

Best of LessWrong: January 2021

1. [Bets, Bonds, and Kindergarteners](#)
2. [Lessons I've Learned from Self-Teaching](#)
3. [Leaky Delegation: You are not a Commodity](#)
4. [Science in a High-Dimensional World](#)
5. [Technological stagnation: Why I came around](#)
6. [Birds, Brains, Planes, and AI: Against Appeals to the Complexity/Mysteriousness/Efficiency of the Brain](#)
7. [\[Link\] Still Alive - Astral Codex Ten](#)
8. [Covid 1/7: The Fire of a Thousand Suns](#)
9. [Simulacrum 3 As Stag-Hunt Strategy](#)
10. [Covid 1/21: Turning the Corner](#)
11. [Pseudorandomness contest: prizes, results, and analysis](#)
12. [Cryonics signup guide #1: Overview](#)
13. [Fourth Wave Covid Toy Modeling](#)
14. [Covid 1/14: To Launch a Thousand Shipments](#)
15. [Tentative covid surface risk estimates](#)
16. [Catching the Spark](#)
17. [Matt Levine on "Fraud is no fun without friends."](#)
18. [What is going on in the world?](#)
19. [Reflections on Larks' 2020 AI alignment literature review](#)
20. [Imitative Generalisation \(AKA 'Learning the Prior'\)](#)
21. [Covid 1/28: Muddling Through](#)
22. [#3: Choosing a cryonics provider](#)
23. [Exercise: Taboo "Should"](#)
24. [Avoid Unnecessarily Political Examples](#)
25. [Review of Soft Takeoff Can Still Lead to DSA](#)
26. [Taking money seriously](#)
27. [Luna Lovegood and the Chamber of Secrets - Part 13](#)
28. [DALL-E by OpenAI](#)
29. [Eight claims about multi-agent AGI safety](#)
30. [How good are our mouse models \(psychology, biology, medicine, etc.\), ignoring translation into humans, just in terms of understanding mice? \(Same question for drosophila.\)](#)
31. [A vastly faster vaccine rollout](#)
32. [Actually possible: thoughts on Utopia](#)
33. [How to Write Like Kaj Sotala](#)
34. [Covid: The Question of Immunity From Infection](#)
35. [Why I'm excited about Debate](#)
36. [Unnatural Categories Are Optimized for Deception](#)
37. [Retrospective on Teaching Rationality Workshops](#)
38. [Literature Review on Goal-Directedness](#)
39. [A few thought on the inner ring](#)
40. [D&D.Sci II: The Sorceror's Personal Shopper](#)
41. [Discussion on the choice of concepts](#)
42. [Thoughts on being mortal](#)
43. [The impact merge](#)
44. [Voting Phase for 2019 Review](#)
45. [For Better Commenting, Avoid PONDS](#)
46. [COVID-19: home stretch and fourth wave Q&A](#)
47. [Everything Okay](#)
48. [Unpopularity of efficiency](#)
49. [Grokking illusionism](#)
50. [What currents of thought on LessWrong do you want to see distilled?](#)

Best of LessWrong: January 2021

1. [Bets, Bonds, and Kindergarteners](#)
2. [Lessons I've Learned from Self-Teaching.](#)
3. [Leaky Delegation: You are not a Commodity](#)
4. [Science in a High-Dimensional World](#)
5. [Technological stagnation: Why I came around](#)
6. [Birds, Brains, Planes, and AI: Against Appeals to the Complexity/Mysteriousness/Efficiency of the Brain](#)
7. [\[Link\] Still Alive - Astral Codex Ten](#)
8. [Covid 1/7: The Fire of a Thousand Suns](#)
9. [Simulacrum 3 As Stag-Hunt Strategy.](#)
10. [Covid 1/21: Turning the Corner](#)
11. [Pseudorandomness contest: prizes, results, and analysis](#)
12. [Cryonics signup guide #1: Overview](#)
13. [Fourth Wave Covid Toy Modeling](#)
14. [Covid 1/14: To Launch a Thousand Shipments](#)
15. [Tentative covid surface risk estimates](#)
16. [Catching the Spark](#)
17. [Matt Levine on "Fraud is no fun without friends."](#)
18. [What is going on in the world?](#)
19. [Reflections on Larks' 2020 AI alignment literature review](#)
20. [Imitative Generalisation \(AKA 'Learning the Prior'\)](#)
21. [Covid 1/28: Muddling Through](#)
22. [#3: Choosing a cryonics provider](#)
23. [Exercise: Taboo "Should"](#)
24. [Avoid Unnecessarily Political Examples](#)
25. [Review of Soft Takeoff Can Still Lead to DSA](#)
26. [Taking money seriously](#)
27. [Luna Lovegood and the Chamber of Secrets - Part 13](#)
28. [DALL-E by OpenAI](#)
29. [Eight claims about multi-agent AGI safety](#)
30. [How good are our mouse models \(psychology, biology, medicine, etc.\), ignoring translation into humans, just in terms of understanding mice? \(Same question for drosophila.\)](#)
31. [A vastly faster vaccine rollout](#)
32. [Actually possible: thoughts on Utopia](#)
33. [How to Write Like Kaj Sotala](#)
34. [Covid: The Question of Immunity From Infection](#)
35. [Why I'm excited about Debate](#)
36. [Unnatural Categories Are Optimized for Deception](#)
37. [Retrospective on Teaching Rationality Workshops](#)
38. [Literature Review on Goal-Directedness](#)
39. [A few thought on the inner ring](#)
40. [D&D.Sci II: The Sorceror's Personal Shopper](#)
41. [Discussion on the choice of concepts](#)
42. [Thoughts on being mortal](#)
43. [The impact merge](#)
44. [Voting Phase for 2019 Review](#)
45. [For Better Commenting, Avoid PONDS](#)
46. [COVID-19: home stretch and fourth wave Q&A](#)

47. [Everything Okay](#)
48. [Unpopularity of efficiency](#)
49. [Grokking illusionism](#)
50. [What currents of thought on LessWrong do you want to see distilled?](#)

Bets, Bonds, and Kindergarteners

Bets and bonds are tools for handling different epistemic states and levels of trust. Which makes them a great fit for negotiating with small children!

A few weeks ago Anna (4y) wanted to play with some packing material. It looked very messy to me, I didn't expect she would clean it up, and I didn't want to fight with her about cleaning it up. I considered saying no, but after thinking about how things like this are handled in the real world I had an idea. If you want to do a hazardous activity, and we think you might go bankrupt and not clean up, we make you post a bond. This money is held in escrow to fund the cleanup if you disappear. I explained how this worked, and she went and got a dollar:



Then:





When she was done playing, she cleaned it up without complaint and got her dollar back. If she hadn't cleaned it up, I would have, and kept the dollar.

Some situations are more complicated, and call for bets. I wanted to go to a park, but Lily (6y) didn't want to go to that park because the last time we had been there there'd been lots of bees. I remembered that had been a summer with unusually many bees, and it no longer being that summer or, in fact, summer at all, I was not worried. Since I was so confident, I offered my \$1 to her \$0.10 that we would not run into bees at the park. This seemed fair to her, and when there were no bees she was happy to pay up.

Over time, they've learned that my being willing to bet, especially at large odds, is pretty informative, and often all I need to do is offer. Lily was having a rough morning, crying by herself about a project not working out. I suggested some things that might be fun to do together, and she rejected them angrily. I told her that often when people are feeling that way, going outside can help a lot, and when she didn't seem to believe

me I offered to bet. Once she heard the 10:1 odds I was offering her I think she just started expecting that I was right, and she decided we should go ride bikes. (She didn't actually cheer up when we got outside: she cheered up as soon as she made this decision.)

I do think there is some risk with this approach that the child will have a bad time just to get the money, or say they are having a bad time and they are actually not, but this isn't something we've run into. Another risk, if we were to wager large amounts, would be that the child would end up less happy than if I hadn't interacted with them at all. I handle this by making sure not to offer a bet I think they would regret losing, and while this is not a courtesy I expect people to make later in life, I think it's appropriate at their ages.

Comment via: [facebook](#)

Lessons I've Learned from Self-Teaching

In 2018, I was a bright-eyed grad student who was freaking out about AI alignment. I guess I'm still a bright-eyed grad student freaking out about AI alignment, but that's beside the point.

I wanted to help, and so I [started levelling up](#). While I'd read Nate Soares's [self-teaching posts](#), there were a few key lessons I'd either failed to internalize or failed to consider at all. I think that implementing these might have doubled the benefit I drew from my studies.

I can't usefully write a letter to my past self, so let me write a letter to you instead, keeping in mind that good advice for past-me [may not be good advice for you](#).

Make Sure You Remember The Content

TL;DR: use a spaced repetition system like [Anki](#). Put in cards for key concepts and practice using the concepts. Review the cards every day without fail. This is the most important piece of advice.

The first few months of 2018 were a dream: I was learning math, having fun, and remaking myself. I read and reviewed [about one textbook a month](#). I was learning how to math, how to write proofs and read equations fluently and think rigorously.

I had so much fun that I hurt my wrists typing up my thoughts on impact measures. This turned a lot of my life upside-down. My wrists [wouldn't fully heal for two years](#), and a lot happened during that time. After I hurt my wrists, I became somewhat depressed, posted less frequently, and read fewer books.

When I looked back in 2019/2020 and asked "when and why did my love for textbooks sputter out?", the obvious answer was "when I hurt my hands and lost my sense of autonomy and became depressed, perchance? And maybe I just became averse to reading that way?"

The obvious answer was wrong, but its obvious-ness stopped me from finding the truth until late last year. It *felt* right, but my introspection had failed me.

The real answer is: when I started learning math, I gained a lot of implicit knowledge, like how to write proofs and read math (relatively) quickly. However, I'm no Hermione Granger: left unaided, I'm bad at remembering explicit facts / theorem statements / etc.

I gained implicit knowledge *but I didn't remember the actual definitions*, unless I actually used them regularly (e.g. as I did for real analysis, which I remained quite fluent in and which I regularly use in my research). Furthermore, I think I *coincidentally* hit steeply diminishing returns on the implicit knowledge around when I injured myself.

So basically I'm reading these math textbooks, doing the problems, getting a bit better at writing proofs but not really durably remembering 95% of the content. Maybe part of my subconscious noticed that I seem to be wasting time, that when I come back four months after reading a third of a graph theory textbook, I barely remember the new content I had "learned." I thought I was doing things right. I was doing dozens of exercises and thinking deeply about why each definition was the way it was, thinking about how I could apply these theorems to better reason about my own life and my own research, etc.

I explicitly noticed this problem in late 2020 and thought,

is there any way I know of to better retain content?

... gee, what about [that thing I did in college that let me learn how to read 2,136 standard-use Japanese characters in 90 days](#)? you know, Anki spaced repetition, that thing I never tried for math because once I tried and failed to memorize dozens of lines of MergeSort pseudocode with it?

hm...

This was the moment I started feeling extremely silly (the exact thought was "there's no possible way that my hand is big enough for how facepalm this moment is", IIRC), but also extremely excited. *I could fix my problem!*

And a problem this was. In early 2020, I had an interview where I was asked to compute $\int x \log x dx$. I was stumped, even though this was simple high school calculus (just integrate by parts!). I failed the interview and then went back to learning [algebraic topology](#) and [functional analysis](#) and representation theory. You know, nothing difficult like high school calculus.

I was pretty frustrated with myself.

[It's not that I didn't understand it](#). I just didn't remember it, especially on the spot. The worst part was that I had brushed up on calculus *the previous spring*, and I still didn't remember it. Turns out that my brain won't remember material it doesn't use for months on end, even if forgetting that material would be embarrassing.

Enter [Anki](#), an amazing spaced repetition system (\$20 for iOS, free for computer). The way I like to explain Anki is:

Anki is a flashcard application into which you can enter a constant number of cards each day while retaining a constant average daily workload. You can add cards each day, without having to study longer and longer to get through all of the cards.

I currently think that unless you have really good memory or you're not learning content you want to remember months from now, you're making a mistake by not using a spaced repetition system. Read [Gwern](#) for more on spaced repetition.

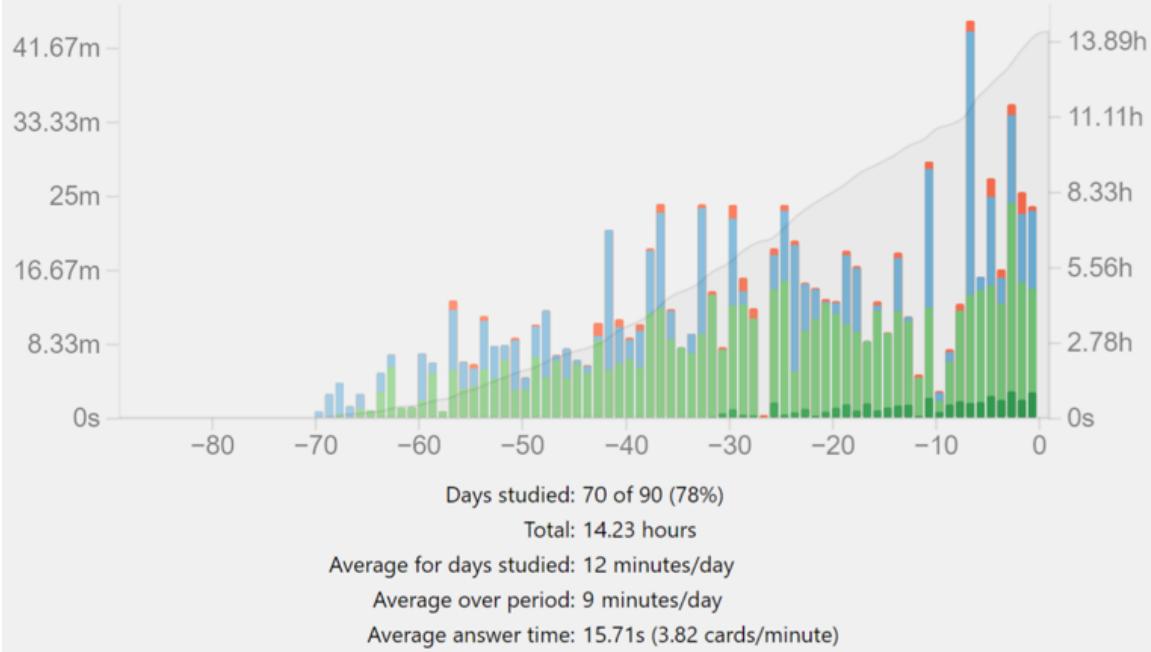
Spaced repetition seems especially useful for students. In college, I ran an experiment: for an upper-level French class, I put things I didn't know how to say into Anki, reviewed daily, and otherwise didn't study at all. I got an A.

How powerful might a bright 6th grader grow, were they to use Anki every day for their whole life? The best time to plant a tree may have been in sixth grade, and the second-best time may have been in seventh grade, but you should still plant the tree now rather than never.

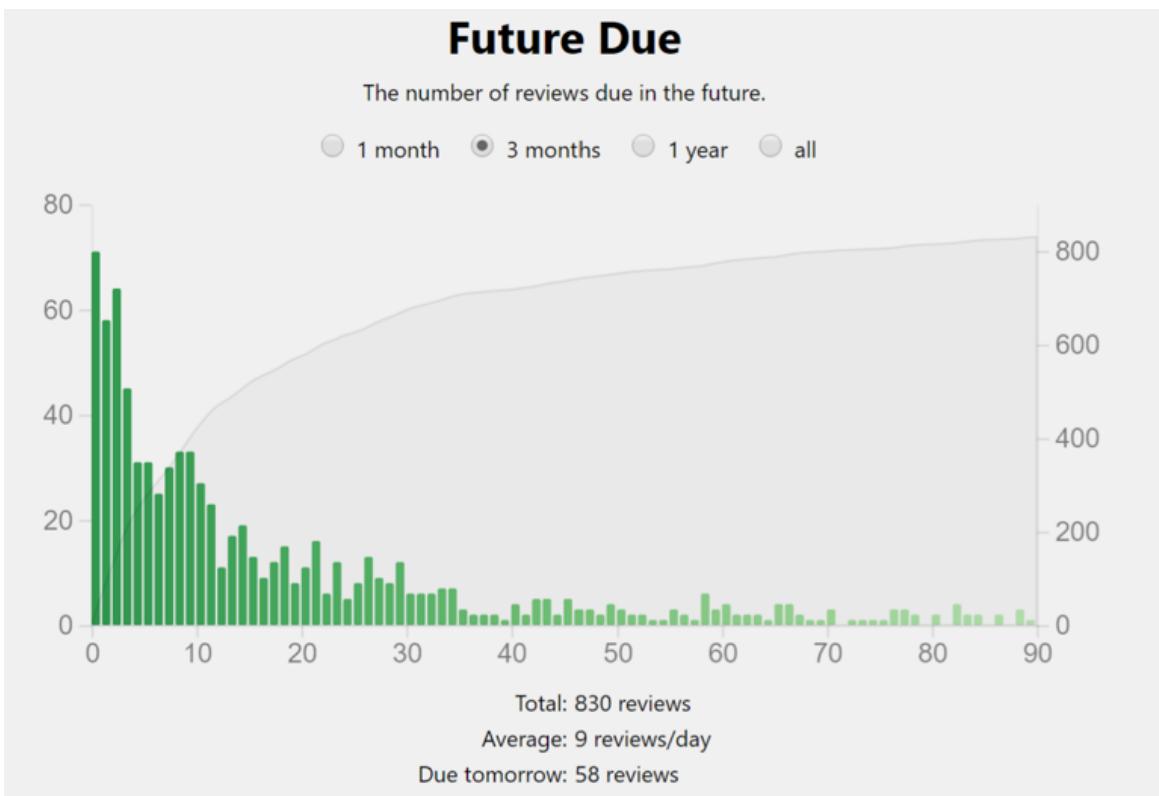
Reviews

The time taken to answer the questions.

Time 1 month 3 months 1 year



I've been using Anki for math for the last 71 days, and I currently have a deck of about 900 cards which I study for ~30 minutes daily. In 2018, I spent about 10 minutes daily reviewing a deck of nearly ten thousand French cards.



If I were to add no more cards, daily reviews drop off quickly.

(Once completed, the reviews in the next few days *will* be pushed into some of the future days, so this projection is slightly optimistic, but you get the point.)

I love Anki, and I was foolish to circumscribe it to language-learning. I [now use Anki](#) to remember key concepts from academic talks, LessWrong blogposts, and yes - textbooks. Which I now love again, and which I read for ~an hour daily again, because I'm *actually retaining the content*.

Measure theory, ring theory, random stuff about taxonomy and epidemiology, quantum mechanics, basic physics, deep RL papers, they all go into Anki, and Anki cards go into my brain - and stay there!

What a wonderful time to be alive.

Random Anki tips

I know quite a bit about how to best use Anki, so if you try this and it doesn't seem to work, please message me instead of banging your head against the wall or giving up!

- **Use cloze deletions on short cards.** Cloze deletions hide a small part of the text, and make you remember it given the rest of the content. They are fast to make and fast to review.
 - The vast, vast majority of my cards are cloze now, and they weren't when I first wrote this post. I think that was a mistake.
- Study every day, preferably at the same time so you aren't always scrambling to get it done before bedtime.
- Sync with AnkiWeb so you don't lose all your cards if your device dies.
- Save time by just screenshotting theorem statements and/or proofs.

- EDIT: [micpie recommends](#) the [Mathpix](#) OCR software, which clips text into MathJax code. This works really well, in my experience.
- [Image occlusion](#) is a great add-on.
- On iPad, I like using MarginNote to read. I can just draw rectangles around key parts of the pdf, make cloze deletions in the app, and then export the cards to an Anki deck.



Memorizing definitions can be useful: when reading a text, it saves you from having to constantly check what the concept is. Make sure to include examples to work through - don't just toss in random definitions you're barely interested in and will never think about again.

- Don't just memorize proofs, focus on the key ideas. Don't just memorize definitions, throw in several example problems which are small enough to actually do in your head (or with a scrap of paper).
 - For example, if I'm trying to really ingrain the concept of an efficient pseudorandom number generator, I have cards where I reason about it by completing short proofs:

Efficient pseudorandom gen imples $\text{BPP} \subseteq \text{P}$

Proof.

