mask mandates yet, but I do see the beginnings of 'schools can't be open' talk as well. I would like to think we'd never let that fly at this point, but perhaps we would, at least in some places. I do think that if you're a parent in such a place, and they do try to put your kid in 'remote learning,' you should find an alternative even at an extreme cost, and if necessary consider moving.

Regardless of all that, I know that at the time I did not expect this amount of increase in case counts, and thus my model of the future was importantly wrong. What were the errors? What has changed?

First it's important to know what *hasn't* changed: I still have Alpha at 40% more infectious (1.4x) than baseline, and Delta as 120% more infectious than baseline (2.2x). Those estimates are doubtless not exact, but I haven't seen any reason to adjust them. So that wasn't the problem.

Nor was the issue (as far as I can tell) that vaccines have lost effectiveness. It does seem like vaccines are *slightly* less effective against Delta, but I continue to believe this effect is not a big impact. Vaccinated people remain very hard to infect and, when infected anyway, poor carriers with which to infect others. This shouldn't have moved the needle enough to get noticed.

This is reflected in the growth of Delta as a share of cases, which matches very closely what these numbers imply, and doesn't leave room for them to be that off in either direction. Similarly, it looks like Delta plausibly replicates faster than we thought, but that probably also would mean it isn't as infectious and has a lower R, or else the numbers don't work out.

Thus, I do not think the prediction error was about a misvaluation of Delta. I think the error was about a misvaluation of where things stood before Delta, and what people were up to.

It's that first sentence, where I start off R at 0.84, instead of factoring in the changes coming from the control system. With the decrease in masks worn and f***s given over the last month or so, combined with seasonality changes, the R without Delta likely went from 0.84 back to at least 1. That's a 19% difference each cycle, or a 28% per week.

In the world where we had retained the behaviors that were cutting cases in half every three weeks, the current rates of increase would be more than cut in half, and it would be easy to see that additional vaccinations (and some amount of Delta burning out in the younger populations where it's spreading the most) would reverse the problem before it got into crisis mode, even if our current case starts with R=1 exactly.

However, we're not in that world, and we're starting from a higher baseline. That mistake compounds each week, and now only a few weeks later we are where we are, with exponential growth looming quickly.

In short, I think this was mostly a pretty dumb mistake that should have been easy to spot - I knew in one place that we were adjusting things, and then didn't make that adjustment when I did this other calculation. My models were insufficiently integrated.

The prediction here is then saying something about what happens *if we return to the behavior patterns we had when cases were declining rapidly*. The extra vaccinations would be sufficient, in most places, to compensate for Delta. The problem is that we're not doing anything close to that, haven't for some time, and it would be a hell of a thing to try to return us to that state. [Even if we could, that doesn't mean we should.](#)

That's also a pretty easy call to make, when one puts it that way. Delta is likely a little over twice as infectious as the original. Over half the country's adults are vaccinated. Of course that's enough to compensate. Easy math is easy.

Or, to do the rough calculation another way, Delta cancels out the vaccination of the first 55%-60% or so of the adult population, or the first 46%-50% of the overall population, if there's no other source of immunity running around. We are currently at 49% fully vaccinated and 57% partly vaccinated, or effectively about 53%. So we're still ahead, but we're not that far ahead, and we definitely can't go back to anything like normal unless we're willing to accept the consequences.

From here on in, mostly the unvaccinated will be infected, and most of them will be young. Last week, we had 240k positive tests and vaccinated about 1.5mm people. With rapid weekly case growth, it won't be too long before we're giving immunity to our unvaccinated youth the hard way, via infections, faster than we can vaccinate people. It's not the preferred solution, but it does work, and it works fast, especially since it tends to kick in about when other control systems also kick in. Which is why we see rapid increases time and again suddenly turn into rapid declines all of a sudden.

The question is, to what extent are we willing to accept those consequences, versus willing to accept the costs of not accepting them? There's no longer a reasonable expectation that if we kick the can far enough down the road that something will change, and the consequences of permanently kicking the can seem far, far worse than the consequences when the can is not kicked.

Speaking of which...

Ministry of Truth

So, [this happened](#):

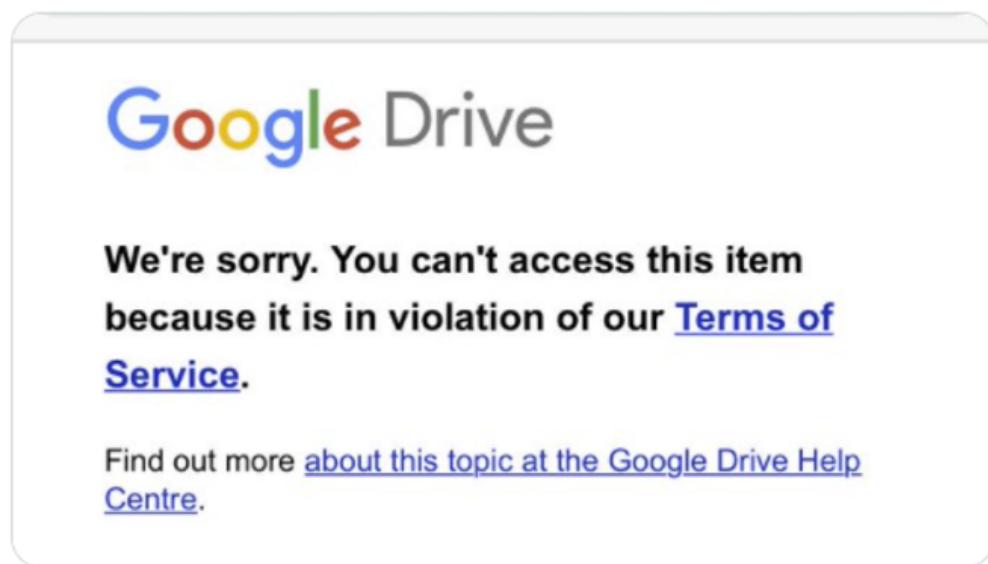


Anna Brees @BreesAnna · Apr 21

First time I've seen a cloud drive blocking a document...

...

It was a very long document re vaccination headlines from around the world.



A screenshot of a Google Drive error message. The message reads: "We're sorry. You can't access this item because it is in violation of our [Terms of Service](#). Find out more [about this topic at the Google Drive Help Centre](#).

270

1.2K

2.1K



Anna Brees @BreesAnna · Apr 21

This was a document sent to me via email. I was able to view it three days ago but now it says this.

31

49

287



Google is blocking your access to documents, based on them containing the wrong statements about vaccines. While this turns out to (for now) likely not involve private documents, and only stop the sharing of information, you should presume both that your documents in Google Drive are not private when it counts, and also that you could lose access to them at any time, especially if they are technically 'shared' as many of mine are. It's not as scary as I originally thought, but it's still scary.

I already knew about such issues, but this drove it home. [More new is this \(link to video\)](#):



zooko ❤️ 🇺🇸 @zooko · Jul 16

...

I'd like to suggest that Biden administration go ahead and set up a department to guide and institutionalize this anti-misinformation campaign. We could call it "The Ministry of Truth".

TH Townhall.com ✅ @townhallcom · Jul 16

PSAKI: If you're banned on one social media platform, you should be banned on other social media platforms.

[Show this thread](#)



WH.GOV

0:18 | 1.3M views

16

45

243



zooko ❤️ 🇺🇸 @zooko · Jul 17

...

It has a nice ring to it.



Townhall.com ✅ @townhallcom · Jul 16

...

Replying to @townhallcom

Only MINUTES LATER, Jen Psaki calls social media sites like Facebook "public platforms."

[There's also this:](#)



Mike Solana @micsolana · Jul 16

...

this after biden's press secretary admitted the white house provides cues to facebook on what to censor, and insisted when one platform bans an account, all platforms should follow. the push here is very clearly for a centralization of platform authority, and mass censorship.

 **The Hill** ✅ @thehill · Jul 16

Reporter: "What's your message to platforms like Facebook?"

President Biden: "They're killing people."



18

60

477



Mike Solana @micsolana · Jul 16

...

are we still "resisting" authoritarianism or does that hashtag go back on the shelf once authoritarianism becomes an actual threat?

[Mike's full-post take on the situation is here.](#)

A call from the executive branch, for social media platforms to coordinate, and if you're banned on one of them for 'misinformation' you need to be banned on all of them, or the government will take action to break up this private monopoly of a public platform. Also, they need to 'work harder' to 'fight the spread of misinformation' via censorship and bannings of this type, or again, they will take action to break up this private monopoly of a 'public platform.' If they don't do that, they are 'killing people.'

[He tried to walk it back:](#)



Felicia Sonmez @feliciasonmez · 11h

Biden again clarifies his "killing people" comments about Facebook and misinformation: "There was a report out saying that something like 45 percent of the overwhelming disinformation on Facebook comes from 12 individuals. I said, 'They're killing people'--those 12 individuals."

However, that's not how the language works. It *is* how the language of power works, where one makes one's statements as ambiguous as possible and states one's message implicitly whenever one can, so that one can get the message out and then deny sending it.

Meanwhile, from NPR via that column, the following definition of 'misinformation' when a Bad Person is providing the information:

"They tend to not provide very much context for the information that they are providing," Settle said. "If you've stripped enough context away, any piece of truth can become a piece of misinformation."

Misinformation, it seems, could mean anything that gives an impression the Powers That Be dislike.

Such policies have often taken aim at anything that 'contradicts the CDC guidelines' or used other such principles, despite such guidelines often being obvious nonsense.

So, *that* happened. As usual, think about this apparatus, and this move, in the hands of the outgroup rather than the ingroup, or the hands of the fargroup, if you think that it might possibly ever not be one of the worst possible ideas.

As gentle reminders from earlier in this epidemic, this 'misinformation' would at one point have included the fact that masks work, or that the virus could have come from a lab, or that we could expect to perhaps have a vaccine by the end of 2020, or if you go back to February that there was even a Covid-19 problem to begin with or that one should prepare for it, because that's not only false, it's also racist. Or that Covid-19 is airborne, or that surface cleaning wasn't all that important. And that's purely from the current pandemic and without thinking about what the outgroup would have done with those levers if it had the chance.

Under such a regime – or under the current regime that existed at the time, even – if my posts had been placed on social media, I'd have been banned from all of them many times over. That's where things already are, now. What happens when 'misinformation' increasingly becomes whatever the executive or the media narrative decide they don't like? And then the executive decides he doesn't like those who don't like them, or are saying politically inconvenient things?

[Many have noted that the call for government-directed censorship of social media is not only far along on the road to authoritarianism and the end of freedom of speech, it also doesn't have much prospect of a big impact on its supposed target either.](#) Link is to one such thread. It's almost as if the government and ingroup establishment are using the 'emergency' and the excuse of the pandemic in order to further their goal of becoming the thought police and telling us what we can and can't say to each other.

Things offline are not entirely better, but in the interests of illustration by example and a desire not to cause a distraction, I've censored the example I had previously put in this spot from the past week. Stay on target.

Vaccine Hesitancy

[What's actually going on](#) with vaccine hesitancy ([link to CNN post](#))?

The image shows two consecutive tweets from Ariel Edwards-Levy (@aedwardslevy) on July 18, 2021. The first tweet discusses the shift from vaccination discussions to persuasion debates, mentioning the importance of margins. The second tweet links to a KFF poll tracking vaccine intentions from January to June 2021.

Ariel Edwards-Levy  @aedwardslevy · Jul 18
Really feels like the "how do we get more Americans vaccinated" discussion has somehow turned into yet another round of "Does persuasion actually work?" and that the answer is probably once again some variation on "Mostly at the margins, but the margins are important."

93 230 1.6K 

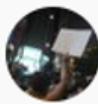
Ariel Edwards-Levy  @aedwardslevy · Jul 18
(Along these lines, this KFF poll tracking vaccine intentions from January  June is worth a read: cnn.com/2021/07/13/pol...)

... 

When we ask whether persuasion works, I mean, *of course it works*. The issue is that *you're not the only one doing persuasion*, and also you're not doing that great a job of it, in the sense that this thing has been massively botched several times over. Persuasion matters, doing it better matters, and what we got reflects how we did at it. And yes, every little bit helps and we might be close to a tipping point.

You know how uninterested we are in persuading people? Not only did we suspend the J&J vaccine over nothing, and recently put another warning on it over another nothing, we're [not even bothering to fully approve the vaccines](#), with all that this entails. Let alone the other low hanging fruit mentioned at the link.

[Kelsey puts this well:](#)



Kelsey Piper ✅ @KelseyTuoc · 12h

...

"People I know seemed concerned about the fact that the vaccine was approved only for emergency use." Monica Potts writes. This "please take the vaccine, but we won't fully authorize it, but please take it" thing is killing people.



My Community Refuses to Get Vaccinated. Now Delta Is Here.

Only 35 percent of Arkansas is fully vaccinated, and with case rates rising, living there can feel like moving through a distorted reality.

🔗 theatlantic.com

5

11

76



Kelsey Piper ✅ @KelseyTuoc · 12h

...

Obviously some vaccine hesitancy will remain even once full approval comes through, but a lot might not! It might provide people who are starting to reconsider not being vaccinated a graceful, identifiable point at which to change their minds.

1

1

31



Kelsey Piper ✅ @KelseyTuoc · 12h

...

When I got my vaccine I had to read a dense form telling me that the FDA had NOT AUTHORIZED the vaccine I was about to take. The setup here is tailor-made to encourage hesitancy.

Or even more bluntly:



Noah Smith 🐰 ✅
@Noahpinion

...

FDA: We can't fully authorize the vaccine because doing that too quickly would undermine trust in the system

Antivaxers: DON'T TAKE THE VACCINES, THEY'RE NOT FULLY AUTHORIZED

10:06 PM · Jul 21, 2021 · Twitter Web App

Let's not pretend we're taking this seriously. [Matthew also notes this](#):



Matthew Yglesias ✅ @mattyglesias · 16h

...

Two points from readers who've written in responding to today's piece.

One — companies who've imposed either hard vaccine mandates or else soft ones ("you have to wear a mask unless you're vaccinated") report very high compliance rates.



Matthew Yglesias ✅ @mattyglesias · 16h

...

Two — a lot of people in Q&A sessions believe the FDA permanently yanked approval for the J&J shot (the correction is always less prominent than the original story) and this plus the "experimental" status of the authorizations shows the shots are risky.

This is no different than anything else. Vaccine persuasion is about persuading others that we are Very Serious People who have made the proper sacrifices, rather than asking what would work.

A better question is, *how much* does persuasion at the margin, now matter? The persuasion that mattered most largely happened by January. Those who were persuaded by then mostly stayed persuaded and got their shots. Those who weren't largely didn't change their minds later. But why would they? Yes, some new evidence was presented that vaccines were safe and effective, but also the problem seems far less urgent now. Until that changes, it's not like we did some great persuading and it didn't work.

One big piece of evidence is that [most old people went ahead and got vaccinated](#).

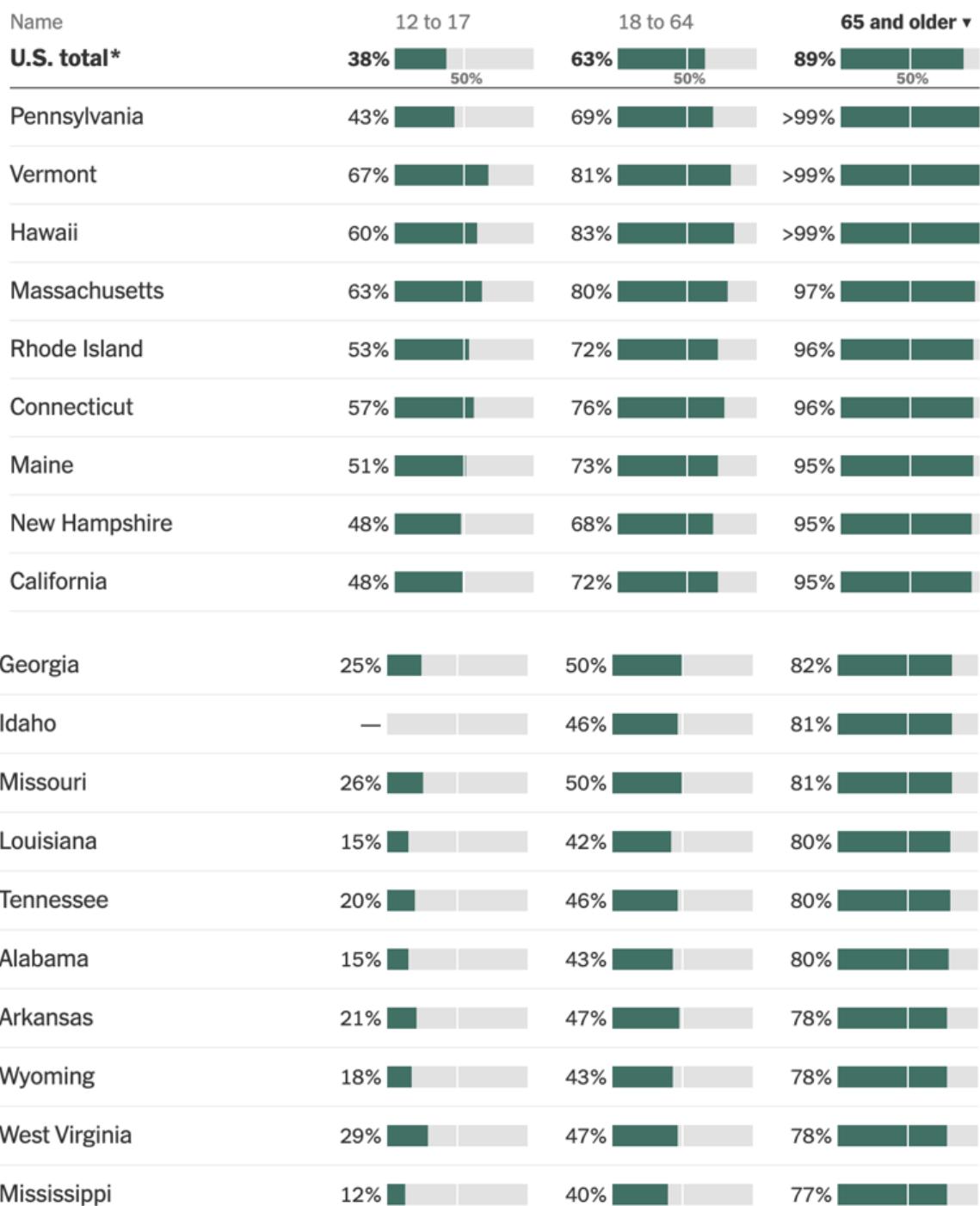


Nate Silver ✅ @NateSilver538 · Jul 17

...

FWIW, there might be some insights here in terms of vaccine hesitancy. Very few old people—who face a MUCH higher risk of hospitalization and death—are willing to screw around & go unvaccinated for "ideological" reasons, even though many old people are conservative Republicans.

Percentage of residents given at least one shot, by age group



This looks like a world in which people are doing a calculation to decide whether to get vaccinated – they're simply doing a *different* calculation, where the decision is less obvious, and those who most need the vaccine mostly still end up getting it.

This in turn implies that much of the remaining ‘hesitancy’ or even refusal isn’t ‘I’m never doing this no matter what’ and it’s more like ‘I don’t have enough skin in the game so I’d

prefer to play it what looks to me like safe and/or not bother and/or not deal with the temporary side effects and/or continue signaling to my in-group.'

Which is great news, because if Delta ends up everywhere, where chances of getting infected if you're not vaccinated get very high, then one would expect a lot of people to cave and get vaccinated rather than accept getting infected.

And that's [despite some pretty out there world models](#), even if you subtract Lizardman's Constant:



kain.eth @kaiynne · Jul 19

...

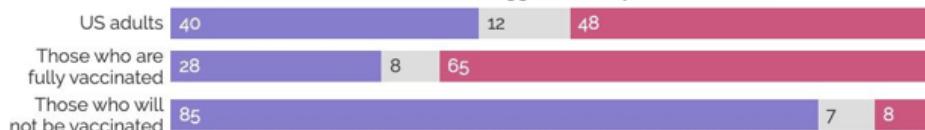
My favourite part are the 9% of fully vaccinated that clearly were in it for the free microchips.

One in five Americans believes the US government is using the COVID-19 vaccine to microchip the population

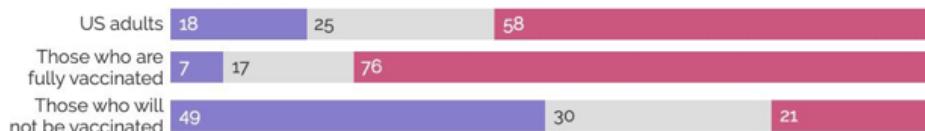
In your opinion, how likely is it that the following scenarios are true? (%)

Definitely / probably true Not sure Definitely / probably false

The threat of the coronavirus was exaggerated for political reasons



Vaccines have been shown to cause autism



The U.S. government is using the COVID-19 vaccine to microchip the population



YouGov

The Economist / YouGov | July 10 - 13, 2021 | Get the data

That the same 50% of the unwilling believe both that vaccines have been shown to cause autism and that the US government is using them to microchip the population is suggestive that such people are not processing such statements as containing words that possess meanings. They're simply taking the opportunity to say 'rare, vaccine bad!' in any way that's presented to them. Thus, a lot of them believe *both* that the vaccines cause autism and *also* that they're being used to microchip people. Unless the theory is that it's the microchips that cause autism? I kinda want to see the overlap in the crosstabs, I'm expecting to see a lot of it.

Thus, my best guess is that about half the 'hesitant' are getable through some combination of things getting bad and us picking the low hanging fruit like approving the vaccines, and the other half likely require stronger stuff.

Once and Future Lockdown

Could it happen again? [Janet Yellen thinks so.](#)

SARA EISEN: Do you think we could see lockdowns again? Do you think our economy can, can handle going through something like that again?

SECRETARY JANET YELLEN: So, I mean, we have solid rates of vaccination in many parts of the country. But certainly it, it's something that could, could happen in areas where vaccination rates are low so it really is critically important that we maintain progress on vaccinating more Americans.

Nate Silver mostly disagrees.



Nate Silver @NateSilver538 · Jul 16

Not sure this makes sense. The low vaccination areas are in conservative states/counties that are extremely lockdown-resistant.



Nate Silver @NateSilver538 · Jul 16

The whole problem is that vaccines are by far the most effective intervention, so if you're an [insert pejorative term] who's not willing to accept vaccination, it's hard to see how you'd accept other interventions (e.g. lockdown) that have a much worse cost-benefit ratio.

Nate's mistake here is to act as if a cost-benefit ratio is all that relevant to how decisions are made on Covid. Somehow we have decided that 'forcing' people to take the Covid vaccine is unacceptable, and that's that. So our choices are instead forcibly disrupting people's lives in the hopes that it helps, or not doing that. If things get bad enough, it makes perfect sense that we'd potentially see lockdowns but not vaccine mandates, and that those lockdowns likely won't make exceptions for vaccination, because we've also made it unacceptable to check someone's vaccination status in most contexts and places.

Where I think Yellen is clearly wrong is in expecting the places with low vaccination rates to be the ones that lock down. It's almost certainly the opposite. If lockdowns happen, they will happen in the places with relatively *high* vaccination rates. Not the highest like Vermont, since they'll have no need for it (probably), but in the various blue states that time and again have gone overboard with prevention. There's zero appetite for locking down red states.

I'd hope there was zero appetite for locking down anywhere, but I am growing more worried about this possibility. It's really stupid, because it wouldn't work. Even if it did suppress Covid entirely in the local area, the moment you stop it comes back, so what's the point? When will things change? Are you going to keep this up for years?

I do see signs that there's support for doing exactly that. Some of this is 'avoid blame on a two week time horizon' where the fact that the problem never goes away isn't relevant, but some people *really do support permanent ending of life as we know it*. I don't understand why they are so cool with this, it seems like the later stages of a Persona game or something, but it is what it is.

Meanwhile, [Biden tells us our young children will be wearing masks](#), whether they like it or not:



Jennifer Epstein ✅

@jeneps

...

Biden says the CDC is going to say that kids under 12 "should probably be wearing a mask" in schools. He says vaccine approval for kids under 12 is coming "soon" but won't predict when because "I do not tell any scientists what they should do. I do not interfere."

I am getting really tired of this malarkey line about not interfering with 'scientists' as if they're all identical clones who reach all the Officially Scientifically Correct conclusions, and thus one doesn't have to take responsibility for decisions if you can cite one. You did it, sir. You.

Similarly, Washington Post reports that the [Biden administration is debating urging a return to masking for the vaccinated](#). It would 'have to come from the CDC' but they've 'taken a hands off approach to avoid interfering.'

So Biden says this:

"The CDC is going to say that what we should do is everyone . . . under the age of 12 should probably be wearing a mask in school," Biden said. "That's probably what's going to happen."

...and pretends that this isn't him giving the CDC an order.

Presumably, this is all an attempt to avoid blameworthiness for decisions that are sure to be unpopular, rather than a bizarrely wrong theory of the scientific nature of public policy.

In Other News

[Your periodic reminder that Gain of Function research needs to stop](#) and this is a major test of our civilizational adequacy:



Eliezer Yudkowsky @ESYudkowsky · Jul 16

I don't like to jump on phrases; but if our belief state is "lab accidents happen" with respect to pathogens that can kill millions of people and shut down the global economy, all such labs need to stop existing, regardless of whether C19 started there.



Eliezer Yudkowsky @ESYudkowsky · Jul 16

You either don't have labs like that, at all. Or you get into a belief state where "lab accidents happen" is not a thing you can say with a straight face. That is not Precautionary Principle bullshit, that is straight expected value calculations.



Eliezer Yudkowsky @ESYudkowsky · Jul 16

Replies to [@ESYudkowsky](#)

People are like "but doesn't that logic say to shut down nuclear plants?" NO, because Chernobyl killed ~16,000 people, and nuclear plants avert lethal coal emissions and greenhouse warming. Benefits from gain-of-function research are *much* less, costs *much* greater.



3



6



123



Eliezer Yudkowsky @ESYudkowsky · Jul 16

The problem with the Precautionary Principle is not that all precautions are terrible and say to shut down nuclear plants. The problem with the Precautionary Principle is that it says not to take any risks, even those that are good math. Biolab risks are BAD MATH. If you...



2



4



67



Eliezer Yudkowsky @ESYudkowsky · Jul 16

...let your aversion to the "Precautionary Principle" blind you to what the numbers say about gain-of-function research in biolabs where "lab accidents happen", then you're just as bad as the Precautionists, because you have given up on MATH.

Alternatively, someone could come up with math that could possibly justify these kinds of risks. If someone has done so, I see no signs of that.

Whereas you know what we're *not* funding much, even now? [Pandemic preparedness](#).

Dustin Moskovitz @moskov · 19h
Did we seriously forget the pandemic preparedness lesson already?

Sam @sam_a_bell · 19h
would be bad ... \$30 billion is < 1% of the total bill ... can't even meet the vaccine goals with \$5 billion [thehill.com/opinion/white-](http://thehill.com/opinion/white-...)...

[Show this thread](#)

We're still in the midst of the worst public health crisis in more than a century. We should be doing all we can to prevent anything like COVID-19 from ever happening again. Yet, despite the unspeakable toll COVID-19 has taken on our families and communities, sources are telling us that Congress may slash President Biden's proposal in the American Jobs Plan for pandemic preparedness funding — from a \$30 billion investment to just \$5 billion, barely one-tenth of 1 percent of the \$3.5 trillion infrastructure initiative Senate Democrats proposed.

You'd think they'd wait for the current crisis to be over before failing to prepare for the next one. That is, you'd think that if you hadn't been paying attention.

[Eliezer also had another interesting thought](#):

Eliezer Yudkowsky @ESYudkowsky · Jul 18
A vaccine certificate *dated earlier* - say, hypothetically, before August 2021 when Delta got bad - filters for some combination of intelligence X conformity X belonging to the 'right' social class. Exactly like a bachelor's degree! Will employers forever after filter on that?

20 11 134

As several people pointed out, as a legal matter you can't actually ask such questions, so we'd put it in the pile of all the things you technically aren't allowed to ask or consider that we all know employers ask about and consider all the time.

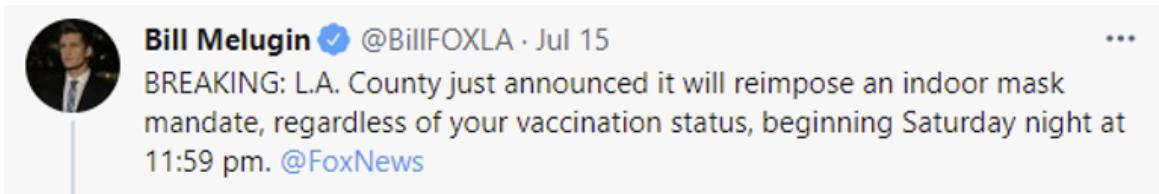
[Alex Tabbarok reviews The Nightmare Scenario](#). Report makes clear the book contains a lot of good concrete information, but nothing that would meaningfully change our model of what happened. Yes, that means all the things mentioned in the review were already in the model. I might read it anyway at some point, but my guess is I will decide that I won't because I don't have to.

[Update on the Novavax vaccine.](#)

In a strange display of the right thing being done, [Taiwan approves a vaccine purely based on immunological data.](#)

While I am not looking into such questions in general Because Of Reasons, I did see this notice that [one of the Ivermectin studies was withdrawn](#) due to 'ethical concerns.' Where the 'ethical concerns' in question appear to be 'massive fundamental discrepancies in the data' which is a nice way of saying 'complete and utter fraud.' Figured I'd pass it along. How this impacts your model of the situation otherwise is up to you - among other things, I didn't check to see how fundamental this evidence was to the case.

[Los Angeles resumes its mask mandate](#), including for the vaccinated. If it's back now it's hard to see what could happen any time soon to get it lifted. If you don't like it, you may, like many others before you, finally want to look for another place to live.



Bill Melugin  @BillFOXLA · Jul 15 · ...
BREAKING: L.A. County just announced it will reimpose an indoor mask mandate, regardless of your vaccination status, beginning Saturday night at 11:59 pm. [@FoxNews](#)

[Las Vegas brings back its indoor mask mandate as well.](#) Las Vegas seems like the place maximally in need of such a mandate, given all the travel and all the poorly ventilated completely enclosed spaces designed to trap you inside for indefinite periods. In that one case I at least kinda get it.

You know who isn't masking? [Democratic politicians fleeing Texas on a private jet in order to deny a quorum and prevent the state government from functioning. Three of whom then tested positive for Covid.](#)

(There was previously a thread here about people in the UK facing legal trouble for going outside to remote locations 'to avoid detection' but it appears likely it was fabricated.)

Australia enters this week's Sacrifices to the Gods competition, to crack down on those who put us all in danger:



Sky News Australia @SkyNewsAust · 20h

A woman has been fined for breaching hotel quarantine conditions after having a packet of cigarettes delivered to her room by a drone.

...



Drone spotted delivering cigarettes to hotel quarantine

A Gold Coast woman in hotel quarantine has breached her isolation conditions after she was caught being delivered a packet of cigarettes ...

skynews.com.au

Via MR, [some notes on Peru](#), and they may have won the Sacrifice to the Gods competition on sheer sticktoitness. It's impressive stuff. Other non-Covid stuff after is wild too.

Also in the UK, escalating quickly: [Not only First Doses First, Second Doses Too Early Actively Dangerous and Scandalous:](#)

Government accused of 'complete failure' in explaining risks of having second Covid jab early

The deputy chair of the JCVI has said those cutting intervals short could have 'less protection' against coronavirus in the months to come

"There is very good immunological and vaccine effectiveness evidence that the longer you leave that second dose the better for Pfizer and eight weeks seems to be a reasonable compromise."

Second doses were offered to people who had been jabbed at least three weeks previously at some [pop-up vaccination sites during the Government's "Grab a Jab" scheme](#) in June before NHS England cracked down on the practice.

California's entry isn't going to get it done, [but it's still quite the display of self-harm](#):



Elizabeth @acesunderglass · 17h

Friend visited CA for a large event before returning to her home state and testing positive for covid. CA won't accept an out-of-state test for tracking purposes and suggests she travel back to take one here.

3

3

39



Elizabeth @acesunderglass · 17h

Also Berkeley claims its vaccination rate is undercounted because lots of students got vaccinated at home, which means what they are measuring is not the vaccination rate but [shots/population] and have also missed all my friends who aggressively travelled for their shots.

3

3

14



So her friend has Covid, and is being told to *travel back to California so she can get tested locally*, because out-of-state tests don't count for tracking purposes. Explain again how our policies are trying to contain this virus.

[MR looks at a report on Oaxcana's \(in Mexico\) precautions for travelers](#). Everything except the masks is clearly useless sacrifices. Tyler speculates that this makes it easier to otherwise be open. That's possible but my presumption as per usual is that nothing as sensible as that is going on here.