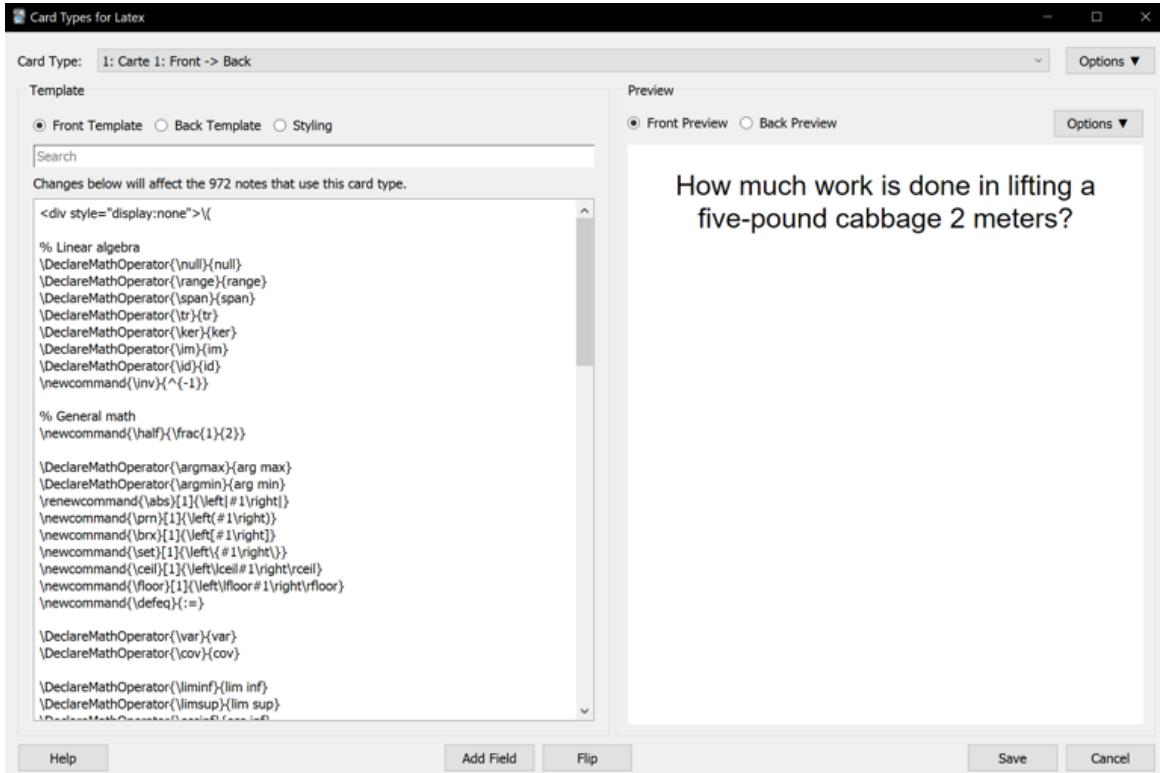
1. Suppose that, for any integer k , you had a way of stretching an $\lfloor \dots \rfloor$ -bit seed to an n -bit output in polynomial time, in such a way that $\lfloor \dots \rfloor$
2. Suppose you had a BPP machine that ran in $\lfloor \dots \rfloor$ time.
3. Loop over all possible seeds (of which there are $\lfloor \dots \rfloor$)
4. feed the corresponding outputs to the BPP machine, and then output the majority answer. $P(\text{accepts} \mid \text{pseudorandom string})$ has to be about the same as $\lfloor \dots \rfloor$ - since otherwise $\lfloor \dots \rfloor$
5. By BPP and the exhaustion of possible strings, majority vote must be right

The math here isn't important. The key thing is, I want to remember the "pseudorandom number generator" concept, and so I find an interesting

result and make myself prove it. The proof isn't too long, which is key—just a few cloze deletions. Don't try to memorize essays with Anki.

Am I ever going to actually use this math for my research? Probably not. Doesn't matter. Anki makes it cheap to learn and retain things.

- If you get a card wrong more than 4 times in the first week, it's a bad card. Remake it.
- Use MathJax instead of Latex, because MathJax renders instantly.



If you want custom commands, use the card type editor.

- My trigger for "I should add a new card" is reading something and thinking, "this is a cool concept!"
 - I recommend adding cards liberally. Don't worry about getting the formatting or phrasing perfect at first. Just add cards and you'll develop a taste for what should be added, and how.

Read Several Textbooks Concurrently

TL;DR study several topics at once so that your brain has time to cement the concepts you're learning, before the text builds on those concepts further.

AllAmericanBreakfast's recent post is great, so I'll refer you to that to make this point:

Wait, what? You want me to make life easier on myself by, instead of studying calculus...studying calculus, linear algebra, and statistics all at once?

~ AllAmericanBreakfast, [The Multi-Tower Study Strategy](#)

The basic idea is that your brain needs time to really cement a new idea in, and so you should study several topics at once in dependency-heavy areas like mathematics. This

advice matches up with both my recent personal experience and with advice I received early on from [Qiaochu Yuan](#), but which I had unfortunately ignored.

For example, right now I'm reading Nielsen and Chuang's [Quantum Computation and Quantum Information](#), Evan Chen's [The Infinite Napkin](#), and a ridiculously easy physics book, Kuhn and Noschese's [A Self-Teaching Guide: Basic Physics](#) (more on this later). Previously, I'd been going back through Wasserman's [All of Statistics](#) and Pearl's [Causality](#) after reading Pearl's [Book of Why](#).

I always feel like I'm learning something new instead of banging my head against the wall. Sometimes you should just read one book, but if you don't need to cram, I recommend diversifying.

Completing The Whole Textbook Is Usually a Big Waste of Time, Please Don't Do It

TL;DR extract the most useful / central concepts and remember them forever via Anki. This doesn't require grasping every arcanum, every detail of a textbook.

When I started reading textbooks, I completed the whole book. I didn't want to miss a crucial concept. But the thing about crucial concepts is that they pop up everywhere. If you missed a crucial concept, you'll know. You're not going to wake up 20 years from now and be like, "OH NO! I forgot to learn about 'force' when I self-studied physics! And I forgot to learn about injectivity in linear algebra!"

As you read more, you'll get a taste for what's probably important, and what's details you can reference later if need be. You can also ask experts if you should study part of a book - this is one key benefit of having an actual teacher.

But don't complete the whole book and all of its exercises, if you just want to become a polymath. Leverage the Pareto principle, get 80% of the benefit out of the key 20/30/40% of the concepts and exercises, and then move on.

Another reason I used to do this a lot was that I wanted to look good by being able to mention on e.g. my resume that I had read the whole book. It's a lot more impressive to say "I read these 10 textbooks" than "I read large parts of this book, and some of this book, and a little of this one, and a lot of this other book." I've found that learning well requires keeping an eye out for these instincts. My brain is not my friend in this battle.

I now often ask myself "am I doing this, at least in part, in order to look good?". Sometimes I answer 'yes', and sometimes I do it anyways - wanting to look good isn't always bad. But [there are sometimes things more important than looking good](#).

Read Easier Textbooks Instead of Struggling Valiantly

TL;DR even though slogging through tough textbooks makes you feel sophisticated and smart, don't.

Imagine I'm learning how to program, but I've never used a computer before. Learning to program *while* learning to operate a computer, will probably take longer than learning to operate a computer and then learning to program. Learning time is superadditive in terms of your ignorance / the dependencies you're missing.

But it's worse than that. Imagine I'm learning quantum mechanics, but I don't know any linear algebra either. I'm now trying to do three things:

1. Learn linear algebra,
2. Learn the formal postulates of quantum mechanics, and
3. Tie all of this into the real world.

Similarly, if I'm trying to learn fluid mechanics without knowing how to manipulate partial differential equations (PDEs), it might look trying to simultaneously

1. Learn PDEs,
2. Learn the physical equations, such as Navier-Stokes, and
3. Tie all of this into the real world to explain what I already know about e.g. water and rivers and blood pressure.

But what if instead I picked up [some dumb book](#) that doesn't even have any calculus, and let it give me approximate explanations via e.g. Archimedes' principle:

Archimedes' principle states that the upward buoyant force that is exerted on a body immersed in a fluid, whether fully or partially, is equal to the weight of the fluid that the body displaces. ([Wikipedia](#))

I breeze through this book no problem, and I can see how to tie in these laws to explain my intuitive models: "logs float more easily than rocks because rock is denser than wood, and so the buoyant force from Archimedes' principle is enough to support the weight of a log." So I'm taking care of point #3, "tie this content into the real world."

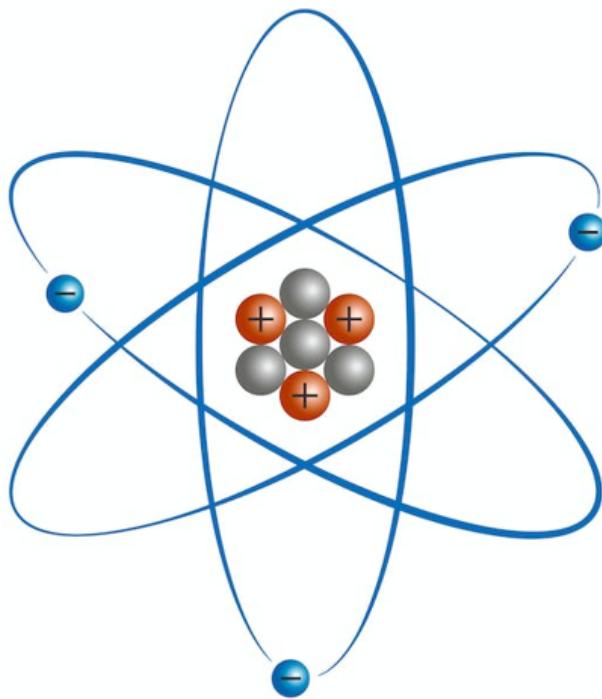
Then, suppose I learn about PDEs and become comfortable with them. Now all I need to do to learn a piece of fluid mechanics is to learn the relevant physical equation, and then think about how it implies things like Archimedes' principle. Crucially, via this method, *I'm only confused about one thing at a time. [Build models in the right order!](#)*

Be Comfortable with Approximate Models

TL;DR allow yourself to learn things in order of comprehensibility: don't try to learn general relativity before Newton's law of gravitation.

In 2019, I wanted to learn a bit of chemistry. I got my hands on a high school chemistry textbook. I got stuck on chapter two because I was being too strict about forming [gears-level models](#).

I was stuck because I was thinking about electron shells. The book acknowledged the [Bohr model](#) of the atom was wrong, electrons aren't *really* discrete particles orbiting the nucleus:



Atom structure

- ⊕ Proton
- ⊗ Neutron
- ⊖ Electron

A wrong model.

But I got nerd-sniped into trying to understand how the electron standing wave only has solutions for certain energy levels, which is a result of quantum mechanics (or so I remember reading). I couldn't understand why, and so the knowledge felt "fake."

It wasn't like I explicitly reasoned "this model is wrong and so I'm not going to keep reading this book", it felt more like... I just wasn't that hungry to learn this, because it wasn't "real." And I gradually stopped returning to that book.

Learn things quickly, note your confusions, and correct them later when the Anki cards show up again. Let yourself learn approximate models of reality.

Conclusion

I knew about the "[read easier textbooks](#)" advice already, but I didn't apply it. Perhaps I just didn't recognize a chance to apply it. The same forces of chaos and entropy and madness which prevented my applying e.g. Luke Muelhauser's advice, may prevent you from applying this post's advice. If you think any of this advice might help you, I recommend setting up a plan *now* for how and when you'll implement it.

Leaky Delegation: You are not a Commodity

Epistemic status: The accumulation of several insights over the years. Reasonably confident that everything mentioned here is an informative factor in decision-making.

Carl is furiously slicing and skinning peaches. His hands move like lightning as slice after slice fills his tray. His freezer has been freshly cleared. Within a day, he will have a new bag of frozen fruit, and can enjoy smoothies for another month.

Stan stands in the kitchen of his college dorm. His hands are carefully placing ingredients on pizza dough: homemade tomato sauce, spiced pork, and mozzarella cheese from a nearby farmer's market. "I don't know why people will pay a restaurant for this," he muses. "So much cheaper to do it yourself."

Michelle is on her way to her job as a software engineer. She tosses a pile of clothes into a bag, and presses a few buttons on her phone. Later that day, someone will come by to pick them up, wash and fold them at a nearby laundromat, and return them the next morning. Less time doing laundry means more time writing code. Her roommate calls her lazy.

An alert flashes on Bruce's screen: "us-east-prod-1 not responding to ping." Almost like a reflex, he pulls up diagnostics on his terminal. The software itself is still running fine, but it looks like his datacenter had a network change. A few more minutes, and everything is functioning again. Hopefully only a few customers noticed the downtime. His mentor keeps asking why he doesn't just run his website on AWS instead of owning his own servers, but Bruce insists it's worth it. His 4-person company has been profitable for 3 years, and keeping server costs low has meant the difference between staying independent and being forced to take outside investment.

The four characters above each take a minority position on outsourcing a task. In the past, I saw the decision as simple: if your time is valuable, then be like Michelle and delegate and outsource as much as you can. Not to do so would be an irrational loss. I silently judged the people I met who inspired Carl and Stan.

Years later, I've found myself cooking daily during a pandemic and appreciating the savings, and just finished arguing online in favor of running one's own servers.

My goal in this post is to share the *perspective shift* that led to me wholly or partially reverse my position on paying a person or company for a good or service (collectively, "delegating" or "outsourcing") in a number of domains, even as I continue to pay for many things most people do themselves. I've noticed hidden factors which mean that, sometimes, the quality *will* be better if you do it yourself, even if the alternative is offered by an expert or company with specialized tools. And sometimes, it *can* be cheaper, even if you value your time very highly and the other person is much faster.

The Internet is [full](#) of [articles](#) on the generic "[buy vs. build](#)" and "[DIY vs. build](#)" decisions. Though some are written from the corporate boardroom and others from the home kitchen or workshop, the underlying analysis is eerily similar: that it's a choice

between spending time (or "in-house resources") or money for a similar value. More sophisticated articles will also consider transaction costs, such as walking to a restaurant or finding your kid a tutor, and costs from principal-agent problems, such as [vetting the tutor](#). In fact, as I've come to realize, the do-or-delegate decision is often not about two alternative ways of getting the same thing, but rather about two options **sufficiently different** that **they're best considered not as replacements for each other, but entirely separate objects with overlapping benefits.**

These differences can be obvious for specific examples, as every home baker can give you an earful about why your local supermarket can't compete with homemade bread. My goal is to extract a generalized model describing the hidden ways in which in-sourced and out-sourced versions may differ. By having this gears-level model, you may wind up doing something yourself when everyone around you is paying for it. Or you may wind up doing the reverse.

Different Product

Have a look at [Instructables](#) or [Etsy](#), or any recipe website, and it's clear that the space of homemade items is vastly larger than the space of storebought items. Something you make yourself is unlikely to match something you could have otherwise acquired. But do these alternatives satisfy the same underlying need?

Naturally, providers do try to fulfill your needs; that's how they get paid, after all. Yet outsourced providers balance the needs of many customers, and face problems you don't. Their offering may be meaningfully different than the homegrown equivalent, sometimes for the worse, even in the face of ruthless competition. And the differences encompass not only the product or service itself, but also the *accompanying intangibles*.

Specialized Preferences

When my housemates and I walked into the restaurant, we were greeted by an exquisite ambience and jazz music. Our hearts sank. "Given its ratings," one explained, "the better the decor, the worse the food."

I found a man whose business is helping descendants of German Jews reclaim their citizenship. The references he gave all spoke glowingly of him. Yet it only took me a few minutes after calling them to decline. "They all said you were invaluable because they didn't speak German. But I do speak German."

You are the only person who wants the exact things you want.

The corollary is that, when you buy something, some of its cost goes into things that you do not want.

Central providers must set their offerings to appeal to a broad base of consumers. When your preferences differ from the median, then you are paying for products not optimized for you.

For instance, on most days, I'd gladly sacrifice the taste of my food for convenience, cost, and health, and so I'd happily pay a local restaurant for a nutritious pile of unappealing glop. But I can't find such a dish nearby — too small a market. And unlike

most customers, I don't eat bread, yet I don't get a discount when I ask them to remove it. But when I cook for myself, I can aggressively optimize for price and convenience against taste. And thus, in spite of the high value I place on my time, I still cook.

Perhaps this is because my particular takes on health put me well outside the norm, and so I am uniquely dissatisfied with local restaurants. But no-one is close to the median in everything, and plenty more go through phases of unusual preferences. For instance, ever notice that most of the recipes in diet books are things you will never find in a restaurant?

Outside of arbitrary tastes, your distinct preferences can also come from your position in life. For instance, common market wisdom is that it is nigh-impossible for an individual to beat market index funds. But an index fund is an off-the-shelf product, selling a standard package of risk. And, [as Buck explained](#), what you actually want is a package of risk that anti-correlates with your career. And so, an American software engineer can "beat the market," in terms of maximizing their personal overall future wealth, simply by not buying US tech stocks.

And while the description of a standard product should fit in a line, or at most a webpage, things you build for yourself or coworkers can have boundaries which are quite complicated. For instance, engineers are permitted to build very janky components for internal use, yet must build with a high degree of polish when producing for the outside world. As I'm writing, I am working on a tool for proving small C programs equivalent. For the time being, I only care about running it on a corpus I've been given as part of a larger project. If I were to release the tool tomorrow, I'd either need to extend it to all of C or write documentation detailing the hundreds of tiny things it doesn't support because they don't appear in my corpus. But I am building it for my coauthors (or, realistically, just me), and so I don't need to. Meanwhile, in a different project 9 years ago, I found that the Java standard library comes with a number of options, which, after searching a large corpus, appear to be literally never used. I wonder how much code is written "for completeness" that will never be used, simply to avoid needing to explain its absence.

Off-the-shelf offerings are designed by distilling the preferences of a large group of people into a few simple components. In contrast, ***you are able to craft a very complicated deal with yourself, containing exactly what you want and no more.***

Bundling and Insurance

Accommodating many preferences often comes in the form of bundles. Anyone who's ever dealt with a TV or phone subscription is aware that products often come with many things they may not use. The economic view is that those who use less are paying for those who use more. Or, put more bluntly: does your plan come with unlimited customer support? Then you're subsidizing the most demanding customers. Did you fly Southwest Airlines and not use your free checked bag? You're subsidizing everyone who did.

But lesser known is that there is a version of this in every physical product you buy, because, when you buy a product, ***you are also paying for the risk of the average user.*** This comes from the [modern doctrine of strict liability](#): that companies are by default liable for damage caused by their products, even without identified

fault. The rationale was given in a famous concurring opinion in [Escola v. Coca-Cola-Bottling Co. \(1944\)](#) (a.k.a. "the exploding Coke bottle case"): companies are much better able to provide "insurance" to their customers by raising the price tag slightly than customers are able to buy insurance themselves.

In other words, some portion of every mass-market product you buy is paying for potential lawsuits that result from that product. This is generally a good thing and a reason to outsource, as you'll be in much better shape financially if your house gets burnt down by a malfunctioning toaster than by your hand-sculpted wood-burning oven. But that also means that, even if your house is fireproof, you're still paying for the risk because someone else's isn't. If you buy a chair, you're paying for the potential injuries because someone else assembled it incorrectly. In general, if you think you'll be better informed or more careful than the average consumer, then you are losing out.

Thankfully, toasters don't often burn houses down, and so this cost is low ([under 1%](#)) for most products. (I'm interested in examples of physical goods for which this is not the case.) But it's alive [in your gym membership](#) (over 1%), decisively high in [risky physical activities](#), and also substantial in, of course, actual insurance. When someone with good oral hygiene chooses to forgo dental insurance, they are in essence insourcing. On a grander scale, I'm now curious how much a company of health nuts could save by underwriting their own health plan, and whether any have done this.

Several of the other factors here have a common theme: you pay more because of information you don't have about the product or service. This one is the reverse: *you pay more because of what the seller doesn't know about you*.

Centralization and Production Technique

In physical products, the constraints of centralized production, distribution, and storage may cause specialized producers to use techniques different from the DIYer. In different circumstances, this can result in a superior or inferior product.

For example, Carl in his kitchen has only disadvantages compared to [commercial producers](#) of frozen fruit. Commercial producers use blast freezers, which both results in a better tasting product (by virtue of fewer ice crystals) and higher throughput. They peel the fruit cutting [with blades](#) not so dissimilar from Carl's knife, and their centralized location gives them access to fresher ingredients than Carl.

In contrast, storebought tomato sauce is also different from the homemade version, but in a negative direction. The tomatoes are peeled either by immersion in a [caustic solution](#) (requiring dye to compensate for the loss of color), or by subjecting them to a [sequence of scalding and physical trauma](#) until the skin falls off. The long storage times require either preservatives or pasteurization. I cannot presently find a source, but I once read about someone who sampled a batch of factory tomato sauce prepared without the additives: a bland, gray glop.

Centralization means scale, and scale creates its own problems. Compared to Stan, the tomato sauce manufacturer pays for scale by degrading health and flavor. The family farm selling spiced pork [pays regulators](#) by building a full parking lot. Even for virtual products, centralization makes a big difference when shared physical resources are used to satisfy many customers. For example, Amazon AWS faces the risk that customers with VMs on the same machine may attempt snoop on each other. They

pay for this by degrading performance: by [expensive mitigations](#) against information-flow attacks, and by wiping disks each time they're connected to a newly-booted VM. Both might be unnecessary if all such VMs served the same company, as is the case for someone running their own [Eucalyptus](#) cluster.

Centralization itself also has a cost, which you pay for in the form of **overhead**. When you buy frozen fruit, you are also paying for a frozen transport chain, shelf space, and the cubicles of designers making packaging and advertising — and also subsidizing unsold product. When buying enterprise software, you are also helping to pay for the vendor's entire sales, HR, and finance departments, thus creating a [pricing deadzone](#) that inflates 4-figure software into the 5-figures.

Differences of production method are likely to be a factor when your homegrown solution is pitted against a centralized provider with expensive capital. It's the reason why Carl's knife-labor is orders of magnitude more expensive than the milliseconds-per-strawberry needed by the frozen fruit factory's machines, and why Bruce's bare-metal machines may be an order of magnitude cheaper than AWS's virtual machines. And the lack of it is why Stan doing cooking and Michelle doing laundry wouldn't be so different from a restaurant or laundromat.

Principal/Agent Problems

Delegation is a coordination problem, and solving it means solving principal/agent problems, of which the main ones are communication and risk. Every manager has stories about hiring an employee or contractor who turned out to suck up more manager time vs. if the manager had simply done it themselves, whether because the hiree couldn't deliver (risk) or required too much help to understand the task (communication). An application to charity is discussed by myself in [Cause Awareness as a Factor against Cause Neutrality](#).

These costs are widely understood, although often underestimated by the inexperienced. Yet even when buying something well-defined, and even when in-sourcing carries the exact same risk of failure, there are extra costs not often considered. These costs involve *information*: you pay to learn things about the provider, and (as in the "Insurance" section above) they charge for their uncertainty about you. **But when you do it yourself, you and the provider both have perfect information.**

Signaling Costs

When I cook a roast for myself, I might toss it in the oven and come back a few hours later, not bothering to set a timer or use a meat thermometer. It might taste worse as a result, but it doesn't give any information about my ability to make a nice steak when I want to, which involves salting it overnight and leaving it in a dry environment.

Now imagine a restaurant that sold both a \$5 roast, prepared as cheaply as possible, and a \$50 steak. A consumer of the first is less likely to share a judgment "The \$5 roast is worth \$5," and more likely to say "The food here sucks."

Every provider must invest in signaling mechanisms which are correlated with quality, but do not directly contribute. A modern business needs a sleek website. McKinsey has [dedicated PowerPoint "Visual Graphics" teams](#) to design the "million-dollar slide

deck" at the end of every project. And personalized-health startup MetaMed found their research would be more trusted with a [grey-haired doctor](#) to deliver it.

Similarly, every provider must avoid visibly cutting corners, even when doing so would benefit the customer. If you tell your house cleaners there's a specific area you don't care about, I predict they'll clean it anyway. This happens a lot also in academic work, where a researcher needs to do work of [questionable utility](#) to placate a reviewer with illegitimate criticisms. I've faced this myself, and I can think of two colleagues who had to invest in *grunt performance engineering* for papers about *proving software correctness guarantees*.

All of these are things that would be useless if the consumer and producer had enough shared knowledge. The upshot of this is that, when buying on the mass market, you are paying for presentation, and the cheapest options might not even be available.

Signaling Benefits

First date: Dress up; make the best first impression possible. Twentieth date: roll out of bed.

But if outsourcing means buying information about the provider, that means there's an extra benefit to outsourcing: the information can be reused in future transactions. While a restaurant still may not sell you a minimal-work \$5 roast even after they've proven to you their quality, this is very true of hiring professionals: **you are also buying a relationship.**

The first time I hired a designer 3 years ago, I wrote a job description and tested a dozen competing candidates. Now, all I do is exchange a couple messages with the winner and get automatically billed a week later. With this low friction, I can outsource tasks as small as sprucing up a *single PowerPoint slide*.

Dark Pattern: Expertise Overkill

In *Deep Work*, Cal Newport writes: for every task you do, think about how much training it takes for someone else to do it, and outsource or eliminate it if it's too low.

Well, skilled professionals routinely do tasks beneath their training and charge full price for it. This comes in the form of the unscrupulous auto shop that bills their minimum 1-hour of mechanic time for a tire change, but also in the form of the honest high-end lawyer who needs to do some low-level contract review in the midst of a larger engagement. For an extreme example (albeit as a TV special rather than a normal sale), see [a world-famous psychologist doing textbook marriage counseling](#).

This clearly occurs because of bundling and coordination costs, but it also occurs because **high-end sellers have an incentive to convince you that everything they do is worth the markup**. This is a "sales dark pattern" akin to the [Dark Patterns](#) in UX design, where a company e.g.: makes the "Buy Now" button look like the "Cancel" button. As an example of both: my first time doing taxes, I got upsold on paying for a service that would help me in the case of an audit, with a full refund if I cancelled within the year. I later learned they actually did little more than be a

middleman for IRS communications (sales dark pattern), and then needed several phone calls to cancel (UX dark pattern).

For physical products, this takes the form of convincing buyers that some superfluous aspect of their product is highly beneficial, or otherwise implying users are buying more than they actually are. This occurs subtly in the techniques restaurants use to make their portions look larger; explicitly by some sellers of "natural" and organic products and by the entire multivitamin industry; and overtly in the case of Red Lobster, which displays a tank of Maine lobster in the front of every restaurant but sells a lobster bisque made of langostino and rock lobster, cheaper shellfish unrelated to "true lobster."

The [apocryphal engineering consultant](#) charges \$10,000 for a chalk mark on a malfunctioning coil: \$1 for the mark, and \$9,999 for knowing where to put it. The [apocryphal Picasso](#) charges 5000 francs for a 2-minute sketch, because "It took me my whole life." But the management consultant who forms the butt of jokes will [tell you what you want to hear](#), "borrow your watch to tell you what time it is," and [sell you on satellite imaging](#) when all you needed was a pair of eyeballs. If you paid for an answer that took either a lifetime of training or an hour of Googling, could you tell the difference? How about a drip marketing channel that was either based on unique insight or copied from HubSpot Academy? If a technology firm [talked you into](#) a more expensive solution than you need?

Providers distinguish their offerings from DIY versions and justify the price by bundling added value, be it a cafe's ambience or the confidence of an expert's eye. The sophisticated buyer is one who can see when a component of that bundle isn't pulling its weight, and is prepared to craft their own to compete. A fool and his money are soon parted, but the reverse is also true. Or, as actually happened to a friend: when your income is large, there will always be people willing to sell you \$12 avocado toast until it's not, and wealth comes from learning to say no.

Value of Learning

A paradox of outsourcing: the better you can do something yourself, the less valuable it is to do it yourself.

There's a simple reason: when you do something yourself, the value is in both having it done and learning how to do it. Doing laundry may be busywork to the habituated, but doing it for the first time also comes with the security of not being dependent on others, and the ability to not be hoodwinked if the laundromat returns wrinkled clothes, claiming "that's just what happens."

The value of the learning differs greatly between individuals. In my case, my career is in software development, and my business is providing training for software engineers who come from a vast array of specialties. I have a lot to benefit from learning to run my own servers, even if I wind up never doing it again. Yet my ambition in cooking ends at "Be able to impress guests a few times a year," and so I'm not likely to benefit from practice handcrafting pizzas except by handcrafting more pizzas. In contrast, a French chef may learn useful things about his suppliers from growing his own vegetables, or gain insights from branching out into making his own tortillas.

You must also ask: what kind of learning is involved? Some kinds of learning may expand your knowledge in the world, while others may be more like muscle memory,

and only let you do specific things faster. That's the main dichotomy, but I'm going to expand on it in the next few paragraphs by with some of my (admittedly scant) knowledge of cognitive science.

My current ability to give precise characterizations of learning comes from my recent study of [ACT-R](#), a cognitive architecture able to, from just a few parameters, reasonably predict millisecond-level timing of human performance on an array of memory and learning tasks.

The ACT-R model offers several mechanisms of learning. Several forms of learning relate to enhancing and growing one's web of concepts, and several more to learning and applying the tiny cognitive operations that act on it. As this web grows, relationships are formed between new and existing concepts, offering enhanced recall of both. These kinds of learning lend themselves to general utility.

Yet there's another form of learning, called "knowledge compilation" or "production compilation," which serves only to specialize knowledge: turning a general operation like "recall the next digit of my phone number and then say it" into a specialized one like "say 'seven.'" I expect that this form of learning provides the most apt modeling of Carl's knifework, refining a general procedure for observing the fruit and choosing a cut with a sequence of preprogrammed motor actions applicable only to peaches. Whereas trying a different food may result in learning new things one can do with a knife, this kind of learning serves but one purpose: cut peaches faster. Knowledge compilation can only take general skills and make one faster at specific applications. Summarizing: **there is more value in basic learning of a new concept or skill than in taking a skill and making it automatic.**

Conclusion

Simple questions are welcoming. "Should we have Mexican or Chinese" is, from a coarse-grained perspective, a request to select from a fixed set of options, and, from a fine-grained perspective, a complex weighting of several logistical, nutritional, and social preferences. As one moves away from the daily minutiae into choosing the objects and processes that control our lives, it is tempting to continue to apply the comfortable coarse lens. Yet, to the eagle-eyed observer, a posed choice between simple alternatives belies substantial differences. And there is value to be had in spotting them.

Cheatsheet: 10 Questions to Ask Before Delegating or Buying

1. Are there aspects that I don't care about but providers take great pains to provide?
2. Do I believe that my preferred set of tradeoffs differs substantially from the median?
3. Do I have reason to believe I'll be savvier than the average consumer of this product or service?
 1. In particular, is there catastrophic risk associated with this offering, and do I have reason to believe I am much less susceptible than average?
4. Do providers have access to some kind of specialized capital or technology for which I lack good substitutes?

5. Do providers have substantial overhead that I would not? This includes regulatory burden and security.
6. How much of the cost is spent on signaling quality?
 1. Do I hope to have a lot of repeat business with the same provider?
7. How much of the price is spent on transaction costs?
8. How do I know whether and when the special expertise claimed by the provider is beneficial?
9. How much do I value the learning from doing it myself?
 1. Do I expect to use the knowledge gained?
 2. Is the learning involved overly specialized?
10. Are there other ways in which the thing to be bought is substantially different from what I'd do for myself?

Science in a High-Dimensional World

Claim: the usual explanation of the Scientific Method is missing some key pieces about how to make science work well in a high-dimensional world (e.g. our world). Updating our picture of science to account for the challenges of dimensionality gives a different model for how to do science and how to recognize high-value research. This post will sketch out that model, and explain what problems it solves.

The Dimensionality Problem

Imagine that we are early scientists, investigating the mechanics of a sled sliding down a slope. What determines how fast the sled goes? Any number of factors could conceivably matter: angle of the hill, weight and shape and material of the sled, blessings or curses laid upon the sled or the hill, the weather, wetness, phase of the moon, latitude and/or longitude and/or altitude, etc. For all the early scientists know, there may be some deep mathematical structure to the world which links the sled's speed to the astrological motions of stars and planets, or the flaps of the wings of butterflies across the ocean, or vibrations from the feet of foxes running through the woods.

Takeaway: there are literally billions of variables which could influence the speed of a sled on a hill, as far as an early scientist knows.



So, the early scientists try to control as much as they can. They use a standardized sled, with standardized weights, on a flat smooth piece of wood treated in a standardized manner, at a standardized angle. Playing around, they find that they need to carefully control a dozen different variables to get reproducible results. With those dozen pieces carefully kept the

same every time... the sled consistently reaches the same speed (within reasonable precision).

At first glance, this does not sound very useful. They had to exercise unrealistic levels of standardization and control over a dozen different variables. Presumably their results will not generalize to real sleds on real hills in the wild.

But stop for a moment to consider the implications of the result. A consistent sled-speed can be achieved while controlling *only a dozen variables*. Out of *literally billions*. Planetary motions? Irrelevant, after controlling for those dozen variables. Flaps of butterfly wings on the other side of the ocean? Irrelevant, after controlling for those dozen variables. Vibrations from foxes' feet? Irrelevant, after controlling for those dozen variables.

The amazing power of achieving a consistent sled-speed is not that other sleds on other hills will reach the same predictable speed. Rather, it's knowing *which variables are needed* to predict the sled's speed. Hopefully, those same variables will be sufficient to determine the speeds of other sleds on other hills - even if some experimentation is required to find the speed for any particular variable-combination.

Determinism

How can we know that *all* other variables in the universe are irrelevant after controlling for a handful? Couldn't there always be some other variable which is relevant, no matter what empirical results we see?

The key to answering that question is determinism. If the system's behavior can be predicted *perfectly*, then there is no mystery left to explain, no information left which some unknown variable could provide. Mathematically, information theorists use the mutual information $I(X, Y)$ to measure the information which X contains about Y. If Y is deterministic - i.e. we can predict Y perfectly - then $I(X, Y)$ is zero no matter what variable X we look at. Or, in terms of correlations: a deterministic variable always has zero correlation with everything else. If we can perfectly predict Y, then there is no further information to gain about it.

In this case, we're saying that sled speed is deterministic *given* some set of variables (sled, weight, surface, angle, etc). So, given those variables, everything else in the universe is irrelevant.

Of course, we can't always perfectly predict things in the real world. There's always some noise - certainly at the quantum scale, and usually at larger scales too. So how do we science?

The first thing to note is that "perfect predictability implies zero mutual information" plays well with approximation: *approximately* perfect predictability implies *approximately* zero mutual information. If we can predict the sled's speed to within 1% error, then any other variables in the universe can only influence that remaining 1% error. Similarly, if we can predict the sled's speed 99% of the time, then any other variables can only matter 1% of the time. And we can combine those: if 99% of the time we can predict the sled's speed to within 1% error, then any other variables can only influence the 1% error except for the 1% of sled-runs when they might have a larger effect.

More generally, if we can perfectly predict any specific variable, then everything else in the universe is irrelevant to that variable - even if we can't perfectly predict all aspects of the system's trajectory. For instance, if we can perfectly predict the *first two digits* of the sled's speed (but not the less-significant digits), then we know that nothing else in the universe is

relevant to those first two digits (although all sorts of things could influence the less-significant digits).

As a special case of this, we can also handle noise using repeated experiments. If I roll a die, I can't predict the outcome perfectly, so I can't rule out influences from all the billions of variables in the universe. But if I roll a die a few thousand times, then I can approximately-perfectly predict the *distribution* of die-rolls (including the mean, variance, etc). So, even though I don't know what influences any one particular die roll, I do know that nothing else in the universe is relevant to the overall distribution of repeated rolls (at least to within some small error margin).

Replication

This does still leave one tricky problem: what if we *accidentally* control some variable? Maybe air pressure influences sled speed, but it never occurred to us to test the sled in a vacuum or high-pressure chamber, so the air pressure was roughly the same for all of our experiments. We are able to deterministically predict sled speed, but only because we accidentally keep air pressure the same every time.

This is a thing which actually does happen! Sometimes we test something in conditions never before tested, and find that the usual rules no longer apply.

Ideally, replication attempts catch this sort of thing. Someone runs the same experiment in a different place and time, a different environment, and hopefully whatever things were accidentally kept constant will vary. (You'd be amazed what varies by location - I once had quite a surprise double-checking the pH of deionized water in Los Angeles.)

Of course, like air pressure, some things may happen to be the same even across replication attempts.

On the other hand, if a variable is accidentally controlled across multiple replication attempts, then it will likely be accidentally controlled outside the lab too. If every lab tests sled-speed at atmospheric pressure, and nobody ever accidentally tries a different air pressure, then that's probably because sleds are almost always *used* at atmospheric pressure. When somebody goes to predict a sled's speed in space, some useful new scientific knowledge will be gained, but until then the results will generally work in practice.

The Scientific Method In A High-Dimensional World

Scenario 1: a biologist hypothesizes that adding hydroxyhypothetical to their yeast culture will make the cells live longer, and the cell population will grow faster as a result. To test this hypothesis, they prepare one batch of cultures with the compound and one without, then measure the increase in cell density after 24 hours. They statistically compare the final cell density in the two batches to see whether the compound had a significant effect.

This is the prototypical Scientific Method: formulate a hypothesis, test it experimentally. Control group, p-values, all that jazz.

Scenario 2: a biologist observes that some of their clonal yeast cultures flourish, while others grow slowly or die out altogether, despite seemingly-identical preparation. What causes this different behavior? They search for differences, measuring and controlling for everything they can think of: position of the dishes in the incubator, order in which samples were prepared, mutations, phages, age of the initial cell, signalling chemicals in the cultures, combinations of all those... Eventually, they find that using initial cells of the same replicative age eliminates most of the randomness.

This looks less like the prototypical Scientific Method. There's probably some hypothesis formation and testing steps in the middle, but it's less about hypothesize-test-iterate, and more about figuring out which variables are relevant.

In a high-dimensional world, effective science looks like scenario 2. This isn't mutually exclusive with the Scientific-Method-as-taught-in-high-school, there's still some hypothesizing and testing, but there's a new piece and a different focus. The main goal is to hunt down sources of randomness, figure out exactly what needs to be controlled in order to get predictable results, and thereby establish which of the billions of variables in the universe are actually relevant.

Based on personal experience and reading lots of papers, this matches my impression of which scientific research offers lots of long-term value in practice. The one-shot black-box hypothesis tests usually aren't that valuable in the long run, compared to research which hunts down the variables relevant to some previously confusing (a.k.a. unpredictable) phenomenon.

Everything Is Connected To Everything Else (But Not Directly)

What if there is no small set of variables which determines the outcome of our experiment? What if there really are billions of variables, all of which matter?

We sometimes see a claim like this made about biological systems. As the story goes, you can perform all sorts of interventions on a biological system - knock out a gene, add a drug, adjust diet or stimulus, etc - and any such intervention will change the level of most of the tens-of-thousands of proteins or metabolites or signalling molecules in the organism. It won't necessarily be a large change, but it will be measurable. Everything is connected to everything else; any change impacts everything.

Note that this is not at all incompatible with a small set of variables determining the outcome! The problem of science-in-a-high-dimensional-world is not to enumerate all variables which have any influence. The problem is to find a set of variables which determine the outcome, so that no other variables have any influence *after* controlling for those.

Suppose sled speed is determined by the sled, slope material, and angle. There may still be billions of other variables in the world which impact the sled, the slope material, and the angle! But none of those billions of variables are relevant *after* controlling for the sled, slope material, and angle; other variables influence the speed only through those three. Those three variables *mediate* the influence of all the billions of other variables.

In general, the goal of science in a high dimensional world is to find sets of variables which mediate the influence of all other variables on some outcome.

In some sense, the central empirical finding of All Of Science is that, in practice, we can generally find *small* sets of variables which mediate the influence of all other variables. Our universe is "local" - things only interact directly with nearby things, and only so many things can be nearby at once. Furthermore, our universe abstracts well: even indirect interactions over long distances can usually be summarized by a small set of variables. Interactions between stars across galactic distances mostly just depend on the total mass of each star, not on all the details of the plasma roiling inside.

Even in biology, every protein interacts with every other protein in the network, but the vast majority of proteins do not interact *directly* - the graph of biochemical interactions is connected, but extremely sparse. The interesting problem is to figure out the structure of that graph - i.e. which variables interact directly with which other variables. If we pick one

particular “outcome” variable, then the question is which variables are its neighbors in the graph - i.e. which variables mediate the influence of all the other variables.

Summary

Let’s put it all together.

In a high-dimensional world like ours, there are billions of variables which could influence an outcome. The great challenge is to figure out *which variables* are directly relevant - i.e. which variables mediate the influence of everything else. In practice, this looks like finding mediators and hunting down sources of randomness. Once we have a set of control variables which is sufficient to (approximately) determine the outcome, we can (approximately) rule out the relevance of any other variables in the rest of the universe, *given* the control variables.

A remarkable empirical finding across many scientific fields, at many different scales and levels of abstraction, is that a *small* set of control variables usually suffices. Most of the universe is not directly relevant to most outcomes most of the time.

Ultimately, this is a picture of “gears-level science”: look for mediation, hunt down sources of randomness, rule out the influence of all the other variables in the universe. This sort of research requires a lot of work compared to one-shot hypothesis tests, but [it provides a lot more long-run value](#): because all the other variables in the universe are irrelevant, we only need to measure/control the control variables each time we want to reuse the model.

Technological stagnation: Why I came around

This is a linkpost for <https://rootsofprogress.org/technological-stagnation>

"We wanted [flying cars](#), instead we got 140 characters," [says Peter Thiel's Founders Fund](#), expressing a sort of jaded disappointment with technological progress. (The fact that the 140 characters have become 280, [a 100% increase](#), does not seem to have impressed him.)

Thiel, along with economists such as Tyler Cowen ([The Great Stagnation](#)) and Robert Gordon ([The Rise and Fall of American Growth](#)), promotes a "stagnation hypothesis": that there has been a significant slowdown in scientific, technological, and economic progress in recent decades—say, for a round number, since about 1970, or the last ~50 years.

When I first heard the stagnation hypothesis, I was skeptical. The arguments weren't convincing to me. But as I studied the history of progress (and looked at the numbers), I slowly came around, and now I'm fairly convinced. So convinced, in fact, that I now seem to be more pessimistic about ending stagnation than some of its original proponents.

In this essay I'll try to capture both why I was originally skeptical, and also why I changed my mind. If you have heard some of the same arguments that I did, and are skeptical for the same reasons, maybe my framing of the issue will help.

Stagnation is relative

To get one misconception out of the way first: "stagnation" does not mean zero progress. No one is claiming that. There wasn't zero progress even before the Industrial Revolution (or the civilizations of Europe and Asia would have looked no different in 1700 than they did in the days of nomadic hunter-gatherers, tens of thousands of years ago).

Stagnation just means *slower* progress. And not even slower than that pre-industrial era, but slower than, roughly, the late 1800s to mid-1900s, when growth rates are said to have peaked.

Because of this, we can't resolve the issue by pointing to *isolated* advances. The microwave, the air conditioner, the electronic pacemaker, a new cancer drug—these are great, but they don't disprove stagnation.

Stagnation is relative, and so to evaluate the hypothesis we must find some way to compare magnitudes. This is difficult.

Only 140 characters?

"We wanted flying cars, instead we got a supercomputer in everyone's pocket and a global communications network to connect everyone on the planet to each other and to the whole of the world's knowledge, art, philosophy and culture." When you put it that way, it doesn't sound so bad.

Indeed, the digital revolution has been absolutely amazing. It's up there with electricity, the internal combustion engine, or mass manufacturing: one of the great, fundamental, transformative technologies of the industrial age. (Although admittedly it's hard to see the effect of computers in the productivity statistics, and I don't know why.)

But we don't need to downplay the magnitude of the digital revolution to see stagnation; conversely, proving its importance will not defeat the stagnation hypothesis. Again, stagnation is relative, and we must find some way to compare the current period to those that came before.

Argumentum ad living room

Eric Weinstein [proposes a test](#): "Go into a room and subtract off all of the screens. How do you know you're not in 1973, but for issues of design?"

This too I found unconvincing. It felt like a weak thought experiment that relied too much on intuition, revealing one's own priors more than anything else. And why should we necessarily expect progress to be visible directly from the home or office? Maybe it is happening in specialized environments that the layman wouldn't have much intuition about: in the factory, the power plant, the agricultural field, the hospital, the oil rig, the cargo ship, the research lab.

No progress except for all the progress

There's also that sleight of hand: "subtract the screens". A starker form of this argument is: "except for computers and the Internet, our economy has been relatively stagnant." Well, sure: if you carve out all the progress, what remains is stagnation.

Would we even expect progress to be evenly distributed across all domains? Any one technology [follows an S-curve](#): a slow start, followed by rapid expansion, then a leveling off in maturity. It's not a sign of stagnation that after the world became electrified, electrical power technology wasn't a high-growth area like it had been in the early 1900s. That's not how progress works. Instead, we are constantly turning our attention to new frontiers. If that's the case, you can't carve out the frontiers and then say, "well, except for the frontiers, we're stagnating".

Bit bigotry?