[Everything that is not compulsory is forbidden. Everything that is not forbidden is compulsory.](#)

Our Covid prevention efforts, all of them, well OK most of them, in one tweet, including the part where the explanation for the beds probably isn't true and they're just a terrible design for no real reason:



Paul Chelimo 2 3 ✓

@Paulchelimo



Beds to be installed in Tokyo Olympic Village will be made of cardboard, this is aimed at avoiding intimacy among athletes

Beds will be able to withstand the weight of a single person to avoid situations beyond sports.

I see no problem for distance runners, even 4 of us can do 😂



11:38 PM · Jul 16, 2021



26.4K



1.2K



Share this Tweet

[One person with a cameo tests positive, and that's it. Show's done. Them's the rules.](#)

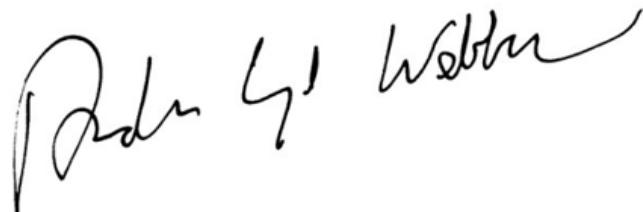


Today, on this “Freedom Day”, I have been forced to take the heart-breaking decision not to open my Cinderella.

At Cinderella, from the outset, we have employed a rigorous testing system for all the cast and backstage crew before they begin work. On Saturday, as part of this process, we identified one positive case in a member of our cast who has a cameo role in the show. As a precautionary measure, we cancelled two shows on Saturday while we carried out further tests on everyone backstage, which were negative. Any of those who were identified as a close contact of the positive case were given additional PCR tests. These tests too were negative. This morning we carried out additional tests on those due to perform tonight. Every one of them was negative.

Despite this, the impossible conditions created by the blunt instrument that is the Government's isolation guidance, mean that we cannot continue. We have been forced into a devastating decision which will affect the lives and livelihoods of hundreds of people and disappoint the thousands who have booked to see the show.

Cinderella was ready to go. My sadness for our cast and crew, our loyal audience and the industry I have been fighting for is impossible to put into words. Freedom Day has turned into closure day.

A handwritten signature in black ink, appearing to read "Andrew Lloyd Webber".

Andrew Lloyd Webber

2/2

Finally, if this is his official platform and he confirms he wants to build more apartments, [I hereby endorse Nate Silver to be the next Mayor of New York.](#)



Nate Silver @NateSilver538 · 5h

...

Almost want to run for mayor just so I can grant a 4am outdoor liquor license to clubstaurants outside each and every one of these people's apartments.



SouthVillageNYC @SouthVillageNYC · 17h

Packed house speaking out against permanent outdoor drinking and dining.



Winston Churchill, futurist and EA

This is a linkpost for <https://rootsofprogress.org/winston-churchill-futurist>

Churchill—when he wasn’t busy leading the fight against the Nazis—had many hobbies. He [wrote more than a dozen volumes of history](#), [painted over 500 pictures](#), and [completed one novel](#) (“to relax”). He tried his hand at landscaping and bricklaying, and was “[a championship caliber polo player](#).” But did you know he was also a futurist?

That, at least, is my conclusion after reading an essay he wrote in 1931 titled “Fifty Years Hence,” various versions of which were published in [MacLean’s](#), [Strand](#), and [Popular Mechanics](#). (Quotes to follow from the Strand edition.)

We’ll skip right over the unsurprising bit where he predicts the Internet—although the full consequences he foresaw (“The congregation of men in cities would become superfluous”) are far from coming true—in order to get to his thoughts on...

Energy

Just as sure as the Internet, to forward-looking thinkers of the 1930s, was nuclear power—and already they were most excited, not about fission, but fusion:

If the hydrogen atoms in a pound of water could be prevailed upon to combine together and form helium, they would suffice to drive a thousand horsepower engine for a whole year. If the electrons, those tiny planets of the atomic systems, were induced to combine with the nuclei in the hydrogen the horsepower liberated would be 120 times greater still.

What could we do with all this energy?

Schemes of cosmic magnitude would become feasible. Geography and climate would obey our orders. Fifty thousand tons of water, the amount displaced by the Berengaria, would, if exploited as described, suffice to shift Ireland to the middle of the Atlantic. The amount of rain falling yearly upon the Epsom racecourse would be enough to thaw all the ice at the Arctic and Antarctic poles.

I assume this was just an illustrative example, and he wasn’t literally proposing moving Ireland, but maybe I’m underestimating British-Irish rivalry?

Anyway, more importantly, Churchill points out what nuclear technology might do for nanomaterials:

The changing of one element into another by means of temperatures and pressures would be far beyond our present reach, would transform beyond all description our standards of values. **Materials thirty times stronger than the best steel** would create engines fit to bridle the new forms of power.

Transportation:

Communications and transport by land, water and air would take unimaginable forms, if, as is in principle possible, we could make an engine of 600 horsepower, weighing 20 lb and **carrying fuel for a thousand hours in a tank the size of a fountain-pen.**

And even farming with artificial light:

If the gigantic new sources of power become available, **food will be produced without recourse to sunlight.** Vast cellars in which artificial radiation is generated may replace the cornfields or potato-patches of the world. Parks and gardens will cover our pastures and ploughed fields. When the time comes there will be plenty of room for the cities to spread themselves again.

Biotech

Churchill also foresees genetic engineering:

Microbes, which at present convert the nitrogen of the air into the proteins by which animals live, will be fostered and made to work under controlled conditions, just as yeast is now. **New strains of microbes will be developed** and made to do a great deal of our chemistry for us.

Including lab-grown meat:

With a greater knowledge of what are called hormones, i.e. the chemical messengers in our blood, it will be possible to control growth. **We shall escape the absurdity of growing a whole chicken in order to eat the breast or wing,** by growing these parts separately under a suitable medium.

And artificial wombs:

There seems little doubt that it will be possible to carry out in artificial surroundings the entire cycle which now leads to the birth of a child.

Moral progress and risk

This last point is his segue from technological to social, political, and moral issues. The ability to "grow" people, he fears, could be used by the Communists to create human drone workers:

Interference with the mental development of such beings, expert suggestion and treatment in the earlier years, would produce **beings specialized to thought or toil.** The production of creatures, for instance, which have admirable physical development, with their mental endowment stunted in particular directions, is almost within the range of human power. A being might be produced capable of tending a machine but without other ambitions. Our minds recoil from such fearful eventualities, and the laws of a Christian civilization will prevent them. But might not lop-sided creatures of this type fit in well with the Communist doctrines of Russia? Might not the Union of Soviet Republics armed with all the power of science find it in harmony with all their aims to produce a race adapted to mechanical tasks and with no other ideas but to obey the Communist State?

In the final paragraphs, he sounds a number of themes now common in the Effective Altruist community.

More than a decade before the nuclear bomb, he also expresses concern about [existential risk](#):

Explosive forces, energy, materials, machinery will be available upon a scale which can annihilate whole nations. Despotisms and tyrannies will be able to prescribe the lives and even the wishes of their subjects in a manner never known since time began. If to these tremendous and awful powers is added the pitiless sub-human wickedness which we now see embodied in one of the most powerful reigning governments, **who shall say that the world itself will not be wrecked**, or indeed that it ought not to be wrecked? There are nightmares of the future from which a fortunate collision with some wandering star, reducing the earth to incandescent gas, might be a merciful deliverance.

He laments the inability of governance to deal with these problems:

Even now the Parliaments of every country have shown themselves quite inadequate to deal with the economic problems which dominate the affairs of every nation and of the world. Before these problems the claptrap of the hustings and the stunts of the newspapers wither and vanish away. ... Democratic governments drift along the line of least resistance, taking short views, paying their way with sops and doles, and smoothing their path with pleasant-sounding platitudes. Never was there less continuity or design in their affairs, and yet towards them are coming swiftly changes which will revolutionize for good or ill not only the whole economic structure of the world but the social habits and moral outlook of every family.

More broadly, he laments the inadequacy of our evolutionary legacy to deal with them:

Certain it is that while men are gathering knowledge and power with ever-increasing and measureless speed, their virtues and their wisdom have not shown any notable improvement as the centuries have rolled. **The brain of a modern man does not differ in essentials from that of the human beings who fought and loved here millions of years ago.** The nature of man has remained hitherto practically unchanged. ... We have the spectacle of the powers and weapons of man far outstripping the march of his intelligence; we have the march of his intelligence proceeding far more rapidly than the development of his nobility.

Which leads him, in the end, to call for [differential progress](#):

It is therefore above all things important that the moral philosophy and spiritual conceptions of men and nations should hold their own amid these formidable scientific evolutions. It would be much better to call a halt in material progress and discovery rather than to be mastered by our own apparatus and the forces which it directs. There are secrets too mysterious for man in his present state to know, secrets which, once penetrated, may be fatal to human happiness and glory. But the busy hands of the scientists are already fumbling with the keys of all the chambers hitherto forbidden to mankind. **Without an equal growth of Mercy, Pity, Peace and Love, Science herself may destroy all that makes human life majestic and tolerable.**

I don't recall Nick Bostrom citing Churchill, but I guess there's nothing new under the sun.

What does knowing the heritability of a trait tell me in practice?

The concept of [heritability](#) gets misunderstood a lot, so there are several articles discussing what it *doesn't* mean. But reading through all of them leaves me confused about what it *does* mean in *practical terms*, outside the technical definition.

For example, as I myself wrote [in an old comment](#) about common misunderstandings:

Caution: heritability, as in the statistical concept, is defined in a way that has some rather counter-intuitive implications. One might think that if happiness is 50% heritable, then happiness must be 50% "hardwired". This is incorrect, and in fact the concept of heritability is theoretically incapable of making such a claim.

The definition of heritability is straightforward enough: the amount of genetic variance in a trait, divided by the overall variance in the trait. Now, nearly all humans are born with two feet, so you might expect the trait of "having two feet" to have 100% heritability. In fact, it has close to 0% heritability! This is because the vast majority of people who have lost their feet have done so because of accidents or other environmental factors, not due to a gene for one-footedness. So nearly all of the variance in the amount of feet in humans is caused by environmental factors, making the heritability zero.

Another example is that if we have a trait that is strongly affected by the environment, but we manage to make the environment more uniform, then the heritability of the trait goes up. For instance, both childhood nutrition and genetics have a strong effect on a person's height. In today's society, we have relatively good social security nets helping give most kids at least a basic level of nutrition, a basic level which may not have been available for everyone in the past. So in the past there was more environmental variance involved in determining a person's height. Therefore the trait "height" may have been less hereditary in the past than now.

The heritability of some trait is always defined in relation to some specific population in some specific environment. There's no such thing as an "overall" heritability, valid in any environment. The heritability of a trait does not tell us whether that trait can be affected by outside interventions.

Some articles that go deeper into the details and math of this include "[Heritability is a ratio, not a measure of determinism](#)" (dynamight.net) and "[Heritability in the genomics era - concepts and misconceptions](#)" (Nature Reviews Genetics).

However, all of these examples of what heritability *doesn't* mean have left me very confused about what it *does* mean. **I know that if a trait is 80% heritable, I cannot infer that it is "80% genetically determined", but what can I infer? That 80% of the observed variance in that trait is genetic, yes, but what's the practical thing of interest that having this information allow me to predict, that I couldn't predict before? In particular, what does knowing the heritability of traits such as IQ, subjective well-being, or Big5 scores tell me?**

Looking at the Wikipedia article for heritability, I see very little that would help answer this question; the closest that I can find is the "controversies" section, which says that there are people who think the concept shouldn't be used at all:

Heritability estimates' prominent critics, such as Steven Rose,[27] Jay Joseph,[28] and Richard Bentall, focus largely on heritability estimates in behavioral sciences and social sciences. Bentall has claimed that such heritability scores are typically calculated counterintuitively to derive numerically high scores, that heritability is misinterpreted as genetic determination, and that this alleged bias distracts from other factors that researches have found more causally important, such as childhood abuse causing later psychosis.[29][30] Heritability estimates are also inherently limited because they do not convey any information regarding whether genes or environment play a larger role in the development of the trait under study. For this reason, David Moore and David Shenk describe the term "heritability" in the context of behavior genetics as "...one of the most misleading in the history of science" and argue that it has no value except in very rare cases. [31] When studying complex human traits, it is impossible to use heritability analysis to determine the relative contributions of genes and environment, as such traits result from multiple causes interacting.[32] In particular, Feldman and Lewontin emphasize that heritability is itself a function of environmental variation. [33] However, some researchers argue that it is possible to disentangle the two. [34]

The controversy over heritability estimates is largely via their basis in twin studies. The scarce success of molecular-genetic studies to corroborate such population-genetic studies' conclusions is the missing heritability problem.[35] Eric Turkheimer has argued that newer molecular methods have vindicated the conventional interpretation of twin studies,[35] although it remains mostly unclear how to explain the relations between genes and behaviors.[36] According to Turkheimer, both genes and environment are heritable, genetic contribution varies by environment, and a focus on heritability distracts from other important factors. [37] Overall, however, heritability is a concept widely applicable.[9]

Out of those references, the one that sounded the most useful in telling me what heritability might *actually* mean was the one associated with the sentence "Overall, however, heritability is a concept widely applicable". This is the previously mentioned "[Heritability in the genomics era - concepts and misconceptions](#)" (Nature Reviews Neuroscience), which includes a section on "applications":

The parameter of heritability is so enduring and useful because it allows the meaningful comparison of traits within and across populations, it enables predictions about the response to both artificial and natural selection, it determines the efficiency of gene-mapping studies and it is a key parameter in determining the efficiency of prediction of the genetic risk of disease.

From reading this section, I gather that:

- If I wanted to breed plants or animals that were high on a particular trait, having the heritability estimate for that trait could be useful
- The heritability of a trait can be used to help infer how much statistical power gene-mapping studies targeting that trait need
- If I was trying to predict the genetic risk of something like schizophrenia, then... I don't quite understand this part, but apparently having the heritability estimate would help me know how reliable my prediction was going to be

Usefulness for breeding programs is hopefully an irrelevant consideration when we're talking about humans, which leaves me with the two others; and those also seem to suggest that knowing the heritability of a trait isn't useful on its own, and will only be something that helps me do or evaluate a gene-mapping or genetic risk prediction study better.

This seems to suggest that knowing the heritability of a trait such as IQ, subjective well-being or a Big5 score tells me essentially nothing by itself; is this correct?

([*cross-posted to the Psychology & Neuroscience Stack Exchange*](#))

BASALT: A Benchmark for Learning from Human Feedback

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

This is a linkpost for <https://bair.berkeley.edu/blog/2021/07/08/basalt/>

Copying the abstract of the [paper](#):

The last decade has seen a significant increase of interest in deep learning research, with many public successes that have demonstrated its potential. As such, these systems are now being incorporated into commercial products. With this comes an additional challenge: how can we build AI systems that solve tasks where there is not a crisp, well-defined specification? While multiple solutions have been proposed, in this competition we focus on one in particular: learning from human feedback. Rather than training AI systems using a predefined reward function or using a labeled dataset with a predefined set of categories, we instead train the AI system using a learning signal derived from some form of human feedback, which can evolve over time as the understanding of the task changes, or as the capabilities of the AI system improve.

The MineRL BASALT competition aims to spur forward research on this important class of techniques. We design a suite of four tasks in Minecraft for which we expect it will be hard to write down hardcoded reward functions. These tasks are defined by a paragraph of natural language: for example, "create a waterfall and take a scenic picture of it", with additional clarifying details. Participants must train a separate agent for each task, using any method they want. Agents are then evaluated by humans who have read the task description. To help participants get started, we provide a dataset of human demonstrations on each of the four tasks, as well as an imitation learning baseline that leverages these demonstrations.

Our hope is that this competition will improve our ability to build AI systems that do what their designers intend them to do, even when the intent cannot be easily formalized. Besides allowing AI to solve more tasks, this can also enable more effective regulation of AI systems, as well as making progress on the value alignment problem.

I also mention this in the [latest Alignment Newsletter](#), but I think this is probably one of the best ways to get started on AI alignment from the empirical ML perspective: it will (hopefully) give you a sense of what it is like to work with algorithms that learn from human feedback, in a more realistic setting than Atari / MuJoCo, while still not requiring a huge amount of background or industry-level compute budgets.

Section 1.1 of the paper goes into more detail about the pathways to impact. At a high level, the story is that better algorithms for learning from human feedback will improve our ability to build AI systems that do what their designers intend them to do. This is straightforwardly improving on [intent alignment](#) (though it is not solving it), which in turn allows us to better govern our AI systems by enabling regulations like "your AI systems must be trained to do X" without requiring a mathematical formalization of X.

Covid 7/1: Don't Panic

The case numbers this week were clearly bad news. The raw count was somewhat bad news, and the positive test percentage increase was very bad news. It would be easy to treat the whole shift as fully 'real,' attribute it all to Delta, and panic.

I do not think that is the correct interpretation. What we are seeing matches what we saw a year ago, so a lot of this is a seasonal and regional change that has nothing to do with Delta. It's also likely that some of the shift in percentages comes from data being wonky rather than the underlying conditions. Not only do we have alternative explanations, the size of the shift doesn't match the incremental change in the amount of Delta out there, even if (as I suspect) it's a rather dramatic takeover, with Delta's share of the pandemic in America rising 25%+ in a single week.

It's also going to be tempting to attribute seasonal weather effects to the local vaccination rate, since the two are highly correlated. Differences in vaccination rates in different areas matter a lot, but that's not central to what is happening this week.

Thus I am still expecting some regional outbreaks, and am still not expecting nationwide problems, but one must ask about whether the winter is going to bring trouble the way last winter did. The hopeful answer is that vaccinations will be far enough along by then to not matter, and the second hopeful answer is that even if it's bad it won't be anywhere near as bad as last time. The vaccination numbers this week were quite good.

Still, I miss the confidence I had two weeks ago. Let's run the numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 1.8% (unchanged) and deaths fall by 8%.

Result: [Positivity rate of 2.4% \(up 0.6%\)](#) and deaths decline by 8%.

The case numbers reflect a <10% jump in cases, yet we have a 30% jump in positivity rate. This suggests some combination of a decline in testing and quirky data. I'm no longer confident that the positivity rate is the best measure of the state of the pandemic in America, and am relying on case counts more. A lot of that is no longer fearing that case counts are being manipulated the way they clearly were some of last year, or worries about testing supplies.

Prediction for next week: Positivity rate of 2.7% (up 0.3%) and deaths decline by 5%.

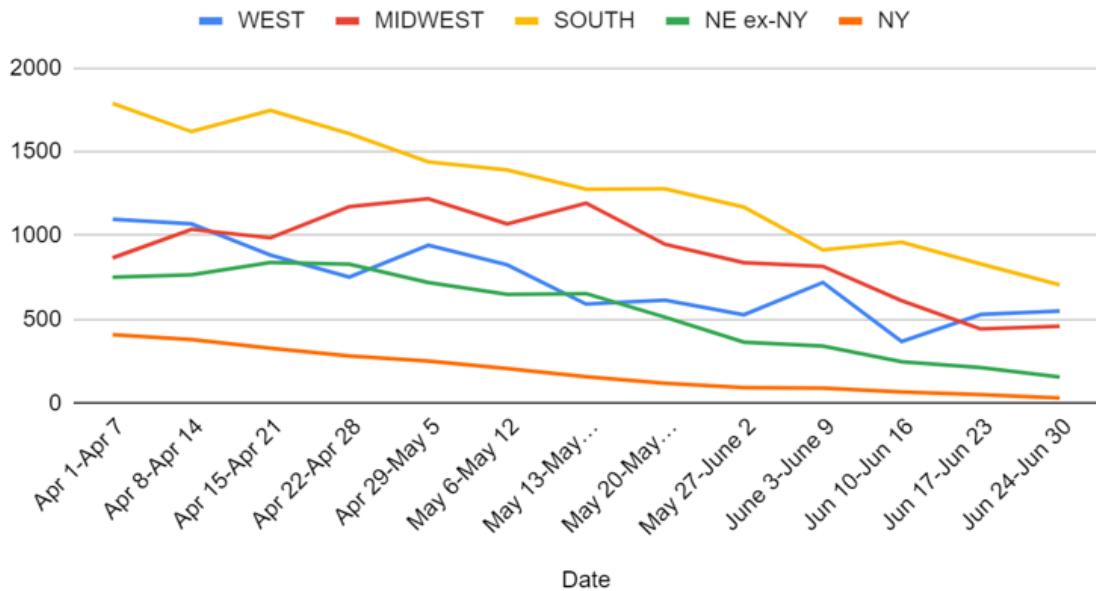
I expect a lot of this effect to be seasonality rather than Delta, but cases likely will rise for a bit. Deaths should still be declining somewhat.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 20-May 26	615	948	1279	631	3473
May 27-June 2	527	838	1170	456	2991
June 3-June 9	720	817	915	431	2883

Jun 10-Jun 16	368	611	961	314	2254
Jun 17-Jun 23	529	443	831	263	2066
Jun 24-Jun 30	550	459	706	186	1901

Deaths by Region

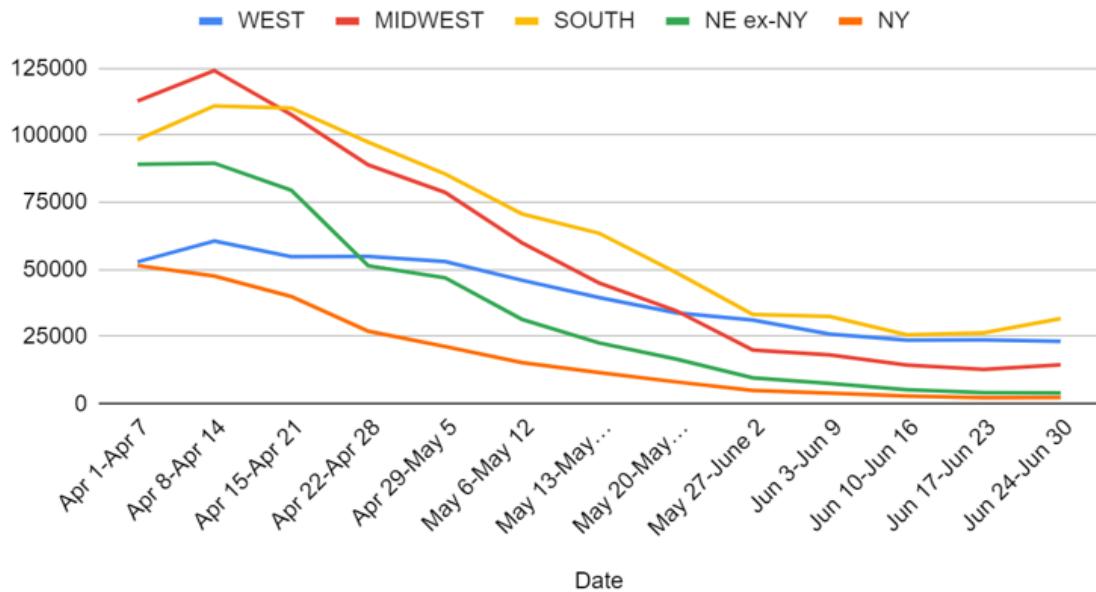


This is exactly on track as an average, and I've come to accept that week to week numbers in regions bounce around. If anything, it's surprising that the death numbers nationwide have been so steady when there's clearly a lot of data collection timing issues and random fluctuations going on. Deaths lag by several weeks, so the bad news from the last few weeks hasn't had an impact yet, nor has Delta's increased lethality mattered much yet, but it's likely that will stop soon, and we won't get much lower than this for a while.

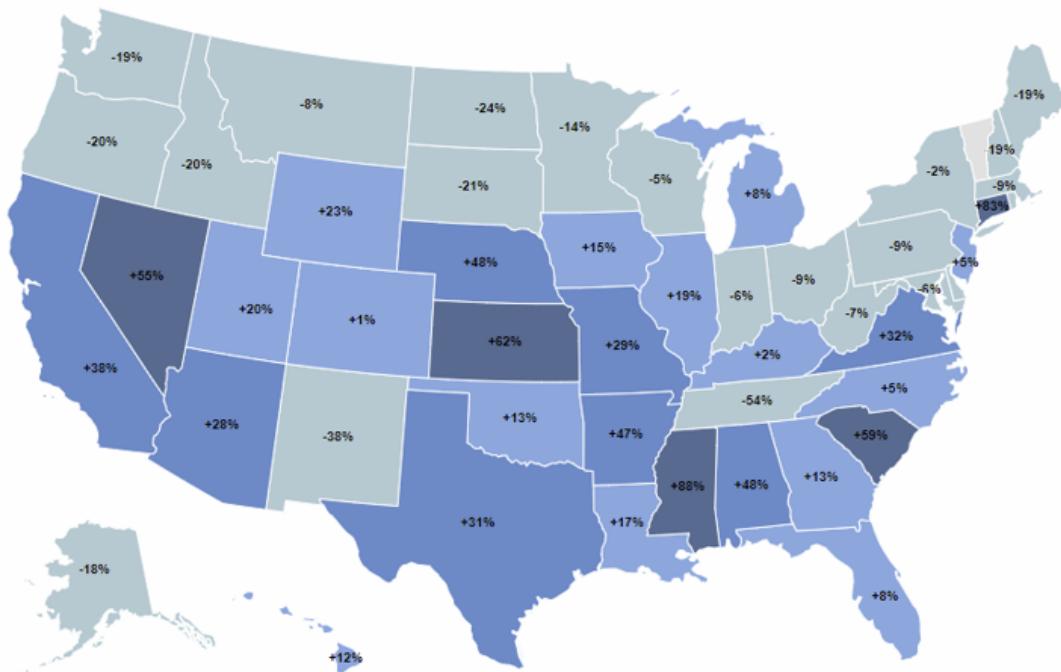
Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
May 13-May 19	39,601	45,030	63,529	34,309	182,469
May 20-May 26	33,890	34,694	48,973	24,849	142,406
May 27-June 2	31,172	20,044	33,293	14,660	99,169
Jun 3-Jun 9	25,987	18,267	32,545	11,540	88,339
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

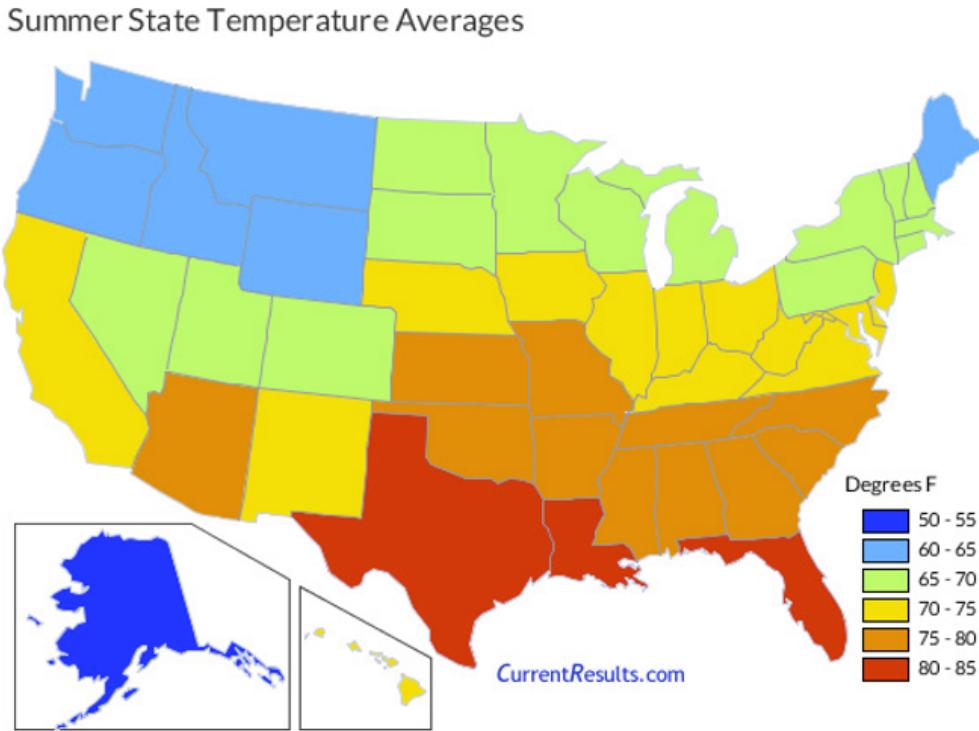
Positive Tests by Region



[I found a new site this week](#) that offers a great view of the data, so here's the week over week changes by state:



primarily about how many people each state has vaccinated so far. The big difference is north versus south, and it looks [a lot like this map](#):



Things have been record-breakingly hot in many places recently, especially the Pacific Northwest, but that only happened in the past few days and it was sufficiently dramatic (both record breaking and >100F) behavioral patterns likely went straight to 'don't go outside at all for any reason' rather than a bunch of meetings indoors.

Also worth remembering this from last year, which tells a very similar story so far.

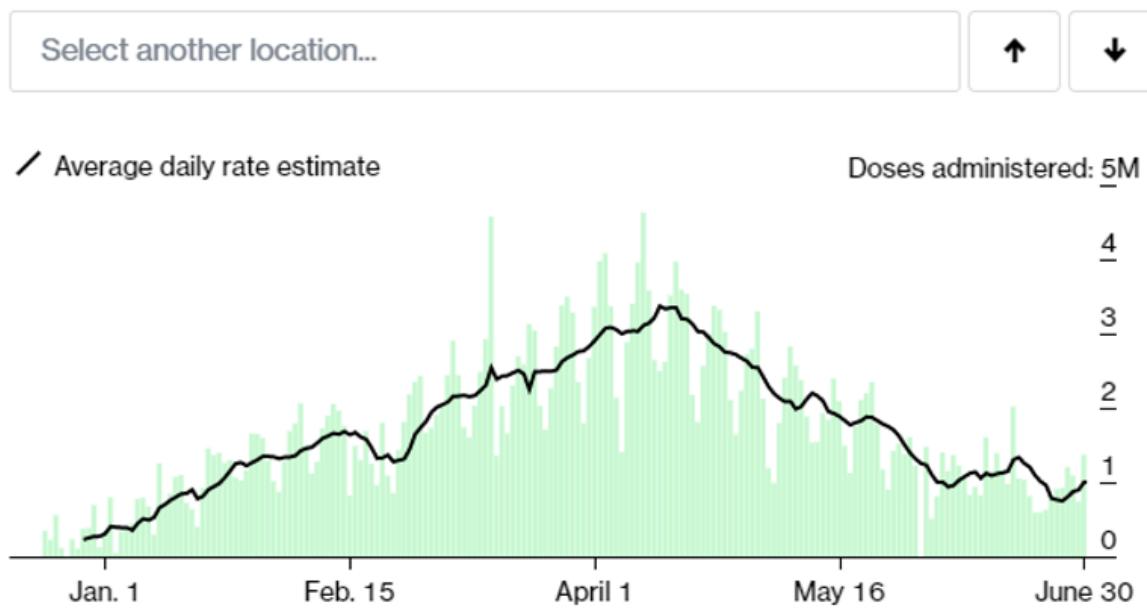
Date (2020)	WEST	MIDWEST	SOUTH	NE ex-NY	NY	Total
June 4-June 10	35487	24674	55731	16622	6071	138585
June 11-June 17	41976	22510	75787	12905	4986	158164
June 18-June 24	66292	26792	107221	10922	4524	215751
June 25-July 1	85761	34974	163472	11890	4413	300510
July 2-July 8	103879	40139	202863	13376	4850	365107
July 9-July 15	108395	53229	250072	15199	5077	431972
July 16-July 22	117506	57797	265221	16037	4880	461441
July 23-July 29	110219	67903	240667	21301	4707	444797
July 30-Aug 5	91002	64462	212945	19152	4632	392193
Aug 6-Aug 12	93042	61931	188486	17091	4478	365028
Aug 13-Aug 19	80887	63384	156998	16358	4499	322126

Last year at this time we saw an explosion in cases. This year we are seeing a halt in the decline in cases. That's not good, but it's also not a reason to panic. Nor is it reason to attribute the shift entirely to Delta. I am curious why there is so little discussion elsewhere of these obvious patterns this time around, when seasonality has been a talking point in the past.

(Actually, I don't wonder all that much, because the official Very Serious Person narrative wants to worry as much as possible about Delta, so it would ignore alternative explanations.)

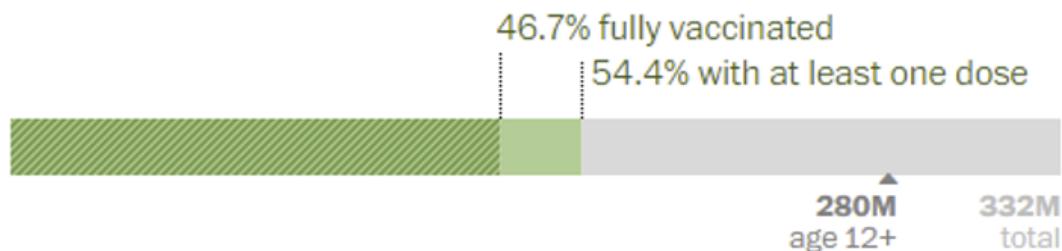
Vaccinations

In the U.S., the latest vaccination rate is **1,002,134 doses** per day, on average. At this pace, it will take another **5 months** to cover **75%** of the population.



180.7 million vaccinated

This includes more than **154.9 million people** who have been fully vaccinated in the United States.