In an [interview with Cowen](#), Thiel says stagnation is "in the world of atoms, not bits":

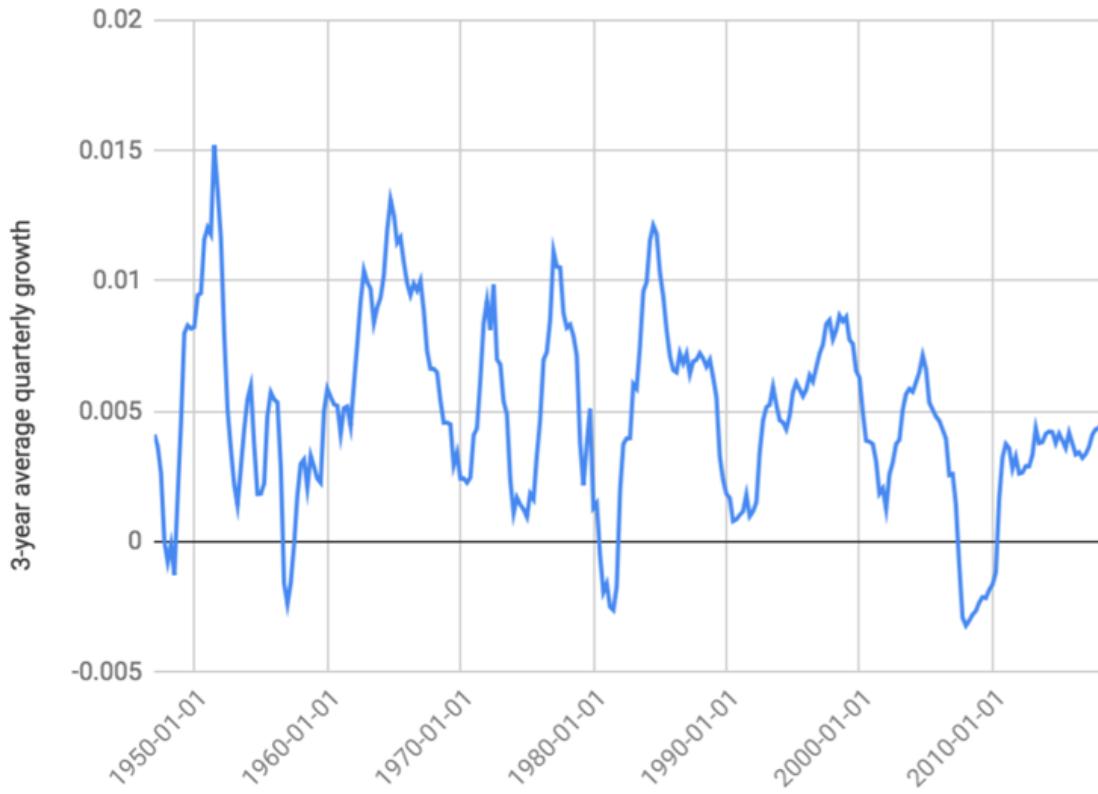
I think we've had a lot of innovation in computers, information technology, Internet, mobile Internet in the world of bits. Not so much in the world of atoms, supersonic travel, space travel, new forms of energy, new forms of medicine, new medical devices, etc.

But again, why should we expect it to be different? Maybe bits are just the current frontier. And what's the matter with bits, anyway? Are they less important than atoms? Progress in any field is still progress.

The quantitative case

So, we need more than isolated anecdotes, or appeals to intuition. A more rigorous case for stagnation can be made quantitatively. [A paper by Cowen and Ben Southwood](#) quotes Gordon: "U.S. economic growth slowed by more than half from 3.2 percent per year during 1970-2006 to only 1.4 percent during 2006-2016." Or look at this chart from the same paper:

3-year average quarterly annual growth USA



Growth in US GDP per capita. [Cowen & Southwood](#)

Gordon's own book points out that growth in output per hour has slowed from an average annual rate of 2.82% in the period 1920-1970, to 1.62% in 1970-2014. He also analyzes TFP (total factor productivity, a residual calculated by subtracting out increases in capital and labor from GDP growth; what remains is assumed to represent productivity growth from technology). Annual TFP growth was 1.89% from 1920-1970, but less than 1% in every decade since. ([More detail in my review of Gordon's book.](#))

Analyzing growth quantitatively is hard, and these conclusions are disputed. GDP is problematic (and these authors acknowledge this). In particular, it does not capture consumer surplus: since you don't pay for articles on Wikipedia, searches on Google, or entertainment on YouTube, a shift to these services away from paid ones actually *shrinks* GDP, but it represents progress and consumer benefit.

Gordon, however, points out that GDP has *never* captured consumer surplus, and there has been plenty of surplus in the past. So if you want to argue that unmeasured surplus is the cause of an apparent (but not a real) decline in growth rates, then you have to argue that GDP has been *systematically increasingly mismeasured* over time.

So far, I've only heard one only argument that even hints in this direction: the shift from manufacturing to services. If services are more mismeasured than manufactured products, then in logic at least this could account for an illusory slowdown. But I've never seen this argument fully developed.

In any case, the quantitative argument is not what convinced me of the stagnation hypothesis nearly as much as the qualitative one.

Sustaining multiple fronts

I remember the first time I thought there might really be something to the stagnation hypothesis: it was when I started mapping out a timeline of major inventions in each main area of industry.

At a high level, I think of technology/industry in six major categories:

- Manufacturing & construction
- Agriculture
- Energy
- Transportation
- Information
- Medicine

Almost every significant advance or technology can be classified in one of these buckets (with a few exceptions, such as finance and perhaps retail).

The first phase of the industrial era, sometimes called “the first Industrial Revolution”, from the 1700s through the mid-1800s, consisted mainly of two fundamental advances: [mechanization](#), and the [steam engine](#). The factory system evolved in part out of the former, and the locomotive was based on the latter. Together, these revolutionized manufacturing, energy, and transportation, and began to transform agriculture as well.

The “second Industrial Revolution”, from the mid-1800s to the mid-1900s, is characterized by a greater influence of science: mainly chemistry, electromagnetism, and microbiology. Applied chemistry gave us better materials, from [Bessemer steel](#) to plastic, and [synthetic fertilizers](#) and [pesticides](#). It also gave us processes to refine petroleum, enabling the oil boom; this led to the internal combustion engine, and the vehicles based on it—cars, planes, and oil-burning ships—that still dominate transportation today. Physics gave us [the electrical industry](#), including generators, motors, and the light bulb; and electronic communications, from the telegraph and telephone through radio and television. And biology gave us the germ theory, which dramatically reduced infectious disease mortality rates through [improvements in sanitation](#), new [vaccines](#), and towards the end of this period, antibiotics. So every single one of the six major categories was completely transformed.

Since then, the “third Industrial Revolution”, starting in the mid-1900s, has mostly seen fundamental advances in a single area: electronic computing and communications. If you date it from 1970, there has really been nothing comparable in manufacturing, agriculture, energy, transportation, or medicine—again, not that these areas have seen zero progress, simply that they’ve seen less-than-revolutionary progress. Computers have completely transformed all of information processing and communications, while there have been no new types of materials, vehicles, fuels, engines, etc. The closest candidates I can see are containerization in shipping, which revolutionized cargo but did nothing for passenger travel; and genetic engineering, which has given us a few big wins but hasn’t reached nearly its full potential yet.

The digital revolution has had echoes, derivative effects, in the other areas, of course: computers now help to control machines in all of those areas, and to plan and optimize processes. But those secondary effects existed in previous eras, too, *along with* primary effects. In the third Industrial Revolution we only have primary effects in one area.

So, making a very rough count of revolutionary technologies, there were:

- 3 in IR1: mechanization, steam power, the locomotive
- 5 in IR2: oil + internal combustion, electric power, electronic communications, industrial chemistry, germ theory
- 1 in IR3 (so far): computing + digital communications

It's not that bits don't matter, or that the computer revolution isn't transformative. It's that in previous eras we saw breakthroughs across the board. It's that we went from five simultaneous technology revolutions to one.

The missing revolutions

The picture becomes starker when we look at the technologies that were promised, but never arrived or haven't come to fruition yet; or those that were launched, but aborted or stunted. If manufacturing, agriculture, etc. weren't transformed, then how *could* they have been?

Energy: The most obvious stunted technology is nuclear power. In the 1950s, everyone expected a nuclear future. Today, nuclear supplies [less than 20% of US electricity and only about 8% of its total energy](#) (and about half those figures in the world at large). Arguably, we should have had [nuclear homes, cars and batteries by now](#).

Transportation: In 1969, Apollo 11 landed on the Moon and Concorde took its first supersonic test flight. But they were not followed by a thriving space transportation industry or broadly available supersonic passenger travel. The last Apollo mission flew in 1972, a mere three years later. Concorde was only ever available as a luxury for the elite, was never highly profitable, and was shut down in 2003, after less than thirty years in service. Meanwhile, passenger travel speeds are unchanged over 50 years (actually slightly reduced). And of course, [flying cars](#) are still the stuff of science fiction. Self-driving cars may be just around the corner, but haven't arrived yet.

Medicine: Cancer and heart disease are still the top causes of death. Solving even one of these, the way we have mostly solved infectious disease and vitamin deficiencies, would have counted as a major breakthrough. Genetic engineering, again, has shown a few excellent early results, but hasn't yet transformed medicine.

Manufacturing: In materials, carbon nanotubes and other nanomaterials are still mostly a research project, and we still have no material to build a [space elevator](#) or a [space pier](#). As for processes, [atomically precise manufacturing](#) is even more science-fiction than flying cars.

If we had gotten even a few of the above, the last 50 years would seem a lot less stagnant.

One to zero

This year, the computer turns 75 years old, and the microprocessor turns 50. Digital technology is due to level off in its maturity phase.

So what comes next? The only thing worse than going from five simultaneous technological revolutions to one, is going from one to zero.

I am hopeful for genetic engineering. The ability to fully understand and control biology obviously has enormous potential. With it, we could cure disease, extend human lifespan, and augment strength and intelligence. We've made a good start with recombinant DNA technology, which gave us synthetic biologics such as insulin, and CRISPR is a major advance on top of that. The rapid success of two different mRNA-based covid vaccines is also a breakthrough, and a sign that a real genetic revolution might be just around the corner.

But genetic engineering is also subject to many of the forces of stagnation: research funding via a centralized bureaucracy, a hyper-cautious regulatory environment, and a cultural perception of something scary and dangerous. So it is not guaranteed to arrive. Without the right support and protection, we might be looking back on biotech from the year 2070 the

way we look back on nuclear energy now, wondering why we never got a genetic cure for cancer and why life expectancy has plateaued.

Aiming higher

None of this is to downplay the importance or impact of any specific innovation, nor to discourage any inventor, present or future. Quite the opposite! It is to encourage us to set our sights still higher.

Now that I understand what was possible around the turn of the last century, I can't settle for anything less. We need breadth in progress, as well as depth. We need revolutions on all fronts at once: not only biotech but manufacturing, energy and transportation as well. We need progress in bits, atoms, cells, *and* joules.

Birds, Brains, Planes, and AI: Against Appeals to the Complexity/Mysteriousness/Efficiency of the Brain

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

[Epistemic status: Strong opinions lightly held, this time with a cool graph.]

I argue that an entire class of common arguments against short timelines is bogus, and provide weak evidence that anchoring to the human-brain-human-lifetime milestone is reasonable.

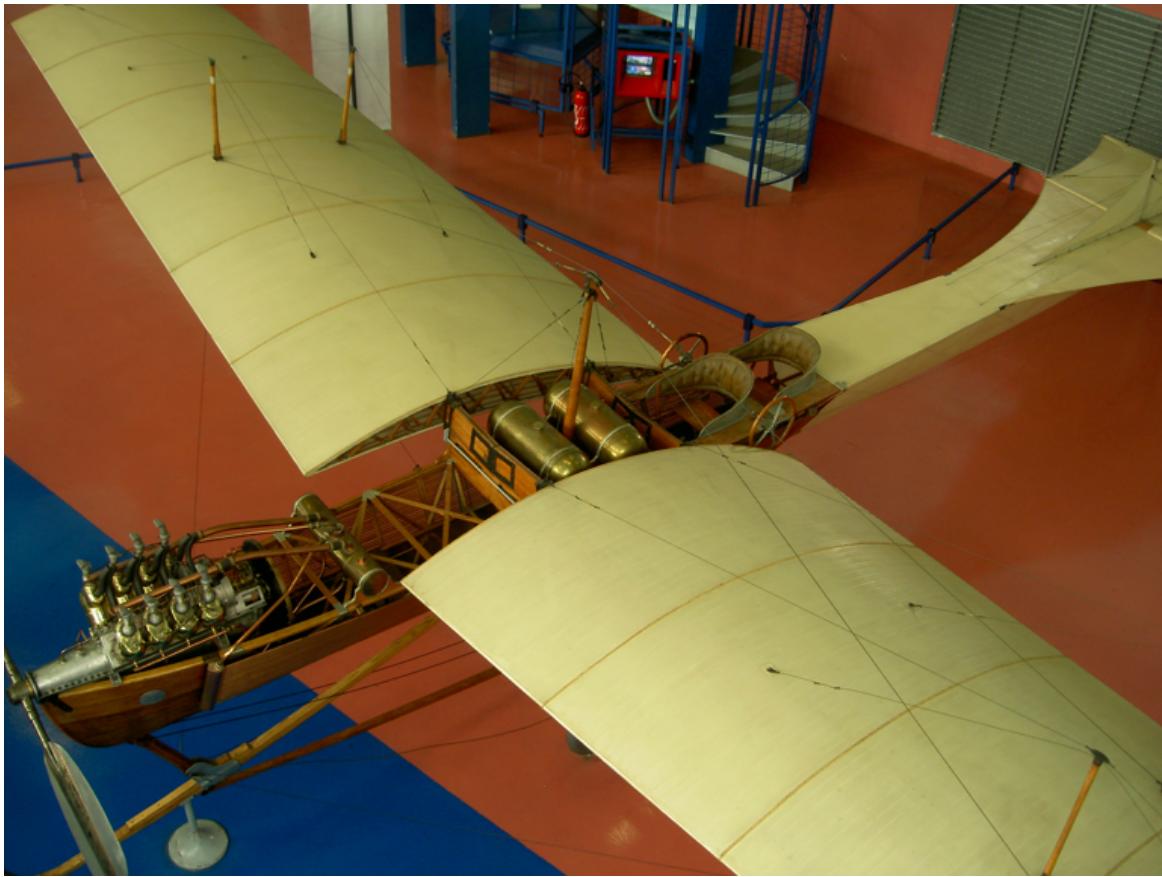
In a sentence, my argument is that the complexity and mysteriousness and efficiency of the human brain (compared to artificial neural nets) is *almost zero evidence* that building [TAI](#) will be difficult, because evolution typically makes things complex and mysterious and efficient, even when there are simple, easily understood, inefficient designs that work almost as well (or even better!) for human purposes.

In slogan form: ***If all we had to do to get TAI was make a simple neural net 10x the size of my brain, my brain would still look the way it does.***

The case of birds & planes illustrates this point nicely. Moreover, it is also a precedent for several other short-timelines talking points, such as the human-brain-human-lifetime (HBHL) anchor.

Plan:

1. Illustrative Analogy
2. Exciting Graph
3. Analysis
 1. Extra brute force can make the problem a lot easier
 2. Evolution produces complex mysterious efficient designs by default, even when simple inefficient designs work just fine for human purposes.
 3. What's bogus and what's not
 4. Example: Data-efficiency
4. Conclusion
5. Appendix



1909 French military plane, the Antionette VII.

By Deep silence (Mikaël Restoux) - Own work (Bourget museum, in France), CC BY 2.5, <https://commons.wikimedia.org/w/index.php?curid=1615429>

Illustrative Analogy

AI timelines, from our current perspective	Flying machine timelines, from the perspective of the late 1800's:
<p>Shorty: Human brains are giant neural nets. This is reason to think we can make human-level AGI (or at least AI with strategically relevant skills, like politics and science) by making giant neural nets.</p>	<p>Shorty: Birds are winged creatures that paddle through the air. This is reason to think we can make winged machines that paddle through the air.</p>
<p>Longs: Whoa whoa, there are loads of important differences between brains and artificial neural nets: <i>[what follows is a direct quote from the objection a friend raised when reading an early draft of this post!]</i></p> <ul style="list-style-type: none"> - During training, deep neural nets 	<p>Longs: Whoa whoa, there are loads of important differences between birds and flying machines:</p>

<p>use some variant of backpropagation. My understanding is that the brain does something else, closer to Hebbian learning. (Though I vaguely remember at least one paper claiming that maybe the brain does something that's similar to backprop after all.)</p> <p>- It's at least possible that the wiring diagram of neurons plus weights is too coarse-grained to accurately model the brain's computation, but it's all there is in deep neural nets. If we need to pay attention to glial cells, intracellular processes, different neurotransmitters etc., it's not clear how to integrate this into the deep learning paradigm.</p> <p>- My impression is that several biological observations on the brain don't have a plausible analog in deep neural nets: growing new neurons (though unclear how important it is for an adult brain), "repurposing" in response to brain damage, ...</p>	<p>- Birds paddle the air by flapping, whereas current machine designs use propellers and fixed wings.</p> <p>- It's at least possible that the anatomical diagram of bones, muscles, and wing surfaces is too coarse-grained to accurately model how a bird flies, but that's all there is to current machine designs (replacing bones with struts and muscles with motors, that is). If we need to pay attention to the percolation of air through and between feathers, micro-eddies in the air sensed by the bird and instinctively responded to, etc. it's not clear how to integrate this into the mechanical paradigm.</p> <p>- My impression is that several biological observations of birds don't have a plausible analog in machines: Growing new feathers and flesh (though unclear how important this is for adult birds), "repurposing" in response to damage ...</p>
---	---

<p>Shorty: The key variables seem to be size and training time. Current neural nets are tiny; the biggest one is only one-thousandth the size of the human brain. But they are rapidly getting bigger.</p> <p>Once we have enough compute to train neural nets as big as the human brain for as long as a human lifetime (HBHL), it should in principle be possible for us to build HLAGI. No doubt there will be lots of details to work out, of course. But that shouldn't take more than a few years.</p>	<p>Shorty: The key variables seem to be engine-power and engine weight. Current motors are not strong & light enough, but they are rapidly getting better.</p> <p>Once the power-to-weight ratio of our motors surpasses the power-to-weight ratio of bird muscles, it should be in principle possible for us to build a flying machine. No doubt there will be lots of details to work out, of course. But that shouldn't take more than a few years.</p>
---	---

<p>Longs: Bah! I don't think we know what the key variables are. For example, biological brains seem to be able to learn faster, with less data, than artificial neural nets. And we don't know why.</p>	<p>Longs: Bah! I don't think we know what the key variables are. For example, birds seem to be able to soar long distances without flapping their wings at all, and we still haven't figured out how they do it. Another example: We still don't know how birds manage</p>
---	---

Besides, “there will be lots of details to work out” is a huge understatement. It took evolution billions of generations of billions of individuals to produce humans. What makes you think we’ll be able to do it quickly? It’s plausible that actually we’ll have to do it the way evolution did it, i.e. meta-learn, i.e. evolve a large population of HBHLs, over many generations. (Or, similarly, train a neural net with a big batch size and a horizon length of a lifetime).

And even if you think we’ll be able to do it substantially quicker than evolution did, it’s pretty presumptuous to think we could do it quickly enough that the HBHL milestone is relevant for forecasting.

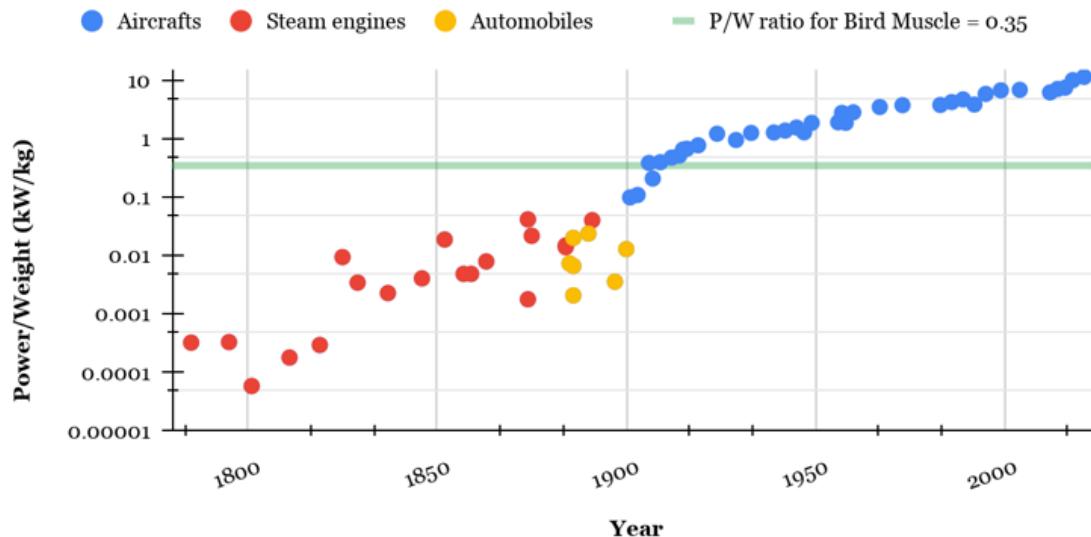
to steer through the air without crashing (flight stability & control).

Besides, “there will be lots of details to work out” is a huge understatement. It took evolution billions of generations of billions of individuals to produce birds. What makes you think we’ll be able to do it quickly? It’s plausible that actually we’ll have to do it the way evolution did it, i.e. meta-design, i.e. evolve a large population of flying machines, tweaking our blueprints each generation of crashed machines to grope towards better designs.

And even if you think we’ll be able to do it substantially quicker than evolution did, it’s pretty presumptuous to think we could do it quickly enough that the date our engines achieve power/weight parity with bird muscle is relevant for forecasting.

Exciting Graph

Power to Weight ratios (kW/kg) of Engines from the late 18th century to present



This data shows that Shorty was entirely correct about forecasting heavier-than-air flight. (For details about the data, see appendix.) Whether Shorty will also be correct about forecasting TAI remains to be seen.