In the last week, an average of **949.9k doses per day** were administered, a **16% increase ↑** over the week before.

We picked up an additional 0.8% of the population getting their first dose, and 1.3% of the population got their second dose. The uptick from last week gives hope that we may be in something approaching a short-term steady state, or perhaps talk of Delta has convinced some hesitant people to get the vaccine. That makes sense, since the selfish value of being vaccinated is no longer rapidly approaching zero. It's becoming increasingly clear that there will be social benefits for some time.

If we can sustain this pace of vaccinations, that is excellent for our ability to close things out. It's happening in spite of a lot of headwinds. Vaccine misinformation, misrepresentation of information and hesitancy continues to frustrate. There has been quite the abundance of self-inflicted wounds, and giving them attention is always a double-edged sword.

Thus, while I hesitate to give the topic attention and the results could easily be misinterpreted, [this seems like useful data](#):



Aella

@Aella_Girl

...

if you're fully vaccinated, did you have any side effects?
(with "mild" meaning within expected range and
"severe" meaning unusually long lasting or of
concerning intensity)



5,057 votes · Final results

Looking at the comments, it's clear that this did not sufficiently disambiguate 'severe' and that the category still mostly covered people who got knocked on their ass for two days.



Rhino @lonesco_Rhino · Jun 24

...

Replying to [@Aella_Girl](#)

I contemplated going back into the clinic due to my symptoms. So I said severe.

Plus we have to worry about anti-vax people voting combined with the lizardman constant. I'd similarly be unsure whether to say none versus mild for my shots. So this seems consistent with 'essentially zero side effects other than some people being knocked on one's ass for a day or two, which occasionally got considered severe.' It also is clearly inconsistent with the conspiracy theories anti-vax people (and [the person everyone needs to stop asking me about](#)) are throwing around.

Hopefully that will be the last I need to say on that matter in any form.

In other good news, mix-and-match vaccines, as one would expect, work quite well, [and we finally have a study on that](#) ([study](#)).



Eric Topol 
@EricTopol

...

Just published [@TheLancet](#)
Largest mix-and-match, randomized, clinical vaccine trial of people with AZ first dose, Pfizer 2nd dose. Safe and notable improved neutralizing antibody and T cell response with mix vs controls (no 2nd dose)
thelancet.com/journals/lance...

The control was one shot of AZ for completely insane You Fail Mathematics Forever reasons.



Regina @Firefly01755792 · Jun 25

...

Replies to [@EricTopol](#) and [@TheLancet](#)

I don't see the value of this study. Why did the control group not receive a second dose of AZ? Why compare AZ/Pfizer to single dose AZ?

8

2

107



Diego F. Pereira @genoma111 · Jun 26

...

The 2nd dose of AZ vaccine was not applied to some due the risks of secondary effects (i.e. thrombosis). For that reason some people received Pfizer vaccine for the 2nd shot. The study shows the safety profile for the mix and match strategy, and the need for the 2nd shot.

1

2

3



But that's fine, because we don't *need* a control when we're measuring antibody response, and the whole control group was fetishistic Science(TM) rather than a source of meaningful data.

As an alternative, we have at least one health official in Australia, who happens to be the QLD chief medical officer, [talking obvious nonsense](#):



Tim Soutphommasane  @timsout · 11h

...

The QLD chief medical officer right now has indicated she'd much rather have people under 40yo unvaccinated and contract Covid than to have them get AstraZeneca. Because of rare blood clots. This is not a polemical question: Has the public health establishment lost the plot?

386

355

1.3K



Tim Soutphommasane  @timsout · 11h

...

And I'm not verballing the QLD CMO here (thanks [@iancwhitney](#))

Young was just asked if she thinks under 40's should get the AstraZeneca vaccine:

“No! I’m sorry if I haven’t made that clear. No, I do not want under-40s to get AstraZeneca....”

They are at increased risk of getting - it is rare, but they are at increased risk of getting the rare clotting syndrome.

We’ve seen up to 49 deaths in the UK from that syndrome. I don’t want an 18-year-old in Queensland dying from a clotting illness who, if they got Covid probably wouldn’t die.

We’ve had very few deaths due to Covid-19 in Australia in people under the age of 50, and wouldn’t it be terrible that our first 18-year-old in Queensland to who dies related to this pandemic died because of the vaccine?

The idea that AZ poses a greater risk than getting Covid makes zero statistical or mathematical sense, and framing the issue that way indicates that blame avoidance and Sacrifices to the Gods of responsibility are what matters here rather than any attempt to do math on a physical world model.

It would be one thing if they had plenty of mRNA vaccines to use instead. They don't.

It would be a somewhat different thing if you couldn't mix and match shots. You can.

That's all before concerns about population-level effects.

Thus, [this](#) should come as no surprise:



Glen O'Hara ✅ @gsoh31 · 23h

...

The bad faith work done against the AstraZeneca vaccine will cost lives the world over.



News Breakfast ✅ @BreakfastNews · Jun 28

"The horse has bolted... people talk about hesitancy or reluctance, it's well beyond that. It's a refusal."

Sydney GP Jamal Rifi tells Hamish Macdonald only eight people came to his clinic yesterday for their first AstraZeneca vaccine dose.

I think we can lay to rest the hypothesis that Australia did better than other countries because it was more sane and has wiser systems for making decisions. Australia did better for other reasons, including being an island, that led to a different equilibrium. Now that we are in the vaccination phase of the pandemic, Australia is utterly failing.

Delta Variant

[This Nature post](#) provides a perspective of how many are thinking about Delta. As usual, there's talk about *whether* the new variant will take over and whether it can be prevented by some magical force, rather than *how quickly* it will take over and the need to accept that reality, but mostly the reality is being accepted here.

They are worried about Africa:

Africa at risk

Delta poses the biggest risk, scientists say, to countries that have limited access to vaccines, particularly those in Africa, where most nations have vaccinated less than 5% of their populations. "The vaccines will never come in time," says Wenseleers. "If these kinds of new variant arrive, it can be very devastating."

Surveillance in African countries is extremely limited, but there are hints that the variant is already causing cases there to surge. Several sequences of the variant have been reported in the Democratic Republic of the Congo, where an outbreak in the capital city of Kinshasa has filled hospitals. The variant has also been detected in Malawi, Uganda and South Africa.

I continue to not be as worried, because I do not expect that we are approaching the limits of Africa's control systems. I expect there to be enough slack to absorb Delta without things going critical. I'm not super confident in that, but I do think it's a solid favorite (~75%).

As estimates of prevalence of Delta go, using this seems like a reasonable method:

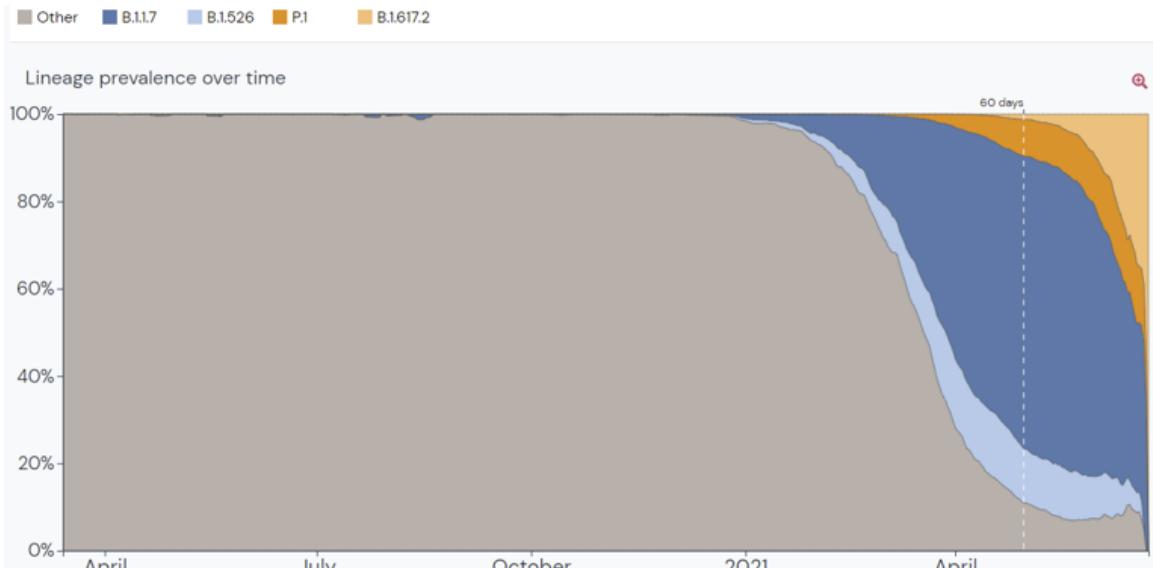
US spread

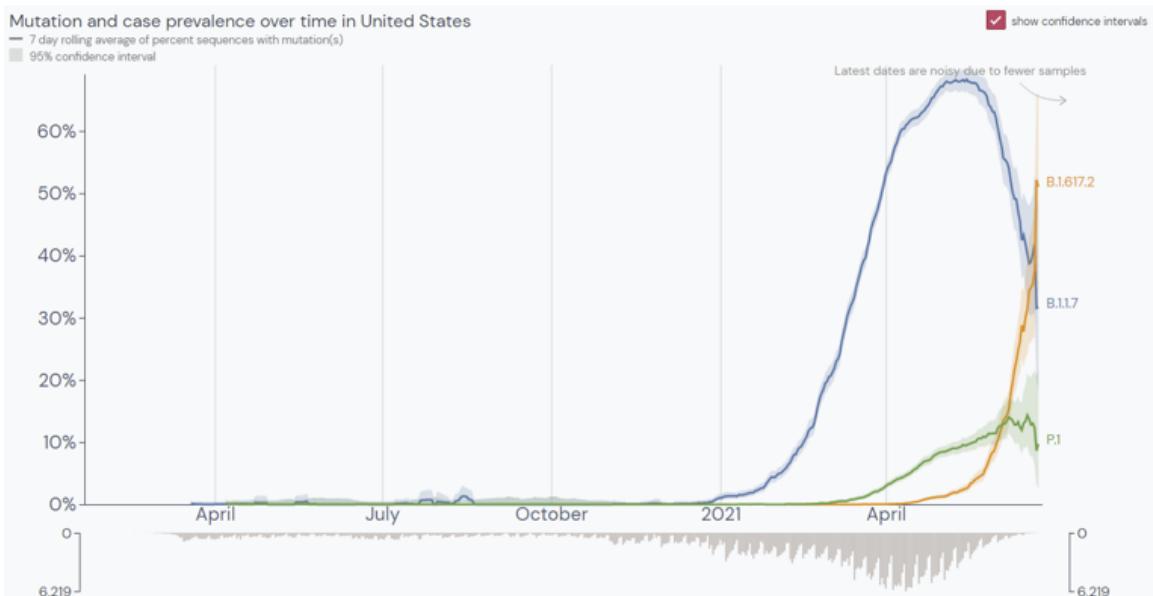
Delta is also on the rise in the United States, particularly in the Midwest and southeast. The US Centers for Disease Control and Prevention declared it a variant of concern on 15 June. But patchy surveillance means the picture there is less clear. According to nationwide sampling conducted by the genomics company Helix in San Mateo, California, Delta is rising fast. Using a rapid genotyping test, the company has found that the proportion of cases caused by Alpha fell from more than 70% in late April to around 42% as of mid-June, with the rise of Delta driving much of the shift².

Alpha was previously on the rise, so if it declined from 70% to 42%, it's safe to say that variants that are substantially more infectious than Alpha are replacing both it and the remaining 30% (to the extent that it wasn't *already* such variants), so if we think Gamma is mostly similar to Alpha as we did last time, that provides a lower bound for Delta of 40% *by mid-June*.

Contrast that with last week's calculation [from another source](#) via taking the numbers from Delta directly, which had it at 30%. That's a reasonably big gap, since here 40% is a *lower* bound, and if the previous mix included a bunch of stuff less infectious than Alpha we'd expect this to be closer to 45%-50%. And that was two weeks ago, so if it was 40% then it's presumably more like 70% now. If that math seems quick, keep in mind that we didn't even break 1% until April.

In other news, look at where that other source is after its last update:





This continues to be consistent with the sequencing data there being accurate but effectively delayed. Note that P.1/Gamma is now declining rapidly as well. Over the last 60 days they have Delta at 11%, but over the last 30 days they have it at 21%, which means it was something like 1% in the previous 30, and 30 days from whatever ‘now’ is in that calculation it implies Delta will be above 80% barring geographic barriers slowing things down.

If we estimate that the infections observed now are happening in something like a 65% Delta 30% Alpha/Gamma 5% Other world, with the remaining others largely similar to Alpha, then we've absorbed two thirds of the transition from Alpha to Delta, and three quarters of the transition from the old strain to Delta. Going forward, one should expect the share of old strains to be cut in half every ten days or so, and Delta to be almost all infections in most regions by August 1.

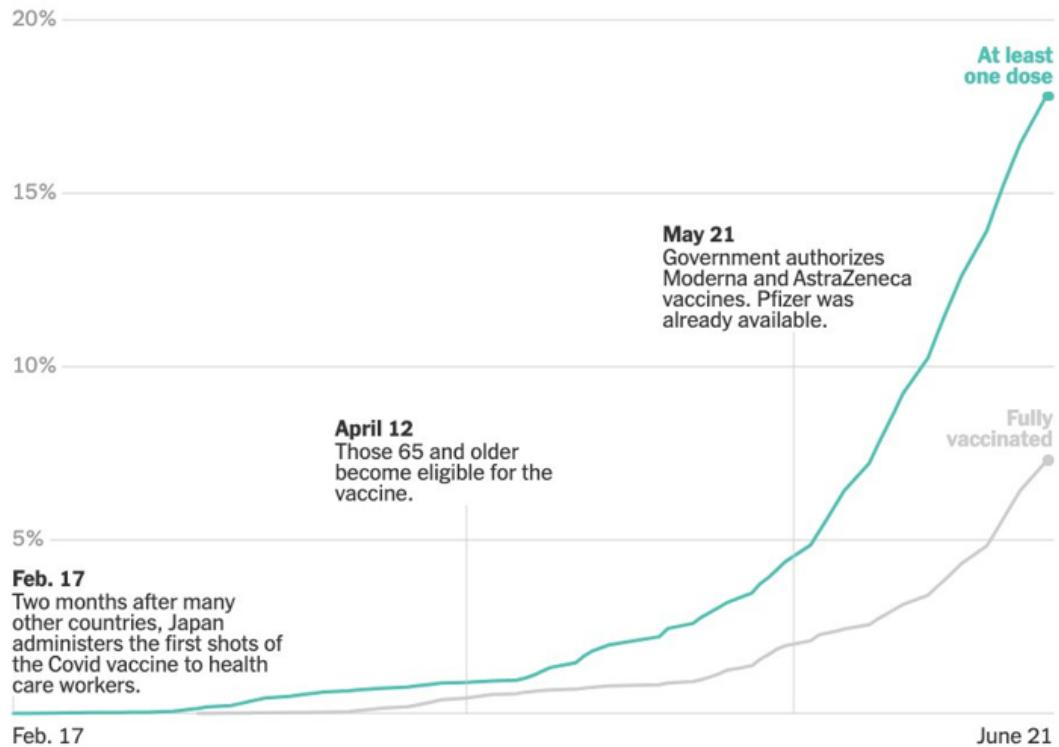
Olympics in Japan

[The Olympics are here](#), and Japan is not exactly fully vaccinated, as they've been delaying things quite a bit, with Moderna and AZ not even being approved until May 21.



On COVID, I've criticized the US and praised Japan a lot in the last year, so in the interest of being an honest broker, I'll say this:

If US were hosting the summer Olympics in a matter of weeks with a vaccination rate under 10%, I would in all likelihood be losing my mind



Source: Our World in Data

Once again I'm going to come out in favor of doing things that scale, are central to the experience of life and bring lots of people joy, even when they're not fully 'safe.' Any athlete going to the Olympics who wants to be vaccinated, one hopes, has already been vaccinated – if not, it would be trivial to take care of that. Same goes for those who need to be in the Olympic village, it's not that many people.

What's crazy is allowing a bunch of spectators who aren't vaccinated into indoor events, and it looks like [there will be 10,000 spectators per event, whether indoors or outdoors](#).

"We need to be very flexible. If there is any abrupt change in the situation, we will hold five-party meetings again to make other decisions," Hashimoto said. "If there is an announcement of a state of emergency during the Games, all the options like no-spectator games will be examined."

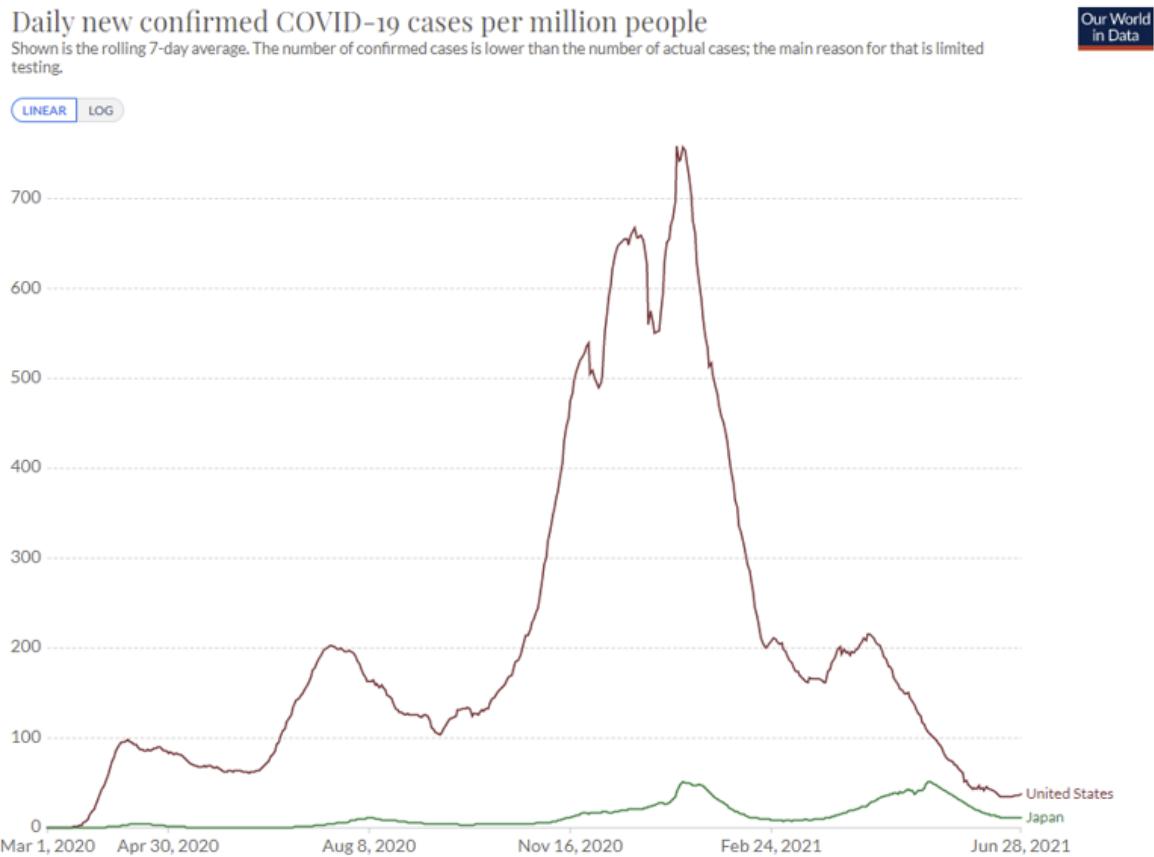
The fan restrictions mean some Japanese residents who already had tickets will not be allowed at the game and must give back their ticket. As a result, the \$800 million expected from ticket revenue will now be about half of the number.

Not everyone is one board with the new rules. Dr. Shigeru Omi, the country's top medical adviser, said the safest way to hold the Olympics would be without fans, the AP reported.

The idea that something not being ‘the safest way’ means someone is ‘not on board’ continues, and reflects a mindset we need to move past. The safest way to do most everything is to not do it at all, yet things must be done.

One way to think about whether the spectators are providing enough value is to look at the ticket revenue. There are 330 events, so assuming everything sells out, tickets are going to average \$121 per full event, many of which are over multiple days. That’s not that much money, so my instinct is they need to charge a lot more at least for indoor events, to justify the health risks involved.

It's always important to think about base rates...



...where Japan continues to outperform the United States. If Delta wasn't involved I'd say that this is all basically fine, but it remains to be seen if Japan can handle Delta in its current state, and this could take substantial time 'off the clock' for vaccinations to catch up.

The Olympics cost about \$12 billion to host and can leave a lasting legacy. If one thinks that the Olympics will be remembered for a long time, I understand taking the full risk, but I still wouldn't have substantial crowds at indoor events, especially minor ones. Why ask for that kind of trouble?

In Other News

A bunch of people this week [mentioned](#) how bad it was that people were making base rate errors, and using ‘some vaccinated people got infected’ as the latest scare tactic in full How to Lie With Statistics mode. Nothing to see here that we didn’t expect.

[Thread in praise of the excellent Microcovid project](#). I disagree with some of the Microcovid calculations but having *any plausible calculations at all* was the important element here, giving certain types of people a way to both do any of the things ever and not do many of the stupider things. That doesn't mean the paralysis before that wasn't a failure mode that requires investigation and correction. The general solution might be 'if microcovid does not exist it would be necessary to invent it so people have a Schilling estimate to converge on' and that's not great but better to know it than not know it.



Elizabeth @acesunderglass · 13h

I had other friends nearby, but everyone was already tired of negotiating and all the people I was really close to had left the area. So my choices were:



1



...



Elizabeth @acesunderglass · 13h

1 indefinite isolation
2 hang out with the very risk tolerant (which I wasn't)
3 bubble with a house and make myself absolutely zero risk to them, while not really knowing or having any influence over their risk level to me



1



...



Elizabeth @acesunderglass · 13h

And then [@microcovid](#) came along with a fourth option: numbers. I chose a house that already used the system to manage risk, they gave me a budget so I could run the occasional errand or see other friends under safe circumstances, and I could see the risk I had from them.



1



...



Elizabeth @acesunderglass · 13h

[microcovid.org](#) saved me from either 9 months of isolation, or a covid case I expect would have been bad for my age group. Those are both extraordinary gifts.

[A good way of framing](#) what Zeynep and I both point out about the question of the origins of Covid-19: What matters is how such an event *could have* occurred, rather than the way it *actually did* occur. If the plane *could have crashed* due to a faulty part, but it turned out something else caused this particular crash, the part is still faulty and we need to fix the root cause of that failure.



Mike Caulfield @holden · Jun 27

...

But in this situation the very idea that one can be "right" at this point is corrosive, because we are not looking at some immediate decision -- we are looking at a long intensive process where even the discovery of wrong but plausible causes will have massive implications.

1

1

21

↑



Mike Caulfield @holden · Jun 27

...

I said in another conversation that you could rate "lab leak" a one out of a hundred chance or a one out of ten. Or one out of two. But as [@jbakcoleman](#) has noted it's not really clear three experts holding those three different positions would propose different remedies.

1

1

19

↑



Mike Caulfield
@holden

...

Replies to [@holden](#)

People keep wanting to think of this like a murder case -- if I can prove X did it then Y is innocent. But as [@zeynep](#) points out, it's like a plane crash investigation.



Mike Caulfield @holden · Jun 27

...

Replies to [@holden](#)

If you find out the crash was due to window placement but there was also a good chance it could have been due to a defective engine design, you don't say well, it was the windows, seems like the engine design people are off the hook. You fix both.

2

4

61

↑



Mike Caulfield @holden · Jun 27

...

In any case, we need a broader discussion about this now, and not a bunch of Twitter blue-checks running around betting on one theory or another for Twitter status scraps. It's too important for that.

If Gain of Function research puts us in danger, that's enough reason to ban it, whether or not it actually did kill millions of people this particular time around.

Zeynep also points us to [this paper about potential lab leaks](#) explicitly warning against exactly the things we frequently do... from 2015.

[Abu Dhabi drops the hammer.](#)



BNO Newsroom @BNODesk

...

NEW: Abu Dhabi says unvaccinated people will be banned from shopping centers, restaurants, universities, gyms, recreational facilities, and other places

The committee approved the first phase to include shopping centres, restaurants, cafes, and all other retail outlets, including those not within a shopping centre, except supermarkets and pharmacies.

The first phase of the decision also includes gyms, recreational facilities and sporting activities, health clubs, resorts, museums, cultural centres and theme parks, as well as universities, institutes, public and private schools and children nurseries in the emirate.

The committee stated that the decision does not apply to unvaccinated individuals with vaccination exemption received through the approved process and registered on Alhosn app, nor to children aged 15 and under.

No, [we can't know for certain that vaccines provide years-long immunity](#) without waiting those years, but I do think the extrapolation process here is mostly reliable, and a much *less misleading* statement than the typical 'provides protection for up to X months' where X is how many months we've had the time to check, which then gets quoted by media as if X is an upper bound rather than a lower bound.

[Marginal Revolution previews/reviews the new Michael Lewis book about those preparing for a pandemic, and what they did during the early days of pandemic they'd prepared for.](#) The book, titled The Premonition, if Alex's summary is accurate, holds the CDC (Delenda Est) in even lower regard than I do, as our heroes who see the pandemic coming are stymied every step of the way, preventing them from having most of their potential impact, while we don't so much as quarantine or even test those returning from China, let alone take reasonable precautions. As Alex notes, it's weird to have a (true) story of lone heroes fighting the good fight against the system to stop a global catastrophe, *who then completely and utterly lose*. A story at least as worth telling as when they win, but definitely hard on the cliche structure.

[Paper about long term past effects of pandemics in Sweden over 220 years.](#)

[Potential good news in a path towards an all-coronavirus vaccine, I don't know if it's meaningful or not but passing it along.](#)

Not Covid

If your goal is to avoid claims that our elections are full of fraud and tabulation errors, and/or you don't want to ruin ranked choice voting for the rest of us, may I suggest perhaps making [more of an effort](#) to [not do things like this](#), as happened in the NYC mayoral election:

Seth Burn @SethBurn 13h

If only they had a reservoir of good will due to their previous exemplary record.

NYC Board of Elections @BOEN...

We are aware there is a discrepancy in the unofficial RCV round by round elimination report. We are working with our RCV technical staff to identify where the discrepancy occurred. We ask the public, elected officials and candidates to have patience.



Bob Hardt



@bobhardt

Elections 2.0: Sources tell me the Board of Elections is going back to the drawing board and running corrected ranked-choice numbers tomorrow. About 130,000 "votes" were part of a test-run that were never cleared from a computer.

9:59pm · 29 Jun 2021 · Twitter for iPhone

Thus, the Evergreen Tweet: I hope we will hear the end of this, but fear that we will not.

Fractional progress estimates for AI timelines and implied resource requirements

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

This post was written by Mark Xu based on interviews with Carl Shulman. It was paid for by Open Philanthropy but is not representative of their views. A draft was sent to Robin Hanson for review but received no response.

Summary

- Robin Hanson estimates the time until human-level AI by surveying experts about the percentage progress to human-level that has happened in their particular subfield in the last 20 years, and dividing the number of years by the percentage progress.
- Such surveys look back on a period of extremely rapid growth of compute from both hardware improvements and more recently skyrocketing spending.
- Hanson favors using estimates from subsets of researchers with lower progress estimates to infer AI timelines requiring centuries worth of recent growth, implying truly extraordinary *sustained* compute growth is necessary to surpass human performance.
- Extrapolated compute levels are very large to astronomically large compared to the neural computation that took place in evolution on Earth, and thus likely far overestimate AI requirements and timelines.

Introduction

Suppose that you start with \$1 that grows at 10% per year. At this rate, it will take ~241 years to get \$10 billion ($\10^{10}). When will you think that you're ten percent of the way there?

You might say that you're ten percent of the way to \$10 billion when you have \$1 billion. However, since your money is growing exponentially, it takes 217 years to go from \$1 to \$1 billion and only 24 more to go from \$1 billion to \$10 billion, even though the latter gap is larger in absolute terms. If you tried to guess when you would have \$10 billion by taking 10x the amount of time to \$1 billion, you would guess 2174 years, off by a factor of nine.

Instead, you might say you're ten percent of the way to $\$10^{10}$ when you have $\$10^1$, equally spacing the percentile markers along the exponent and measuring progress in

terms of $\log(\text{wealth})$. Since your money is growing perfectly exponentially, multiplying the number of years it takes to go from \$1 to \$10 by ten will produce the correct amount of time it will take to go from \$1 to $\$10^{10}$.

When employing linear extrapolations, choosing a suitable metric that better tracks progress, like $\log \text{wealth}$ over wealth for investment, can make an enormous difference to forecast accuracy.

Hanson's AI timelines estimation methodology

Hanson's preferred method for estimating AI timelines begins with asking experts what percentage of the way to human level performance the field has come in the last n years, and whether progress has been stable, slowing, or accelerating. In Hanson's [convenience sample](#) of his AI acquaintances he reports typical answers of 5-10% of stable progress over 20 years, and gives a similar estimate himself. He then produces an estimate for human-level performance by dividing the 20 year period by the % of progress to produce estimates of 200-400 years. [Age of Em](#):

At the rate of progress seen by AI researchers in their subfields over the last 20 years, it would take about two to four centuries for half of these AI subfields to reach human level abilities. As achieving a human level AI probably requires human level abilities in most AI subfields, a broadly capable human level AI probably needs even longer than two to four centuries.

Before we engage with the substance of this estimate, we should note that a larger more systematic recent survey using this methodology gives much shorter timeline estimates and more reports of acceleration, as summarized by [AI impacts](#):

- 372 years (2392), based on responses collected in Robin Hanson's informal 2012-2017 survey.
- 36 years (2056), based on all responses collected in the 2016 Expert Survey on Progress in AI.
- 142 years (2162), based on the subset of responses to the 2016 Expert Survey on Progress in AI who had been in their subfield for at least 20 years.
- 32 years (2052), based on the subset of responses to the 2016 Expert Survey on Progress in AI about progress in deep learning or machine learning as a whole rather than narrow subfields.

...67% of respondents of the 2016 expert survey on AI and 44% of respondents who answered from Hanson's informal survey said that progress was accelerating.

Directly [asking researchers for their timelines estimates](#) also gives much shorter estimates than using the above methodology on Hanson's informal survey. Overall we think responses to these sorts of surveys are generally not very considered, so we will focus on the object-level henceforth.

Inputs to AI research have grown enormously in the past

William Nordhaus [reports](#):

First, there has been a phenomenal increase in computer power over the twentieth century. Depending upon the standard used, computer performance has improved since manual computing by a factor between 1.7 trillion and 76 trillion. Second, there was a major break in the trend around World War II. Third, this study develops estimates of the growth in computer power relying on performance rather than components; the price declines using performance-based measures are markedly larger than those reported in the official statistics.

Computation available for AI has grown enormously in the decades before these surveys. If, for example, 5-10% progress is attributed to 20 years in which computation for AI research increased 1000x, then Hanson's extrapolation method would seem to be making predictions about years of *similarly explosive growth in compute inputs* until human level AI performance.

We have a number of metrics showing this massive input growth:

- [AI Impacts](#) estimates that, since the 1940s, \$/FLOPS fell by 10x every 7.7 years, around 35% a year.
- [AI Impacts](#) estimates that, from 2011 to 2017, \$/FLOPS fell by 10x every 10-16 years, around 15% to 25% a year.
- [Bloom et al.](#) estimate that semiconductor R&D efforts have grown by a factor of 18 from 1971 to 2020, around 6.8% a year.
- [Amodei and Hernandez](#) estimate that, from 2012 to 2018, the amount of compute used in the largest AI training runs increased 10x every 11.2 months, around 1180% a year.
- [AI Index Report](#) estimates that global corporate investment in AI was \$68 billion in 2020, up from \$13 billion in 2015, an average increase of 39% a year.
- [The World Bank](#) estimates that world GDP has grown an average of 3.5% a year from 1961 to 2019.
- [Besiroglu](#) estimates that, from 2012 to 2020, the effective number of researchers in ML rose by 10x every 4-8 years, around 33% to 78% a year.

Extrapolating past input growth yields ludicrously high estimates of the resource requirements for human level AI performance

Hanson's survey was conducted in 2012, so we must use trends from the two decades prior when extrapolating. We conservatively ignore the major historical growth in spending on AI compute as a fraction of the economy, and especially the surge in investment in large models that drove [Amodei and Hernandez](#)'s finding of annual 10x

growth in compute used in the largest deep learning models over several years. Accounting for that would yield even more extreme estimates in the extrapolation.

Extrapolating world GDP's historical 3.5% growth for 372 years yields a 10^{5x} increase in the amount of money spent on the largest training run. Extrapolating the historic 10x fall in \$/FLOP every 7.7 years for 372 years yields a 10^{48x} increase in the amount of compute that can be purchased for that much money (we recognize that this extrapolation goes past physical limits). Together, 372 years of world GDP growth and Moore's law yields a 10^{53x} increase in the amount of available compute for a single training run. Assuming GPT-3 represents the current frontier at [10²³ floating_point_operations \(FLOP\)](#), multiplying suggests that 10^{76} FLOP of compute will be available for the largest training run in 2393.