In some sense, Shorty has already made two successful predictions: I started writing this argument before having *any* of this data; I just had an intuition that power-to-weight is the key variable for flight and that therefore we probably got flying machines shortly after having comparable power-to-weight as bird muscle. Halfway through the first draft, I googled and confirmed that yes, the Wright Flyer's motor was close to bird muscle in power-to-weight. Then, while writing the second draft, I hired an RA, Amogh Nanajjar, to collect more data and build this graph. As expected, there was a trend of power-to-weight improving over time, with flight happening right around the time bird-muscle parity was reached.

I had previously heard from a friend, who read a book about the invention of flight, that the Wright brothers were the first because they (a) studied birds and learned some insights from them, and (b) did a bunch of trial and error, rapid iteration, etc. (e.g. in wind tunnels). The story I heard was all about the importance of insight and experimentation--but this graph seems to show that the key constraint was engine power-to-weight. Insight and experimentation were important for determining who invented flight, but not for determining which decade flight was invented in.

Analysis

Part 1: Extra brute force can make the problem a lot easier

One way in which compute can substitute for insights/algorithms/architectures/ideas is that you can use compute to search for them. But there is a different and arguably more important way in which compute can substitute for insights/etc.: Scaling up the key variables, so that the problem becomes easier, so that fewer insights/etc. are needed.

For example, with flight, the problem becomes easier the more power/weight ratio your motors have. Even if the Wright brothers didn't exist and nobody else had their insights, eventually we would have achieved powered flight anyway, because when our engines are 100x more powerful for the same weight, we can use extremely simple, inefficient designs. (For example, imagine a u-shaped craft with a low center of gravity and helicopter-style rotors on each tip. Add a third, smaller propeller on a turret somewhere for steering. EDIT: Oops, lol, [I'm actually wrong about this.](#) Keeping center of gravity low doesn't help. Welp, this is embarrassing.)

With neural nets, we have plenty of evidence now that bigger = better, with theory to back it up. Suppose the problem of making human-level AGI with HBHL levels of compute is really difficult. OK, 10x the parameter count and 10x the training time and try again. Still too hard? Repeat.

Note that I'm *not* saying that if you take a particular design that doesn't work, and make it bigger, it'll start working. (If you took Da Vinci's flying machine and made the engine 100x more powerful, it would not work). Rather, I'm saying that the problem of finding a design that works gets qualitatively easier the more parameters and training time you have to work with.

Finally, remember that human-level AGI is not the only kind of TAI. Sufficiently powerful R&D tools would work, as would sufficiently powerful [persuasion tools](#), as might something that is agenty and inferior to humans in some ways but vastly superior in others.

Part 2: Evolution produces complex mysterious efficient designs by default, even when simple inefficient designs work just fine for human purposes.

Suppose that actually all we have to do to get TAI is something fairly simple and obvious, but with a neural net 10x the size of my (actual) brain and trained for 10x longer. In this world, does the human brain look any different than it does in the actual world?

No. Here is a nonexhaustive list of reasons why evolution would evolve human brains to look like they do, with all their complexity and mysteriousness and efficiency, even if the same capability levels could be reached with 10x more neurons and a very simple architecture. Feel free to skip ahead if you think this is obvious.

1. In general, evolved creatures are complex and mysterious to us, even when simple and human-comprehensible architectures work fine. Take birds, for example: As mentioned before, all the way up to the Wright brothers there were a lot of very basic things about birds that were still not understood. From [this article](#): “They watched buzzards glide from horizon to horizon without moving their wings, and guessed they must be sucking some mysterious essence of upness from the air. Few seemed to realize that air moves up and down as well as horizontally.” I don’t know much about ornithology but I’d be willing to bet that there were lots of important things discovered about birds *after* airplanes already existed, and that there are *still* at least a few remaining mysteries about how birds fly. (Spot check: Yep, the [history of ornithopters](#) page says “...the development of comprehensive aerodynamic theory for flapping remains an outstanding problem...”). And of course evolved creatures are often more efficient in various ways than their still-useful engineered counterparts.
2. Making the brain 10x bigger would be enormously costly to fitness, because it would cost 10x more energy and restrict mobility (not to mention the difficulties of getting through the birth canal!) Much better to come up with clever modules, instincts, optimizations, etc. that achieve the same capabilities in a smaller brain.
3. Evolution is heavily constrained on training data, perhaps even more than on brain size. It can’t just evolve the organism to have 10x more training data, because longer-lived organisms have more opportunities to be eaten or suffer accidents, especially in their 10x-longer childhoods. Far better to hard-code some behaviors as instincts.
4. Evolution gets clever optimizations and modules and such “for free” in some sense. Since it is evolving millions of individuals for millions of generations anyway, it’s not a big deal for it to perform massive search and gradient descent through architecture-space.
5. Completely blank slate brains (i.e. extremely simple architecture, no instincts or finely tuned priors) would be unfit even if they were highly capable because they wouldn’t be aligned to evolution’s values (i.e. reproduction.) Perhaps most of the complexity in the human brain--the instincts, inbuilt priors, and even most of the modules--isn’t for capabilities at all, [but rather for alignment](#).

Part 3: What's bogus and what's not

The general pattern of argument I think is bogus is:

The brain has property X, which seems to be important to how it functions. We don't know how to make AI's with property X. It took evolution a long time to make brains have property X. This is reason to think TAI is not near.

As argued above, if TAI *is* near, there should still be *many* X which are important to how the brain functions, which we don't know how to reproduce in AI, and which it took evolution a long time to produce. So rattling off a bunch of X's is basically zero evidence against TAI being near.

Put differently, here are two objections any particular argument of this type needs to overcome:

1. TAI does not actually require X (analogous to how airplanes didn't require anywhere near the energy-efficiency of birds, nor the ability to soar, nor the ability to flap their wings, nor the ability to take off from unimproved surfaces... the list goes on)
2. We'll figure out how to get property X in AIs soon after we have the other key properties (size and training time), because (a) we can do search, like evolution did but much more efficient, (b) we can increase the other key variables to make our design/search problem easier, and (c) we can use human ingenuity & biological inspiration. Historically there is plenty of precedent for the previous three factors being strong enough; see e.g. the case of powered flight.

This reveals how the arguments could be reformulated to become non-bogus! They need to argue (a) that X is probably necessary for TAI, *and* (b) that X isn't something that we'll figure out fairly quickly once the key variables of size and training time are surpassed.

In some cases there are decent arguments to be made for both (a) and (b). I think efficiency is one of them, so I'll use that as my example below.

Part 4: Example: Data-efficiency

Let's work through the example of data-efficiency. A bad version of this argument would be:

Humans are much more data-efficient learners than current AI systems. Data-efficiency is very important; any human who learned as inefficiently as current AI would basically be mentally disabled. This is reason to think TAI is not near.

The rebuttal to this bad argument is:

If birds were as energy-inefficient as planes, they'd be disabled too, and would probably die quickly. Yet planes work fine. (See Table 1 from [this AI Impacts page](#)) Even if TAI is near, there are going to be lots of X's that are important for the brain, that we don't know how to make in AI yet, but that are either unnecessary for TAI or not too difficult to get once we have the other key variables. So even if TAI is near, I should expect to hear people going around pointing out various X's and claiming that this is reason to think TAI is far away. You haven't done anything to convince me that this isn't what's happening with X = data-efficiency.

However, I do think the argument can be reformulated and expanded to become good. Here's a sketch, inspired by Ajeya Cotra's argument [here](#).

We probably can't get TAI without figuring out how to make AIs that are as data-efficient as humans. It's true that there are some useful tasks for which there is plenty of data--like call center work, or driving trucks--but AIs that can do these tasks won't be transformative. Transformative AI will be doing things like managing corporations, leading armies, designing new chips, and writing AI theory publications. Insofar as AI learns more slowly than humans, by the time it accumulates enough experience doing one of these tasks, (a) the world would have changed enough that its skills would be obsolete, and/or (b) it would have made a lot of expensive mistakes in the meantime.

Moreover, we probably won't figure out how to make AIs that are as data-efficient as humans for a long time--decades at least. This is because 1. We've been trying to figure this out for decades and haven't succeeded, and 2. Having a few orders of magnitude more compute won't help much. Now, to justify point #2: Neural nets actually do get more data-efficient as they get bigger, but we can plot the trend and see that they will still be less data-efficient than humans when they are a few orders of magnitude bigger. So making them bigger won't be enough, we'll need new architectures/algorithms/etc. As for using compute to search for architectures/etc., that might work, but given how long evolution took, we should think it's unlikely that we could do this with only a few orders of magnitude of searching--probably we'd need to do many generations of large population size. (We could also think of this search process as analogous to typical deep learning training runs, in which case we should expect it'll take many gradient updates with large batch size.) Anyhow, there's no reason to think that data-efficient learning is something you need to be human-brain-sized to do. If we can't make our tiny AIs learn efficiently after several decades of trying, we shouldn't be able to make big AIs learn efficiently after just one more decade of trying.

I think this is a good argument. Do I buy it? Not yet. For one thing, I haven't verified whether the claims it makes are true, I just made them up as plausible claims which would be persuasive to me if true. For another, some of the claims actually seem false to me. Finally, I suspect that in 1895 someone could have made a similarly plausible argument about energy efficiency, and another similarly plausible argument about flight control, and both arguments would have been wrong: Energy efficiency turned out to be insufficiently necessary, and flight control turned out to be insufficiently difficult!

Conclusion

What I am not saying: I am not saying that the case of birds and planes is strong evidence that TAI will happen once we hit the HBHL milestone. I do think it is evidence, but it is weak evidence. (For my all-things-considered view of how many orders of magnitude of compute it'll take to get TAI, see future posts, or ask me.) I would like to see a more thorough investigation of cases in which humans attempt to design something that has an obvious biological analogue. It would be interesting to see if the case of flight was typical. Flight being typical would be strong evidence for short timelines, I think.

What I am saying: I am saying that many common anti-short-timelines arguments are bogus. They need to do much more than just appeal to the complexity/mysteriousness/efficiency of the brain; they need to argue that some property X is both necessary for TAI and not about to be figured out for AI anytime soon, not even after the HBHL milestone is passed by several orders of magnitude.

Why this matters: In my opinion the biggest source of uncertainty about AI timelines has to do with how much “special sauce” is necessary for making transformative AI. As [jylin04 puts it](#),

A first and frequently debated crux is whether we can get to TAI from end-to-end training of models specified by relatively few bits of information at initialization, such as neural networks initialized with random weights. OpenAI in particular seems to take the affirmative view[^3], while people in academia, especially those with more of a neuroscience / cognitive science background, seem to think instead that we'll have to hard-code in lots of inductive biases from neuroscience to get to AGI [^4].

In my words: Evolution clearly put lots of special sauce into humans, and took millions of generations of millions of individuals to do so. *How much special sauce will we need to get TAI?*

Shorty is one end of a spectrum of disagreement on this question. Shorty thinks the amount of special sauce required is small enough that we'll “work out the details” within a few years of having the key variables (size and training time). At the other end of the spectrum would be someone who thought that the amount of special sauce required is similar to the amount found in the brain. Longs is in the middle. Longs thinks the amount of special sauce required is large enough that the HBHL milestone isn't particularly relevant to timelines; we'll either have to brute-force search for the special sauce like evolution did, or have some brilliant new insights, or mimic the brain, etc.

This post rebutted common arguments against Shorty's position. It also presented weak evidence in favor of Shorty's position: the precedent of birds and planes. In future posts I'll say more about what I think the probability distribution over amount-of-special-sauce-needed should be and why.

Acknowledgements: Thanks to my RA, Amogh Nanjajjar, for compiling the data and building the graph. Thanks to Kaj Sotala, Max Daniel, Lukas Gloor, and Carl Shulman for comments on drafts.

Appendix

Some footnotes:

1. I didn't say anything about why we might think size and training time are the key variables, or even what “key variables” means. Hopefully I'll get a chance in the comments or in subsequent posts.

2. I deliberately left vague what “training time” means and what “size” means. Thus, I’m not committing myself to any particular way of calculating the HBHL milestone yet. I’m open to being convinced that the HBHL milestone is farther in the future than it might seem.
3. Persuasion tools, even very powerful ones, wouldn’t be TAI by the standard definition. However they would constitute a [potential-AI-induced-point-of-no-return](#), so they still count for timelines purposes.
4. This “How much special sauce is needed?” variable is very similar to Ajeya Cotra’s variable “how much compute would lead to TAI given 2020’s algorithms.”

Some bookkeeping details about the data:

1. This dataset is not complete. Amogh did a reasonably thorough search for engines throughout the period (with a focus on stuff before 1910) but was unable to find power or weight stats for many of the engines we heard about. Nevertheless I am reasonably confident that this dataset is representative; if an engine was significantly better than the others of its time, probably this would have been mentioned and Amogh would have flagged it as a potential outlier.
2. Many of the points for steam engine power/weight should really be bumped up slightly. This is because most of the data we had was for the weight of the entire locomotive of a steam-powered train, rather than just the steam engine part. I don’t know what fraction of a locomotive is non-steam-engine but 50% seems like a reasonable guess. I don’t think this changes the overall picture much; in particular, the two highest red dots do not need to be bumped up at all (I checked).
3. The birds bar is the power/weight ratio for the muscles of a particular species of bird, reported by [this source](#), which reports the power/weight for a particular species of bird. Amogh has done a bit of searching and doesn’t think muscle power/weight is significantly different for other species of bird. Seems plausible to me; even if the average bird has muscles that are twice (or half) as powerful-per-kilogram, the overall graph would look basically the same.
4. I attempted to find estimates of human muscle power-to-weight ratio; it gets smaller the more tired the muscles get, but at peak performance for fit individuals it seems to be about an order of magnitude less than bird muscle. ([This chart](#) lists power-to-weight ratio for human cyclists, which according [to this](#) are probably about half muscle, so look at the left-hand column and double it.) Interestingly, this means that the engines of the first flying machines were possibly the first engines to be substantially better than human flapping/pedaling as a source of flying-machine power.
5. EDIT Gaaah I forgot to include a link to the data! [Here's the spreadsheet](#).

[Link] Still Alive - Astral Codex Ten

This is a linkpost for <https://astralcodexten.substack.com/p/still-alive>

I.

*This was a triumph
I'm making a note here, huge success*

No, seriously, it was awful. I [deleted my blog](#) of 1,557 posts. I wanted to protect my privacy, but I ended up with articles about me in *New Yorker*, *Reason*, and *The Daily Beast*. I wanted to protect my anonymity, but I Streisand-Effectd myself, and a bunch of trolls went around posting my real name everywhere they could find. I wanted to avoid losing my day job, but ended up quitting so they wouldn't be affected by the fallout. I lost a five-digit sum in advertising and Patreon fees. I accidentally sent about three hundred emails to each of five thousand people in the process of trying to put my blog back up.

I had, not to mince words about it, a really weird year.

The [first post](#) on Scott Alexander's new blog on Substack, [Astral Codex Ten](#).

Covid 1/7: The Fire of a Thousand Suns

I'll summarize everything up front.

Might as well [get on with it](#).

Let's do the numbers.

The Numbers

Predictions

Prediction last week: 14.3% positive rate on 9.7 million tests, and an average of 2,500 deaths, again with wide error bars.

Results: 16.4% positive rate on 9.3 million tests, and an average of 2,657 deaths.

The phase shift on 12/30, in the wake of Christmas, seems to have been real, giving us the clear holiday bump we did not see from previous holidays.

That was both the very bad outcome for infections I was worried about, and also not high on my list of things to be concerned or furious about this week.

For deaths, my estimate was lower than it should have been and I should have assumed a full reversion, so on reflection it's my mistake rather than especially bad news.

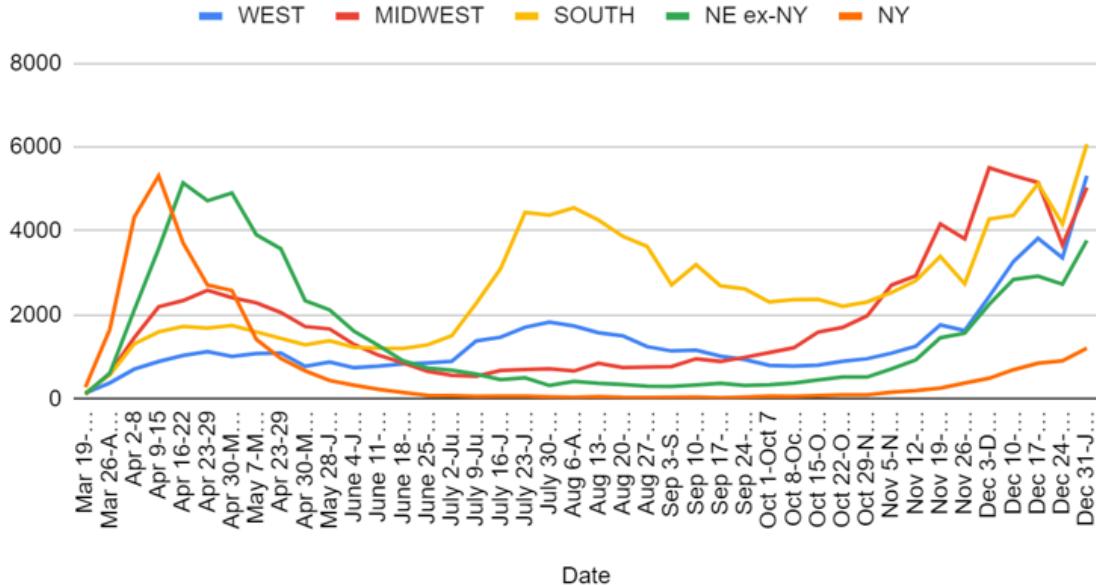
The new strain is not yet prevalent enough to be noticeably impacting the numbers.

Prediction: 17.0% positive rate on 9.5 million tests, and an average of 2,800 deaths. The holidays are over, there will be some fallout, with things getting slightly worse, but with the main boost in deaths from Christmas mostly coming later.

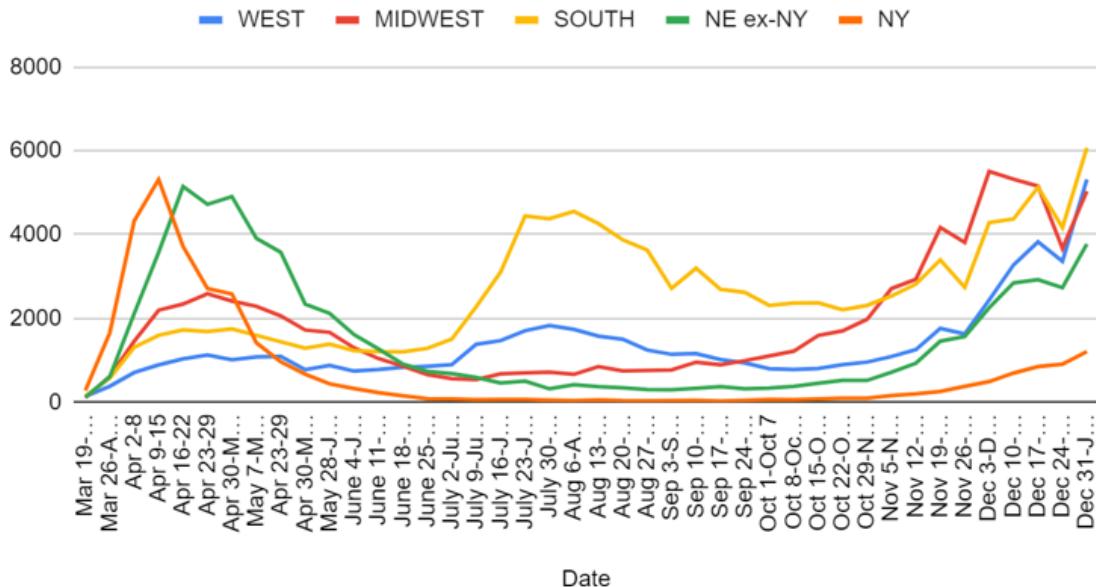
Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 29-Nov 4	956	1977	2309	613
Nov 5-Nov 11	1089	2712	2535	870
Nov 12-Nov 18	1255	2934	2818	1127
Nov 19-Nov 25	1761	4169	3396	1714
Nov 26-Dec 2	1628	3814	2742	1939
Dec 3-Dec 9	2437	5508	4286	2744
Dec 10-Dec 16	3278	5324	4376	3541
Dec 17-Dec 23	3826	5158	5131	3772
Dec 24-Dec 30	3363	3668	4171	3640
Dec 31-Jan 6	5320	5036	6072	4986

Deaths by Region



Deaths by Region

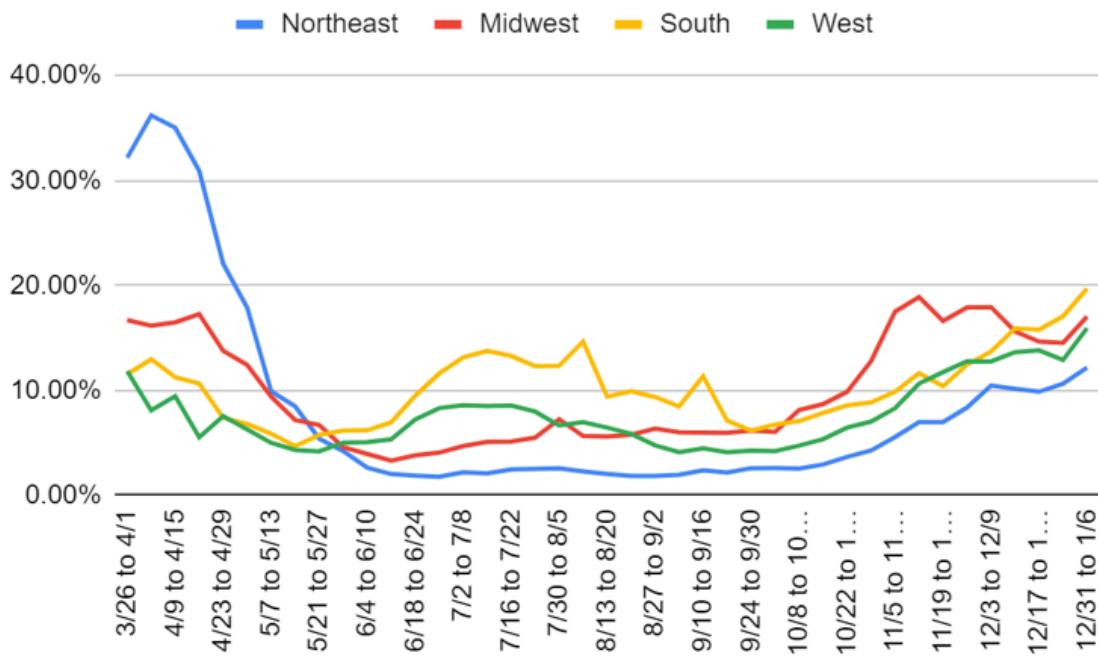


This is the one place it's not as bad as it looks. The data source for these numbers is Wikipedia, which shows a relatively large amount of shifting of deaths from last week into this week.

If we assume that a lot of this week's deaths were actually last week, that explains much of the increase. It's not *good* news or anything, it's definitely bad news, but it is not full on terrifying like it would be if we didn't know about Christmas.

We still should expect further increases in the next few weeks before the tide likely temporarily turns.

Positive Test Percentages

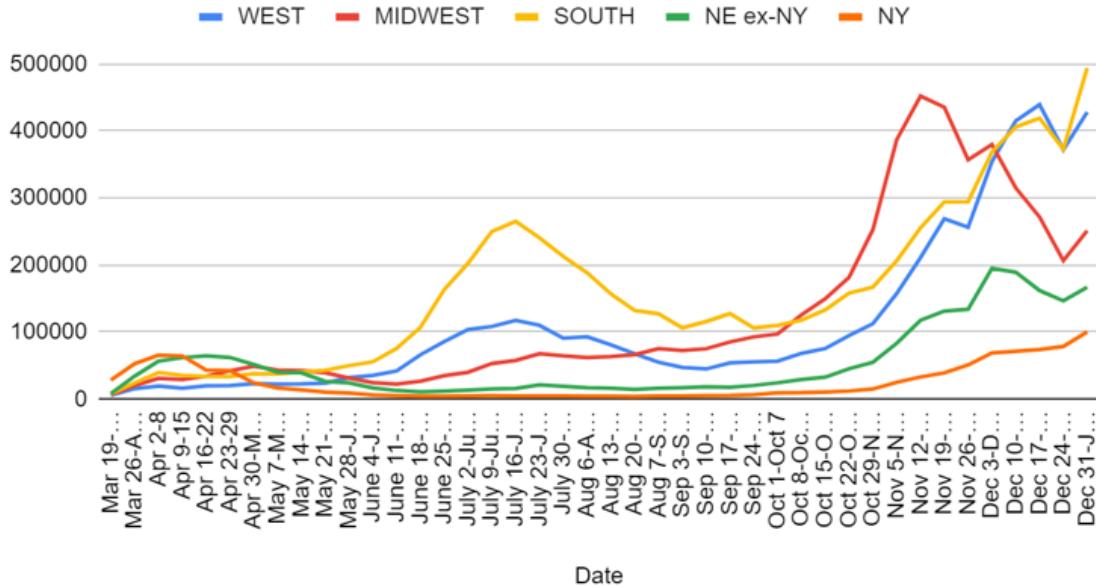


Percentages	Northeast	Midwest	South	West
11/5 to 11/11	5.56%	17.51%	9.89%	8.31%
11/12 to 11/18	6.99%	18.90%	11.64%	10.66%
11/19 to 11/25	7.00%	16.62%	10.41%	11.75%
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%
12/3 to 12/9	10.47%	17.94%	13.70%	12.76%
12/10 to 12/16	10.15%	15.63%	15.91%	13.65%
12/17 to 12/23	9.88%	14.65%	15.78%	13.82%
12/24 to 12/30	10.65%	14.54%	17.07%	12.90%
12/31 to 1/6	12.18%	17.03%	19.69%	15.94%

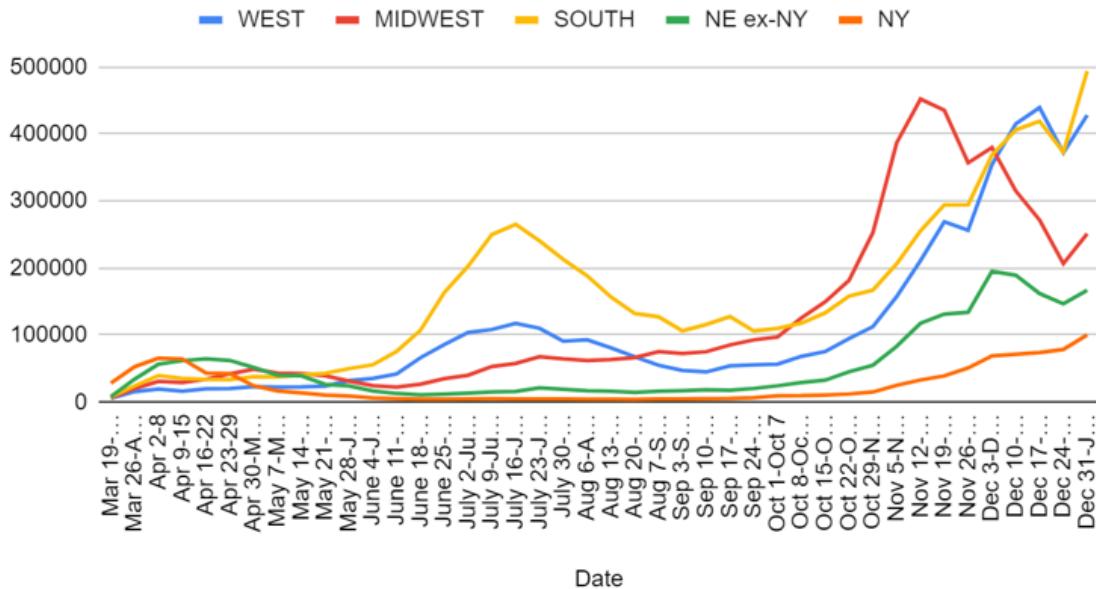
No way to sugar coat this one. Everything is pointed in the wrong direction in a clear phase shift. The good news is that it could be a one-time move from Christmas, and thus we can still plausibly expect things to remain stable from this point forward without waiting for further control systems to kick in. I hope to start seeing slow improvements soon, but I must admit that the overall graph is not as encouraging of that as I thought a few weeks ago, especially outside of the Midwest, and the turning point may be farther off than we'd like. The Midwest is still probably past its peak for this wave, even if everyone else has to wait a bit.

Positive Tests

Positive Tests by Region



Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Nov 12-Nov 18	211,222	452,265	255,637	150,724
Nov 19-Nov 25	269,230	435,688	294,230	170,595
Nov 26-Dec 2	256,629	357,102	294,734	185,087
Dec 3-Dec 9	354,397	379,823	368,596	263,886
Dec 10-Dec 16	415,220	315,304	406,353	260,863

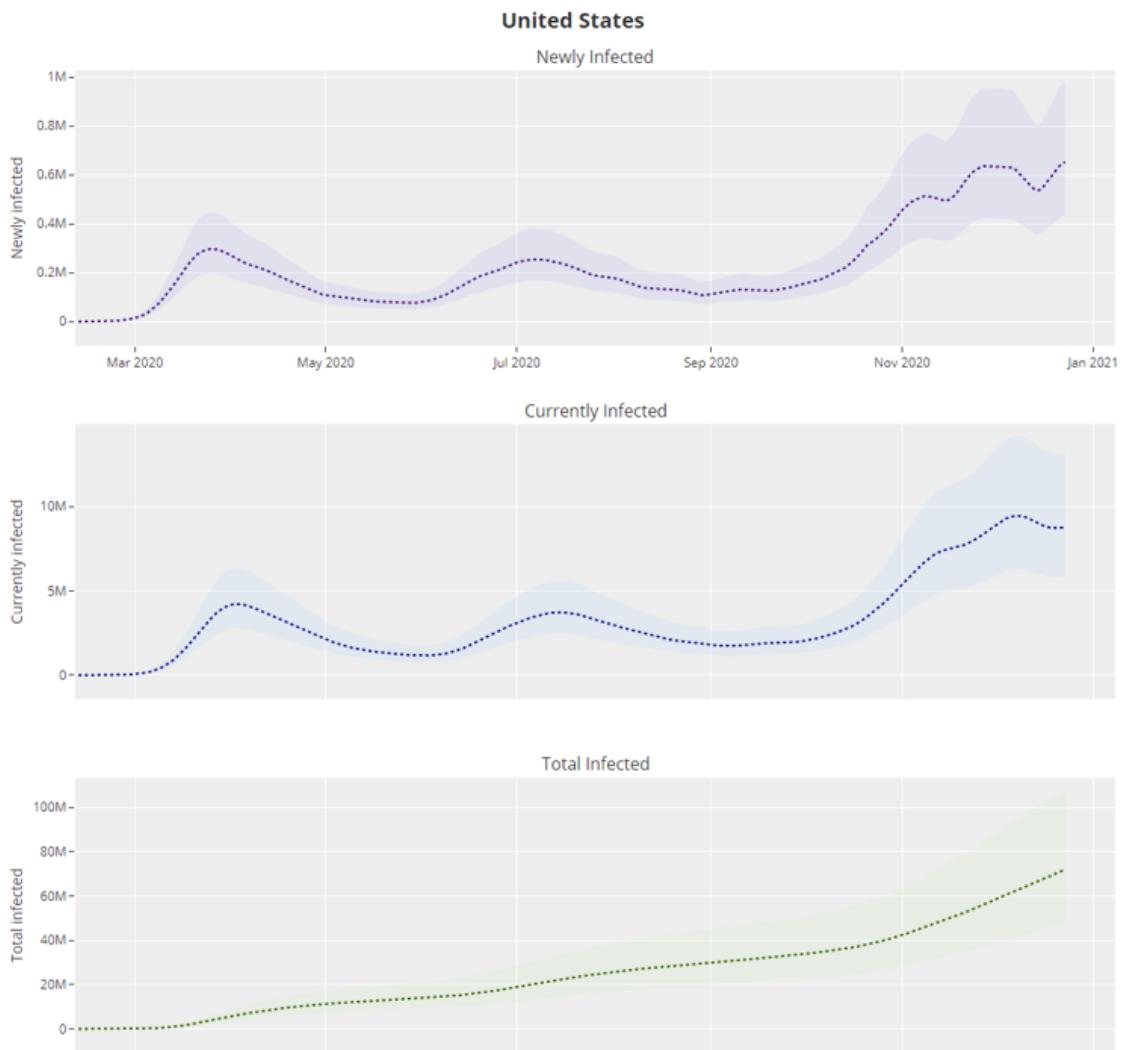
Dec 17-Dec 23	439,493	271,825	419,230	236,264
Dec 24-Dec 30	372,095	206,671	373,086	225,476
Dec 31-Jan 6	428,407	251,443	494,090	267,350

Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Nov 5-Nov 11	8,290,417	10.8%	1,059,559	2.4%	3.16%
Nov 12-Nov 18	9,040,426	12.4%	1,155,670	2.9%	3.50%
Nov 19-Nov 25	10,419,059	11.8%	1,373,751	2.9%	3.88%
Nov 26-Dec 2	9,747,026	11.8%	1,287,010	4.0%	4.23%
Dec 3-Dec 9	10,465,254	13.9%	1,411,142	4.9%	4.67%
Dec 10-Dec 16	10,701,134	13.9%	1,444,725	4.9%	5.12%
Dec 17-Dec 23	10,716,189	13.7%	1,440,770	5.1%	5.57%
Dec 24-Dec 30	9,082,257	13.9%	1,303,286	6.0%	5.95%
Dec 31-Jan 6	9,317,985	16.4%	1,365,473	7.3%	6.42%

The failure of test counts to recover here is maddening. We need to be adding more capacity, not taking multiple weeks off. The positive test percentages speak for themselves.

Covid Machine Learning Project



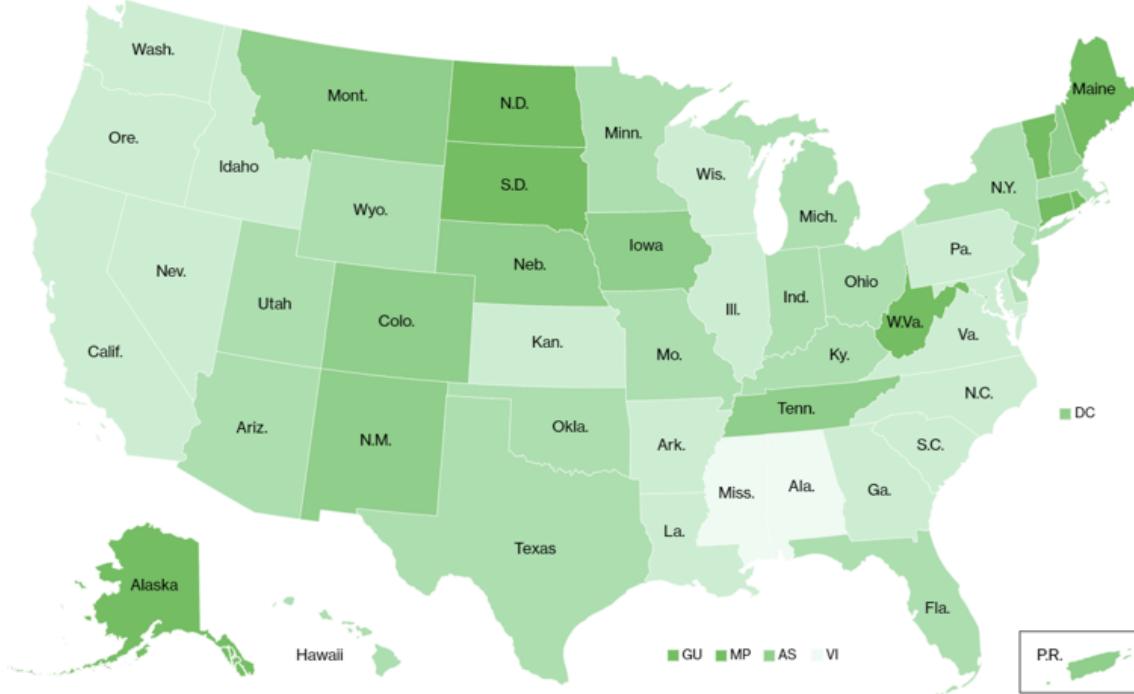
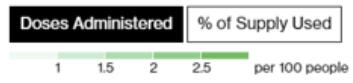
The code doesn't know about Thanksgiving so it has a dip in infections in late November that clearly is not real, and now it is back to thinking R_0 is over 1.1. Meanwhile, it had total infections on December 23 at 21.7%, versus a belief of 20.3% on December 16 last week, which it now thinks was 20.4% back then. Last week's belief that things were steadily improving has been fixed by additional data to smooth out Thanksgiving.

Vaccinations

Vaccinations in the U.S. began Dec. 14 with health-care workers, and so far **5.48 million doses** have been given, according to a state-by-state tally by Bloomberg and data from the Centers for Disease Control and Prevention.

Vaccines Across America

Across the U.S., 1.7 doses have been administered for every 100 people, and 32% of the shots distributed to states have been administered



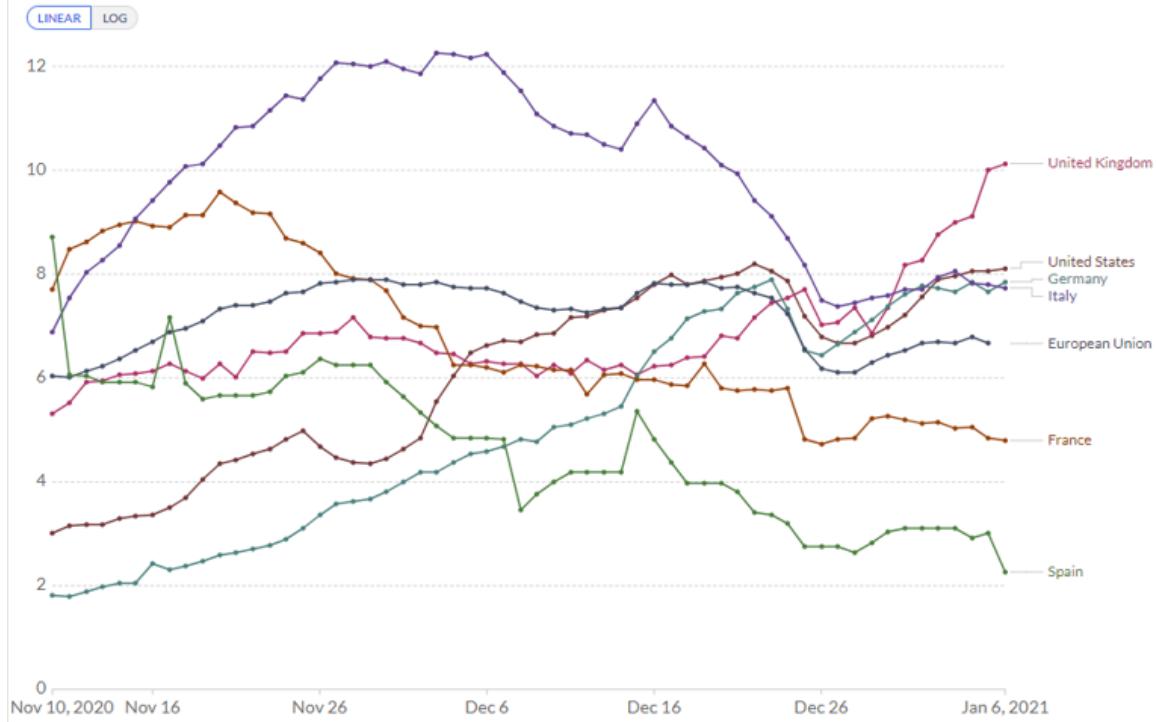
5.48 million doses so far, after almost a month. Pathetic. Scandalous. Insane. Shots. In. Arms. Now.

Europe

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.

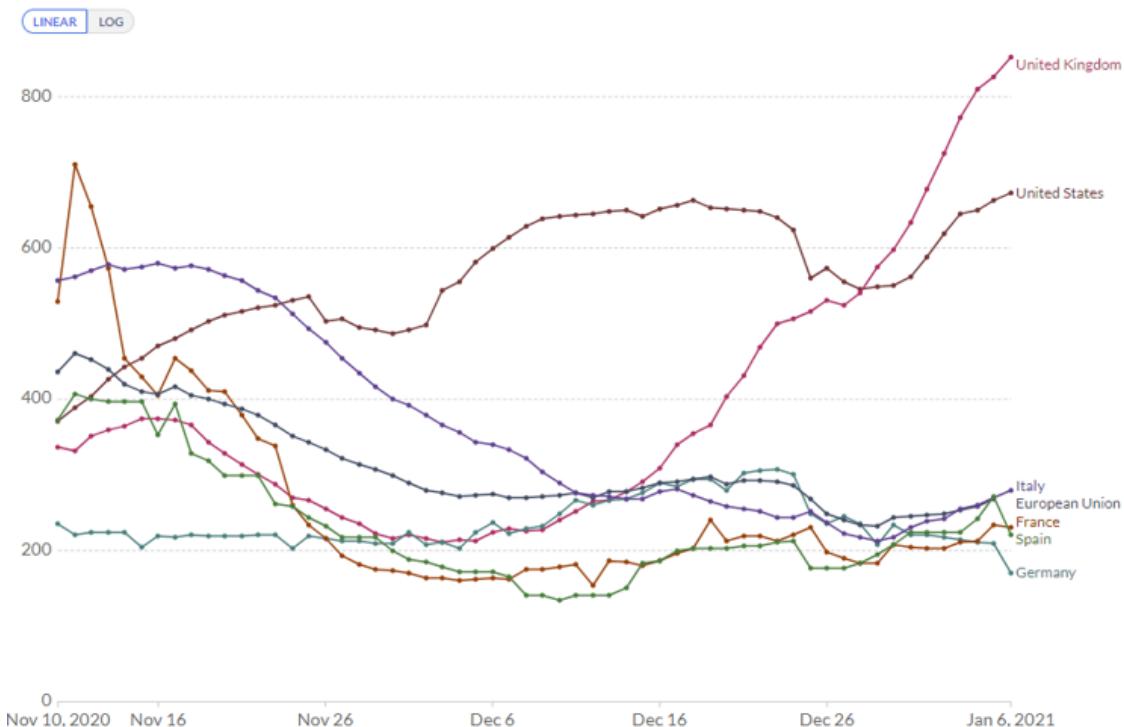
Our World
in Data



Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World
in Data



One of these lines is not like the others. The control system in the United Kingdom is failing.

The English Strain: Are We F*ed? Is it Over?**

I found this comment very helpful, so I'll quote it in full, it's the best way I know of to estimate exactly how far along we are:

M

Mitchell Powell says:

January 2, 2021 at 7:29 pm (Edit)

When it comes to estimating the number of cases of Covid-UK in the US right now, here's my best attempt. I have very little confidence that it's in the right ballpark, but maybe folks here could let me know if there's a better way to do this. Basically I'm multiplying two numbers: an estimate of what fraction of US cases are "S gene dropouts", and then an estimate of what fraction of S gene dropouts are Covid-UK.

I start with the Helix Research Team's blog, at a post December 23rd (https://blog.helix.com/sars-cov2_uk-variant/), where they report:

"We observe a rise in S gene dropout starting in early October, with 0.25% of our daily COVID-19-positive tests exhibiting this pattern during the first week. The rate of positive samples with S gene dropout has grown slowly over time, with last week exhibiting the highest level yet at 0.5% of COVID-19-positive tests that are consistent with the H69del/V70del variant."