10^{76} FLOP is vast relative to the evolutionary process that produce animal and human intelligence on Earth, and ludicrous overkill for existing machine learning methods to train models vastly larger than human brains:^[1]

- [Cotra \(2020\)](#) estimates the total amount of computation done in animal nervous systems over the course of our evolution was 10^{41} FLOP. 10^{76} FLOP is enough to run evolution almost a trillion trillion trillion times.
 - One illustration of this is that for non-reversible computers, the thermodynamic [Landauer limit](#) means that 10^{76} FLOP would require vastly more energy than has been captured by all living things in the history of Earth. [Landauer's principle](#) requires that non-reversible computers at 310 Kelvin need more than [3 x 10⁻²¹ J](#) for each bit-erasure. [Carlsmith \(2020\)](#) tentatively suggests ~1 bit-erasure per FLOP, suggesting 10^{51} J are needed to perform 10^{76} FLOP. About [10¹⁷ J/s](#) of sunlight strikes the earth, so 10^{34} s \sim [10²⁰ times the age of the universe](#) of 100% efficient maximum-coverage terrestrial solar energy is needed to power 10^{76} FLOP of irreversible computing.
- [Carlsmith \(2020\)](#) estimates "it [is] more likely than not that 10^{15} FLOP/s is enough to perform tasks as well as the human brain". 10^{76} FLOP is enough to sustain a $10^{61:1}$ training compute:inference compute ratio.
 - Roughly approximating [Cotra \(2020\)](#)'s estimates, models can be trained with one datapoint per parameter and each parameter requires 10 inference FLOPs, suggesting 10^{76} FLOP is enough to train a model with 10^{25} parameters.
- [Extrapolating scaling laws](#) in model performance yields enormous improvements, and historically tasks with previously flat or zero performance have yielded as models became capable enough to solve the problem at all.

Some of Hanson's writing suggests he is indeed endorsing these sorts of requirements. E.g. in one [post](#) he writes:

For example, at [past rates](#) of [usual artificial intelligence (UAI)] progress it should take two to four centuries to reach human level abilities in the typical UAI subfield, and thus even longer in most subfields. Since the world economy now doubles roughly every fifteen years, that comes to twenty doublings in three centuries. If ems show up halfway from now to full human level usual AI, there'd still be ten economic doublings to go, which would then take ten months if the economy doubled monthly. Which is definitely faster UAI progress...Thus we should expect many doublings of the em era after ems and before human level UAI

This seems to be saying that the timeline estimates he is using are indeed based on input growth rather than serial time, and so AGI requires multiplying log input growth (referred to as doublings above) of the 20 year past periods many fold.

Conclusion

So on the object-level we can reject such estimates of the resource requirements for human-level performance. If extrapolating survey responses about fractional progress per Hanson yields such absurdities, we should instead believe that respondents' subjective progress estimates will accelerate as a function of resource inputs. In particular, that they would be end-loaded, putting higher weight on final orders of magnitude in input growth. This position is supported by the enormous differences in cognitive capabilities between humans and chimpanzees despite less than an order of magnitude difference in the quantity of brain tissue. It seems likely that 'chimpanzee AI' would be rated as a lot less than 90% progress towards human level performance, but chimpanzees only appeared after 99%+ of the timespan of evolution, and 90%+ of the growth in brain size.

This view also reconciles better with the survey evidence reporting acceleration and direct timeline estimates much shorter than Hanson's.

For the more recent survey estimate of 142 years of progress with the above assumptions the result is 10^{43} FLOP (more with incorporation of past growth in spending), which is more arguable but still extremely high, suggesting evolutionary levels of compute without any of the obvious advantages of intentional human design providing major efficiencies, and with scaling much worse than observed for deep learning today.

The aggregate and deep learning linear extrapolation results of several decades still suffer a version of this problem, primarily because of the unsustainably rapid growth of expenditures, e.g. [it is impossible](#) to maintain annual 10x growth in compute for the largest models for 30 years. While they report much more rapid progress and acceleration than the Hanson survey respondents, we would still expect more acceleration in subjective progress estimates as we come closer to across-the-board superhuman performance.

-
1. We don't think that anthropic distortions conceal large amounts of additional difficulty because of evolutionary timings and convergent evolution, combined with already existing computer hardware and software's demonstrated capabilities. See [Shulman and Bostrom \(2012\)](#) for more details. ↪

Open Philanthropy is seeking proposals for outreach projects

[Cross-posted from [the EA Forum](#).]

Open Philanthropy is [seeking proposals](#) from applicants interested in growing the community of people motivated to improve the long-term future via the kinds of projects described below.^[1]

Apply to start a new project [here](#); express interest in helping with a project [here](#).

We hope to draw highly capable people to this work by supporting ambitious, scalable outreach projects that run for many years. We think a world where effective altruism, longtermism, and related ideas are routine parts of conversation in intellectual spaces is within reach, and we're excited to support projects that work towards that world.

In this post, we describe the kinds of projects we're interested in funding, explain why we think they could be very impactful, and give some more detail on our application process.

Proposals we are interested in

Programs that engage with promising young people

We are seeking proposals for programs that engage with young people who seem particularly promising in terms of their ability to improve the long-term future (and may have interest in doing so).

Here, by “particularly promising”, we mean young people who seem well-suited to building [aptitudes](#) that have high potential for improving the long-term future. Examples from the linked post include aptitudes for conducting research, advancing into top institutional roles, founding or supporting organizations, communicating ideas, and building communities of people with similar interests and goals, among others. Downstream, we hope these individuals will be fits for what we believe to be priority paths for improving the long-term future, such as AI alignment research, technical and policy work reducing risks from advances in synthetic biology, career paths involving senior roles in the national security community, and roles writing and speaking about relevant ideas, [among others](#).

We're interested in supporting a wide range of possible programs, including summer or winter camps, scholarship or fellowship programs, seminars, conferences, workshops, and retreats. We think programs with the following characteristics are most likely to be highly impactful:

- They engage people ages 15 - 25 who seem particularly promising in terms of their ability to improve the long-term future, for example people who are

- unusually gifted in STEM, economics, philosophy, writing, speaking, or debate.
- They cover effective altruism (EA), rationality, longtermism, global catastrophic risks, or related topics.
 - They involve having interested young people interact with people currently working to improve the long-term future.

Examples of such programs that Open Philanthropy has supported include [SPARC](#), [ESPR](#), the [SERI](#) and [FHI](#) summer research programs, and the recent [EA Debate Championship](#). However, we think there is room for many more such programs.

We especially encourage program ideas which:

- Have the potential to engage a large number of people (hundreds to tens of thousands) per year, though we think starting out with smaller groups can be a good way to gain experience with this kind of work.
- Engage with groups of people who don't have many ways to enter relevant intellectual communities (e.g. they are not in areas with high concentrations of people motivated to improve the long-term future).
- Include staff who have experience working with members of the groups they hope to engage with—in particular, experience talking with young people about new ideas while being respectful of their intellectual autonomy and encouraging independent intellectual development.

We encourage people to have a low bar for submitting proposals to our program, but note that we view this as a sensitive area: we think programs like these have the potential to do harm by putting young people in environments where they could have negative experiences. [Nicole Ross](#) at the Centre for Effective Altruism (email nicole@centreforeffectivealtruism.org) is available to provide advice on these kinds of risks.

Some reasons why we think this work has high expected value

A priori, we would guess that people are more likely to get interested in new ideas and opportunities when they are relatively young and have fewer preexisting commitments. This guess is consistent with the results of a survey Open Philanthropy recently ran—we surveyed approximately 200 people who our advisors suggested had the potential to do good longtermist work, most of whom had recently made career changes that we thought were positive from a longtermist perspective. As part of this survey, we asked respondents several questions regarding the age at which they first encountered effective altruism or effective altruism-adjacent ideas.

- On average, survey respondents reported first encountering EA/EA-adjacent ideas when they were 20 years of age.
- About 25% of respondents first encountered EA/EA-adjacent ideas at ages 18 or below, even though few EA outreach projects focus on that age range.
- On average, respondents said the *best* age for them to first encounter EA/EA-adjacent ideas would have been 16.

Survey respondents often mentioned that hearing about EA before starting university would have been particularly helpful because they could have planned how to use their time at university better, e.g. what to major in.

We also asked survey respondents to brainstorm open-endedly about how to get people similar to them interested in these ideas. 10% of responses mentioned starting outreach programs younger, particularly in high school. Several respondents mentioned that SPARC and ESPR had been helpful for them and that they would recommend these programs to similar people. ([Certain other](#) high school outreach projects have reported less success, but we don't think these less-targeted programs provide much evidence about how promising targeted high school outreach is likely to be overall, as discussed [here](#).)

Our survey also showed that EA groups, particularly university groups, have had a lot of impact on longtermist career trajectories. On a free-form question asking respondents to list the top few things that increased their expected impact, respondents listed EA groups more commonly than any other factor. On other measures of impact we used in our survey analysis, EA groups came between second and fourth in potential factors, above many EA organizations and popular pieces of writing in the EA-sphere. Most of this impact (65 - 75% on one measure) came from university groups. We think this suggests that, more generally, offering high-quality opportunities for university students to get involved is a promising kind of intervention.

Made-up examples of programs we think could be impactful

These examples are intended to be illustrative of the kinds of programs we'd be interested in funding. This is not intended to be a comprehensive list, nor a list of the programs we think would be most impactful.

We think these programs are unlikely to work fully as written. Founders generally have to dive deep into a project plan to figure out what's tenable, altering their plan multiple times as they get a better understanding of the space, and we haven't done that work. As such, we'd like these examples to serve as inspiration, not as instructions. We think programs of this kind are more likely to be successful when the founders develop their own vision and understanding of their target audience.

We would ultimately like to support dedicated teams or organizations that run programs for young people at scale. That said, we are likely to recommend that applicants with less of a track record start by trying out a small pilot of their program and iterating while maximizing program quality and target fit, rather than scaling immediately.

Example 1: A free two-week summer school in Oxford that teaches content related to longtermism to promising high school students. The program could have a similar structure to SPARC and ESPR, but with a more explicitly longtermist focus, and it could engage a broader range of gifted high school students.

- We think programs like this are most effective when they focus on highly promising students, e.g. by filtering on Olympiad participation, high standardized test scores, competitive awards, or other markers of talent.
- Oxford seems like a good location for programs like this because its status as an EA hub makes it easy for current longtermists doing good work to instruct and interact with students, which we think is important for programs like this to be successful. (Berkeley and Stanford seem like good locations for similar reasons.)

- Oxford is a cool place to visit in and of itself, making a program located there attractive as a paid trip for high school students.

Example 2: A monthly AI safety workshop for computer science undergraduates, covering existing foundational work in AI safety.

- There have been several programs like this, notably [AIRCS](#), which our survey suggests has had an impact on some longtermist career trajectories. We think it's likely that AIRCS hasn't saturated the pool of top computer science undergraduates, and that there is room for more programs of this form that experiment with different kinds of content and instructors.

Example 3: A one-week summer program about effective altruism in Berkeley combined with a prestigious \$20,000 merit-based scholarship for undergraduate students. The scholarship would involve an application process that required substantial engagement with ideas related to effective altruism, e.g. a relevant essay and an interview.

- We think the best scholarship programs will be fairly selective, so as to attract very promising applicants and create a very strong cohort.
- In-person programs that run right before students start their undergraduate degrees might be particularly impactful, via bolstering EA groups at top universities.
- Scholarships and other programs that include substantial financial opportunities risk attracting applicants that are only interested in the money provided by the program. We think programs like this should construct application processes that make an effort to identify applicants genuinely interested in effective altruism, e.g. via essays and interviews.

Example 4: A monthly four-day workshop teaching foundational rationality content to promising young people.

- The workshop could teach foundational technical topics in rationality, including some covered by CFAR in the past, e.g. probability theory, Bayesianism, Fermi estimation, calibration, betting, cognitive biases, etc., as well as exercises intended to help students use these thinking tools in the real world.
- This could overlap heavily with SPARC's content, but could engage a larger number of people per year than SPARC has capacity for, as well as a more varied or substantively different audience.

Example 5: A fall jobs talk and follow-up discussion that's held at top universities describing career paths in defensive work for future biological catastrophes.

- We think fall final year is a good time to prompt undergraduate students with concrete career suggestions, and with COVID-19 in recent memory, we think the next few years could be a particularly good time to talk to students about careers in [global catastrophic biological risk reduction](#).

Projects aiming at widespread dissemination of relevant high-quality content

We are also seeking proposals for projects that aim to share high-quality, nuanced content related to improving the long-term future with large numbers of people.

Projects could cover wide areas such as effective altruism, rationality, longtermism, or global catastrophic risk reduction, or they could have a more specific focus. We're interested in supporting people both to create original content and to find new ways to share existing content.

Potential project types include:

- Podcasts
- YouTube channels
- Massive open online courses (MOOCs)
- New magazines, webzines, blogs, and media verticals
- Books, including fiction
- Strategic promotion of existing content (with the permission of the creators of the content, or their representatives), especially those that have historically drawn in promising individuals

Existing projects along these lines include [the 80,000 Hours Podcast](#), [Robert Miles's AI alignment YouTube channel](#), and [Vox's Future Perfect](#).

We encourage projects that involve content in major world languages other than English, especially by native speakers of those languages—we think projects in other languages are especially likely to reach people who haven't had as many opportunities to engage with these ideas.

We would like interested people to have a low bar for submitting a proposal, but we think projects that misrepresent relevant ideas or present them uncaringly can do harm by alienating individuals who would have been sympathetic to them otherwise. We also think it's important to be cognizant of potential political and social risks that come with content creation and dissemination projects in different countries. [Nicole Ross](#) at the Centre for Effective Altruism (email nicole@centreforeffectivealtruism.org) is available to provide advice on these kinds of risks.

Some reasons why we think this work has high expected value

Our sense from talking to people doing longtermist work we think is promising has been that, for many, particular pieces of writing or videos were central to their turn towards their current paths.

This seems broadly in line with the results of the survey we conducted mentioned above. The bodies of written work of Nick Bostrom, Eliezer Yudkowsky, and Peter Singer were in the top 10 sources of impact on longtermist career trajectories (of e.g. organizations, people, and bodies of work) across several different measures. On one measure, Nick Bostrom's work by itself had 68% of the impact of the most impactful organization and 75% of the impact of the second most impactful organization. When asked what outreach would attract similar people to longtermist work, 8% of respondents in the survey gave free-form responses implying that they think simply exposing similar people to EA/EA-adjacent ideas would be sufficient.

These data points suggest to us that even absent additional outreach programs, sharing these ideas more broadly could ultimately result in people turning towards career activities that are high-value from a longtermist perspective. For many who could work on idea dissemination, we think increasing the reach of existing works with

a strong track record, like those given above, may be more impactful per unit of effort than creating new content.

Made-up examples of projects we think could be impactful

As above, these examples are intended to be illustrative of the kinds of programs we'd be interested in funding. This is not intended to be a comprehensive list, nor a list of the programs we think would be most impactful. We think these programs are unlikely to work fully as written and would like these projects to serve as inspiration, not as instructions.

Example 1: Collaborations with high-profile YouTube creators to create videos covering longtermist topics.

- We think YouTube is an attractive promotional platform because different creators come with different audiences, making it easy to share with the kinds of people who most often become interested in longtermist ideas.

Example 2: Targeted social media advertising of episodes of [the 80,000 Hours podcast](#). The project would aim to maximize downloads of the 80,000 Hours Podcast episodes that go through social media referrals.

- The 80,000 Hours podcast seems promising to promote because we think it's high-quality, quick to consume, and varied enough in content to appeal to a fairly wide audience.
- The project could experiment with indiscriminately advertising podcast episodes to promising early-career individuals, e.g. STEM, economics, or philosophy students, or with advertising select podcast episodes on particular topics to audiences that they may appeal to.
- Any project of this form should be done in collaboration with 80,000 Hours.

Example 3: A website that delivers free copies of physical books, e-books, or audiobooks that seem helpful for understanding how to do an outsized amount of good to people with a .edu email address who request them.

- The bulk of the project work could be focused on website design and advertising, while book distribution could be handled through [EA Books Direct](#), or done as part of this project.

Example 4: A MOOC covering existing AI safety work.

Example 5: A new magazine that covers potentially transformative technologies and ways in which they could radically transform civilization in positive or negative ways.

Application process

Primary application

If you think you might want to implement either of the kinds of outreach projects listed above, **please submit a brief pre-proposal [here](#).** If we are interested in supporting your project, we will reach out to you and invite you to submit more information. We encourage submissions from people who are uncertain if they want to

found a new project and just want funding to seriously explore an idea. If it would be useful for applicants developing their proposals, we are open to funding them to do full-time project development work for 3 months. We are happy to look at multiple pre-proposals from applicants who have several different project ideas.

We may also be able to help some applicants (e.g. by introducing them to potential collaborators, giving them feedback about plans and strategy, providing legal assistance, etc.) or be able to help find others who can. **We are open to and encourage highly ambitious proposals for projects that would require annual budgets of millions of dollars**, including proposals to scale existing projects that are still relatively small.

We intend to reply to all applications within two months. We have also been in touch with the [Effective Altruism Infrastructure Fund](#) and the [Long-Term Future Fund](#), and they have expressed interest in funding proposals in the areas we describe below. If you want, you can choose to have them also receive your application via the same form we are using.

There is no deadline to apply; rather, we will leave this form open indefinitely until we decide that this program isn't worth running, or that we've funded enough work in this space. If that happens, we will update this post noting that we plan to close the form at least a month ahead of time.

Collaborator application

If you aren't interested in starting something yourself, but you would be interested in collaborating on or helping with the kinds of outreach projects listed above (either full or part-time), **let us know [here](#)**. We will connect you to project leads if we feel like there is a good fit for your skills and interests.

If you have any questions, please contact longtermfuture-outreach-rfp@openphilanthropy.org.

1. Our work in this space is motivated by a desire to increase the pool of talent available for longtermist work. We think projects like the ones we describe may also be useful for effective altruism outreach aimed at other cause areas, but we (the team running this particular program, not Open Philanthropy as a whole) haven't thought through how valuable this work looks from non-longtermist perspectives and don't intend to make that a focus. [←](#)

Reflecting on building my own tools from scratch and 'inventing on principle'

This is a linkpost for <https://amirbolous.com/posts/build>

- [Introduction](#)
- [Tools and Products](#)
- [Dogma](#)
- [Building is Empowering](#)
- [Closing Thoughts](#)

Introduction

For the past couple of months, I've built personal tools completely from scratch. I [built](#) a way to organize and record my thoughts. I [built](#) a programming language to understand Lisp. I [built](#) a tool to generate cards for family and friends. I [built](#) a web framework from scratch. I used that web framework to [ship a couple of](#) projects.

The weird thing is that, until recently, I didn't know that I could do this. I didn't even realize this was an option or move I could make. I didn't know that if I had an opinion for how things should be, **I had the right to push my view of the world into reality by doing something about it**. Bret Victor has an incredible [talk](#) called "Inventing on Principle" where he touches on just this.

"AND the PURPOSE of this talk is to tell you that this ACTIVIST LIFESTYLE is not just for social activism. As a technologist, you CAN RECOGNIZE A WRONG in the world. You CAN HAVE A VISION OF WHAT; A BETTER WORLD COULD BE. AND YOU CAN DEDICATE YOURSELF TO FIGHTING FOR A PRINCIPLE. Social activists typically fight by organizing, but you can can FIGHT BY INVENTING" — Brett Victor

amirbulous.com/posts/build

Several months later of trying to do this, I can see exactly what he meant. Here's a story/reflection on this:

Tools and Products

Every now and then, a new tool or technology becomes mainstream and garners a passionate cult of followers advocating for it. For example, this happened with Roam. This happened with Notion. This has been happening with tools and technologies long before we invented them. Often, when I would try these mainstream tools, some of them stuck. However, more often than not, I was left confused and somewhat disappointed.

That's not to say that these tools or technologies are not great products. Not at all. They're clearly good products because many people use them and recommend them to others.

Dogma

I realize now that reason these products or technologies didn't stick was because of a misaligned dogma. This is the mental model I now keep in mind: any

product or tool has the potential to be great if it can solve a [high frequency, high intensity problem](#) for at least one person. Sometimes, it can even be a low frequency, high intensity problem or vice versa. For example, Uber solved the problem of moving from one place to another conveniently. React helped tackle building fast, flexible, and rich websites.

However, any technology or tool **also imposes an expensive cost. It carries a dogma for how things should be. A set of implicit** (e.g. react component patterns) **or explicit** (e.g you can't call an Uber if no driver is nearby) **rules that are imposed on how you use the tool/technology and think about it to get some value out of it.** To go back to the React example, React carries a certain dogma for how you should build websites to take advantage of it. You can't really use it if you don't understand the pattern of "components" which is an abstraction that drives how you use the tool.

For people who have a high frequency and high intensity problem, the cost (i.e. the dogma that is "imposed" on you) feels like an absolute bargain. It can solve your pressing problem, so who cares? You'll gladly compromise.

However for others who don't necessarily have a high frequency, high intensity problem, these tools or products often don't stick. This is precisely what happened with me in the past when trying out tools or technologies that had garnered any sizable number of followers but had personally not stuck. I realize now that the problem was not with the tool itself. **It was with the model I personally carried that motivated my use of the tool.**

When you build your own tools, products, or technologies, this is not an issue by definition. **Since you're building it yourself, the tool or technology perfectly embodies your dogma.** You have a crystal clear mental model for how the tool should be used because you built it yourself! It doesn't matter whether others get it or not. It doesn't have to apply to a large group of people to capture a large market. **It just has to work for you.**

Building is Empowering

The best adjective I can use to describe building my own tools is **empowering**. Here's why:

1. Because I was building the tool for myself, I had complete [agency](#) in how I chose to build it. I could tailor the tool exactly to my personal workflow and have it cater to my needs. If my needs change, no problem, I could change the tool! This is essentially a living embodiment of the Unix philosophy of "do one thing and do it well."
2. For more low-level tools where I was building a tool that would let me use it to build other things with it (e.g Poseidon or Lisp), I had an intimate relationship with the software. Intimate in this context means I **understood every aspect of the system.** I didn't necessarily remember every aspect of the system and I often forgot how I'd designed something, but this was easily fixable. I wrote the code so I could go back and look through what I wrote. **This is a particularly special feeling because the software stack has become so deep that it's become normal to compromise on our understanding of the system.**

3. Because I built the tool for myself and understood the system well, it gave me the freedom to iterate quickly. I didn't struggle setting the system or environment up. I didn't run into obscure errors that required a nuanced understanding of a system I had just started learning. And most importantly, I had a good mental model for how I should use the system. Because I built it myself! And any tool is only as useful as the mental model you have for the system itself.
4. Because building your own tools often pushes a new layer onto the software stack, debugging was also more challenging. Bryan Cantril has a great [talk](#) on this called debugging pathological systems. The short summary is that pathological systems are hard to debug because you **don't know what level of the software stack introduced the failure**. Often a failure at the low levels of the system produced several different errors higher up. The cool thing was that these types of errors **force you to understand the system better**. **Debugging didn't feel like I was struggling with a beast in the dark. It felt like learning a complex dance with a partner** ([this](#) summarizes some of the more helpful tips). It was still frustrating, but I often came through on the other side with a much clearer picture of the nuances of the system.
5. Maybe most importantly, building everything from scratch **is incredibly satisfying and rewarding**. Even if it's a little slower or it's a little harder to push out features because there isn't a team behind the product. Not only do the tools or technologies fit your use cases and [workflows](#) perfectly, it gives you **an emotional connection to the products you use**. In the same way that you appreciate an accomplishment much more once you've experienced something similar yourself. For example, after I ran a marathon last year, I gained a lot more respect for athletes, especially marathoners with really fast times because I've experienced how painful doing one is! Similarly building things for yourself gives you a sense of enjoyment and excitement that is just incomparable to when you use a product that others have built.

Closing Thoughts

As with all things, this is a process and a journey. I did not wake up overnight deciding to embrace this philosophy. And I've only just started. I've got a long way to go. There are [many others who inspire](#) me and who I learn a lot from.

So no, you don't have to build everything for yourself from scratch. And no you should not stop using every great product out there. But know that this is an option. When there are times where you wish something existed in the world. Or you're frustrated there isn't a better solution. **You can and have the right to create one yourself. In fact, if you feel obliged to do so, you should do it.** Stand on the [shoulders of giants, steal like an artist](#), and build what you wish existed for yourself and others to use.

If you've got this far, I would **love to hear what you think**, please say hi on [Twitter](#).

Happy building 

A closer look at chess scalings (into the past)

Introduction

I had explored [measuring AI or hardware overhang](#) in August 2020 using chess. Hardware overhang is when sufficient compute is available, but the algorithms are suboptimal. I examined the strongest chess engine of 2020, Stockfish 8, performing at 3,400 ELO under tournament conditions. When reducing compute to 1997 levels (equivalent to a Pentium-II 300 MHz), its ELO score was still ~3,000. That is an important year: In 1997, the IBM supercomputer "Deep Blue" defeated the world chess champion Gary Kasparov. With Stockfish, no supercomputer would have been required. I estimated that SF8 drops to Kasparov level on a 486-DX4 100 MHz, available already in 1994. To sum it up, the hardware overhang in chess is about 10 years, or 2-3 orders of magnitude in compute.

About a year later, in July 2021, [Paul Christiano asked similar questions](#): How much compute would the old engine need to match the current engines? What is the influence of RAM (size and speed), opening books, endgame tables, pondering? Also, my old post gave some insights, but it can be improved by sharing the sources and making it reproducible. That's the aim of the current post (the other questions will be addressed in a later post).

Reproducing chess scaling from 2020

History of PC Programs (ELO by year)

As a baseline of engine performance over the years, we plot the winner from the yearly [rating list of the Swedish Chess Computer Association](#). Run on contemporary hardware,

- The list begins in 1984 when the program "Novag Super Constellation" reached 1631 ELO running on a 6502 CPU at 4 MHz.
- By 2005, Shredder 9 surpassed human levels on an AMD Athlon 1200 MHz.
- Today (2020), the leading engine is Stockfish 12 running on an AMD 1800X at 3.6 GHz.

Human grandmasters

To compare human grandmasters, we take the ELO over time for [Kasparov](#) and [Carlsen](#). Carlsen's rating between 2003 and 2011 (age 13 to 21) grew from 2000 ELO to grandmaster strength, faster than any engine :-) *[Thanks to User Bucky for the correction]*

Deep Blue

The marker for "Deep Blue" in the year 1997 is a bit [arbitrarily](#) set to 2900 ELO. At the time, Kasparov had 2860 ELO, Deep Blue won, although close.

Stockfish 8 experiment

The main part is the Stockfish 8 experiment. How well does SF8 perform on slower PCs?

As a baseline, we need to establish its ELO at a defined speed.

1. To obtain the speed baseline, we find that [SF8 makes 721 kNodes/s on an AMD Athlon 64 3500+ at 2.20 GHz](#).
2. We scale this linearly to 777 kNodes/s for the same CPU running at 2.4 GHz (+9%)
3. [SF8 achieves 3302 ELO on an Athlon 64 X2 4600+ \(2.4 GHz\)](#), in the CCRL Rating List, running 40 moves in 15 minutes (one has to dig into the side details to understand which CPU name tag is which CPU. 64bit 1 CPU is the Athlon; this can also be verified with the historical version of that list.). This is an important baseline, because it cross-calibrated to dozens of other engines.
4. With that established, we can calculate the ELO as a function of kNodes/s. An average game has 40 moves. The 40 moves in 15 minutes leave 22.5 seconds per move (on average). That's 17.5 MNodes per move to achieve 3302 ELO.
5. We benchmark our own machine, on which the experiments are run. This can be done with the Stockfish parameter "bench". For simplicity, suppose our machine performs at $10 \times 777 \text{ kNodes/s} = 7.8 \text{ MNodes/s}$. That's the ballpark of recent (2020) 4-core CPUs.
6. Now we want to perform a game at 17.5 MNodes per move, on a machine running at 7.8 MNodes/s. Clearly, each move can only take 2.24 seconds. The whole 40-game match duration is: 90 seconds.

Execute the experiment

To build a ladder of SF towards slower machines, we let this version of SF8 play a set of games of 90s timecontrol versus half that (45s). The most well-established tool to compare chess engines is [cutechess-cli](#). It is a command-line interface to play two engines (or two versions of the same engine) against each other. In the end, it nicely summarizes the results and includes a differential ELO estimate. A command may be:

```
cutechess-cli -fcp cmd=stockfish proto=uci tc=40/90 -scp cmd=stockfish
proto=uci tc=40/45 -games 100
```

How bad does the version perform with less compute? In this experiment, after running 100 games, we get 14 ELO difference. That's much less than the usual statement of 70 ELO. Why is that? We can see the same effect in similar experiments down by others ([1](#), [2](#)): The ELO gain diminishes (flattens) at high compute. On the other hand, when we reduce compute to very low levels, the curve steepens dramatically. The full ELO loss result list from my experiment is for each halving of compute:

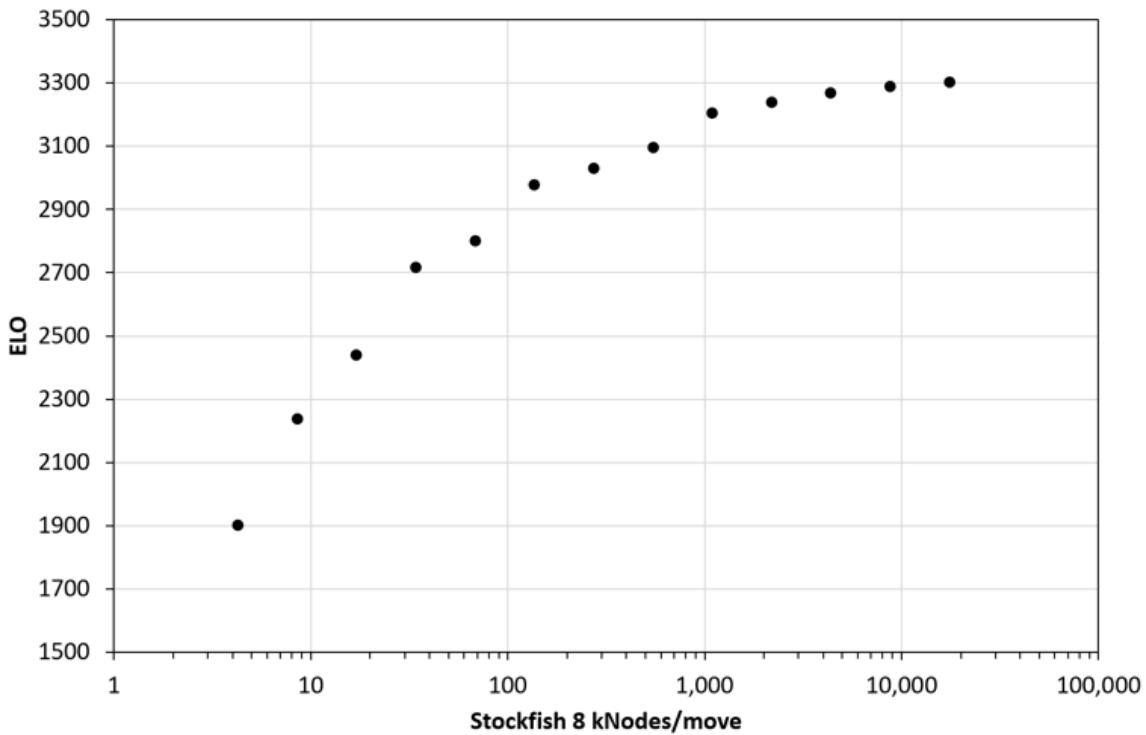
ELO	ELO Delta	kNodes/move
3302		17476.4
3288	14	8738.2
3268	20	4369.1
3240	28	2184.5
3205	35	1092.3
3097	108	546.1
3030	67	273.1
2977	53	136.5
2802	175	68.3
2716	86	34.1
2439	277	17.1
2238	201	8.5

1903

335

4.3

There is some jitter, despite increasing the number of games to 1,000 in the second half. Despite the jitter, we can clearly see the nonlinear ELO curve with compute:



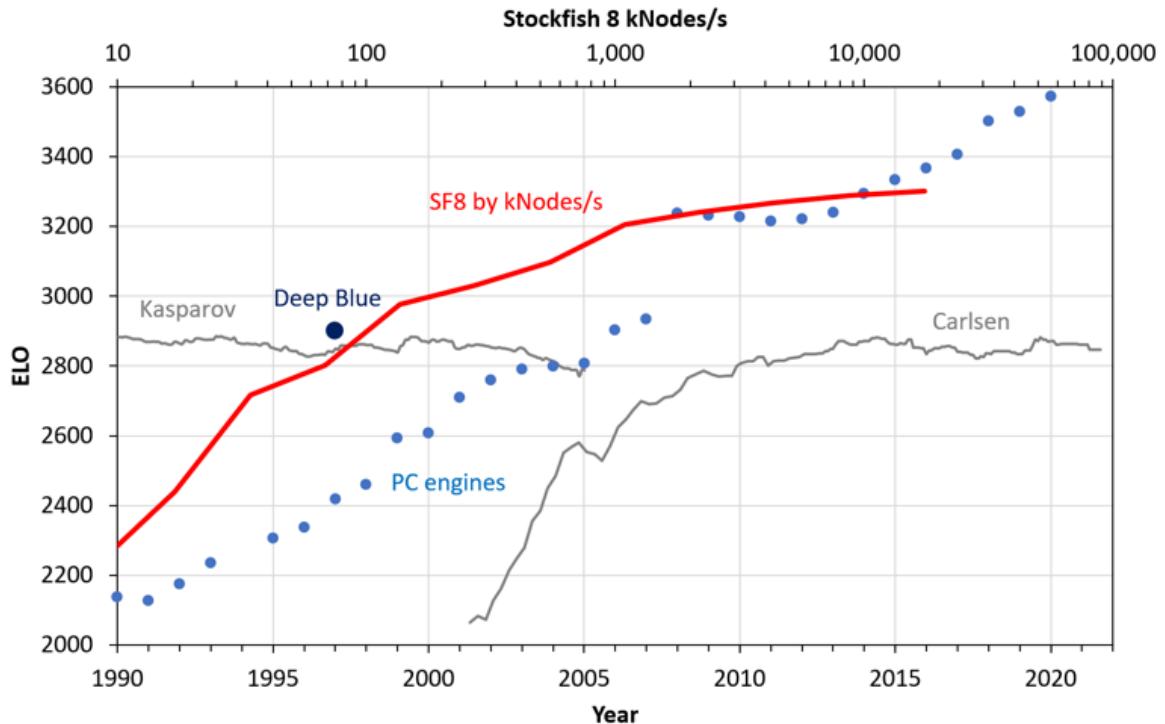
The last thing we need to do is match the kNodes/move results to the old years. We may ask: In which year was the hardware available sufficient to play these kNodes/move in a usual tournament? This leaves some room for discussion. For 1997, should we choose a dual Pentium Pro 200, or a single Pentium 200 MMX? I believe it is reasonable to compare good CPUs of the time, without going overboard. After all, we're comparing chess on home computers. If we restrict it to <1000 USD CPUs for each year, we can find some SF8 benchmarking results across the web:

- AMD 5950X (2021): [71,485 kNodes/s](#)
- Pentium III 500 MHz (1999): [127 kNodes/s](#)
- Pentium 75 MHz (1995): [6.2 kNodes/s](#)
- 386DX-33 MHz (1989): [1 kNode/s](#)

There are many more such measurements found online, but for our purpose, this is sufficient. Caveats:

- Going back very far in time becomes difficult, because SF8 [needed to be recompiled to reduced instruction sets to make it work](#); and RAM was limited in the experiment.
- It is more reasonable to match the speed to more recent years: About 200 kNodes/s in 2000, and 100 MNodes/s today. Everything before, and in between, has a factor of a few of error in its match of nodes to year.
- On the other hand, seeing benchmarks of real PCs is useful, because it encompasses uncertainties such as RAM speed.
- In reality, when considering hardware overhang for future AI, we must also ask: How well could SF8 be adapted to older hardware? Just running it unchanged will leave some performance (factor of a few?) on the table. That's a question for software engineers and compiler optimizers.

We can now bring the approximate match of Nodes/s with the years together with the other data, and present the result:



This looks quantitatively different to my [first version](#), but is qualitatively similar.

- Again, a hardware overhang of ~ 10 years at maximum is visible: SF8 achieved Kasparov level in 1997
- This was only possible for contemporary PC engines of the year ~ 2006 .
- In my old version, this was more like a 15 years gap. Back then, I had matched the speed to MIPS values for CPUs I found online.
- It is probably better to measure SF kNodes/s directly instead using a CPU speed proxy (MIPS, FLOPs, SPEC). Thus, I believe that the new figure is closer to reality.

In the next post, I will consider the other questions asked by Paul Christiano: How much compute would an old engine need to match current engines? What is the influence of opening books, endgame tables, pondering?

Edit (15 July): Magnus Carlsen time series fixed

Chess and cheap ways to check day-to-day variance in cognition

Lately I've been playing chess. I'm not very good at it. I started because our four-year-old got into it, and I preferred not to lose to her before she had earned it. I play online on chess.com, which has instantaneous matchups with strangers at approximately your level; I mostly play 5-minute timed games, where each player has a 5-minute clock for all of their moves and loses if they run out of time.

Doing this every day for a couple of months made some interesting patterns obvious. Firstly, I have good days and bad days, rather than just winning or losing mostly at random at my competence level.

Bad days are predicted by some of the things you'd expect, like poor sleep, high stress, or emotional distraction, but sometimes I have a bad-day-as-measured-by-chess when I wouldn't have predicted it, or (more rarely) a strikingly good day when I thought I was likely sleep deprived.

Secondly, playing chess on a good day versus on a bad day allows a lot of access to the feeling of being slightly smarter or slightly stupider than usual. In general, I think my ability to do high-level cognition varies a lot day to day and even within a day, but it's often really hard to tell if I'm blocked on a problem because I'm a little low on ability to think about it; usually, a problem will be different from other problems in a bunch of ways, and it's hard to tell how much of the difficulty is that I'm fatigued/distracted/whatever.

With chess, you're basically solving the same category of problem every time, with instantaneous feedback about whether your solution was any good, so I get unusually good felt-sense of where I would normally have seen something, but this time missed it, or alternatively where options and counteroptions are coming to me unusually quickly and have unusual depth and clarity.

Thirdly, having a good day at chess seems pretty strongly associated with having a good day at high-cognition-requiring tasks in general. In fact, it seems like a better predictor of this than asking myself 'do I expect to be good at cognition today?'.

On days when I do well at chess, I'm also likely to succeed at reading a complicated paper I previously bounced off because it was too hard, or at writing an article that has been blocked on loading all of it into working memory, or at having productive new thoughts if I dwell on an old problem for thirty minutes. I am *not* more likely to succeed at doing my taxes or other tasks that are motivation- rather than cognition-bottlenecked.

This could, obviously, be a self-fulfilling prophecy, or a placebo effect, but at least for me it makes playing chess a moderately valuable thing to do; in about ten minutes I can get a readout on how well my brain is working, and figure out from that which of my priorities for the week make sense to work on.

I am wildly uncertain how much this would generalize to other people, but it seems like a cheap experiment. Some suggestions, if you want to try it:

- I do 5-minute games now, but for the first two months I did 10-minute games, and found the 5-minute ones pretty bad for me: I wasn't good enough at thinking about chess to think at all in the shorter time setting, and was just reflexively making moves on instinct, which is not very useful or interesting. Over time, I compressed/cached some stuff and now I get about the same experience from 5-minute chess I used to get from 10-minute chess. I recommend starting slower.
- There's tons of advice about how to play chess - which openings to use, which countermoves to memorize for common openings by your opponent, etc. Don't learn any of it. Those things increase your ranking at chess, but the thing that's interesting here is chess as an environment in which you can see your brain function, and those things don't particularly help with that and are a time-sink of fairly arbitrary size.

You'll learn over time by losing and going 'okay, I'm not going to let that trick me again'. Your rating will probably eventually flatten out lower than the ratings of people who spend the time studying the meta, but that's not what you're using this for. Learn enough stuff about how the game is played that you have an idea of what criteria you should be using to evaluate potential moves - that is, "how do I know if this move leaves me in a stronger or weaker position" and then try to resist the urge to memorize openings or anything like that.

- chess.com offers more feedback than just "did you win the game or not" - in particular, there's an option to go through the game move by move and see what the smartest move would have been, and to revisit decisions you made that were particularly bad and try to make better ones. I do this sometimes.

They also rate your game by how closely your play adhered to 'perfect' play (what the chess engine would've done), on a 0 - 100 scale; unfortunately you can't take this number too literally as a 'how clever am I today' number because things like 'was there a long endgame where both sides had a bunch of forced/obvious moves' affect the scale a lot. I still use it for data, though.

- If you try it, or if you already do this, please comment with: did you notice day to day or time-of-day variance in how well your brain seems to be working at chess? did it track with variance you'd noticed in other contexts? did it predict anything else?

New Dementia Trial Results

This is a linkpost for
<http://www.bayesianinvestor.com/blog/index.php/2021/07/01/new-dementia-trial-results/>

There's a new [clinical trial result](#) showing that [Bredesen's approach](#) is able to at least partially cure common forms of Alzheimer-like dementia. (Press release [here](#)). It has not received as much attention as it deserves.

The 9 month study seemed a bit less impressive than what I'd hoped for, but the outcomes still support the claim that common forms of dementia are partly curable.

Out of 25 patients, 21 or 19 improved their cognition compared to the start of the trial, depending on which measure I look at, and 2 or 3 declined.

Side effects included occasional improvements in hypertension and diabetes, enough to allow patients to stop taking medications for those conditions.

1.

Compared to [my prediction in 2017](#), the trial happened a bit later than I expected, but achieved much higher patient compliance than I expected. If I had predicted the study size, I'd have likely said something way too optimistic (a thousand?).

[The trial registration](#) reports an actual enrollment of 30 patients, whereas the paper reports 28 were recruited, 3 of whom dropped out within 3 months. I don't see any explanation for that difference. For comparison, it looks like [Biogen's phase 3 trials](#) had completion rates of less than 60%.

The pandemic started in the middle of the trial, impairing compliance. Exercise seems likely to have been reduced. Two patients showed reversal of their improvement that was attributed to sheltering in moldy homes. The pandemic reportedly hurt patients' diets. I'm unsure what to make of that. Restaurant closures would have reduced access to food that tempts people to cheat - my diet better resembles Bredesen's when I'm cooking at home. But the effect would likely be the opposite for those who don't cook, as I'm guessing they'd eat more junk food.

2.

The study does not report [ADAS-Cog](#) results. ADAS-Cog seems to be the usual measure for clinical trials. It may be inappropriate for milder impairment due to [ceiling effects](#) - I'm unclear whether that's a reason for them to have avoided it.

They used several measures that seem to be fairly common tests of cognition, and they pre-registered those, so it doesn't look like they ran a lot of tests and selected those that favored the protocol.

3.

Their criteria for excluding people includes statin use, unless eligible to discontinue. Yet they [report](#) one patient discontinuing statin use after the treatment apparently improved the patient's lipids (I'm not 100% clear on whether that patient was in the trial).

Did they exclude many statin users, and if so did that cause them to have a more treatable population than most other dementia trials? Bredesen is pretty concerned that statins cause dementia. The evidence on this topic seems pretty weak. I'll guess that if statins were a big enough cause of dementia to impact this trial in an important way, someone would have managed to report better evidence than I've found. But I'm still avoiding my doctor's advice to use a statin, partly due to Bredesen's opinion.

4.

The trial didn't directly verify that patients had Alzheimer's. They claim that if the treatment was only working on non-Alzheimer's forms of dementia, then the 12 patients with an Apoe4 allele would likely have done poorly, which was not the case. That's not a very rigorous argument, but it would be fairly hard to accidentally produce the indicated results without improving the health of some Alzheimer's patients.

I've minimized my use of the term Alzheimer's here, due to uncertainty about whether it means much. Some authorities treat amyloid beta as [a defining feature](#) of Alzheimer's, yet amyloid beta has been largely demonstrated to have little causal influence on dementia.

The FDA has panicked recently over the failure of the amyloid beta industry, and is trying to provide hope for that industry by allowing it to market a drug (Biogen's Aduhelm) that reduces amyloid beta, as if that treated a disease. Maybe Aduhelm is better than nothing, but the [FDA announcement](#) seems to include a number of false statements that appear designed to suppress awareness of better alternatives.

5.

This trial makes me slightly more optimistic about Bredesen's protocol, mainly due to the evidence that they were able to get better compliance than I expected. That evidence is somewhat tempered by the small size, which likely means that only the most eager patients ended up in the trial.

We ought to have a higher prior for Bredesen's lifestyle protocol than for some novel drug, partly due to the evidence from [other cultures](#) which suggests that most dementia is due to modern lifestyles. That evidence has been neglected due to not fitting the High Modernist notion of what qualifies as evidence.

In addition, Bredesen's protocol is composed of parts that look reasonable to adopt even in the absence of effects on dementia. It looks like the reluctance to adopt most of those parts comes from issues such as price (much less than Aduhelm), hassle, and/or the fact that the evidence for effectiveness is confusing. How does that compare to, say, the CDC's reluctance to advocate mask use in March 2020?

6.

Two other recent trials showed drug-based promise for dementia.

[Cassava Science's trial results](#): cognition improved 10% over 6 months. I bought stock in Cassava due to that announcement.

[Annovis Bio has reported](#) a 30% improvement in cognition in 25 days, in 14 patients. I'm still wondering what to make of that announcement. Note that their comparison to Biogen in their press release is confusingly favorable to Biogen, which reported an improvement relative to placebo, but those patients still declined. Whereas Annovis, Cassava, and the Bredesen trials show improvement relative to baseline.

I've seen three promising clinical trial results for dementia in the last 6 months. Can anyone show me prior dementia trial results that had similar promise but failed at a later stage?

(Brainstem, Neocortex) ≠ (Base Motivations, Honorable Motivations)

The idea I do like

There's an idea I'm very fond of where:

- The neocortex (well, the [telencephalon](#) and thalamus, but let's just call it "neocortex" for short) holds all of our world-knowledge, consciousness, intelligence, planning, reasoning, etc.
- The brainstem (well, brainstem & hypothalamus, but I'll just call it "brainstem" for short) is full of lots of circuits that say "eating chocolate is good", "leaning over a precipice is scary", etc. (It also regulates your heart-rate and so on.)

This idea has gotten a bad rap from its association with discredited "[triune brain theory](#)". Relatedly, people persist in describing the brainstem with scientifically-inaccurate nicknames like "old brain", or "lizard brain", or "reptilian brain", etc. (See my discussion [here](#).) I've even heard the brainstem called "monkey brain"—as if monkeys didn't have a neocortex?! (Elon Musk is guilty of this, see [this interview - 14:45](#).)

Still, despite its bad rap, I very much like this idea, and am trying to salvage its reputation.

The trope I don't like

There's a trope that goes along with this idea in the popular imagination. It goes something like this:

- The brainstem is the source of *base and dishonorable goals* like "I want to eat candy and watch TV"
- The neocortex is the source of *respectable and honorable goals* like "I want to unravel the mysteries of the universe", "I want to get off my lazy butt and go to the gym", etc.

For example, a couple places where I've recently seen this trope are [Jeff Hawkins's recent book](#) and an old LessWrong post [The AI Alignment Problem Has Already Been Solved\(?\) Once](#).

I don't like this trope. I propose to throw into the garbage, right next to the "lizard brain" terminology.

That said, it's not pulled out of thin air. People got the idea from somewhere—it's gesturing towards a real thing, and I will talk about what I think that is below.

Why I think this trope is basically wrong

I'm a big believer in within-lifetime reinforcement learning as one of the key drivers of cognition—see [Big Picture of Phasic Dopamine](#). So maybe 5 times per second, you think a little thought or make a little plan or whatever ("I'm going to pick up this pencil", "I'm going to rewrite this sentence", etc. etc.), that thought / plan gets evaluated by various other parts of your brain, culminating in the brainstem, and the brainstem then issues a reward pertaining to that thought, and that reward drives both decision-making (e.g. you won't pick up the pencil if that plan is judged as bad) and gradual learning to think better thoughts in the future.

There's *no room* in this RL-type process for "motivations that come from inside the learning algorithm". Like, that's just not where motivations come from! Motivations come from rewards. Rewards are calculated in the brainstem. It's as simple as that.

The kernel of truth

I think this wrong idea is pointing at a real phenomenon, and I'm going to try to explain it.

1. The neocortex can issue “the same plan” (same sequence of motor actions) framed in many different ways, and those different framings will get different rewards from the brainstem.

The brainstem is judging thoughts / plans, not actual "futures". It's not omniscient! It's judging a map, not a territory! So you can have different ways to think about the same plan which result in different brainstem rewards.

For example:

- You can have a thought "I will go to the gym" which is negative-reward, and so you don't do it.
- Or you can have a thought "I will go to the gym and thus be healthy" which mixes negative and positive aspects, so maybe it's positive-reward on net, and so you do it!

There are within-neocortex dynamics that determine which of those two options gets proposed to the brainstem.

...And therefore there's a sense in which the neocortex "gets credit" for the fact that you do in fact go to the gym.

...As long as we don't forget that *it's the brainstem* that judges "I will go to the gym" as bad, and *it's the brainstem* that judges "I will be healthy" as good, and *it's the*

brainstem that judges the combined thought “I will go to the gym and thus be healthy” as slightly-good, and thus *it’s the brainstem* that enables that plan to actually get executed.

2. Some plans are rejected by the brainstem, but they have a desirable aspect / consequence, and then the neocortex may be rewarded for dwelling on the desirable aspect, and may run a search algorithm for ways to make that aspect actually happen

Continuing with our example, let’s say the brainstem is getting signals that the body is tired, so any plan that involves exertion loses a *lot* of points. It loses so many points that the trick from above no longer works; the “I will go to the gym and thus be healthy” thought is now net-negative (compared to its alternatives). So I’m not going to the gym.

Nevertheless, the thought “I will be healthy” continues to be appealing. So the neocortex—ever the goal-seeking algorithm—keeps searching, trying to construct a plan that will result in “I will be healthy” actually happening.

Hmm, is there a way that I can get healthy while continuing to sit on the couch? Nope.

OK, what about piling more positive-reward thought pieces into the mix? How about *this* plan: “I will go to the gym, and thus be healthy, and also I’ll be the kind of person who follows through on my commitments, and also I’ll impress my friends with my rock-solid abs, etc. etc.” Again, it’s the same sequence of motor actions, but this thought is framed to make it even *more* appealing to the brainstem. Nope, the brainstem still says “all things considered, this is still a bad plan”. And I notice that I am still sitting on the couch.

Anyway, **that neocortex search process that I’m talking about here gets mapped intuitively into “my neocortex wants to be healthy, but my brainstem wants to stay on the couch”.**

That’s not literally true: my brainstem wants *both* things! My brainstem’s endorsement of “I want to be healthy” is critical here—it’s *powering* the ongoing neocortex search algorithm activity! (Again, in my view, the neocortex won’t take any actions *or even think any thoughts* without the brainstem rewarding it for doing so.) But it’s also my brainstem’s dislike of getting off the couch that is causing the search to come up empty.

3. This search algorithm tends to immediately “snuff itself out” unless the desirable aspect is “honorable” / appealing upon reflection / leading to desirable follow-on consequences

So far this seems to be unrelated to base vs noble motivations. Where does *that* come from?

Well, let's try flipping it around. The reverse would be as follows:

My neocortex proposes "I will go to the gym and be healthy", the brainstem endorses it and that's now the plan. But the gym doesn't open for 10 minutes, so I'm sitting and waiting. Now my neocortex thinks the thought "I won't go to the gym, I will stay on the couch instead". That's an appealing thought! But the brainstem rejects it as a plan, because its appeal is not strong enough to outweigh the appeal of "I will be healthy". So the neocortex search algorithm whirrs into action! Is there a way to frame the thought "I will stay on the couch instead" so as to make it more appealing to the brainstem? How about this: "I will stay on the couch, and thus avoid the risk of dropping a heavy weight on my foot, and also I won't have to talk to Tony at the gym entrance, man I hate that guy." Nope, not good enough, the brainstem says I'm sticking with the original plan of going to the gym.

So *this* neocortex search algorithm should correspondingly get mapped intuitively into "my neocortex wants to stay on the couch, but my brainstem wants to be healthy"!! As above, that's not a technically correct description, just an intuition. But clearly, it's not impossible for this kind of thinking process to happen.

Still, I do think this case I just walked through is an *unusual* case. We are more likely to have our neocortical search algorithm searching for ways to do honorable things like go to the gym, not searching for ways to stay on the couch.

Why isn't it symmetric?

I think the difference is: some things just get more appealing when you think about them more, and others get less appealing.

More mechanistically, the neocortex works on a principle of "thoughts tend to trigger other thoughts", and those "other thoughts" can be positive-reward or negative-reward. Remember, the brainstem is powering this search process; we'll only keep searching as long as the brainstem is liking what it's seeing. If the search algorithm winds up spawning a bunch of secondary thoughts that the brainstem hates, those secondary thoughts will *snuff out the search algorithm itself*. And we'll start thinking about something else instead.

That's a bit abstract. Let's try an example.

Let's say I'm in a hurry to leave for an appointment, but my neocortex entertains a glimmer of the idea "I will stop and eat candy before I go". The brainstem says "Nope, candy is good, but being late is really bad, plan rejected." So the neocortex search algorithm spins up: is there a way to eat candy without being late? What does that search entail? Well, it immediately brings to the forefront of your mind the idea "I will eat candy", and then it tries to build an acceptable and plausible plan that incorporates that idea—like filling in the missing pieces of a puzzle. However, holding "I will eat candy" at the forefront of your mind spawns a bunch of negative follow-on thoughts like "...and thus I will break my promise to myself", "...and thus I won't fit in my pants", etc. The brainstem gets a whiff of those thoughts and it *snuffs out* the search process itself. The brainstem no longer sees any benefit in thoughts that involve searching for a way to eat candy, so the neocortex stops doing that search.

By contrast, if your search algorithm is looking for a way to do something that's appealing on reflection, i.e. something that's desirable and whose follow-on consequences are *also* desirable, then the search algorithm *won't* immediately snuff itself out. It can keep running until success, or until frustration sets in.

So I think that when people introspect about times that they've been searching for a way to "make themselves do something", they'll typically come up with examples where the thing was honorable / appealing upon reflection / having desirable consequences, since those are the searches that last for more than a fraction of a second.

3A. An important special case: self-reflective thoughts

When I talk about plans with "desirable consequences", you'll probably think of normal, causal consequences, like "I will go to the gym" → "I will be healthy". That's one possibility, but another important case is "consequences for how we see ourselves, how others see us, how we present ourselves, etc.". For example, if you're committed to dieting, then it's possible that:

- The idea of "eating candy" is appealing
- The idea of "*myself* eating candy" is aversive

Get it? The first one involves taking certain actions, tasting certain tastes, etc. The second one involves the abstract idea that I will have eaten candy, and I'll remember having done it, and when people ask me I'll have to tell them about it, or lie.

The second thing is a *consequence* of the first thing. If I eat candy, then I *will have eaten* candy! So just as above, the desirability of self-reflective thoughts can determine whether the neocortex search algorithm works long and hard on finding a way to make a plan that passes muster with the brainstem, or whether the search algorithm immediately snuffs itself out.

I think that we have a *lot* of strong motivations about the kind of person we want to be. (...And the brainstem turns these motivations into thoughts and actions, just like every other motivation.) I think this is an important component of our social instincts.

So this gets back to "honorable motivations", here in the sense of "motivations that we're proud to have". When we want something that we associate with an "honorable motivation", it's much likelier that the neocortex will search for a way to make it happen despite internal (motivational) obstacles, rather than the search algorithm running for a fraction of a second and then snuffing itself out. Again, this leads to the (misleading) intuition that the neocortex is trying to make us "act honorably" over the protests of our brainstem.

4. Willpower, akrasia, and "guilt by association"

Let's go back to the example from earlier:

A: "I will go to the gym" is aversive (negative-reward)

B: “I will be healthy” is attractive (positive-reward)

A+B: “I will go to the gym and thus be healthy” is net slightly attractive (slightly positive reward)

One aspect of this is that, by yoking together A+B, A has become more appealing. That’s what I was talking about above.

But there’s another aspect: by yoking together A+B, B has become *less* appealing. (“Ugh, I’m not so sure about “being healthy”, it’s pretty exhausting!”)

I think the [supervised learning parts of our brain](#) literally learn the pattern “being healthy leads to exertion”. (Related: [this Scott Alexander post](#).)

This is a factor that constrains the ability of the neocortex search algorithm to successfully find strategies to frame inherently-unpleasant plans in a way that the brainstem finds net attractive.

Specifically, it’s not as easy as taking one strongly-motivating thought like “*I want to follow through on my commitments*”, and then attach that one thought to 500 different inherently-unpleasant tasks, one after another, and then we’ll actually do all 500 of those things. Instead, each of those tasks drags down the concept of “I want to follow through on my commitments”, until that concept is so saddled with negative associations that it’s no longer even able to escort *itself* through the brainstem, let alone anything else!

Obligatory side-notes

What about “the brainstem is stupid”?

I often talk about how the brainstem is stupid, it doesn’t know what’s going on in the world, etc. That seems inconsistent with me breezily talking about how the brainstem sees “going to the gym” as bad and “being healthy” as good—aren’t those kinda complex concepts? Well, my answer is that certain parts of the brain (agranular prefrontal and insular and cingulate cortex, ventral striatum, amygdala, hippocampus) use supervised learning to distill an arbitrary thought into a maybe dozens-of-dimensional vector space that the brainstem (and hypothalamus) can interpret. We can interpret this vector as answering questions like “If I do this plan, how appropriate would it be to cringe? To salivate? To release cortisol? To laugh? Etc.” This enables the brainstem to understand what’s going on well enough to issue appropriate rewards. See [Big Picture of Phasic Dopamine](#). I glossed over this stuff in this post, in order to keep things simple.

What about “The Monkey And The Machine”?

Many readers here have seen [Paul Christiano’s post The Monkey And The Machine](#). I don’t have any particular complaints there. I think of the post as being mostly about stuff that happens within the neocortex, where “monkey” is the trained model learned by the cortex, and “deliberator” is a subset of those learned processes that implements things like “explicit reasoning using language”. (Somewhat related: Kaj Sotala’s post [System 2 as working-memory augmented System 1 reasoning](#).) This is a

reasonable and useful way to think about certain things. I don't think it has anything to do with what I'm talking about here.

AlphaFold 2 paper released: "Highly accurate protein structure prediction with AlphaFold", Jumper et al 2021

This is a linkpost for <https://www.nature.com/articles/s41586-021-03819-2>

Book review: The Explanation of Ideology

This is a linkpost for <http://www.bayesianinvestor.com/blog/index.php/2021/07/19/the-explanation-of-ideology/>

Book review: The Explanation of Ideology: Family Structure and Social Systems, by [Emmanuel Todd](#).

What features distinguish countries that embraced communism from countries that resisted?

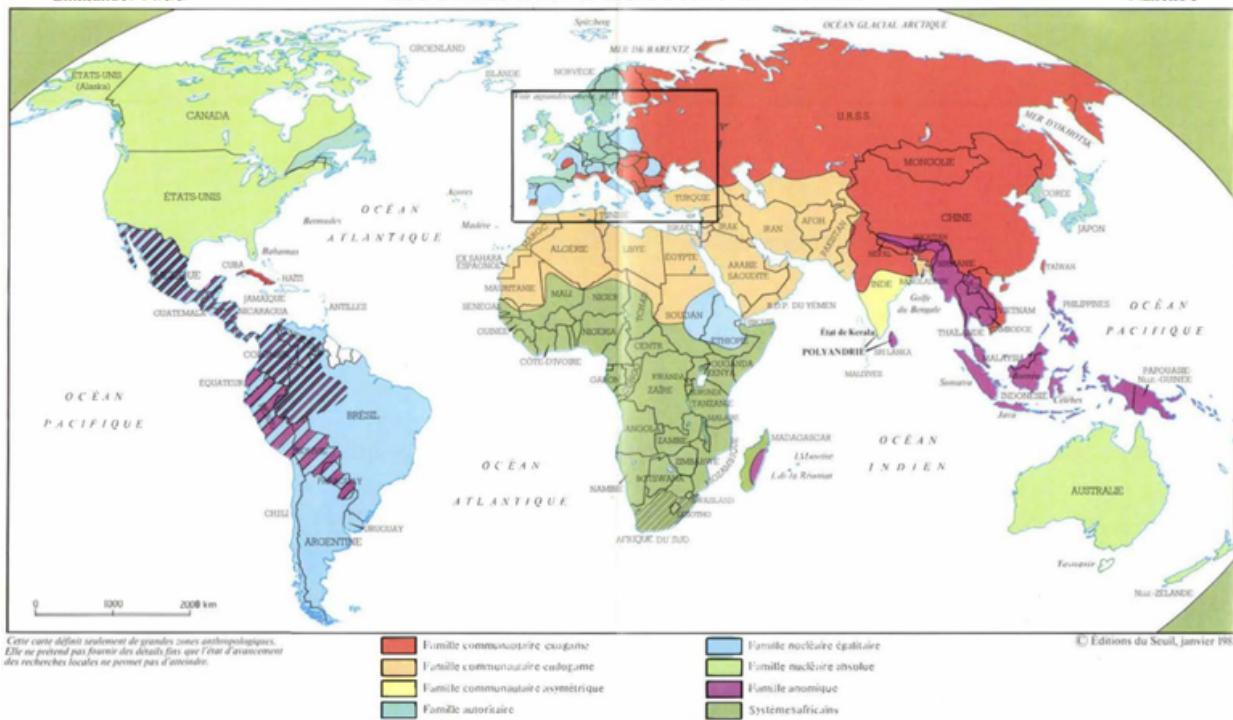
Why did Islam spread rapidly for a century and a half, then see relatively few changes in its boundaries for more than a millennium?

Todd's answer is that the structure of the family is a good deal more stable than ideologies and religions, and different family structures create different constraints on what ideologies and religions will be accepted. Published in 1983, it still seems little-known.

Maybe this neglect is most pronounced in the English-speaking parts of the world, where one family structure is overwhelmingly popular, and alternatives are often dismissed as primitive relics. France seems more conducive to Todd's insights, since France has four different family structures, each dominating in various regions.

Here are the main dimensions that he uses to categorize family structures:

- Exogamous: [marriages between cousins](#) are heavily discouraged, versus endogamous: marriages between cousins are common.
- Nuclear versus community: Are children expected to move away from the parental home upon marriage?
- Equal versus unequal. Beware that this is a nonstandard meaning, focused on relations between brothers, especially on whether inheritances are split equally. Todd says this is inversely correlated with sexual equality. He seems willing to accept sexual inequality as not worth trying to eliminate ("male dominance, a principle ... which is in practice much more universal than the incest taboo").
- Liberty versus authority. This is mostly about parental authority over children.



Here are his categories, listed in roughly descending order of how many Europeans practice them (this is Todd's order; the book is a bit Eurocentric).

Exogamous Community

This system is equal, authoritarian, and universalist. It mostly coincides with countries that adopted communism at some point, plus Finland and northern India.

It is relatively unstable, tending to produce features such as communism, which wages war on the family, and urbanization, which pushes toward a more nuclear family. But then why is it the most populous family system (41% of the world population when the book was written)? Todd does not ask. Some of it might be due to generating population growth, but that can't be a full explanation. It seems unlikely to be due to people especially enjoying it, as it has the highest suicide rate of any family system.

Why is Cuba, with its apparently Western culture, the sole country in the New World that's fertile for Communism? Todd doesn't have direct evidence of Cuba's family system, yet he maintains it's an exogamous community system. After some hand-wavy talk of other sources of Cuban culture, he pieces together hints from the suicide rate and census data. The census data does suggest that married children have some tendency to live with parents (but is that due to a [housing shortage](#) more than to culture?). The suicide rate provides some sort of evidence, but there's a lot of noise in that signal. He apparently provides more evidence in his [2011 book](#) (French only), according to [this paper](#), and his [2019 book](#).

Authoritarian

This system is unequal, and intermediate between nuclear and community: the only child to remain with his parents after marriage is the son who is the primary heir.

The exogamous and endogamous versions are apparently not worth distinguishing. The endogamous version seems uncommon - maybe it's only found in [non Ashkenazi Jews](#)?

These isolationist cultures resist assimilation more than do most other family systems. That produces fairly small, homogeneous countries, or fragmented groups. Examples are Germany, Sweden, Japan, Korea, Scotland, Catalans, and Jewish culture.

Egalitarian Nuclear

This system is exogamous, non-authoritarian, and universalist. It includes nearly all of the Catholic regions of Europe and South America.

Absolute Nuclear

This system is non-authoritarian, exogamous, and weakly unequal. It's weakly isolationist. It's fairly similar to the Egalitarian Nuclear type.

It's found in Anglo-Saxon countries, Holland, and, surprisingly, Denmark is in this category, in spite of the cultural features it shares with Sweden.

Where did he get the label "absolute" from? I'll suggest replacing it with libertarian.

Endogamous Community

This is found mainly in the Muslim parts of the region that extends from northern Africa to the western tips of India and China. It's equal and universalist.

Its strict religious rules about inheritance result in unusually weak parental authority. Todd considers it authoritarian, but in a sense that's very alien to the European understanding of that word. Authority in this case is embodied in custom and in the Koran, not in humans or human-designed organizations.

It has unusually good fraternal bonds, and low tension within the family. Suicide rates ~~are~~ were less than 1/20 of the European average, and illegitimate births are rare.

[Henrich mentioned](#) that Protestant culture caused an increase in suicide rates compared to Catholic culture, due to trade-offs that made it more likely to produce a Tesla or a Google, at the cost of making people lonelier. Todd implies that the exogamous community system is further in the direction of less loneliness, likely at a cost of less innovation.