So, for the moment, let's say there's something like 0.5% of US Covid cases showing S gene dropout in December.

Next, we can look at their testing of a subset of S gene dropout cases, in which they tested 31 cases and found Covid-UK in 4 of them (<https://blog.helix.com/sars-cov-2-uk-variant-b-1-1-7-in-the-us-four-cases-identified-in-florida-and-california/>):

"Of the initial 31 samples with high quality sequencing data, four match the B.1.1.7 strain."

Now, ignoring for a moment any problems with how truly random any of these samples are, let's multiply 0.5% by 4/31, and we get 0.065% of US Covid cases having the UK strain.

Next, there's the fact that those numbers seem to come from about mid-to-late December. Given a bit less than two weeks, and the previously observed pattern of the new B.1.1.7 strain doubling in prevalence about once a week in the UK, we can then gingerly work our way forward and say that perhaps 0.2% of new cases in the US as of the beginning of 2021 are Covid-UK.

If that's the case, we're about 9 doublings away from Covid-UK reaching equal prevalence with Covid-Classic, and if it were to double once a week, that would result in things getting really hairy starting in early March.

Now, I know that there are possible objections galore to the calculation that 0.2% of current US cases are of the B.1.1.7 strain. But is there a better way to calculate an estimate available at the moment?

[Reply](#)

This is not a comforting picture. In my very quick and dirty first highly abstract model, with 50% more infectiousness in the new strain than the old, this would have us be 0.22% of the newstrain during the year's first cycle, 2% by the end of January, 16% by the end of

February, 68% by the end of March. In the simplest possible model that you should not in any way take seriously, the end of May would be the peak at roughly 225% of the infections per day we have now, July 1 we'd be back to where we are now, then a very rapid decline after that.

What evidence do we have about the infectiousness of the new strain?

Denmark sequences the new strain, as a reason not to worry, a play in three acts:

Act One (posted 12/31):

[...] dotchart 3d Ø < 6 >

About the English strain part. This seems very one-sided to me. Only the evidence in favor of it being more infectious is being shared by Zvi.

1. In Denmark 12.5% of all positive COVID are randomly sequenced. The English strain has been observed in the samples since 14 November. However it remains stable at 0.5-1% of samples. This makes me update in the direction that it is not 70% more infectious compared to other strains. Otherwise we should see the strain make up a larger proportion of positive samples over time. (https://files.ssi.dk/engelsk_virusvariant-28122020)

Asking "[What evidence filtered evidence?](#)" in situations like the new strain is important. I do my best to include all the evidence that's worth including, but the degree to which I look for it might vary, and I could plausibly be doing things like not mentioning places where there is an absence of evidence that constitutes at least some evidence of absence. Please do keep calling me out on such mistakes by pointing out what's missing, with or without the principle of charity.

Act Two (posted 12/31 in reply):

[...] jsteinhardt 3d Ø < 14 >

I don't think it's correct to say that it remains stable at 0.5-1% of samples in Denmark. There were 13 samples of the new variant last week, vs. only 3 two weeks ago, if I understood the data correctly. If it went from 0.5% to 1% in a week then you should be alarmed. (Although 3 and 13 are both small enough that it's hard to compute a growth rate, but it certainly seems consistent with the UK data to me.)

(I hadn't posted the Denmark data last week because I hadn't seen anything that led me to update substantially in either direction, sample sizes and magnitudes seemed too low.)

Act Three:



Kai Kupferschmidt ✅
@kakape

...

Replies to @kakape @_nickdavies and @LSHTM

Additional data from Denmark (which like UK is doing a lot of sequencing), suggests they are observing a similar pattern of the new variant spreading fast locally.



Mads Albertsen @MadsAlbertsen85 · Jan 2

86 cases of B.1.1.7 identified in Denmark since November (11% of all cases sequenced in the period). Last 4 weeks we have sequenced 1500-2000 genomes pr. week broadly representing DK. The percentage of B.1.1.7 has been 0.2, 0.5, 0.9, and 2.3 (week 52). ssi.dk/aktuelt/nyhede...

Show this thread

So that looks... exactly like what you would see with a 65% more infectious strain, which doubles about once per week, while other evidence points me closer to about 50% more infectious than 65%. I'm essentially assuming at this point that the new strain is at least substantially (35%+) more infectious.

I don't know what the likelihood ratio on this was, but it seems pretty damn high. And that's the one country I had heard was doing a bunch of sequencing other than the United Kingdom, on one of only two strains we are worried about. So it seems not very selected. But I could easily be missing awareness of others, if so please do share.

Did I mention most European schools are still open? Well, [here's some news](#):



BNO Newsroom ✅ @BNODesk · 20h

...

NEW: Netherlands reports 50 cases of coronavirus mutation first found in the UK, including 30 cases linked to an elementary school

43

581

749



The San Diego case of the new strain gets contact traced, [the results are in](#) and [32 more cases have been found](#):



Kristian G. Andersen ✅ @K_G_Anderesen · Jan 5

...

Additional cases of B.1.1.7 in San Diego. While worrying, these were found via targeted measures and not representative - we think the current prevalence is ~1%.

This will increase in coming weeks and the genomics suggest substantial local transmission.

If he's right about the current prevalence being ~1%, that's not comforting. Quite the opposite. That's *two weeks ahead of my previous estimated schedule*, and the one I used for the spreadsheet modeling. Which gives us two less weeks to vaccinate, and for temperatures to improve, before we get hit hard. The good news is that the naive model doesn't think this does that much additional damage, mostly moving infections forward in time, although it does make the last peak a little bigger and the total infection count a little higher.

In other evidence ([source 1](#), [source 2](#)):



Jeffrey Barrett
@jcbarret

...

Out today: two academic publications (not yet peer reviewed) that formally test whether the new B.1.1.7 variant is more transmissible. Both conclude yes, about 50% more. 



Jeffrey Barrett @jcbarret · Dec 31, 2020

...

Replies to [@jcbarret](#)

First, a pre-print led by [@erikmvolz](#) and [@neil_ferguson](#) at Imperial, which applied a variety of different models using both genome sequence data and the S-gene dropout data I've mentioned before.



Jeffrey Barrett @jcbarret · Dec 31, 2020

...

Using different models (lots more detail in the paper), spatial resolutions, and either the genomes or S-gene dropout to track B.1.1.7, the conclusion is very consistent: about a 0.5 additive increase to R.



Jeffrey Barrett @jcbarret · Dec 31, 2020

...

Second, a preliminary report led by [@MoritzGerstung](#) and [@harald_voeh](#), who adapted a hierarchical Bayesian model they made to study the rate of positive tests around England to consider separately B.1.1.7 and other lineages.



Jeffrey Barrett @jcbarret · Dec 31, 2020

...

Finally, as in the other paper, we can model the R number. If > 1 the epidemic is growing, < 1 shrinking. And consistently (a) B.1.1.7 is 50% higher than others and (b) it was often > 1 during the November lockdown in England.

At this point, if this were outside the range of 40%-75% more infectious, I would be very

surprised.

The South African Strain: Are We Even More Fucked? Is It Even More Over?

Don't look now, actually do look, we might actually truly be [Malcom Reynolds "It's worse than you think"](#) level fucked [by the South African strain](#), in which case it might be really, really over:



Aaron Sibarium
@aaronsibarium

...

The bad news: the SA strain may evade vaccines.

The good news: we can just tweak the vaccines so that they still work against the SA strain. And since the West is doing such a great job with vaccine distribution, this is totally relevant and reassuring.

[cnbc.com/2021/01/04/sou...](https://www.cnbc.com/2021/01/04/south-african-coronavirus-mutation.html)

Regius Professor of Medicine at Oxford University John Bell said on Sunday that the variant identified in South Africa was worrying in this regard, however.

“They both have multiple, different mutations in them, so they’re not a single mutation,” he told Times Radio. “And the mutations associated with the South African form are really pretty substantial changes in the structure of the (virus’ spike) protein.”

He said there were questions as to whether the Pfizer/BioNTech and Oxford University/AstraZeneca vaccines would be “disabled” in the presence of such mutations.

The team behind the Oxford University jab was investigating the effect of the variants on its vaccine, he said, adding that his gut feeling was that it would still be effective against the strain identified in the U.K., but he was more uncertain about the one identified in South Africa.

However, he told the radio station that if the vaccine did not work on this variant then it was likely the vaccines could be adapted and that would not take as long as a year.

[Here's another source](#) that seems to be further along in its analysis, [linked paper here](#):



Bloom Lab
@jbloom_lab

ooo

We mapped how all mutations to #SARSCoV2 receptor-binding domain (RBD) affect recognition by convalescent polyclonal human sera (biorxiv.org/content/10.110...).

Among implications: E484K (South African lineage) worrying for immune escape; RBD mutations in UK lineage less so (1/n).



Comprehensive mapping of mutations to the SARS-CoV-2 rec...
The evolution of SARS-CoV-2 could impair recognition of the virus by human antibody-mediated immunity. To facilitate ...
biorxiv.org

8:02 AM · Jan 5, 2021 · Twitter Web App

This is the kind of thing you see all the time from such sources, and such sources are capable of being a lot louder than this on such things even on very false alarms, so I don't put too much weight on them yet.

The problem is, I asked my father, who taught immunology at Columbia University and had a business doing due diligence on pharmaceutical companies, to look into the situation, and what he told me was very much not encouraging.

Here is what he came back with:

"A variant isolate of SARS-CoV2 from South Africa is causing concern, because convalescent sera from some patients recovered from infection with the predominant strain, while effective in neutralizing the same strain, were not effective in neutralizing the South African variant. Sera from other such patients did neutralize the variant effectively.

The concern is whether the current mRNA vaccines (coding for the spike protein from the predominant strain) will be effective against the variant. Unfortunately, the authors did not test sera from vaccinated individuals against the variant, so we don't know for sure. It may be widely effective, effective in some individuals but not in others, or widely ineffective in provoking synthesis of neutralizing antibody to the variant.

A wrinkle: The current vaccines may offer protection from serious illness even in individuals in whom it fails to provide neutralizing antibody, but unlikely to prevent infection.

One of the advantages of the mRNA vaccines is that their manufacture can be modified quickly to accommodate changes. One would hope that the new vaccines could be granted emergency approval quickly, since they would be very similar to the approved ones."

In conversation, he put the probability that the new variant would interfere with the neutralizing agent of the antibodies generated by those who get the mRNA vaccines at 80%. This is something we can find out quickly through experimentation, and it's a sign of how bad we are at running experiments that the news has not yet (to my knowledge) reached us. There's also the chance that other forms of protection could help out, or that the neutralizing agent could still have substantial effects sufficient to prevent illness or reduce severity. But if it comes back that the worst case scenario is indeed the case, and this mutation allows mRNA-vaccinated people to become sick, and the strain is indeed otherwise more infectious than the old one, then we have a very big problem.

So the good news is that mRNA vaccines can quickly be modified to handle the new strain. And if we had a benevolent well-functioning government, we could recognize that it is the same vaccine still, and not require a new round of trials. I have very little hope that we would do this, and thus very little hope we would be able to make the necessary adjustments in a timely manner. I imagine the FDA and company holding up approvals of the new versions for several months, and then the exact same botched roll-out to happen again, and by then for it to be very much too late, even if we don't face the mutation problem again.

The other good news is that it seems highly unlikely the new strain can do much reinfection of previously infected individuals, and that even if the trend of noticing more 'reinfections' is real, it isn't going to move the bottom lines much. And also that it is possible we get actively lucky with all this, and the folks who are vaccinated can get infected but don't get sick, which results in them having much more robust protection against future infection or more mutations than they got from the vaccine alone.

Anyone who knows more about this, please do chime in.

Now that we know we are fucked, and for the moment assuming for purposes of analysis that we're not *that* fucked, the question then shifts to: Is it over?

But What Do We Do Now?

[The same thing we do every night, Pinky. Tell everyone to Be Afraid.](#) Be Very Afraid.



Dr. Tom Frieden  @DrTomFrie... 16h
Bottom line about new variant B.1.1.7? Makes it all the more urgent to vaccinate, mask up, close places where spread occurs, learn more about the virus.

26 178 856 ...

([That first part about vaccinations](#) (MR) is completely right, of course.)

This “be more scared” is both people’s natural message to others, and also people’s natural reaction for themselves, for overlapping but also different reasons. When shown evidence that it’s going to get worse later, people ramp up their internal “this is bad” meter, and act more scared now. I know of one non-rationalist person who was shown my post on the new strain, and decided on the spot to head home, and that they were going to start staying home now and not see friends anymore.

I keep pointing out that *on a personal level, in terms of your personal incentives, this is wrong.*

If things are going to get much worse later, including a possible hospital system collapse, the new strain makes protecting yourself now *less* valuable to you, not more valuable. It’s now going to cost you much more to stay safe the whole way, and action later is going to be *more risky* than action now rather than less risky. So if you have risks you need to take to stay sane and/or stay solvent, or things that need attention like other medical concerns, better to take care of that now rather than later.

If you are thinking on a society-wide level, it’s complicated. Slowing spread also slows B.1.1.7, giving us a little more time to vaccinate before we need to fall back on our control systems. We might not entirely waste that time. It also means that when the control systems start to notice the new strain (which will happen at the same time regardless, because as I understand the math the strain takes over equally quickly regardless) it will be from a lower baseline level, so depending on exactly how the control systems function, we could get substantially better outcomes when the crisis hits.

On the other hand, you’re burning your remaining ammunition in terms of getting people to listen, lock down and play it safe. You’re also doing so in time for control systems to undo a lot of your work before the crisis hits. And if the hospitals are going to be overwhelmed in the future, to a much worse degree than they are now, then extra infections now could plausibly make the peak less bad, and thus reduce both deaths and the amount of overshoot.

None of which are questions anyone involved is thinking about. It’s one knob. Danger rising, sound alarms for danger, get people more scared, hope they act safer.

What Should You Do Personally?

What should you actually do on a personal level? I have been trying to get a better handle on that. I have [a post from yesterday with some toy modeling attempts](#). It's not my strong suit and it's not going to be right, but so far I've seen essentially nothing, so giving it a go should hopefully inspire something better to follow.

What is clear is that there is a big difference between 50% and 65% more infectious. If we are looking at only 50%, the new strain will have a big impact, but if we can do a reasonable job (relative to current expectations, mind you) with vaccinations I'm growing more hopeful the final stage may not be too bad, and a range of outcomes seem plausible. If it's 65%, I don't see a way out of at least a brief period with a lot more infections than now, and our hope for keeping deaths down and hospitals intact is based on getting enough of the vulnerable vaccinated by then.

Therefore, keep a close eye on which scenario we are looking at, and how things are playing out otherwise. It really does change what behaviors make sense.

If the cost of protecting yourself for long enough is high, that makes the value of staying safe now lower. Getting infected now is not good, the hospitals are already not in a great place, but things could get much worse.

I won't go over in detail what 'stay safe' means, but it means all the things you'd think: Socially distance, wear masks, **take Vitamin D**, do things and meet people outdoors, check for symptoms, check air ventilation. I continue to think surfaces are a trivial worry beyond washing one's hands, although if one is trying to be actively super paranoid (e.g. in a hospital collapse) precautionary principle is reasonable.

As a reminder, death rates from Covid-19 for young healthy people are quite low, but there are potential long-term consequences and medium-term consequences that we have large uncertainty about with regard to both frequency and magnitude (because experimentation and data gathering are illegal) and those consequences can be very not fun. Losing your sense of smell for months is much worse than it might sound. You do *not* want to get Covid-19 if you can avoid it.

Once again, these are the core things to know when forming your plan:

If we are in one of the very bad scenarios, getting the virus during the later peak will be very, very bad, and also rather likely for anyone not either immune or taking lots of precautions.

If you can work from home and otherwise take 'reasonably careful' precautions, you will probably not get Covid-19 even in those bad scenarios. Even in the worst case where 75% or more of the population ends up infected. But you need to prepare to take those precautions and sustain them for months, and prepare to pay that price.

It is worth securing your supplies in case the supply chain becomes an issue. The earlier you do that, the more measured you can be doing it, and the better it is for everyone.

Unless you are immune, at some point, whether or not you decide that point is already here, there is a high chance you will want to start taking extreme precautions. Be ready. If you have to choose, better to take your necessary risks now, rather than later.

I've said what I feel that I can. I will note that the incentives involved strongly restrict the things I can say here, whether or not I believe them - so you'll need to work that out for yourselves.

As always, I encourage you to form your own model of the world, think for yourself, and do what you believe is right.

Two Dose, One Dose, Who Knows, You Knows

This week's crisis is getting shots into arms before the shots spoil, but soon the limiting factor will (*I hope*) be not having shots to put into arms. Luckily, there are several good ways to help with that, the tricky part is getting those in charge to listen.

[From Marginal Revolution:](#)

The British approved the Pfizer vaccine, they approved the AstraZeneca vaccine, they moved to [first doses first](#) and now they are allowing (not yet encouraging) [they are running a trial\)](#) [mix and match](#). Under the present circumstances, the British focus on doing what it takes to save lives is smart, admirable, and impressive.

As I wrote on Dec. 10, in [Herd Immunity is Herd Immunity](#):

Mix and matching has two potentially good properties. First, mix and matching could make the immune system response stronger than either vaccine alone because different vaccines stimulate the immune system in different ways. Second, it could help with distribution. It's going to be easier to scale up the AZ vaccine than the mRNA vaccines, so if we can use both widely we can get more bang for our shot.

Addendum: The CDC is projecting [80,000 COVID deaths in the United States over the next three weeks](#).

(I included the last line because I wanted to note that this is 3,809 deaths per day and the average had not at the time been over 2,700 for any 7-day period so far, that that projection was made on December 31, *so that seems like a rather bold prediction* if it's meant to reflect a physical reality rather than being a scare tactic. I can only conclude that it isn't.)

I do think there is a legitimate worry with people's faith in the vaccine process if you start mixing and matching, even though I expect mixing and matching to work at least as well as not doing so, and probably better. If you can pull it off without that issue, wonderful, definitely do that. Hell, at this point, *anything* that is clearly oriented to life saving based on modeling the physical world is something I would welcome even if I think it's wrong on the merits (up to a point, if it's sufficiently wrong then no that's bad and please stop). It's rough out there.

First doses first went from crazy two weeks ago, to the British are doing it a week ago, to now first doses first plus mix and matching.

It gives me hope, you love to see it.

I'd also like to point out that the British did a great but hard thing, closing schools exactly one day after opening them [even though doing that made them look superficially like the complete and total idiots they are](#):



anomalyuk @anomalyuk · Jan 5

This is bad, obviously, but it's *less bad* than if they went "Well, we can't take them out of school after just one day, because then we'd look stupid"

As much damage has been done by people trying to mask their past mistakes as by the mistakes in the first place.



Laura McInerney ✅ @miss_mcinerney · Jan 4

So, to be clear, we just sent 3 million children into primary school FOR ONE DAY, so they could all mix around the virus, and *then* go into lockdown? That's what's actually just happened, right? My brain isn't making this up?

2

1

10



anomalyuk @anomalyuk · Jan 5

If the vaccine rollout isn't a chaotic mess with screwups and reporting errors, then they're not going fast enough.



Quite right. Changing your mind and publicly admitting you are wrong, opening yourself up to blame from all sides, is hard, and should be applauded, and also there should totally be lots of screwups and reporting errors if we are moving fast enough. Good calls all around.

This is [the most recent Marginal Revolution argument](#) for first doses first. A key insight worth repeating is that there have been zero known expected value calculations in support of not doing second doses. There's a lot of haphazard arguments and fear, uncertainty and doubt, but nothing I find remotely compelling, or that comes with any plausibly compelling numbers attached.

Allow us to present [a very clean intuition pump](#):



Homer Texas

@helicopterdrops

...

Replying to [@donaldknewell](#) [@ATabarrok](#) and 2 others

To paraphrase Alex: if your parents are 70 years old and you have two doses of vaccine, how are you distributing them?

7:22 PM · Jan 4, 2021 · Twitter Web App

5 Retweets 1 Quote Tweet 20 Likes

Even Operation Warp Speed is feeling enough pressure to get creative. Since there are huge low hanging gains everywhere, there's room for different countries to find different fruit. For example, did you know that the [studies seem to show half doses work as well as full doses for the non-elderly?](#)

Study mRNA-1273-P201

Study Design

Study mRNA-1273-P201 is an ongoing phase 2a, randomized, observer-blind, placebo-controlled, dose-confirmation study to evaluate the safety, reactogenicity, and immunogenicity of mRNA-1273 in healthy adults 18 years and older. The study enrolled 600 participants, consisting of 300 participants 18 to <55 years old and 300 participants 55 years and older, who were randomized equally to receive either 2 doses of 50ug of mRNA-1273, 100ug of mRNA-1273, or saline placebo given 28 days apart. Participants will be followed for safety and immunogenicity for 12 months post last vaccination.

The immune response as assessed by bAb and nAb after 2 doses were comparable in the 50- μ g and 100- μ g dose groups, with an overall geometric mean fold rise (GMFR)>20-fold in bAB as measured by ELISA and >50-fold in nAb as measured by microneutralization assay at 28 days post-dose 2. In the 100- μ g dose group, the older age cohort (\geq 55 years) had slightly lower bAb response when compared to the younger age cohort (18 to <55 years) at 28 days post-dose 2, but the nAb response was similar between both age groups

So, crazy idea, maybe give them half doses?

A top official of Operation Warp Speed floated a new idea on Sunday for stretching the limited number of coronavirus vaccine doses in the United States: Halving the dose of each shot of Moderna's vaccine to potentially double the number of people who could receive it.

[Data from Moderna's clinical trials](#) demonstrated that people between the ages of 18 and 55 who received two 50-microgram doses showed an "identical immune response" to the standard of two 100-microgram doses, said the official, Dr. Moncef Slaoui.

Dr. Slaoui said that Operation Warp Speed was in discussions with the Food and Drug Administration and the pharmaceutical company Moderna over implementing the half-dose regimen. Moderna did not respond immediately to a request for comment.