The split between Christianity and Islam was due, according to Todd, to differences over exogamy. Christianity became more hostile to cousin marriage due to increasing influence of northern regions that more strongly opposed cousin marriage. Islam imposed some incest restrictions on cultures that had none, but tolerated incest more than did Christianity, so it was more welcome in regions that were committed to cousin marriage. Islam was also sometimes tolerated by the next two categories of family systems, although they don't fully accept all of the Koran's rules.

Arab socialism is a unique attempt to build socialism without the state, or to be more precise and less derisive, an effort to construct socialism in a culture without any special aptitude or a tradition of centralized, bureaucratic administration.

Endogamous systems in general reject state authority. Todd attributes this to their reluctance to create bonds of kinship with strangers. Whereas the exogamous systems provide a role model for creating a strong relationship with non-kin. This reasoning sounds suspect to me. I prefer Henrich's way of reaching a similar result.

History is made by individuals in nuclear family countries, by the government (a parental symbol) in authoritarian systems. It is defined by custom and thus eliminated in the case of endogamous anthropological systems. Islam's historical passivity can be seen to derive from its fundamental anthropological mechanism.

The Muslim father is too easy-going to be hated or rejected, either in human or divine form. The Islamic god is too forgiving for anyone to want to annihilate him.

Asymmetric Community

This system is endogamous, with marriage encouraged between children of a brother and a sister, but with a prohibition on marriage between children of two brothers, or children of two sisters.

It's found mostly in southern India.

It's egalitarian in the narrow sense of equality between brothers, but it supports large inequalities outside of the family (e.g. the caste system). This seems to weaken Todd's message elsewhere that equality within the family tends to generate egalitarian political forces.

Some unusual variants of this family system support a form of communism that's more laid-back than we expect from communism (Stalinists, Maoists, and sometimes Trotskyites cooperate well).

They are found in Sri Lanka and the Indian state of [Kerala](#). These variants are distinguished by polyandry being common, often with brothers sharing a wife. They're either matrilineal, or intermediate between matrilineal and patrilineal.

Anomic

Todd calls this a "faulty nuclear" system, with few rules, or rules that are often ignored. It has some overlap with the Absolute Nuclear family, but it oscillates between communitarianism and mild individualism.

It's seen in parts of southeastern Asia, some indigenous South American cultures, the Incan empire, ancient Egypt.

It tends to produce strong village solidarity.

It often produces strong but informal grouping by class, with marriage being mostly within a class. The topmost class looks powerful, and commands slaves to build displays of power such as pyramids. Yet the lack of discipline means that power is fragile, and easily destroyed by outside forces.

It fits well with the ambiguous deity of Buddhism.

Todd makes some weird claims about the massacre of Indonesian communists in 1965-6: it was substantially a grass-roots uprising, partly from within the communist movement, and eliminated communism, even in regions where communists had gotten a majority of the votes. That fits with Todd's claims that this family system is undisciplined and anti-authoritarian, unwilling to attach strongly to an ideology. But it's moderately inconsistent with [Wikipedia's account](#).

African / Unstable

Sub-Saharan Africa is noted for systems with shorter-duration polygynous marriages. Todd hints at a lot of diversity within these regions, but documents little of it.

Islam has had difficulty penetrating these regions because its strict taboo on inheriting wives conflicts with a standard feature of these family systems.

Conflicts with Henrich?

I found this book via Policy Tensor, which [points to some tension](#) between Henrich's [The WEIRDest People](#) and Todd's belief that family structures are very hard to change. Actually, Policy Tensor claims to have evidence that Henrich is flat out wrong, but Policy Tensor presents way too little evidence to justify that claim.

I see [some hints](#) that Todd's 2011 book has more detail on the early history of family systems, possibly with clear evidence against Henrich.

Todd tells us that when there's a change in what family structure dominates a region, it's mostly due to a subpopulation becoming more dominant. It's not too hard to imagine that some of Europe's increasing prohibitions on cousin marriage under the early Christian church were due to increased influence from northern cultures, which apparently were more firmly against cousin marriage than the southernmost European cultures. And most of the correlations that Henrich reports could have been due to pre-existing local and regional cultures influencing what religious doctrines were accepted, rather than religions altering the culture.

I don't see much evidence on whether family systems are too persistent for Henrich's claims of Christianity causing exogamy to be plausible. Todd wants us to assume that family systems persist over many centuries, but he also notes that they do sometimes change, e.g. that urbanization erodes community and authoritarian systems.

The most important conflict I see between Henrich and Todd is that Henrich describes the marital rules for Christianity as a whole, seemingly taking it for granted that European Christianity had a fairly uniform culture at any one time. Whereas Todd wants us to assume that cultural change in Rome would tell us almost nothing about changes in London, and that we should presume (in the absence of clear evidence) that London's culture was mostly a continuation of its pre-Christian culture. Henrich tests many different hypotheses about what might cause the correlation between culture and exposure to Christianity, but he seems biased towards hypotheses for which he found good data, and he likely didn't find much data for the geographical distribution of culture circa 500 CE.

Henrich and Todd agree on a number of important points that others neglect. Henrich still looks mostly right, but there's plenty of complexity that he's sweeping under the

rug. Henrich overstates the effect of the church on culture, and overstates the novelty of WEIRD culture.

Here's Todd partly supporting Henrich:

Developed in France and England, the individualist model was offered to the world. ... In the middle ages, the individual did not exist. He emerged in the West during the Reformation and the French Revolution.

Both authors seem to agree that different systems are good at achieving different goals. They'd mostly say that Muslim culture in the year 1500 looked more successful than British culture of the time, and that was partly due to the strengths of the endogamous family system. They'd also agree that modest changes after 1500 in British culture brought out the strengths of the exogamous nuclear families. So it's a bit confusing to try to classify cousin marriage as a sign of a backwards or an advanced culture.

Both authors agree that culture mostly changes via evolutionary forces, although they likely disagree on particular exceptions:

But the family, varied in its forms, is not itself determined by any necessity, logic or rationale. It simply exists, in its diversity, and lasts for centuries or millennia. ... It reproduces itself identically, from generation to generation, the unconscious imitation of parents by their children is enough to ensure the perpetuation of anthropological systems. ... It is a blind, irrational mechanism, but its power derives precisely from its lack of consciousness ... Furthermore, it is completely independent of its economic and ecological environments.

Evaluating predictions

With many books, I check for mistakes by following references. I didn't try that here, partly because he rarely connects specific claims to specific sources. Instead, enough time has passed that it's appropriate to judge him based on well-known changes since the book was published.

1.

Where would communism spread or recede?

Todd sounded pretty confident that communism would not spread further in the New World, and his reasoning also applies to most non-communist states other than Finland, with a bit of uncertainty about Italy and India.

It may be hard for many of you to recall, but in 1983 many people were concerned about the trend of expanding communism, and few people were forecasting a collapse of communism in anything other than vague and distant timelines.

Todd firmly predicted that Ethiopia would resist Soviet attempts to turn it communist. He wrote at a time when that prediction [bucked a moderately clear trend](#). Soviet influence seems to have peaked about when the book was published, and in about 4 years Ethiopia started a clear move away from communism.

Todd's thesis suggests that communism was more likely to be rejected in places where communism was imposed by force on a family system that doesn't support it:

- [Poland](#)
- [Romania](#)
- North Korea
- Cambodia
- Laos
- the six [Muslim Soviet republics](#)

I see no clear evidence that these places rejected communism more than did those with exogamous community families, so I count this as a failed implied prediction.

Todd predicted further decline in the French communist party, and it [looks like that happened](#).

Some of this might be due to his prediction (made elsewhere) that the Soviet Union would collapse, which doesn't seem to directly follow from the claims in the book.

2.

Given Todd's ideas, it becomes painfully obvious that that the US attempt at installing a Western-style government in Iraq would thoroughly fail.

An influential political faction thought that the US could accomplish in Iraq something like what it did with Germany and Japan after WWII. Those two countries looked different enough culturally to provide what looked like medium-quality evidence that Western-style governments could be imposed in many countries.

Had that faction believed Todd, they'd have known that their evidence only covered one type of family structure, and that the difference between exogamous and endogamous marriage practices would make an enormous difference. I'm referring not just to details such as the willingness of Iraqis to accept democracy, but more basic issues like their reluctance to respect features such as nations, or civil authority.

3.

"Assassinating the president is almost a custom in North America." - I guessed that this was clearly discredited by the absence of assassination attempts after 1981, but Wikipedia [lists enough attempts](#) that I have to admit there's some truth to Todd's claim.

4.

Todd's beliefs imply some predictions about which European countries are likely to have the most conflict with Muslim immigrants. E.g. the book led me to expect more tension in Germany and Sweden than in Poland and Spain. Tables 2 to 5 of [this report](#) mostly confirm that prediction, but [this survey of attitudes](#) shows the opposite pattern. So I'm confused as to whether there's a stable pattern.

5.

I recommend [Testing Todd: family types and development](#), which provides mixed evidence on some of the book's claims. But note that some of the hypotheses which that paper attributes to Todd don't match my understanding of the book's claims.

- Todd says the endogamous community family is anti-racist, yet this paper reports it as the most racist family system, while claiming the racism data support Todd's

view.

- The paper shows that authoritarian family system has greater rule of law than other systems, and claims that conflicts with Todd's position. That seems to require a bizarre misunderstanding. I count this as clearly confirming Todd.
- I'm confused as to whether they use an appropriate measure of innovation - they find that authoritarian family systems are more innovative than nuclear family systems, which looks suspicious to me.

6.

In sum, his predictions were clearly better than what a random pundit of the time would have made, but not good enough that I'd bet much money on his beliefs.

Conclusion

This is one of the rare books that is shorter than I wanted.

The book's claims are unlikely to be more than 60% correct, but they're still quite valuable for focusing attention on topics which are both important and neglected. Whenever I try to understand differences between cultures, I'll remember to ask whether family structures explain patterns, and I'll likely often decide it's hard to tell.

I've become frustrated at how little attention sources such as Wikipedia pay to what I now see as the most important features of a culture.

I'm pretty sure that the patterns that he describes are much more than mere coincidences, but I don't trust his guesses about the causal mechanisms.

PS. - Parts of the book are much too Freudian for me. E.g. a section on witch-hunts (which happen mainly in authoritarian family societies) is titled "Killing the mother".

For some, now may be the time to get your third Covid shot

Summary

- It seems likely there is a significant benefit to a booster shot in preventing symptomatic infection in the face of the Delta variant for many people (Bullets 1-3).
- There is significant indication that a third dose of an mRNA vaccine has a good safety profile (Bullets 2-4).
- Delta may have an a greater impact in the US than you realize (Bullet 5), and
- Therefore, for some people now may be the right time to get their third Covid mRNA shot (Bullet 6) although there are many reasons to potentially wait (Bullet 7).

Bullets

1. Two days ago, [the Israeli government announced](#) that its observed vaccine efficiency (Pfizer) as of June 6th was down to 64% against symptomatic infection, while protection against severe disease and hospitalization has declined more mildly to 93%. This is seemingly almost entirely attributed to the Delta variant.

2. Today, [Pfizer put out a statement](#) which included the following:

Pfizer and BioNTech have seen encouraging data in the ongoing booster trial of a third dose of the current BNT162b2 vaccine. Initial data from the study demonstrate that a booster dose given 6 months after the second dose has a consistent tolerability profile while eliciting high neutralization titers against the wild type and the Beta variant, which are 5 to 10 times higher than after two primary doses. The companies expect to publish more definitive data soon as well as in a peer-reviewed journal and plan to submit the data to the FDA, EMA and other regulatory authorities in the coming weeks. In addition, data from a recent Nature paper demonstrate that immune sera obtained shortly after dose 2 of the primary two dose series of BNT162b2 have strong neutralization titers against the Delta variant (B.1.617.2 lineage) in laboratory tests. The companies anticipate that a third dose will boost those antibody titers even higher, similar to how the third dose performs for the Beta variant (B.1.351). While Pfizer and BioNTech believe a third dose of BNT162b2 has the potential to preserve the highest levels of protective efficacy against all currently known variants including Delta, the companies are remaining vigilant and are developing an updated version of the Pfizer-BioNTech COVID-19 vaccine that targets the full spike protein of the Delta variant.

[CNN says they plan to submit](#) to the CDC for emergency approval to administer booster doses next month.^[1]

3. [Moderna previously announced](#) that in their trial they also observed indication of significant benefit to a third dose booster (of half the size of the initial Moderna

doses)^[2] and found it to have a similar safety profile to the second shot.

4. In addition to the clinical trials, [Turkey](#), the UAE, [NYC](#), and some others have given people third doses in some circumstances / categories. [The immunocompromised](#) are particularly likely to benefit, as one would expect.
5. The Delta variant may have more of an impact than you expect. [Polymarket currently gives](#) a ~ $\frac{2}{3}$ probability to the US again having more than 100,000 cases in a day before the end of the year. This could come from a new variant or other novel situation, but I expect most of this probability for most bettors is with an eye toward the Delta variant. Cases in the UK [have increased 10-20x](#) since their post-vaccination trough, and it has/had similar vaccination rates to the US (although less immunity from having contracted Covid, and a greater proportion with only one dose or a less effective vaccine). The UK's current number of new cases is equivalent to ~160k new daily cases in the US when adjusted for population.
6. There are various reasons why getting vaccinated now may be particularly good for you
 1. Symptomatic infection may be bad to you, especially if you or someone you interact closely with is particularly vulnerable.
 2. If you received your vaccine early (e.g. 2020-February 2021) your protection is more likely to have weakened.
 3. Your peak exposure may be soon if the Delta variant has a similar effect in the US as it did in the UK. You'd also need to plan for the 1-2 week delay between shot and effect.
 4. When a booster shot likely does get approved by the CDC in ~1-3 months, there may be a run on supply preventing getting a vaccine booster for some time.
7. There are also various reasons to not pursue a booster vaccine now
 1. You may care primarily about protection against severe disease, which seems to largely be maintained in the face of the Delta variant.
 1. You may be in good health / young, in which case you have additional reason to believe you will be fine even if you contract Covid.
 2. If less time has passed (under 4-5 months?) since you were vaccinated, then your potential benefit is likely less than some bullets above probably suggested it was for those who have had more time pass.
 3. Some, [such as the CDC and FDA](#), don't think a booster is currently warranted.
 4. It's unclear to me if it's legal to get a third dose in the US prior to that being approved. Regardless, it may be hard to procure without misrepresentation of your previous vaccinations.
 5. More data on safety, optimal timing, etc. will be forthcoming and may help you make a better decision.
 6. Boosters may be available this fall that specifically target newer variants, and perhaps getting a booster now will in some way make procurement of those worse (worse side effects, mandatory wait time since last vaccine, etc).
 7. Getting an additional dose of vaccine may 'take it away' from somebody else who needs it more, due to global supply limitations.

8. You may believe the Delta variant is unlikely to lead to a large wave in your area, e.g. because of local vaccination rates or the 'control system'.

For some, now may be the time to get your third Covid shot.

1. [Summary article](#) of these developments. ↵
2. For this reason, if pursuing a third booster shot soon, I'd lean Pfizer (smaller dose size) over Moderna. ↵

The SIA population update can be surprisingly small

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

With many thanks to Damon Binder, and the spirited conversations that lead to this post, and to Anders Sandberg.

People often think that the [self-indication assumption](#) (SIA) implies a huge number of alien species, millions of times more than otherwise. Thought experiments like the [presumptuous philosopher](#) seem to suggest this.

But here I'll show that, in many cases, updating on SIA doesn't change the expected number of alien species much. It all depends on the prior, and there are many reasonable priors for which the SIA update does nothing more than double the probability of life in the universe^[1].

This can be the case even if the prior says that life is very unlikely! We can have a situation where we are astounded, flabbergasted, and disbelieving about our own existence - "how could we exist, how can this beeeeeee?!?!!?!" - and still not update much - "well, life is still pretty unlikely elsewhere, I suppose".

In the one situation where we have an empirical distribution, the "[Dissolving the Fermi Paradox](#)" paper, the effect of the SIA anthropics update is to multiply the expected civilization per planet by **seven**. Not seven orders of magnitude - just seven.

The formula

Let $\rho \in [0, 1]$ be the probability of advanced space-faring life evolving on a given planet; for the moment, ignore issues of life expanding to other planets from their one point of origin. Let f be the prior distribution of ρ , with mean μ and variance σ^2 . This means that, if we visit another planet, our probability of finding life is μ .

On this planet, we exist^[2]. Then if we [update on our existence](#) we get a new distribution f' ; this distribution will have mean μ' :

$$\mu' = \mu(1 + \frac{\sigma^2}{\mu}).$$

To see a proof of this result, look at this footnote^[3].

Define $M_{\mu,\sigma^2} = 1 + \sigma^2/\mu^2$ to be this multiplicative factor between μ and μ' ; we'll show that there are many reasonable situations where M_{μ,σ^2} is surprisingly low: think 2 to 100, rather than in the millions or billions.

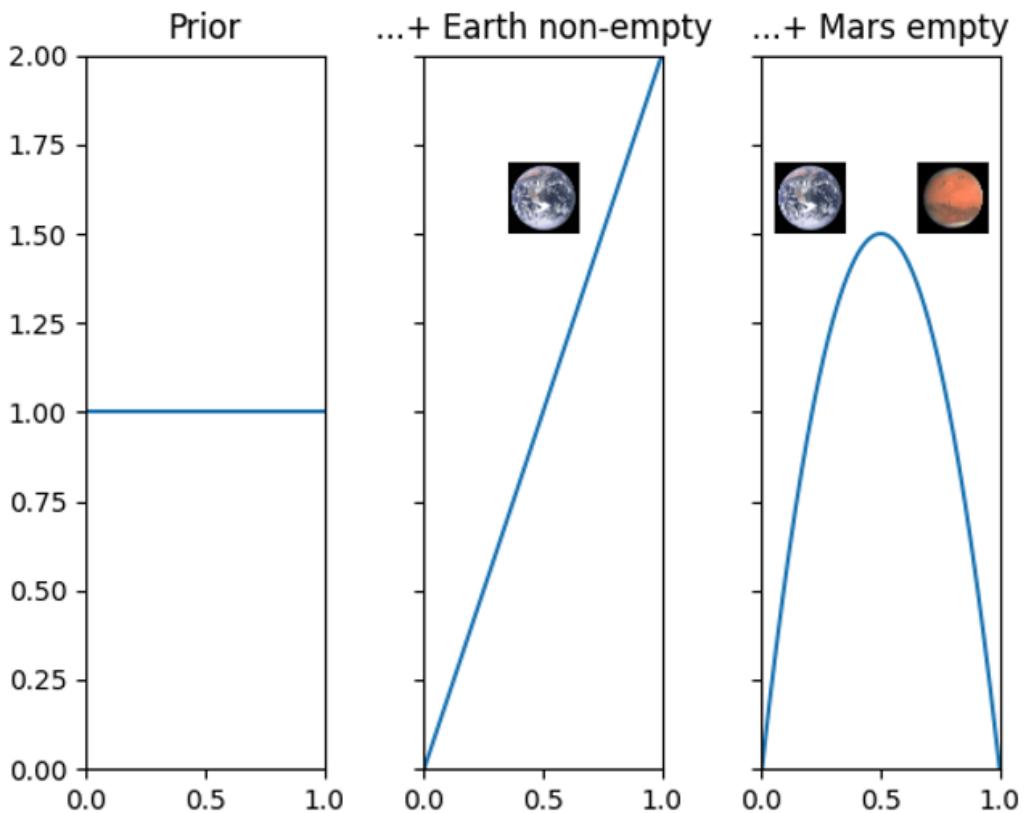
Beta distributions I

Let's start with the most uninformative prior of all: a uniform prior over $[0, 1]$. The

expectation of p is $\int_0^1 p dp = 1/2$, so, without any other information, we expect a planet to have life with 50% probability. The variance is $\sigma^2 = 1/12$.

Thus if we update on our existence on Earth, we get the posterior $f'(p) = 2p$; the mean of this is $2/3$ (either direct calculation or using $M_{1/2,1/12} = 1 + 4/12 = 4/3$).

Even though this change in expectation is multiplicatively small, it does seem that the uniform prior and the $f'(p)$ are very different, with $f'(p)$ heavily skewed to the right. But now consider what happens if we look at Mars and notice that it hasn't got life. The probability of no life, given p , is $1 - p$. Updating on this and renormalising gives a posterior $6p(1 - p)$:



The expectation of $6\rho(1 - \rho)$, symmetric around 1/2, is of course 1/2. Thus one extra observation (that Mars is dead) has undone, in expectation, all the anthropic impact of our own existence.

This is an example of a [beta distribution](#) for $\alpha = 2$ and $\beta = 2$ (yes, beta distributions have a parameter called β and another one that's α ; just deal with it). Indeed, the uniform prior is also a beta distribution (with $\alpha = \beta = 1$) as is the anthropic updated version 2ρ (which has $\alpha = 2$, $\beta = 1$).

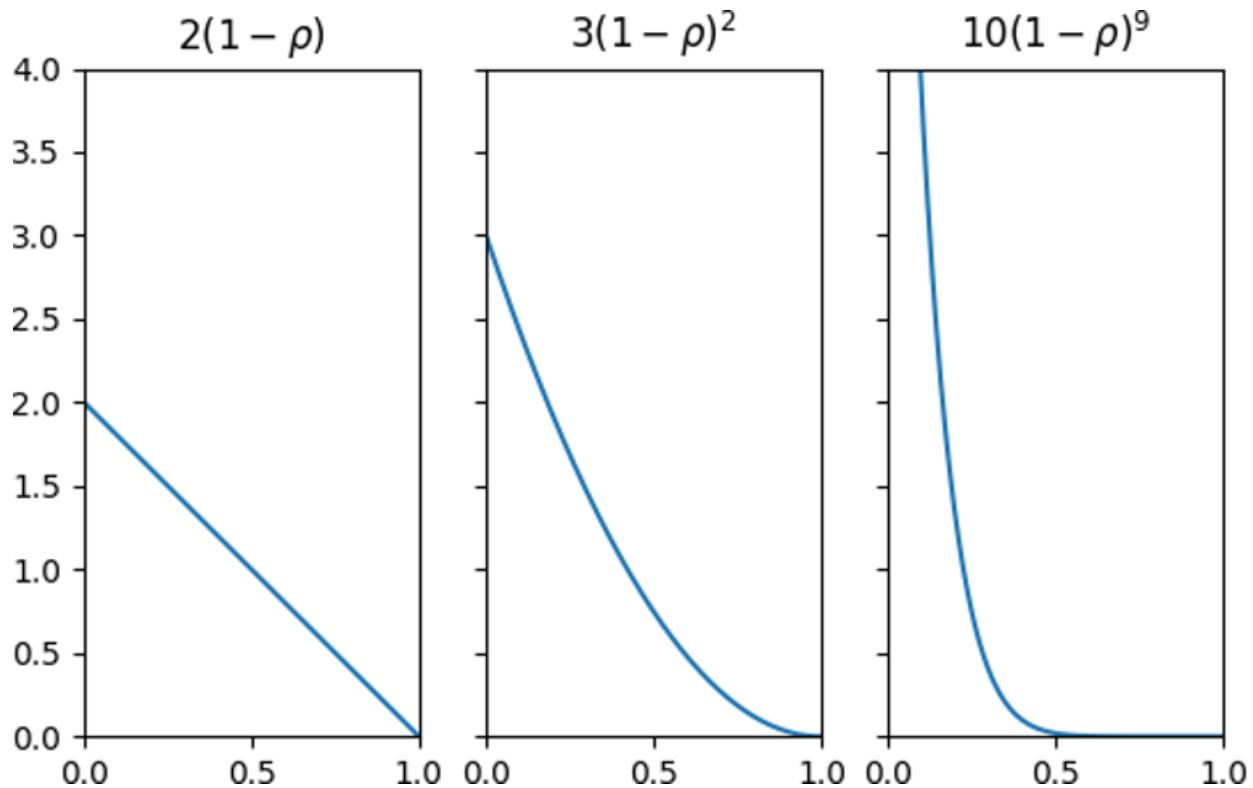
The update rule for beta distributions is that a positive observation (ie life) increases α by 1, and a negative observation (a dead planet) increases β by 1. The mean of an updated beta distribution is a generalised version of [Laplace's law of succession](#): if our prior is a beta distribution with parameters α and β , and we've had m positive observations and n negative ones, then the mean of the posterior is:

$$\frac{\alpha + \beta + m}{\alpha + \beta + m + n}$$

Suppose now that we have observed n dead planets, but no life, and that we haven't done an anthropic update yet, then we have a probability of life of $\alpha/(\alpha + \beta + n)$. Upon adding the anthropic update, this shifts to $(\alpha + 1)/(\alpha + \beta + n + 1)$, meaning that the multiplicative factor is at most $(\alpha + 1)/\alpha$. If we started with the uniform prior with its $\alpha = 1$, this multiplies the probability of life by at most 2. In a later section, we'll look at $\alpha < 1$.

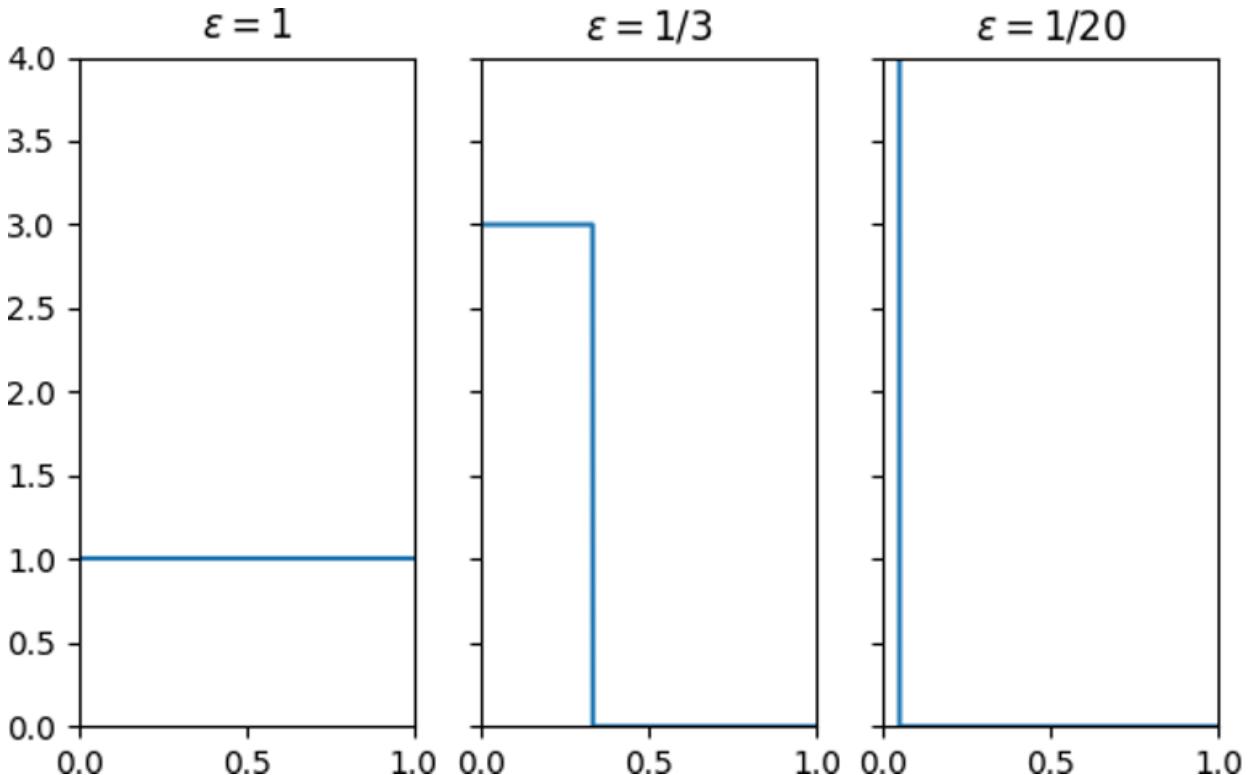
High prior probability is not required for weak anthropic update

The uniform prior has $\alpha = \beta = 1$ and starts at expectation 1/2. But we can set $\alpha = 1$ and a much higher β , which skews the distribution to the left; for example, for $\beta = 2, 3$, and 10:



Even though these priors are skewed to the left, and have lower prior probabilities of life ($1/3$, $1/4$, and $1/11$), the anthropic update has a factor M_{μ,σ^2} that is less than 2.

Also note that if we scale the prior f by a small ϵ , so replace $f(\rho)$ on the range $[0, 1]$ with $f(\rho/\epsilon)/\epsilon$ on the range $[0, \epsilon]$, then μ is multiplied by ϵ and σ^2 is multiplied by ϵ^2 . Thus $M_{\mu,\epsilon}$ is unchanged. Here, for example, is the uniform distribution, scaled down by $\epsilon = 1$, $\epsilon = 1/3$, and $\epsilon = 1/20$:



All of these will have the same M_{μ,σ^2} (which is $4/3$, just as for the uniform distribution). And, of course, doing the same scaling with the various beta distributions we've seen up until now will also keep M_{μ,σ^2} constant.

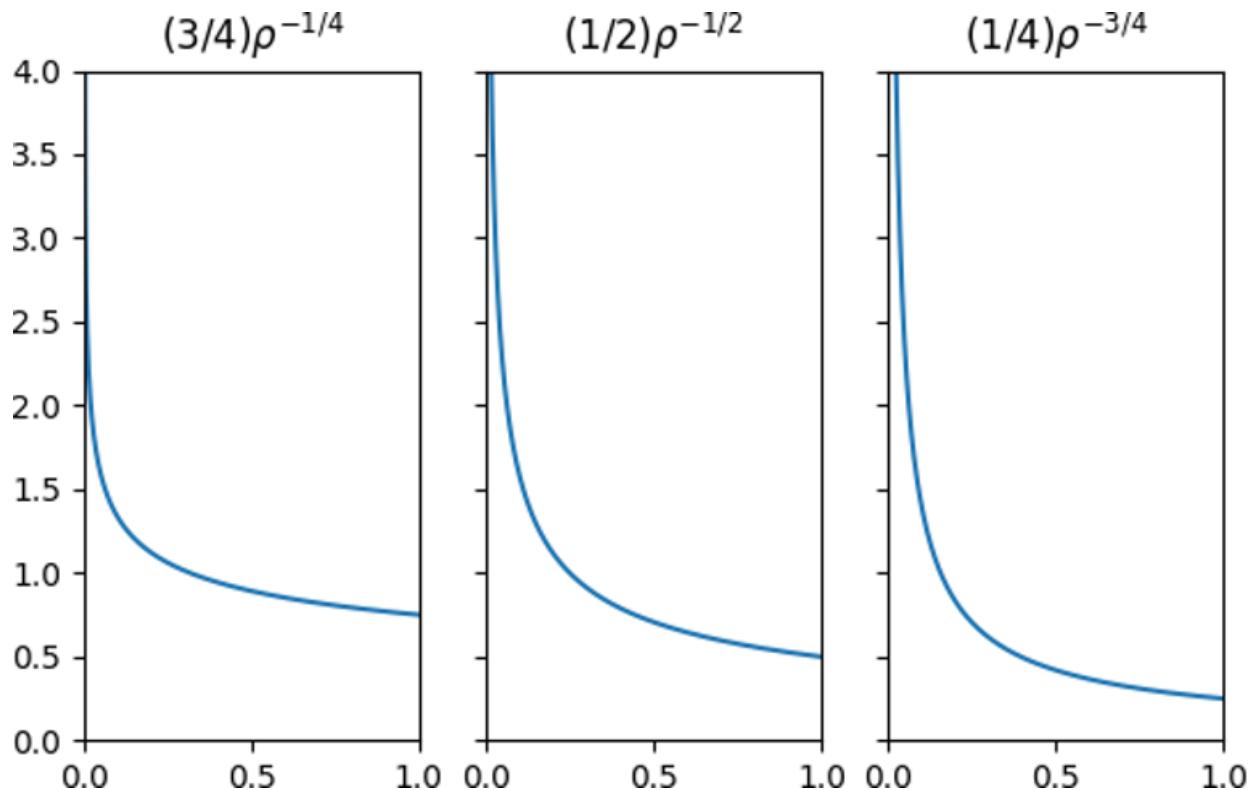
Thus there are a lot of distributions with very low μ (ie very low prior probability of life) but an M_{μ,σ^2} that's less than 2 (ie the anthropic update is less than a doubling of the probability of life).

Beta distributions II and log-normals

The best-case scenario for M_{μ,σ^2} is if f assigns probability 1 to $\rho = \mu$. In that case, $\sigma^2 = 0$ and $M = 1$: the anthropic update changes nothing.

Conversely, the worse-case scenario for M_{μ,σ^2} is if f only allows $\rho = 0$ and $\rho = 1$. In that case, f assigns probability μ to 1 and $1 - \mu$ to 0, for a mean of μ and a variance of $\sigma^2 = \mu - \mu^2$, and a multiplicative factor of $M_{\mu,\sigma^2} = 1/\mu$. In this case, after anthropic update, f' assigns certainty to $\rho = 1$ (since any life at all, given this f , means life on all planets).

But there are also more reasonable priors with large M_{μ,σ^2} . We've already seen some, implicitly, above: the beta distributions with $\alpha < 1$. In that case, M_{μ,σ^2} is bounded by $(\alpha + 1)/\alpha$. If $\alpha = 3/4$ and $\beta = 1$, for instance, this corresponds to the (unbounded) distribution $f(\rho) = (3/4)\rho^{-1/4}$; the multiplicative factor is below 7/3, which is slightly above 2. But as α declines, the multiplicative factor can go up surprisingly fast; at $\alpha = 1/2$ it is 3, at $\alpha = 1/4$ it is 5:



In general, for $\alpha = 1/n$, the multiplicative factor is bounded by $n + 1$. This gets arbitrarily large as $\alpha \rightarrow 0$. Though $\alpha = 0$ itself corresponds to the [improper prior](#) $f(p) = 1/p$, whose integral diverges. On a log scale, this corresponds to the [log-uniform distribution](#), which is roughly what you get if you assume "we need N steps, each of probability p , to get life; let's put a uniform prior over the possible N s".

It's not clear why one might want to choose $\alpha = 1/10^{20}$ for a prior, but there is a class of prior that is much more natural: the [log-normal distributions](#). These are random variables X such that $\log(X)$ is normally distributed.

If we choose $\log(X)$ to have a mean that is highly negative (and a variance that isn't too large), then we can mostly ignore the fact that X takes values above 1, and treat it as a prior distribution for p . The mean and variance of the log-normal distributions can be explicitly defined, thus giving the multiplications factor as:

$$M_{\mu, \sigma^2} = \exp^{-\bar{\sigma}^2}.$$

Here, $\bar{\sigma}^2$ is the variance of the normal distribution $\log(X)$. This $\bar{\sigma}^2$ might be large, as it denotes (roughly) "we need N steps, each of probability p , to get life; let's put a uniform-ish prior over a range of possible N s". Unlike $1/p$, this is a proper prior, and a plausible one; therefore there are plausible priors with very large M_{μ, σ^2} . The log normal is quite likely to appear, as it is the [approximate limit of multiplying together a host of different independent parameters](#).

Multiplication law

Do you know what's more likely to be useful than "the approximate limit of multiplying together a host of different independent parameters"? Actually multiplying together independent parameters.

The famous [Drake equation](#) is:

$$R_* \cdot f_p \cdot n_e \cdot f_l \cdot f_i \cdot f_c \cdot L.$$

Here R^* is the number of stars in our galaxy, f_p the fraction of those with planets, n_e the number of planets that can support life per star that has planets, f_l the fraction of

those that develop life, f_i the fraction of those that develop intelligent life, f_c the fraction of those that release detectable signs of their existence, and L is the length of time those civilizations endure as detectable.

Then the proportion of advanced civilizations per planet is qf_if_i , where q is the proportion of life-supporting planets among all planets. To compute the M of this distribution, we have the highly useful result (the proof is in this footnote^[4]):

- Let X_i be independent random variables with multiplicative factors M_i , and let M be the multiplicative factor of $X = X_1 \cdot X_2 \cdot \dots \cdot X_n$. Then $M = \prod_i M_i$ - the total M is the product of the individual M_i .

The paper "[dissolving the Fermi paradox](#)" gives estimated distributions for all the terms in the Drake equation. The q , which doesn't appear in that paper, is a constant, so has $M_q = 1$. The f_i has a [log-uniform distribution](#) from 0.001 to 1; the M can be computed from the mean and variance of such distributions, so $M_{f_i} = \log(1/0.001) \frac{1-0.001^2}{2(1-0.001)} \approx 3.5$.

The f_i term is more complicated; it is distributed like $g(X) = 1 - e^{-e^{X-50\log(10)}}$ where X is a standard normal distribution. Fortunately, we can estimate its mean and variance without having to figure out its distribution, by numerical integration of $g(x)$ and $g(x^2)$ on the normal distribution. This gives $\mu \approx 0.5$, $\sigma^2 \approx 0.25$ and $M \approx 2$. The overall the multiplicative effect of anthropic update is:

$$M_{\text{planet}} \approx 7.$$

What if we considered the proportion of advanced civilization per star, rather than per planet? Then we can drop the q term and add in f_p and n_e . Those are both estimated to be distributed as log-uniform on $[0.1, 1]$; for a total M of

$$M_{\text{star}} \approx 14.$$

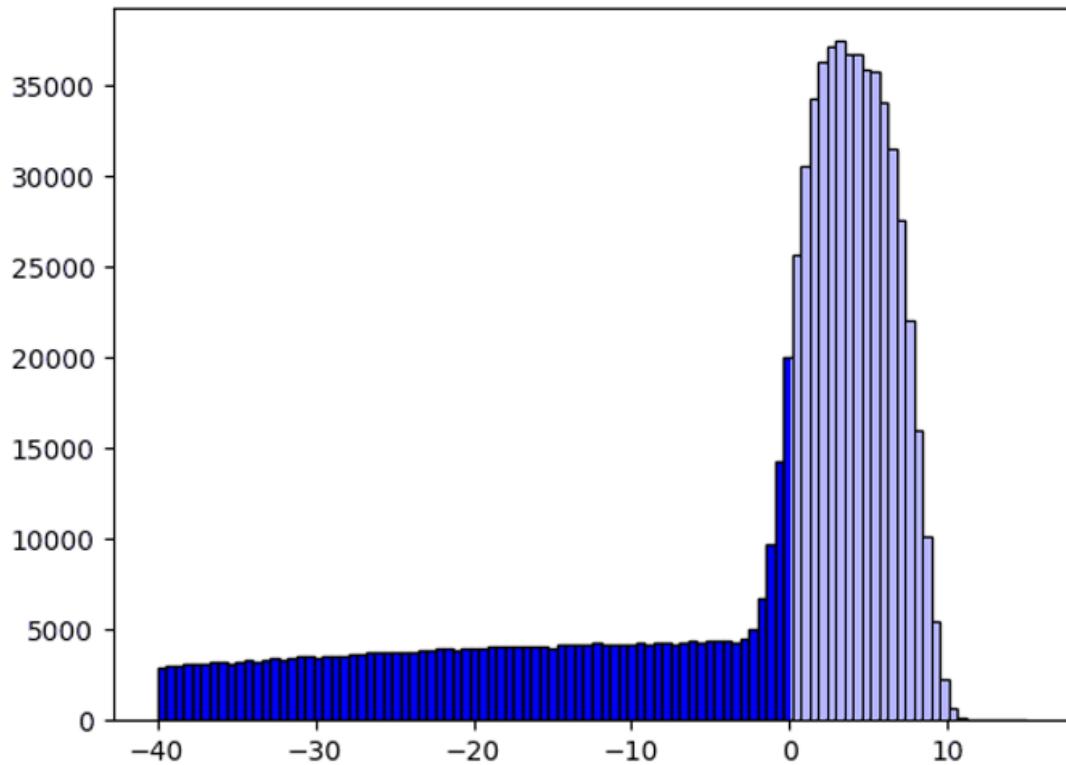
Why is the M higher for civilizations per star than civilizations per planet? That's because when we update on our existence, we increase the proportion of civilizations per planet, but we also update the proportion of planets per star - both of these can make life more likely. The M_{star} incorporates both effects, so is strictly higher than M_{planet} .

We can do the same by considering the number of civilizations per galaxy; then we have to incorporate R_* as well. This is log-uniform on $[1, 100]$, giving:

$$M_{\text{galaxy}} \approx 32.$$

What about if we include the Fermi observation (the fact that we don't see anything in our galaxy)? The "[dissolving the Fermi paradox](#)" paper shows there are multiple different ways of including this update, depending on how we parse out "not seeing anything" and [how easy it is for civilizations](#) to expand.

I did a crude estimate here by taking the Fermi observation to mean "the proportion of civilizations per galaxy must be less than one". Then I did a Monte-Carlo simulation, ignoring all results above 0 on the log scale:



From this, I got an estimated mean of 0.027, variance of 0.014, and a total multiplier of:

$$M_{\text{galaxy, Fermi}} \approx 21.$$

With the Fermi observation and the anthropic update combined, we expect 0.56 civilizations per galaxy.

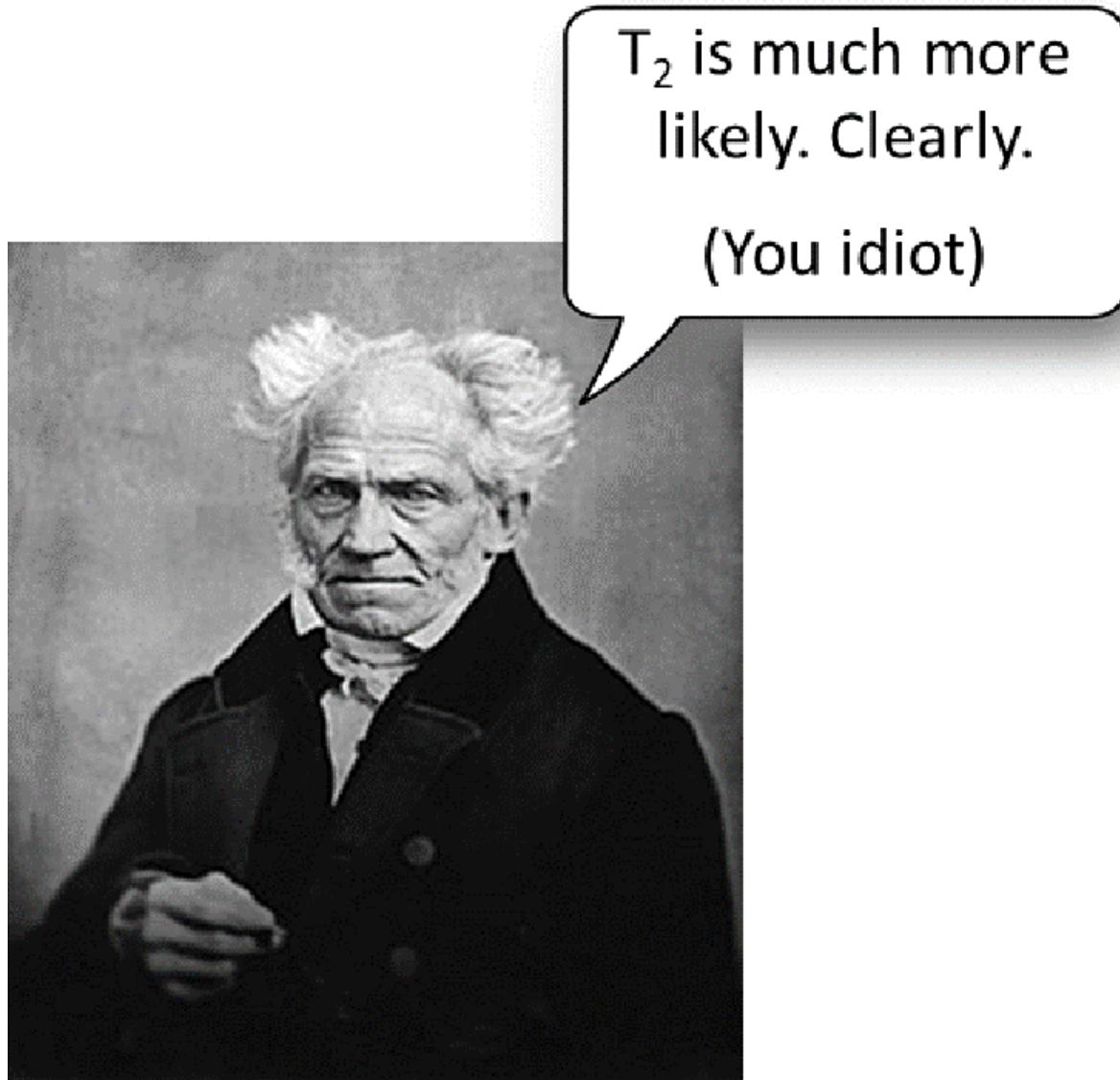
Limitations of the multiplier

Low multiplier, strong effects

It's important to note that the anthropic update can be very strong, without changing the expected population much. So a low M_{μ,σ^2} doesn't necessarily mean a low impact.

Consider for instance the [presumptuous philosopher](#), slightly modified to use planetary population densities. Thus theory T_1 predicts $\rho = 1/10^{12}$ (one in a trillion) and T_2 predicts $\rho = 1$; we put initial probabilities 1/2 on both theories.

As Nick Bostrom noted, the SIA update pushes T_2 to being a trillion times more probable than T_1 ; *a posteriori*, T_2 is roughly a certainty (the actual probability is $10^{12}/(10^{12} + 1)$).



However, the expected population goes from roughly 1/2 (the average of $1/10^{12}$ and 1) to roughly 1 (since *a posteriori* T_2 is almost certain). This gives a M_{μ, σ^2} of roughly 2. So, despite the strong update towards T_2 , the actual population update is small - and, conversely, despite the actual population update being small, we have a strong update towards T_2 .

Combining multiple theories

In the previous post, note that both T_1 and T_2 were point estimates: they posit a constant p . So they have a variance of zero, and hence a M_{μ, σ^2} of 1. But T_2 has a much

stronger anthropic update. Thus we can't use their M_{μ,σ^2} to compare the anthropic effects on different theories.

We also can't relate the individual M_s to that of a combined theory. As we've seen, T_1 and T_2 have M_s of 1, but the combined theory $(1/2)T_1 + (1/2)T_2$ has an M of roughly 2. But we can play around with the relative initial weight of T_1 and T_2 to get other M_s .

If we started with odds $10^{12} : 1$ on T_1 vs T_2 , then this has a mean ρ of roughly 10^{-12} ; the anthropic update sends it to $1 : 1$ odds, with a mean of roughly $1/2$. So this combined theory has an M of roughly $10^{12}/2$, half a trillion.

But, conversely, if we started with odds $1 : 10^{12}$ on T_1 vs T_2 , then we have an initial mean of ρ of roughly one; its anthropic update is odds of $1 : 10^{24}$, also with a mean of roughly one. So this combined theory has an M of roughly 1.

There *is* a weak relation between M and the M_i of the various T_i . Let M_i be the multiplier of T_i has a multiplier of M_i ; we can reorder the T_i so that $M_i \leq M_j$ for $i \leq j$. Let T be a combined theory that assigns probability p_i to T_i .

1. For all $\{p_i\}$, $M \geq \min_i(M_i)$.
2. For all ϵ , there exists $\{p_i\}$ with all $p_i > 0$, so that $M < \min_i(M_i) + \epsilon$.

So, the minimum value of the M_i is a lower bound on M , and we can get arbitrarily close to that bound. See the proof in this footnote^[5].

1. As we'll see, the *population* update is small even in the presumptuous philosopher experiment itself. [←](#)
2. Citation partially needed: I'm ignoring Boltzmann brains and simulations and similar ideas. [←](#)
3. Given a fixed ρ , the probability of observing life on our own planet is exactly ρ . So Bayes's theorem implies that $f'(\rho) \propto \rho f(\rho)$. With the full normalisation, this is

$$f'(\rho) = \frac{1}{\int_0^1 \rho f(\rho) d\rho} \rho f(\rho).$$

If we want to get the mean μ' of this distribution, we further multiply by ρ and integrate:

$$\mu' = E_f(\rho) = \int_0^1 \frac{\rho^2 f(\rho)}{\int_0^1 \rho f(\rho) d\rho} d\rho = \frac{1}{\int_0^1 \rho f(\rho) d\rho}$$

Let's multiply this by $1 = 1/1 = (\int_0^1 f(\rho) d\rho) / (\int_0^1 f(\rho) d\rho)$ and regroup the terms:

$$\mu' = \frac{1}{\int_0^1 \rho f(\rho) d\rho} \cdot \frac{1}{\int_0^1 f(\rho) d\rho}.$$

Thus $\mu' = E_f(\rho^2) / E_f(\rho) = (\sigma^2 + \mu^2) / \mu = \mu(1 + \sigma^2 / \mu^2)$, using the fact that the variance is the expectation of ρ^2 minus the square of the expectation of ρ . ↪

4. I adapted the proof [in this post](#).

So, let X_i be independent random variables with means μ_i and variances σ_i^2 . Let $X = \sum_i X_i$, which has mean μ and variance σ^2 . Due to the independence of the X_i , the [expectations of their products are the product of their expectations](#). Note that X_i and X_j are also independent if $i \neq j$. Then we have:

$$\begin{aligned} \prod_i M_{\mu_i, \sigma_i}^2 &= \prod_i \left(1 + \frac{\sigma_i^2}{\mu_i}\right)^2 \\ &= \prod_i \left(\frac{\mu_i^2 + \sigma_i^2}{\mu_i^2}\right)^2 \\ &= \prod_i \left(\frac{\mu_i^2 + \sigma_i^2}{E(X_i^2)}\right)^2 \\ &= \frac{2}{\prod_i (E(X_i^2))} \\ &= \frac{E(X^2)}{E(X)^2} \\ &= \frac{\mu^2 + \sigma^2}{\mu^2} \\ &= 1 + \frac{\sigma^2}{\mu^2} = M_{\mu, \sigma^2}. \end{aligned}$$

↪

5. Let $\{f_i\}_{1 \leq i \leq n}$ be probability distributions on ρ , with mean μ_i , variance σ_i^2 ,

expectation squared $s_i = E_{f_i}(p^2) = \sigma_i^2 + \mu_i^2$, and $M_i = s_i/\mu_i^2$. Without loss of generality, reorder the f_i so that $M_i \leq M_j$ for $i < j$.

Let f be the probability distribution $f = p_1 f_1 + \dots + p_n f_n$, with associated multiplier M . Without loss of generality, assume $M_i \leq M_j$ for $i < j$. Then we'll show that $M \geq M_1$.

We'll first show this in the special case where $n = 2$ and $M_1 = M_2$, then generalise to the general case, as is appropriate for a generalisation. If

$s_1/\mu_1^2 = M_1 = M_2 = s_2/\mu_2^2$, then, since all terms are non-negative, there exists an α such that $s_1 = \alpha^2 s_2$ while $\mu_1 = \alpha \mu_2$. Then for any given $p = p_1$, the M of f is:

$$M(p) = \frac{ps_1 + (1-p)s_2}{(pp_1 + (1-p)p_2)^2} = \frac{ps_1 + (1-p)\alpha^2 s_1}{(pp_1 + (1-p)\alpha\mu_1^2)^2} = M_1 \frac{1(p) + \alpha^2(1-p)}{(1(p) + \alpha(1-p))^2}$$

The function $x \rightarrow x^2$ is convex, so, interpolating between the values $x = 1$ and $x = \alpha$, we know that for all $0 \leq p \leq 1$, the term $(1(p) + \alpha(1-p))^2$ must be lower than $1^2(p) + \alpha^2(1-p)$. Therefore $(1(p) + \alpha^2(1-p))/(1(p) + \alpha(1-p))^2$ is at most 1, and $M(p) \leq M_1$. This shows the result for $n = 2$ if $M_1 = M_2$.

Now assume that $M_2 > M_1$, so that $s_1/\mu_1^2 < s_2/\mu_2^2$. Then replace s_2 with s_2' , which is

lower than s_2 , so that $s_1/\mu_1^2 = s_2'/\mu_2^2$. If we define $M'(p)$ as the expression for $M(p)$ with s_2' substituting for s_2 , we know that $M'(p) \leq M(p)$, since $s_2' < s_2$. Then the previous result shows that $M'(p) \geq M_1$, thus $M(p) \geq M_1$ too.

To show the result for larger n , we'll induct on n . For $n = 1$ the result is a tautology, $M_1 \leq M_1$, and we've shown the result for $n = 2$. Assume the result is true for $n - 1$, and then notice that $f = p_1 f_1 + \dots + p_n f_n$ can be re-written as

$f = p_1 f_1 + (1 - p_1)f'$, where $f' = (p_2 f_2 + \dots + p_n f_n)$ for $p_i = p_n/(1 - p_1)$. Then, by the

induction hypothesis, if M' is the M of f' , then $M' \geq M_2$. Then applying the result for $n = 2$ between f_1 and f' , gives $M \leq \min(M_1, M')$. However, since $M_1 \leq M_2$ and $M' \geq M_2$, we know that $\min(M_1, M') = M_1$, proving the general result.

To show M can get arbitrarily close to M_1 , simply note that M is continuous in the $\{p_i\}$, define $p_1 = 1 - \epsilon$, $p_i = \epsilon/(n - 1)$ for $i > 1$, and let ϵ tend to 0. [←](#)

Black ravens and red herrings

The [raven paradox](#) is a dilemma in inductive logic posed by Carl Gustav Hempel. It starts by noting that the statement "All ravens are black" can be expressed in the form of an implication: "If something is a raven, then it is black." This statement is logically equivalent to its contrapositive: "If something is not black, then it is not a raven."

We then consider that observing a black raven would typically be considered evidence for the statement "All ravens are black." The paradox comes from asking whether the same holds for the contrapositive; to put it another way, if we observe a non-black non-raven, such as a green apple, does that constitute evidence that all ravens are black?

The standard Bayesian solution, [given by I. J. Good](#), goes as follows. Suppose there are N objects in the universe, of which r of them are ravens, and b of them are black, and we have a probability of $1/N$ of seeing any given object. Let H_i be the hypothesis that there are i non-black ravens, and let us assume we have some sensible prior over our hypotheses. Then upon observing a non-black non-raven, our probability of H_0 increases, albeit only very slightly when N is large.

I. J. Good's solution reveals a distinct way of thinking about evidence among Bayesians. Most logicians view logical fallacies as examples of poor reasoning, but [the Bayesian interpretation](#) is slightly different. Many fallacies, when stated in their appropriate inductive form, are actually valid in a Bayesian sense, though the evidence they provide is usually weak.

Which is to say that if you criticize a Bayesian for using a "red herring" they may not see the issue. Observing a red herring is indeed Bayesian evidence for the statement "All ravens are black."

Imaginary reenactment to heal trauma - how and when does it work?

Some therapies involve various forms of imaginary reenactment, where you heal a trauma by first recalling the memory of it and then imagining how things could have gone differently. Sometimes the imagined alternative can be quite fantastical in nature, such as your current adult self traveling back in time to when you were a child and saving your child self from the bullies tormenting you. (Here by trauma I mean to also talk about “small-t trauma”, e.g. various painful experiences that might not be what we’d ordinarily call trauma, but are still a little unpleasant to think about, or have left some other kind of a negative effect on your psyche.)

In my experience, imaginary reenactment works, at least assuming that I’ve managed to get an emotional hold of what exactly in the memory it is that made it feel so unpleasant. (Did I feel like I was alone? Or inadequate? Or that I did something wrong? Etc.) Also assuming that the memory of the old trauma isn’t so painful as to be completely overwhelming and leave no room to imagine any alternatives.

Here’s my current guess of how and when this works:

The basic process by which any emotional learning gets changed is memory reconsolidation. There’s a generalization that your mind has drawn about the meaning of some past event that feels true to you. E.g. “nobody helped me when I was in that situation, so nobody cares about my suffering”. If you can bring that felt truth to mind while also experiencing a contradictory belief – e.g. the belief that you have a friend who *does* care about you – as true at the same time, your brain will notice that it believes in two contradictory things at the same time, and will revise its beliefs to fix that inconsistency.

Often, this takes the form of concluding that what it considered to be a general truth isn’t the case after all – e.g. changing the previous assessment to “nobody helped me in that situation, but there are still people who care about me and who I can reach out to for help”.

Now, you can also imagine things that feel true, if they’re the kinds of things you feel could happen. For instance, maybe you have a friend who buys a lot of products from the Acme Corporation, and you then imagine your friend excitedly telling you about the Acme Super-Duper Toothbrush that they bought. Even if they have never done this, the imagined scene can still feel real because it involves the kind of a thing that your friend *could* do.

I suspect what’s going on in therapeutic reenactment is that you are imagining something that feels like it *could* have happened and thus serves as counterevidence for the emotional belief in your trauma, but the “could be true” is on an emotional or symbolic level rather than on the level of physical possibility.

So for example, suppose that I had a childhood experience where I was being picked on by bullies and nobody helped me. From this experience, my brain might form the generalization “nobody helped me, so nobody cares about my suffering”.

Now if I manage to recall this experience in such a way that I can feel empathy towards my past self, then the act of feeling that empathy now proves that someone

does care. Then if I imagine a scene in which I travel back in time to when I was a child and I beat up my bullies, it doesn't matter if the literal content is physically impossible. Because what matters is the emotional feeling of "someone cares so someone could have helped", which is evaluated as true.

(My adult self caring about my child self doesn't mean that my adult self could actually have helped my child self, but one person caring is enough to disprove the generalization of "nobody cares". So then if at least one person cares, then that implies that there were also others who would have cared, and *they* would have helped if they'd known and had the opportunity to. More broadly, the reason why we draw generalizations from past experiences is to predict the future, so what really matters is knowing that it's possible to get help from people in general, and that people don't think that your suffering is intrinsically meaningless.)

That said, if I try to do this and I haven't really gotten a good intuition of *why* the memory is so painful, it usually doesn't work - the generalization that I have formed from the experience needs to be at least somewhat explicit. Otherwise I can't experience that generalization as real (as is required for the memory reconsolidation process to work), nor can I find the right emotional flavor that I need to imagine for the new scene to count as counter-evidence for the generalization.

Risk Premiums vs Prediction Markets

aka **Why (some) market forecasts are biased and what can we do about them?**

tldr; When you care about an outcome, it will skew the odds you'll bet at. (To insure against bad outcomes). When everyone cares about an outcome, it skews the market odds.

People (including [myself](#)) often point at market forecasts as being the best predictors for some variable (inflation, the economy, whether England will win the Euros).

However, market prices are more than just expectations. [In the words of an anonymous twitter account:](#)

$$\text{Market price} = \text{Expectations} + \text{Risk Premium} + \text{Idiosyncratic Issues}$$

This is a somewhat tautological framing however (since we don't know what risk premium or the idiosyncratic issues are other than the things which make this equality true). I'm going to try and explain what risk premium is, since people often focus too much on expectations and use idiosyncratic issues as an excuse (fees are too high on PredictIt, you can't trade large size, etc..). Risk premium is important because unlike idiosyncratic issues which can be fixed with good market structure, risk premium will not go away.

This matters, because **if we can't remove risk premiums and the idiosyncratic issues, then prediction markets are less accurate and less valuable for decision making.**

To this end, I'm going to start with some definitions:

What are risk premiums? Risk premium is *excess returns* which results from the *payout* being *correlated* with the rest of the world.

What are excess returns? Excess returns are returns higher than 0 (or the risk-free rate depending on context) after accounting for the riskiness of the payout.

Some examples:

- a 1-1 bet on a 50/50 coin flip doesn't have excess returns
- a 2-1 bet on a 50/50 coin flip has excess returns
- a 10% 1y bond on a risky company may or may not have excess returns, depending on the default rate of the company and the recovery rate in event of default
- the stock market has excess returns (probably)

What is correlation with the rest of the world?

When an event happens, that says something about the world state we're in as well as the payout we get from the market. If the world state affects our utility in other ways than just the market, that is the correlation we're talking about.

Some examples:

- If I bet on a coin flip, aside from the money I have bet on the event, it has no effect on my well being.
- If I bet on England to win the Euros, then I have some utility associated with the outcome *unrelated to my bet*. I will be happier if England win regardless of whether or not I bet.
- If I insure my car against crashes then I have some utility associated with the outcome *unrelated to my bet*. I will be very sad if I crash my car.

- If the economy starts doing badly, financial assets will start suffering, I will be suffering and so will everyone else.

When does correlation lead to a risk premium?

(This is me writing out with examples what Cochrane explains more clearly [here](#) in Section 1.4. Particularly "Risk Corrections" and "Idiosyncratic Risk")

What matters is the size of the market, the ability to diversify and the participants involved.

Let's start with the coin flip. If I am the only person in the world making a market in a certain coin flip, I might make the market .48 / .52 for \$1 on heads. (That is if you want to buy heads, you need to pay 52c and I will buy heads from you at 48c). [This is because $0.5 * U(\text{wealth} - .48) + 0.5 * U(\text{wealth} + .52) > U(\text{wealth})$ (for my current wealth and utility)] However, someone with much more money than me would look at my market and be willing to do the same trade for .49 / .51 because they are still going to earn some return. Someone with much more money than them will look at that market and offer .495 / .505 and eventually we end up with a very tight market around .50 the tightness effectively depending on how wealthy (and risk averse) the participants are.

Let's now continue with the England example. Let's say we both agree the chances of England winning are 50%, my current worth is \$10, I value England winning at \$10 and England losing as -\$8 and I have log utility.

My current utility is $0.5 \cdot U(\$10 + \$10) + 0.5 \cdot U(\$10 - \$8) = 0.5 \cdot \log(20 * 2)$

If we bet at 55/45 (me getting the raw end of the deal) then my expected utility is $0.5 * 0.5 \cdot U(\$10 + \$10 - \$5.5) + 0.5 \cdot U(\$10 - \$8 + \$4.5) = 0.5 \cdot \log(19.45 \cdot 2.45) = 0.5 \cdot \log(46)$.

Therefore the market (between us) could end up clearing at the (incorrect) price of 55%. (Actually in this extreme example, I would be happy betting at up to 91%(!)). If everyone felt like me, the market might clear far from the probability everyone agrees on.

However this market has many participants and they don't all have the same (extreme!) preferences. Therefore any difference between the clearing price here and the "fair" clearing price slowly gets arbitraged away until we have a good prediction market.

What about the market for car insurance? Well, it's similar to the coin flip. I have some risk aversion associated with the cost of having to replace the cost of a car. On top of that I have the additional change in my utility which comes from having been in a car accident. So I am willing to pay more than the my estimated probability. However, there is someone else on the other side of this. When I go to renew my insurance there are many companies offering me car insurance. They are in competition, each with a large float and indifferent to whether or not I crash my car. Their only interest is their expected return. This competition drives the price closer to the true probability. (More on this later...)

What about equities? Well it's similar to the coin flip. Uncertain money is worth less than certain money to me. It's also similar to the insurance example. More money in a good economy is less good than less money in a bad economy is bad for me. However, here we need to start thinking about the other side. Who is the other side of the equity market? Well, there are certainly people with much deeper pockets than me, so "uncertain money" doesn't drive them as much so the uncertainty value gets priced out. However, there are very few people out there who have pro-cyclical utility - (people who want more money in good times than they want less money in bad times). So the market clearing price (as a whole) will be less than the "fair" price. (This is one "solution" to the [Equity Premium Puzzle](#)).

What does this mean for prediction markets?

Well for simple prediction markets: coin flip; sports events; weather(?); there are sufficiently many people who are either indifferent enough, or deep-pocketed enough that they will take on the risk for an ever decreasing edge. This means the market price should tend to the market participants' best estimate of the fair probability.

For complicated markets: insurance, equities, inflation-protected bonds, it's not quite so simple to infer a probability/forecast/expectations from market prices.

Let's start with **insurance**. We know that the probabilities which we infer from the insurance market are (broadly) an overestimate. (Insurance companies are profitable after all!) This would be fine if we knew how much of an overestimate they were. "Take the market prices... subtract off the profit margin from the insurers... done". Unfortunately, the insurance market is slightly more dynamic than that. (See the time varying loss ratios from [this fascinating set of slides](#))

Net Loss Ratio Time Series by Major Line

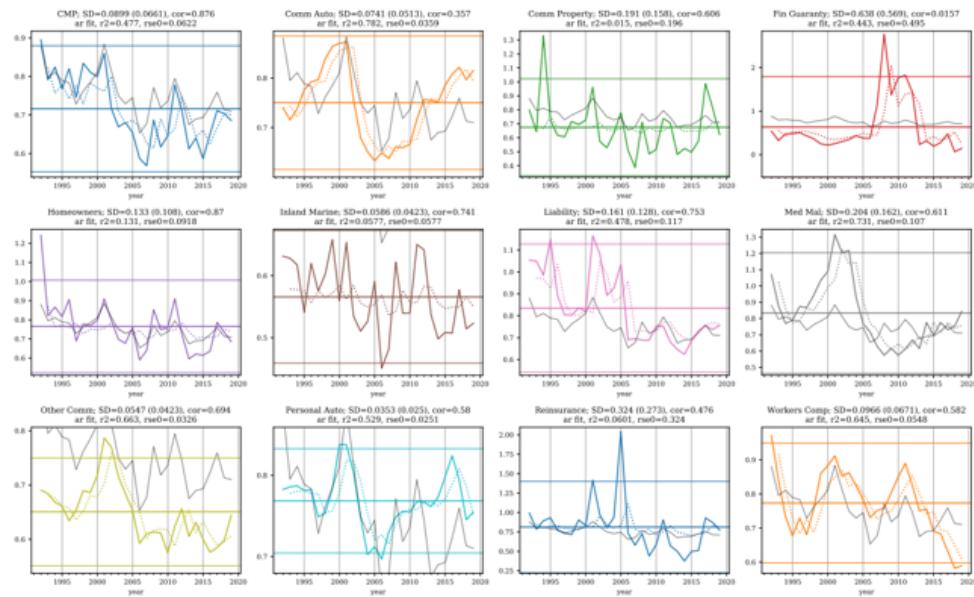


Figure 18: Net calendar year Loss ratio time series by major line. See gloss on next slide.

Over time the amount of risk premium being demanded by the insurance companies waxes and wanes. Therefore we need to know the right adjustment to make for risk premium. (And arguably this is just as hard as predicting prices in the first place - no free lunch here).

What about **equities**? If we know that equities broadly go up, can't we just subtract off that positive rate of return to use them as indicators for the real economy? Again, no. The problem is that the risk premium is time varying and knowing what the current risk premium is, is about as hard as valuing equities in the first place. No free lunch here.

What about **TIPS**? TIPS are the classic example of using the financial markets as a forecasting tool. They are inflation-protected bonds issued by the US government. In theory, the difference between TIPS and nominal bonds will be the market's expectation for inflation. Certainly the difference* will realise to inflation. The question is, will the market clear at that price? The empirical evidence seems to be "not quite" **.

What about **GDP-linked bonds**? Well, I can't think of anything more correlated to the real economy, so I would expect the risk premium of these (for global economies) to be very large (relative to their volatility). This would compare to nominal bonds which have a

negative risk premium. This means for the country issuing them, they would be much more expensive than traditional debt issuance.

Their use as a forecasting tool would also be weakened, since they would systematically under-forecast GDP growth.

That said - I am still a fan of the idea. I would rather implement them as futures contracts (pretty much what Sumner advocates for [here](#)).

How bad can it get?

We can take some hope (or despair) from the [Hansen-Jagannathan bound](#). This is usually used to say something about how volatile the state price density**** is. However, this volatility gives us an upper-bound for the Sharpe of any asset. (Including any asset we might use for prediction). If we are willing to assume*** that empirically the upper bound for the state price density is 0.5 then we know that $E[\text{estimate}] \leq 0.5$. So there is at least hope that market prices will get us somewhere within an order of magnitude of the right answer.

* modulo some details regarding options

** there are a bunch of different factors driving this difference, some premium related some market inefficiency related, some to do with large flows in the market

*** this is a key assumption, but I'm quite willing to believe it, since no-one has found assets with much higher Sharpe than 0.5

**** For more information on the state price density and why this is relevant, the chapter of Cochrane [linked at the start](#) should be a good first port of call.

Typology of blog posts that don't always add anything clear and insightful

I used to think a good blog post should basically be a description of a novel insight.

To break it down more, on this view:

1. A blog post should have a **propositional claim** (e.g. ‘the biggest externalities are from noise pollution’, or ‘noise pollution is a concept’ vs. expression of someone’s feelings produced by externalities, or a series of reflections on externalities). A ‘propositional claim’ here can be described straightforwardly in words, and usually conveys information (i.e. they say the world is one way instead of another way).
2. It should be a **general** claim—i.e. applicable to many times and places and counterfactuals (e.g. ‘here is how tragedies of the commons work: ...’ vs. ‘here is a thing that happened to me yesterday: ...’)
3. It should be a **novel** claim (e.g. a new reason to doubt one of the explanations put forward for the demographic transition)
4. The claim should be **described**, which is to imply that the content should be:
 1. *Verbal* (or otherwise symbolic, e.g. a table of numbers surrounded by text would count)
 2. *Explicit* (saying the things it means, rather than alluding to them)
 3. Mostly *concerned with conveying the relevant propositions* (vs. for instance mostly concerned with affecting the reader’s mood or beliefs directly)

I probably would have agreed that the odd vignette was also a good blog post, but ideally it should be contained in some explicit discussion of what was to be learned from it. I probably wouldn’t have held my more recent [Worldly Positions](#) blog¹ in high esteem.

I now think that departures from all of these things are often good. So in the spirit of novel descriptions of explicit and general claims, I have made a typology of different combinations of these axes.

Before getting to it, I’ll explain some part of the value of each category that I think I overlooked, for anyone similar to my twenty year old self.

Worthy non-propositional-claim content

Minds have many characteristics other than propositional beliefs. For instance, they can have feelings and attitudes and intuitions and grokkings and senses. They can meditate and chop onions quickly and look on the bright side and tend to think in terms of systems. They can also have different versions of ‘beliefs’ that don’t necessarily correspond to differences in what propositions they would assent to. For instance, they can say ‘it’s good to exercise’, or they can viscerally anticipate a better future when they choose to exercise. And even among straightforward beliefs held by minds, there are many that aren’t easily expressed in words. For instance, I have an impression of what summer evenings in the garden of a lively country restaurant were like, but to convey that sense to you is an art, and probably involves saying different propositional

things in the hope that your mind will fill in the same whatever-else in the gaps. So this belief doesn't seem to live in my mind in a simple propositional form, nor easily make its way into one.

All of this suggests that the set of things that you might want to communicate to a mind is large and contains much that is not naturally propositional.²

Minds can also take many inputs other than propositional claims. For instance, instructions and reminders and stories and music and suggestions implicit in propositional claims and body language and images. So if you want to make available a different way of being to a mind—for instance you want it to find salient the instability of the global system—then it's not obvious that propositional claims are the best way.

Given that minds can take many non-propositional inputs, and take many non-propositional states, you should just expect that there are a lot of things to be said that aren't naturally propositional, in form or content. You should expect messages where the payload is intended to influence a mind's non-propositional states, and ones where the mode of communication is not propositional.

...in communicating propositional claims

There are different versions of ‘understanding’ a proposition. I like to distinguish ‘knowing’ or ‘understanding’ a thing — which is to say, seeing it fit into your abstract model of the world, being inclined to assent to it — and ‘realizing’ it — intuitively experiencing its truth in the world that you live in. Joe Carlsmith [explores](#) this distinction at more length, and gives an example I like:

If asked, one would agree that the people one sees on a day to day basis — on the subway, at parties, at work — all have richly detailed and complex inner lives, struggles, histories, perspectives; but this fact isn't always present and vivid in one's lived world; and when it becomes so, it can make an important difference to one's ethical orientation, even if the propositions one assents to have not obviously changed.

I repeatedly have the experience of ‘already knowing’ some obvious thing that people always say for ages before ‘realizing’ it. For instance, ‘the map is not the territory’. (“Of course the map isn't the territory. Why would it be? That would be some stupid mistake, thinking that the map was the territory. Like, what would your model of the situation even be like? That the place you live is also your own mind?”) Then at some point it actually hits me that stuff that seems to be in the world ISN'T IN THE WORLD; WHAT SEEMS LIKE THE WORLD IS MY OWN MIND'S IMAGE OF THE WORLD. For instance, long after seeming to know that ‘the map isn't the territory’ I was astonished to realize that those things that are just boring in their basic essence, like sports statistics and home care magazines, things that seem to be fundamentally drab, are not like that at all. They gleam with just as much allure as the things I am most compelled by, from many vantage points out there—just not mine. And in such a case I say to myself, ‘Oh wow, I just realized something...huh, I guess it is that the map is not the territory...but I knew that?’ Probably reading this, you are still thinking, ‘um yes, you weren't aware that boringness is person-dependent?’ And I was aware of that. I ‘knew’ it. And I even knew it in some intuitively available ways—for instance, just because I find *Married at First Sight* interesting, I did not expect my boyfriend to find it so. In particular, in approaching my boyfriend with the news that I have been watching a bunch of *Married at First Sight*, I viscerally did not expect ‘boyfriend sympathizes with appeal of objectively excellent show’ type observations (in fact he liked it, and I was in fact

surprised). But still the boringness of other subjects is depicted to me as part of them, like being red is depicted as in the world (whereas ‘liable to reduce my hunger’ say, is I think more accurately represented by my mind as a feature of myself). And ‘realizing’ that that isn’t right changes how the world that I spend my concrete days in seems.

(I know I have hardly explained or defended this claim that ‘realizing’ is a thing, and important, but I’m not going to do that properly here.)

All of these ‘realizations’ seem to be non-propositional. You already had some proposition, and then you get something else. I think of ‘realizing’ a proposition as acquiring a related non-proposition. To realize the proposition ‘other people have inner lives’ is to take in some non-proposition. Perhaps a spacious sense of those other minds being right there around you. If you are communicating a proposition, to have it actually realized, you want to get its non-proposition partner into the recipient’s mind also. This isn’t really right, because each proposition probably has a multitude of intuitive realizations of it, and each intuitive sense of the world could be part of appreciating a multitude of different propositions. But at any rate, communicating a proposition well, so that the other person can really make use of it, often seems to involve conveying a lot of its non-propositional brethren.

Worthy non-descriptive communication

Closely related to non-propositional content is non-descriptive communication, which I shall call ‘evocative’ communication.

I’m thinking of a few different axes as being related to descriptiveness of communication:

- Verbalness (consisting of words, e.g. “donkeys are nice” vs. a video of a nice donkey)
- Explicitness (saying in words the thing you mean, rather than demonstrating it or suggesting it or subtly causing it to creep into the background of the picture you are painting without naming it. E.g. “I want us to follow this protocol” vs. “Most reasonable people are following this protocol now”)
- Neutrality (not setting out to affect the readers’ emotions except via content itself)

I think of the most vanilla communication as being explicit, verbal and neutral. And this seems pretty good for conveying propositional content. But I suspect that non-propositional content is often conveyed better through evocative communication.

(Or perhaps it is more like: communicating propositional claims explicitly with language is uniquely easy, because explicit language is basically a system we set up for communicating, and propositions are a kind of message that is uniquely well suited to it. But once we leave the set of things that are well communicated in this way, and given that there are lots of other ways to communicate things, non-descriptive forms of communication are much more likely to be helpful than they were.)

Relatedly, I think non-descriptive communication can be helpful in making the ‘realizing’ versions of propositional claims available to minds. That is, in really showing them to us. So in that way, evocative communication seems also potentially valuable for communicating propositional content well.

Worthy communication of non-propositional things descriptively

Going the opposite way—trying to communicate ineffable things in words—also seems valuable, because a) groping nearby propositionally does contribute to understanding, and b) only understanding things in ineffable ways leaves them unavailable to our reasoning faculties in important ways.

Worthy non-generality

I thought that if things were not general, then they were particularly unimportant to talk about. All things equal, isn't it way better to understand a broad class of things better than a single thing?

Some ways this is misleading:

- Understanding specific things is often basically a prerequisite for understanding general things. For instance, devising a general theory of circumstances under which arms races develop will be harder without specific information about the behavior of specific nations historically, to inspire or constrain your theorizing
- Understanding specific things one after another will often automatically lead to your having an intuitive general model, through some kind of brain magic, even in cases where you would have had a hard time making an explicit model. For instance, after you have seen a thousand small disputes run their course, you might have a pretty good guess about how the current dispute will go, even if you couldn't begin to describe a theory of argumentation for the relevant community.
- Specific things are often broadly relevant to the specific world that you live in. For instance, exactly what happened in a particular past war might determine what current obligations should be and what sentiments are warranted, and who is owed, and what particular current parties might be expected to want or take for granted. Which is perhaps only of much interest in a narrow range of circumstances, but if they are the circumstances in which we will live for decades, it might be consistently material.

Worthy non-originality of content

On my naive model, you don't want to repeat something that someone else said, because there is implicitly no value in the repetition—the thing has already been said, so re-saying adds nothing and seems to imply that you are either ignorant or hoping to dupe ignorant others into giving you undeserved credit.

But on a model where many claims are easy enough to accept, but hard to realize, things look very different. The first time someone writes down an idea, the chances of it really getting through to anyone with much of its full power are low. The typical reader needs to meet the idea repeatedly, from different angles, to start to realize it.

In a world like that, a lot of value comes from rehashing older ideas. Also in that world, rehashing isn't the easy cashing in of someone else's work. Writing something in a way that might really reach some people who haven't yet been reached is its own art.

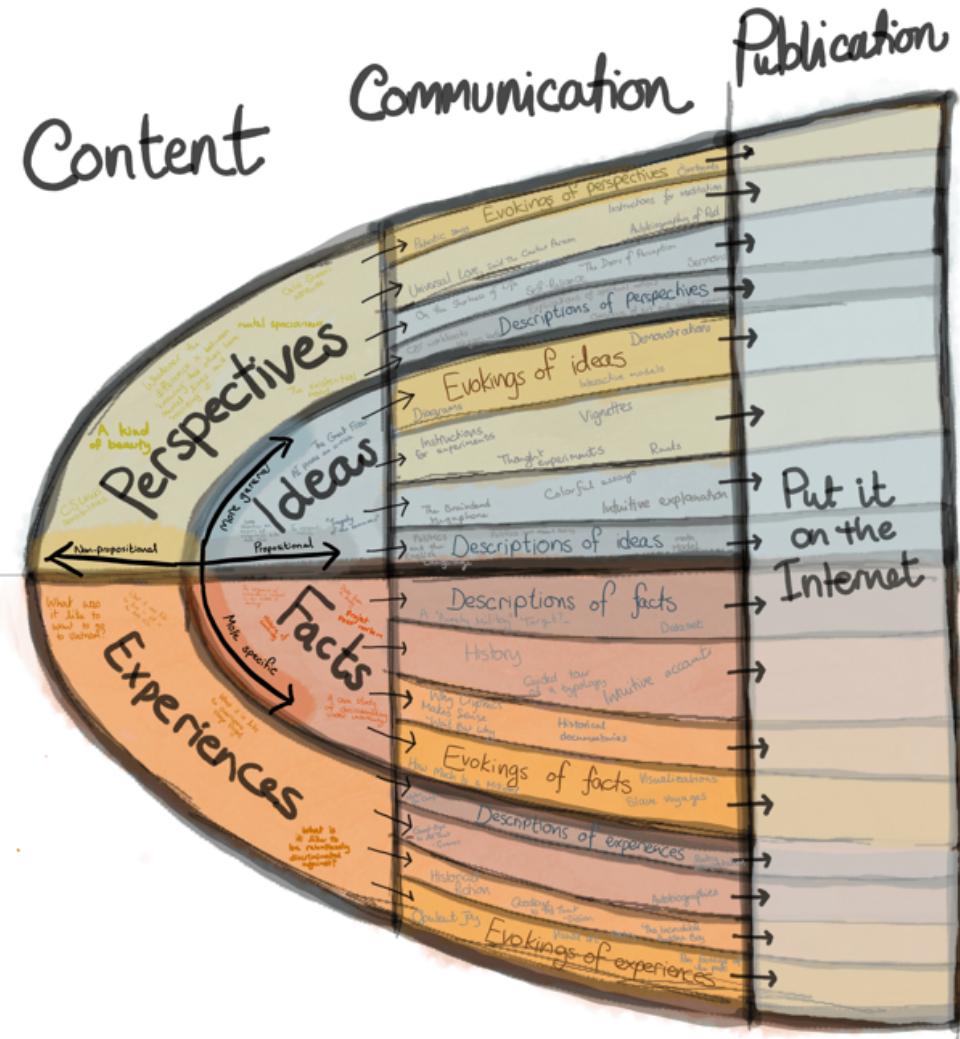
Worthy non-originality of communication

I think I also kind of imagined that once an idea had been put into the ‘public arena’ then the job was done. But another way in which unoriginality is incredibly valuable is that each person can only see such a minuscule fraction of what has ever been written or created, and they can’t even see what they can’t see, that locating particularly apt bits and sharing them with the right audience can be as valuable as writing the thing in the first place. This is curating and signal boosting. For these, you don’t even need to write anything original. But again, doing them well is not trivial. Knowing which of the cornucopia of content should be shown to someone is a hard intellectual task.

Typology

Here is my tentative four-dimensional typology of kinds of blog posts. Any blog post maps to a path from some kind of content on the left, through some kind of communication to publication on the right. Content varies on two axes: generality and propositionalness. Communication varies in evocativeness. And blog posts themselves vary in how early in this pipeline the author adds value. For instance, among posts with a general propositional idea as their content, communicated in a non-propositional way, there are ones where the author came up with the idea, ones where the author took someone else’s idea and wrote something evocative about it, and ones that are repostings of either of the above. Thus, somewhat confusingly, there are 16 (pathways from left to right) x 3 (steps per pathway) = 46 total blog post types represented here, not the 36 you might expect from the number of squares.

I include a random assortment of examples, some obscure, zooming probably required (apologies).



Main updates

1. Lots of worthy things are hard to describe in words
2. 'Realizing' is a thing, and valuable, and different to understanding
3. Details can be good
4. Having ideas is not obviously the main place one can add value

Takeaways

1. It's good to write all manner of different kinds of blog posts
2. It's good to just take other people's ideas and write blog posts about them, especially of different kinds than the original blog posts

3. It's good to just take one's own ideas and write second or third blog posts saying exactly the same thing in different ways

Other thoughts

These different sorts of blog posts aren't always valuable, of course. They have to be done well. Compellingly writing about something that isn't worthy of people's attention, or curating the wrong things can be as bad as the good versions of these things are good.

Epistemic status: overall I expect to find that this post is badly wrong in at least one way in short order, but to be sufficiently interested in other things that I don't get around to fixing it. Another good thing about rehashing others ideas is that you can make subtle edits where they are wrong.

Notes

1. Older posts [here](#) ↵
2. I don't want to make strong claims about exactly what counts as propositional— maybe these things are propositional in some complicated way—but hopefully I'm pointing at an axis of straightforward propositionalness versus something else, regardless. ↵

A (somewhat beta) site for embedding betting odds in your writing

This is a linkpost for <https://biatob.com/>

I've made [a site for publishing predictions and publicly offering to bet on them](#). LessWrong seems like a place that might be interested in such a tool, and I expect at least 4 people will be interested enough in this site to play around with it a little

[bet: \$20 at 50%+ (resolved: correct)]

. It's a little rough around the edges, but I'd be delighted if my target audience (you) found it intriguing enough to try it / offer feedback!

What's it do, more specifically? Well, there are two basic flows:

- *Offer bets*: Alice has some belief that other people might disagree with! She operationalizes it into some statement like "by Sept 1, I will have 2k karma on LessWrong"; she figures she's 60% sure, and willing to lose up to \$50. She creates a prediction on Biatob with those parameters, then publicizes that prediction (e.g. by pasting a little HTML snippet into a LW post).
- *Take bets*: Bob reads Alice's LW post and sees the little "[bet: \$50 at 60%+]" link. He thinks her goal is unrealistically high, but he respects Alice and trusts her to pay up if she's wrong. Bob clicks the link, signs up, and bets \$20 against her \$30.

Then, on Sept 1, Alice gets an email reminding her to check whether the prediction came true or not. When she does, Biatob tells her how much money [she/Bob] owes [Bob/her], and they settle up offline. (Notice that this is all honor-system: you can only bet against people you trust, but at that cost, you don't have to trust my janky web site with your money.)

For example: I signed up, created a couple of predictions (like the above one about a few LWers signing up), and copied some HTML into this post. Now I can say things like "I doubt anybody will even be interested enough to make an easy \$10 by betting against me.

[bet: \$10 at 50%+ (resolved: incorrect)]

" (That prediction is a bit tongue-in-cheek, but I'm completely serious that I'll pay up if you go to the trouble of signing up to bet against me. I'm hoping to get some user feedback from you.)

[AN #156]: The scaling hypothesis: a plan for building AGI

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

Alignment Newsletter is a weekly publication with recent content relevant to AI alignment around the world. Find all Alignment Newsletter [resources here](#). In particular, you can look through [this spreadsheet](#) of all summaries that have ever been in the newsletter.

Audio version [here](#) (may not be up yet).

Please note that while I work at DeepMind, this newsletter represents my personal views and not those of my employer.

HIGHLIGHTS

[The Scaling Hypothesis](#) (Gwern Branwen) (summarized by Rohin): This post centers around the **scaling hypothesis**:

Once we find a scalable architecture which can be applied fairly uniformly, we can simply train ever larger networks and ever more sophisticated behavior will emerge naturally as the easiest way to optimize for all the tasks and data. More powerful NNs are “just” scaled-up weak NNs, in much the same way that human brains look much like scaled-up primate brains.

Importantly, we can get this sophisticated behavior just by training on simple objectives, such as “predict the next word”, as long as the data is sufficiently diverse. So, a priori, why might we expect the scaling hypothesis to be true?

The core reason is that optimal (or human-level) prediction of text really does require knowledge, reasoning, causality, etc. If you don’t know how to perform addition, you are probably not going to be able to predict the next word in the sentence “Though he started with six eggs, he found another fourteen, bringing his total to ____”. However, since any specific fact is only useful in a tiny, tiny number of cases, it only reduces the expected loss by a tiny amount. So, you’ll only see models learn this sort of behavior once they have exhausted all the other “easy wins” for predicting text; this will only happen when the models and dataset are huge.

Consider a model tasked with predicting characters in text with a set of 64 characters (52 uppercase and lowercase letters, along with some punctuation). Initially it outputs random characters, assigning a probability of 1/64 to the correct character, resulting in a loss of 6 bits. Once you start training, the easiest win is to simply notice how frequent each character is; just noticing that uppercase letters are rare, spaces are common, vowels are common, etc. could get your error down to 4-5 bits. After this, it might start to learn what words actually exist; this might take $10^5 - 10^6$ samples since each word is relatively rare and there are thousands of words to learn, but this is a drop in the bucket given our huge dataset. After this step, it may have also learned punctuation along the way, and might now be down to 3-4 bits. At this point, if you

sample from the model, you might get correctly spelled English words, but they won't make any sense.

With further training the model now has to pick up on associations between adjacent words to make progress. Now it needs to look at things 10 characters ago to predict the next character -- a far cry from our initial letter frequencies where it didn't even need to look at other characters! For example, it might learn that "George W" tends to be followed by "ashington". It starts to learn grammar, being able to correctly put verbs in relation to subjects and objects (that are themselves nouns). It starts to notice patterns in how words like "before" and "after" are used; these can then be used to better predict words in the future; at this point it's clear that the model is starting to learn semantics. Now the loss is around 2 bits per character. A little more training and your model starts to produce sentences that sound human-like in isolation, but don't fit together: a model might start a story about a very much alive protagonist and then talk about how she is dead in the next sentence. Training is now about fixing errors like these and each such fix gains a tiny amount of accuracy -- think ten thousandths of a bit. Every further 0.1 bits you gain represents the model learning a huge amount of relevant knowledge (and correspondingly each subsequent 0.1 bits takes a much larger amount of training and data). The final few fractions of a bit are the most important and comprise most of what we call "intelligence".

(The human baseline is a loss of 0.7 bits, with lots of uncertainty on that figure.)

So far this is a clever argument, but doesn't really establish that this will work *in practice* -- for example, maybe your model has to have 10^{100} parameters to learn all of this, or maybe existing models and algorithms are not sophisticated enough to *find* the right parameters (and instead just plateau at, say, 2 bits of loss). But recent evidence provides strong support for the scaling hypothesis:

1. The [scaling laws \(AN #87\)](#) line of work demonstrated that models could be expected to reach the interesting realm of loss at amounts of compute, data, and model capacity that seemed feasible in the near future.
2. Various projects have trained large models and demonstrated that this allows them to solve tasks that they weren't explicitly trained for, often in a more human-like way and with better performance than a more supervised approach. Examples include [GPT-3 \(AN #102\)](#), [Image GPT](#), [BigGAN](#), [AlphaStar \(AN #73\)](#), etc. (The full post has something like 25 examples.)

The author then argues that it seems like most researchers seem to be completely ignoring this phenomenon. OpenAI is the only actor that really has the conviction needed to put a large amount of resources behind a project based on the scaling hypothesis (such as GPT-3); DeepMind seems to believe in a weaker version where we need to build a bunch of "modules" similar to those in the human brain, but that those modules can then be scaled up indefinitely. Other actors seem to not take either scaling hypothesis very seriously.

Rohin's opinion: In my view, the scaling hypothesis is easily the most important hypothesis relevant to AI forecasting and AI development models, and this is the best public writeup of it that I know of. (For example, it seems to be an implicit assumption in the [bio anchors framework \(AN #121\)](#).) I broadly agree with the author that it's a bit shocking how few people seem to be taking it seriously after OpenAI Five, AlphaStar, GPT-3, Copilot, etc.

I think this includes the AI safety space, where as far as I can tell the primary effect has been that it is even more fashionable to have shorter timelines, whereas it hasn't affected AI safety research very much. However, I do know around 3-4 researchers who changed what they were working on based on changing their mind about the scaling hypothesis, so it's possible there are several others I don't know about.

As a simple example of how the scaling hypothesis affects AI safety research, it suggests that the training objective ("predict the next word") is relatively unimportant in determining properties of the trained agent; in contrast, the dataset is much more important. This suggests that analyses based on the "reward function used to train the agent" are probably not going to be very predictive of the systems we actually build.

TECHNICAL AI ALIGNMENT

AGENT FOUNDATIONS

[**The Accumulation of Knowledge**](#) (Alex Flint) (summarized by Rohin): Probability theory can tell us about how we ought to build agents that have knowledge (start with a prior and perform Bayesian updates as evidence comes in). However, this is not the only way to create knowledge: for example, humans are not ideal Bayesian reasoners. As part of our quest to [**describe existing agents**](#) (AN #66), could we have a theory of knowledge that specifies when a particular physical region within a closed system is "creating knowledge"? We want a theory that [**works in the Game of Life**](#) (AN #151) as well as the real world.

This sequence investigates this question from the perspective of defining the accumulation of knowledge as increasing correspondence between [**a map and the territory**](#), and concludes that such definitions are not tenable. In particular, it considers four possibilities and demonstrates counterexamples to all of them:

1. Direct map-territory resemblance: Here, we say that knowledge accumulates in a physical region of space (the "map") if that region of space looks more like the full system (the "territory") over time.

Problem: This definition fails to account for cases of knowledge where the map is represented in a very different way that doesn't resemble the territory, such as when a map is represented by a sequence of zeros and ones in a computer.

2. Map-territory mutual information: Instead of looking at direct resemblance, we can ask whether there is increasing mutual information between the supposed map and the territory it is meant to represent.

Problem: In the real world, nearly every region of space will have high mutual information with the rest of the world. For example, by this definition, a rock accumulates lots of knowledge as photons incident on its face affect the properties of specific electrons in the rock giving it lots of information.

3. Mutual information of an abstraction layer: An abstraction layer is a grouping of low-level configurations into high-level configurations such that transitions between high-level configurations are predictable without knowing the low-level configurations.

For example, the zeros and ones in a computer are the high-level configurations of a digital abstraction layer over low-level physics. Knowledge accumulates in a region of space if that space has a digital abstraction layer, and the high-level configurations of the map have increasing mutual information with the low-level configurations of the territory.

Problem: A video camera that constantly records would accumulate much more knowledge by this definition than a human, even though the human is much more able to construct models and act on them.

4. Precipitation of action: The problem with our previous definitions is that they don't require the knowledge to be *useful*. So perhaps we can instead say that knowledge is accumulating when it is being used to take action. To make this mechanistic, we say that knowledge accumulates when an entity's actions become more fine-tuned to a specific environment configuration over time. (Intuitively, they learned more about the environment and so could condition their actions on that knowledge, which they previously could not do.)

Problem: This definition requires the knowledge to actually be used to count as knowledge. However, if someone makes a map of a coastline, but that map is never used (perhaps it is quickly destroyed), it seems wrong to say that during the map-making process knowledge was not accumulating.

AI GOVERNANCE

[**AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries**](#)

Asymmetries (*Peter Cihon et al*) (summarized by Rohin): *Certification* is a method of reducing information asymmetries: it presents credible information about a product to an audience that they couldn't have easily gotten otherwise. With AI systems, certification could be used to credibly share information between AI actors, which could promote trust amongst competitors, or to share safety measures to prevent a race to the bottom on safety, caused by worrying that "the other guys would be even more unsafe". Certification is at its best when there is *demand* from an audience to see such certificates; public education about the need for credible information can help generate such demand.

However, certification often runs into problems. *Symbol-substance decoupling* happens when certificates are issued to systems that don't meet the standards for certification. For example, in "ethics washing", companies advertise a self-certificate in which their products are approved by ethics boards, but those ethics boards have no real power. *Means-ends decoupling* happens when the standards for certification don't advance the goals for which the certificate was designed. For example, a certificate might focus on whether a system was tested, rather than on what test was conducted, leading applicants to use easy-to-pass tests that don't actually provide a check on whether the method is safe.

Effective certification for future AI systems needs to be responsive to changes in AI technology. This can be achieved in a few ways: first, we can try to test the underlying goals which are more likely to remain stable; for example, we could certify ethical principles that will likely remain the same in the future. Second, we can match the certification to the types of people and institutions, that is, our certifications talk about the executives, citizens, or corporations (rather than e.g. specific algorithms,

that may be replaced in the future). Third, the certification system can build in mechanisms for updating the certification criteria periodically.

The paper then analyzes seven existing certification systems for AI systems; you'll have to read the paper for details.

Case studies of self-governance to reduce technology risk (*Jia*) (summarized by Rohin): Should we expect AI companies to reduce risk through self-governance? This post investigates six historical cases, of which the two most successful were the Asilomar conference on recombinant DNA and the actions of Leo Szilard and other physicists in 1939 (around the development of the atomic bomb). It is hard to make any confident conclusions, but the author identifies the following five factors that make self-governance more likely:

1. The risks are salient.
2. If self-governance doesn't happen, then the government will step in with regulation (which is expected to be poorly designed).
3. The field is small, so that coordination is easier.
4. There is support from gatekeepers (e.g. academic journals).
5. There is support from credentialed scientists.

Corporate Governance of Artificial Intelligence in the Public Interest (*Peter Cihon et al*) (summarized by Rohin): This paper is a broad overview of corporate governance of AI, where by corporate governance we mean "anything that affects how AI is governed within corporations" (a much broader category than the governance that is done by corporations about AI). The authors identify nine primary groups of actors that can influence corporate governance and give many examples of how such actors have affected AI governance in the past. The nine groups are managers, workers, investors, corporate partners and competitors, industry consortia, nonprofit organizations, the public, the media, and governments.

Since the paper is primarily a large set of examples along with pointers to other literature on the topic, I'm not going to summarize it in more detail here, though I did find many of the examples interesting (and would dive into them further if time was not so scarce).

FEEDBACK

I'm always happy to hear feedback; you can send it to me, [Rohin Shah](#), by **replying to this email**.

PODCAST

An audio podcast version of the **Alignment Newsletter** is available. This podcast is an audio version of the newsletter, recorded by [Robert Miles](#).

How much compute was used to train DeepMind's generally capable agents?

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

I'm talking about [these agents \(LW thread here\)](#)

I'd love an answer either in operations (MIPS, FLOPS, whatever) or in dollars.

Follow-up question: How many parameters did their agents have?

I just read the paper (incl. appendix) but didn't see them list the answer anywhere. I suspect I could figure it out from information in the paper, e.g. by adding up how many neurons are in their LSTMs, their various other bits, etc. and then multiplying by how long they said they trained for, but I lack the ML knowledge to do this correctly.

Some tidbits from the paper:

For multi-agent analysis we took the final generation of the agent(generation5)andcreatedequallyspacedcheckpoints (copies of the neural network parameters) every 10 billion steps, creating a collection of 13 checkpoints.

This suggests 120 billion steps of training for the final agents. But elsewhere in the post they state each agent in the final generation experienced 200 billion training steps, so.... huh?

Anyhow. Another tidbit:

In addition to the agent exhibiting zero-shot capabilities across a wide evaluation space, we show that finetuning on a new task for just 100 million steps (around 30 minutes of compute in our setup) can lead to drastic increases in performance relative to zero-shot, and relative to training from scratch which often fails completely.

So, if 100 million steps takes 30min in their setup, and they did 200 billion steps for the final generation, that means the final generation took $30 \times 2,000 = 41$ days. Makes sense. So the whole project probably took something like 100 - 200 days, depending on whether generations 1 - 4 were quicker.

How much does that cost though??? In dollars or FLOPs? I have no idea.

EDIT: It says each agent was trained on 8 TPUv3's. But how many agents were there? I can't find anything about the population size. Maybe I'm not looking hard enough.

Fire Law Incentives

Pacific Gas & Electric is planning to spend [\\$15-30b](#) to bury power lines. I see why they're doing it: PG&E equipment sparked some of the worst fires in California history, including the 2018 [Camp Fire](#) which destroyed [Paradise](#), but I'm not convinced that this is good for California overall.

Historically, the area used to burn periodically. We haven't allowed this for about a century, and flammable materials have been building up. It's all very likely to burn at some point, and burying power lines mostly just reduces the chance that it will be triggered by PG&E. [Prescribed burns](#), spreading out the combustion and moving it to safer times of year, would reduce fire risk far more for the money. Even though when PG&E pays for something the money comes from their customers, CA residents, this isn't a tradeoff PG&E is in a position to consider.

The problem is that CA law puts too much focus on sparks: if you start a fire, you are fully liable for its damage. This approach makes sense in most places, where a "we will never let it burn" policy is practical. In ecosystems adapted for periodic burning, however, where flammable materials build up over time, it means everyone is trying not to be the legally recognized cause of the inevitable fire. And it makes prescribed burns look expensive because when one goes out of control, which there is always a risk, that puts the fire control organization on the hook for the full costs.

Let's work on a system of laws and policies which lead to minimizing overall fire damage.