Each vaccine would still be delivered in two, on-schedule doses four weeks apart, Dr. Slaoui [said in an interview with "CBS's Face the Nation."](#) He said it would be up to the F.D.A. to decide whether to move forward with the plan.

Dr. Slaoui was asked whether the United States would follow Britain's lead on another tactic for getting shots to more people: Delaying second doses of newly authorized vaccines to immunize a larger swath of the population. There is little or no data on dose delays, Dr. Slaoui said, but "injecting half the volume" might constitute "a more responsible approach that will be based on facts and data to immunize more people."

Natalie Dean, a biostatistician at the University of Florida, agreed that there might be more data to support a vaccine strategy that relied on half-doses rather than delayed doses.

"There is a path forward if you can show that two lower doses yield a similar immune response," Dr. Dean said.

As caseloads continue to surge upward around the globe, and concerns mount over a new and potentially more transmissible variant of the coronavirus, "everyone is looking for solutions right now, because there is an urgent need for more doses," Dr. Dean added. "But the dust has not settled on the best way to achieve this."

John Moore, a vaccine expert at Cornell University, pointed out that the approach wouldn't necessarily work for all vaccines. Injections are already doled out in very small volumes, and some might be harder to halve than others, he noted.

While Dr. Moore agreed that halving doses has more scientific backing than dose delays, he noted that "this is not something I would want to see done unless it were absolutely necessary."

[This thread breaks down the data available.](#) Key findings:



Eric Feigl-Ding ✅ @DrEricDing 13h

9) Conclusion—based on the (better/larger) Phase 2 dosing data, it seems the half dose 50 ug Moderna #COVID19 vaccine is about 75-80% as good as the full dose in the first month, ➡ but once 2nd dose is delivered, neutralizing antibody levels are the same for half vs full dose!

14 64 297 ...

 **Eric Feigl-Ding**  @DrEricDing 13h
10) That said, I wish we weren't so supply chain strapped and stuck in such a raging pandemic that we have to ponder how to stretch our vaccine doses to more people this way. This reminds me of the deferred 2nd dose vaccine debate, which is tough discussion to be honest.

11) also this still has to be FDA approved of course in order for the half dose to be considered. They will be reviewing the same data I presented, available here: fda.gov/media/144452/d...

FDA could decide this pretty quickly I expect w/ another virtual expert meeting. #COVID19

12) A caveat is that the Phase 3 mega trial did not use the 50 ug dose, but rather the 100 ug. Thus, it's imputing efficacy for #COVID19 prevention based on the neutralizing antibody data—akin to imputing heart attacks prevented based on LDL cholesterol lowering alone. 😊

13) from the above data, is it safe to infer that 50 ug dose would be similar to 100 ug after 2 doses based on equivalent neutralizing antibody levels from Phase 2? Unclear.

[Here's another thread](#), I'll focus on takeaways.



Prof. Akiko Iwasaki @VirusesImmunity · Jan 1

...

So how effective is a single dose vaccine? We do not know for sure, but for at least a month or more, a single shot mRNA vaccines should provide ~90% protection (>14 d post vaccination). This is from the Moderna VRBPAC Briefing Document. (6/n)



Prof. Akiko Iwasaki @VirusesImmunity · Jan 1

...

Whether a single dose vaccine provide protection from severe COVID is not clear due to small sample size and short follow up duration. (7/n)



Prof. Akiko Iwasaki @VirusesImmunity · Jan 1

...

I am still a proponent of 2 dose vaccine but given the urgency, we can delay the 2nd dose until more vaccines become available. I know many others have been saying this all along, but it was the B.1.1.7 variant transmission rate that did it for me. (8/n)



Prof. Akiko Iwasaki @VirusesImmunity · Jan 1

...

So how long can we safely wait in between 1st and 2nd doses? If you look at the childhood vaccine schedules, some boosters are given months to years apart. For COVID vaccine, booster after a few months might be okay. (10/n)

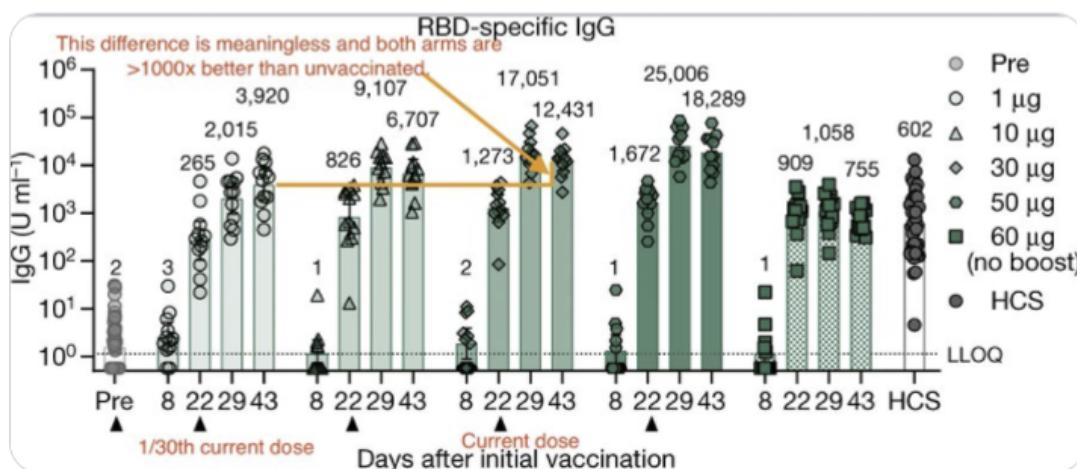
And then there's this, [which suggests that half doses might already be way more than we need](#) although I haven't examined the data involved:

William Gibson
@wgibson

8

The fact that people are furious about half-dose strategies is nonsensical. At the current vaccine-limited rate we would immunize to herd immunity somewhere between 4 and 10 years from now.

But few know that 1/30th the Pfizer dose elicits excellent antibody responses!



I always find it strange that people say things like “we don’t know whether a single dose vaccine provides protection from severe Covid” in spots like this. I don’t see a plausible physical way for it not to.

My conclusion (at least for now) is that the half dose is *probably* all you need if you are not elderly and get the second shot afterwards, with close to full effectiveness, and is *more efficient per vaccine dose* if given as only one dose but less effective than one full dose. There is some danger that the dropoff in effectiveness is bigger than we think, but not so big that we could plausibly come out behind on the deal. If one shot is already 80% effective, I find it highly implausible we wouldn't do better than that with the double half-dose, given the data here.

Meanwhile, on the matter of first doses first, I'd offer to eat my hat as well, except that's a one sided bet and also I don't own a hat:

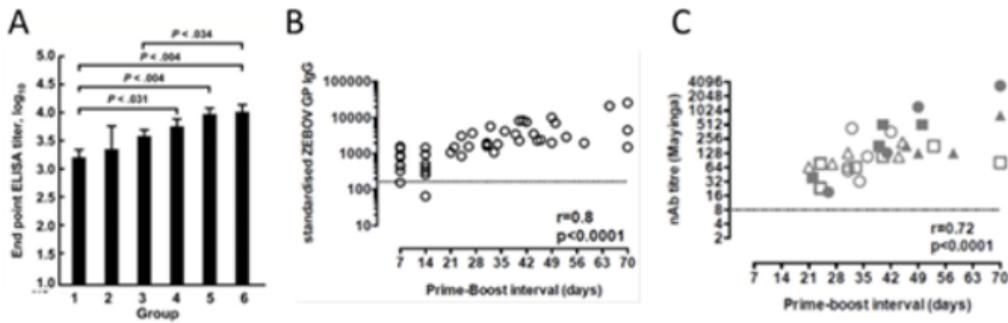


Sandy Douglas
@sandyddouglas

ooo

Replies to @sandyddouglas

Based upon the biology, I'd eat my hat if the Pfizer vaccine is substantially less effective with a longer dose interval. Most vaccines induce stronger immune responses with longer intervals. A couple of examples below. There are more.



A is from Ledgerwood et al., J Infect Dis, 2013. 208(3): p. 418-22.

This was a Phase I trial of DNA prime – inactivated vaccine boost flu regimes. Graph shows ELISA data 14 days post-boost. Group 1 received licensed inactivated vaccine prime-boost. Groups 2-6 had DNA prime – inactivated vaccine boost intervals of 4, 8, 12, 16, 24 weeks.

B & C are from Ewer, K., et al., N Engl J Med, 2016. 374(17): p. 1635-46

This was an Ebola vaccine study, using varying doses and intervals of an adenovirus prime and another viral vector (MVA) as the boost.

And [everyone's entitled to their opinion:](#)



Bob Wachter 
@Bob_Wachter

...

Replies to [@Bob_Wachter](#)

Far better to have 100M people who are 80% protected than 50M people who are 95% protected, particularly as we are facing a foe that is getting smarter and nastier. Or at least it seems that way to me. You? (7/7)

10:32 AM · Dec 31, 2020 · Twitter Web App

That all works for me. Yes, we could do so much better if everyone did all the right things and grabbed all the multiplicative low-hanging fruit, but even grabbing *some* of the fruit, even different fruit, is a huge win. I'll take it.

Yes, We Can Agree Andrew Cuomo Is The Worst

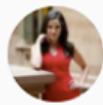
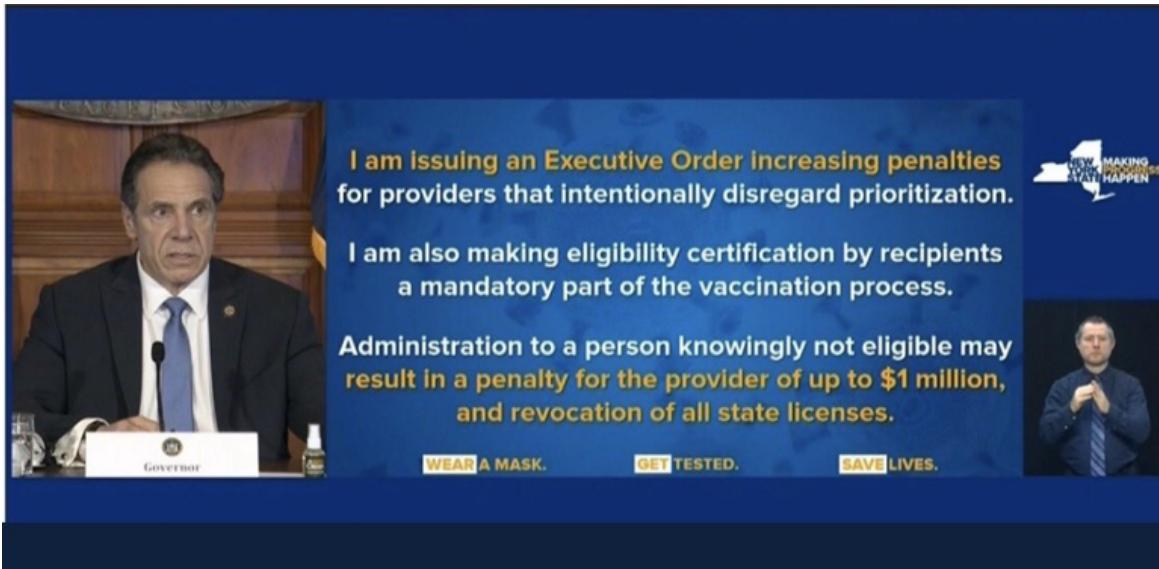
That's less true now than when I wrote this, "because of reasons," but still. Somehow he has topped himself. I'm not saying History's Greatest Villain, or even clearly *this week's* greatest villain – the competition's been stiff lately – but it is not through lack of trying.



Mason   

@webdevMason

If you wanted to make sure that rapidly expiring vaccines distributed in 10-dose vials end up in the trash, this is how you'd do it



Morgan Mckay @morganfmckay · 21h

...

"You will see fraud with this vaccine...it's valuable, it's money. I understand that this is a valuable commodity...so I want to make it criminal," Cuomo said about entities using the vaccine in ways it is not intended or authorized by the state.

2

8

13

↑



Mason 🏳️‍🌈 ♂ @webdevMason

1d

Based on this tweet's favs/RTs it seems the one thing uniting every faction on America's roiling political spectrum is our shared conviction that Andrew Cuomo is the fucking worst. You love to see it

52 163 3k ...



Mason 🏳️‍🌈 ♂ @webdevMason

19h

Not to be outdone, California gov Gavin Newsom tells providers that if they give an elderly friend, teacher or emergency worker that vaccine before they're supposed to, he'll yank their license AND trash their reputation.

Letting the miracle drug expire on the shelf remains A-OK

On January 4, doctors outside of hospitals became eligible. On January 11, home care workers will become eligible. At *some unannounced point after that*, maybe we'll get to phase 1B, which includes those over 75 years old. One hopes.

But, Cuomo says, I also threatened our health care providers to speed things up, doesn't that fix everything?

Business

Cuomo Warns Hospitals to Speed Shots or Face Fines: Virus Update

updated 39 minutes ago

Did you know that the more you yell and threaten, the more everyone does what you want?

Second-worst person New York Mayor [DeBlasio is not happy about the threats](#), calling them 'just arrogance' because he can only think about status implications rather than physical consequences.

Oh, it's so much worse than that. What happens when the central planner combines threats to those who don't distribute all the vaccine doses they get, with other threats to those who let someone 'jump the line'? Care to solve for the equilibrium?

Here's a [new framing to help you out](#):



Aditya Mukerjee, the Otterrific 🐾 🏳️‍🌈 ✅
@chimeracoder

...

Cuomo: if you vaccinate a single person not on the approved list, we will fine you \$1,000,000

Hospitals: okay

Cuomo: and if you have any vaccines left over, we will also fine you

Hospitals: ...



Morgan Mckay @morganfmckay · 21h

NEWS: Cuomo says that hospitals can face fines of up to \$100,000 if they do not use all their vaccines by the end of this week.

Going forward, facilities must use all vaccines within 7 days of receipt

[Show this thread](#)



Eliezer Yudkowsky ✅
@ESYudkowsky

...

Broke: Regulations so narrow they only allow a single course of action, or doing nothing.

Woke: Regulations that rule out all courses of action, including doing nothing.

[Eliezer](#) got this one wrong, because actually there is a particular nothing that will *not* get ruled out.

[Here's the equilibrium:](#)



Aditya Mukerjee, the Otterrific 🐂 🏳️‍🌈 ✅
@chimeracoder

...

Replying to [@chimeracoder](#)

Cuomo is GUARANTEEING vaccine shortages in NY, because now hospitals have a strong incentive to err on the side of ordering only the vaccines they know they can give to eligible people, even if that means running out

[@NYGovCuomo](#) needs to resign.

12:23 PM · Jan 4, 2021 · Twitter for Android

268 Retweets 16 Quote Tweets 1.2K Likes

Things [we will totally do in the future](#), to avoid the risk that someone somewhere is using our existing private infrastructure, but we'll worry about the logistics for this when the time comes:



Morgan Mckay @morganfmckay · 20h

...

Replies to [@morganfmckay](#)

Cuomo: The state will be establishing drive-thrus for public distribution when it is time for the public to receive their vaccine.

State will recruit additional retired personnel, nurses, doctors and pharmacists to administer vaccines. Will use public facilities

4

13

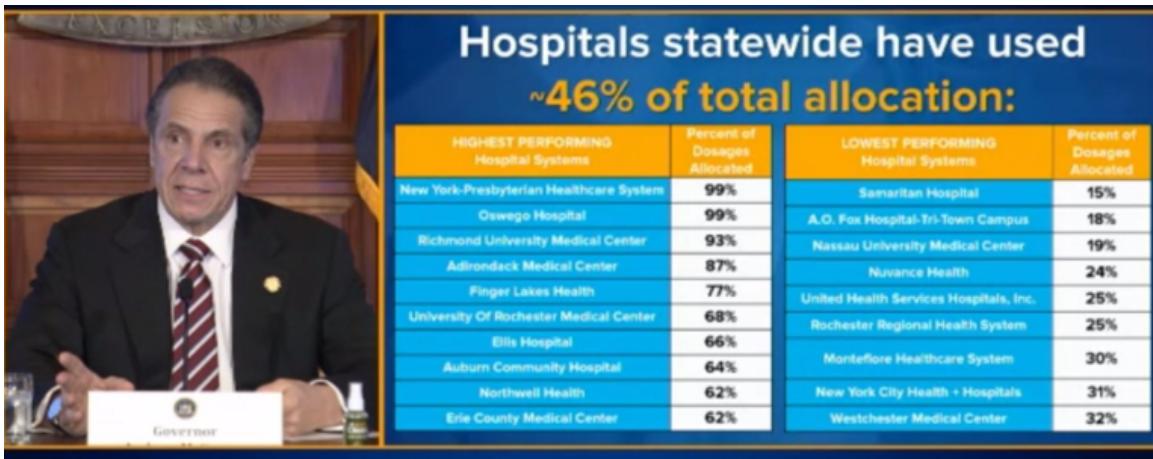
34

↑

Note that Cuomo is not using the present tense. He is not saying “We are recruiting” such people, or that we “are establishing” such locations.

Avoiding blame in the next two weeks, you see, does not extend to problems we will have three weeks from now. Worry about tomorrow, tomorrow.

Even if you pull it off and thread the needle, it does not seem that ‘[use 99% of your allocation](#) while we have more doses than we can use despite obeying all the restrictions’ leads to the next step of *then maybe you should send those places some more*:



From the same thread, I'm trying to figure out the logic behind this one if it isn't purely sacrificial, [and it's a thinker](#):



Morgan McKay @morganfmckay · 20h

Cuomo: Schools can stay open in counties with a positive COVID-19 infection rate over 9% if school testing can show they are below the community average. It will be ultimately up to the school district.

5

17

8

↑

000

You have to be sympathetic, though. These logistics are hard. It's not like there was any warning that such problems were coming. Yes, we did know vaccine shipments were likely, but it's not like there was any time or way to plan for mass vaccinations in advance.

[Oh. Wait..\(you might want to read the whole thing on this one\).](#)

NEWS

New York's mass-vaccination plans are shelved as Cuomo takes different path

Cuomo adviser: 'There has to be a real statewide coordination here, which is what we're doing'

ALBANY – County officials who have for years been planning for a mass vaccination said they are seeing that training and preparation – much of it funded by millions of dollars in federal grants – pushed aside as the administration of Gov. Andrew M. Cuomo has retained control of the state’s coronavirus vaccination program, including having hospitals rather than local health departments administer the doses.

Interviews with multiple county officials over the past week confirm that many are unclear why the governor’s administration has not activated the county-by-county system, a plan that included recent practice sessions in which members of the public received regular flu vaccines at drive-thru sites.

In Albany County, officials have privately said they could vaccinate the population of the southern half of the county in a few days if they were given the coronavirus vaccines and allowed to mobilize their plan.

[What was that, Mr. Reynolds?](#)

[So let me get this straight.](#) New York State has a detailed mass vaccination plan, developed well in advance for exactly this situation. It’s ready to go. There are distribution sites, officials in charge in each locality, every county is standing by. The plans are submitted. And we’re just going to... not use it, because we’d rather be sure that the doses all get distributed in exactly the right order.

[Ross Barkan has similar thoughts](#) about the whole mess, and offers some reminders about Cuomo’s past performance.

Even if you think there’s rhyme or reason to the decisions being made, they’re not only not transparent, they’re not being told to anyone.

No one has any idea what is going on, who has priority over who or when anyone can get vaccinated, or how and where and when one might go about doing that if you aren’t living or working at a location doing the vaccinations. No one can plan.

This is all hitting on a personal level, on multiple fronts.

As a concrete example, my wife is a psychiatrist. On Monday, she was informed (with no warning) via various roundabout methods, that she was eligible for vaccination.

The announcements said nothing about how to go about getting vaccinated, by appointment or otherwise. So we thought about what to do and she started making phone calls. A lot of phone calls. Officials directed her to other officials. Officials said they learned about the new

eligibility that day the same way we had, and they were scrambling and would have to call us back. No one could provide any clarity.

After all the phone calls, she's on one list where they will *maybe* call her back and offer her an appointment somewhere at some point with some amount of notice, none of which can be specified, and then she'll need to drop everything and take that opportunity, since there might not be another. Then there are the other potential locations, which haven't even gotten that far, and pretty much [everyone has a limited supply available and also no idea how to use it.](#)

So despite being eligible for vaccination, and despite there being lots of unused doses, and despite our willingness to get into a car and drive quite far and make a lot of phone calls, still no vaccination and no timeline for getting one.

Meanwhile, I talked to my father, who is both very old and has comorbidities. He is therefore in "Group 1b" which includes 2.5 million people in the city of New York, behind a huge pool of people in 1a who are mostly not at high risk and that are only half done, and to which groups are being added haphazardly – which makes sense, since Cuomo won't let anyone from 1b get vaccinated, so it's vital to expand 1a as much as possible so vaccine doesn't spoil. When we go to group 1b, he fears slash assumes it's going to get prioritized to go to 'harder hit' zip codes and areas, and that he'll effectively be behind a quarter of the city and forced to wait, and that nothing he can do will change that. I don't expect this and told him he was quite wrong, that the system is more than sufficiently a giant mess that being willing to call around and go where you need to go will be quite sufficient to be early in 1b.

But at this point he expects allocation by Politics and Power to do its best to kill him.

Contrast this with the system effectively being used in [Florida](#), featuring such horrible blameworthy actions as finding ways to do things, and to vaccinate seniors who actively want to get vaccinated:



Lachlan Markay

@lachlan

...

Gut-punch of a lede: "Local health officials are turning to online services like Eventbrite to improvise distribution schemes for the COVID-19 vaccine in the absence of federal support or a national plan."



Benjy Sarlin

@BenjySarlin · Jan 5

...

FWIW a couple of relatives got vaccinated via Eventbrite appointment in FL and said it ran very smoothly

[What about teachers?](#) Good question. Who counts as an 'education worker' here?



Jillian Jorgensen ✅ @Jill_Jorgensen · 22h

...

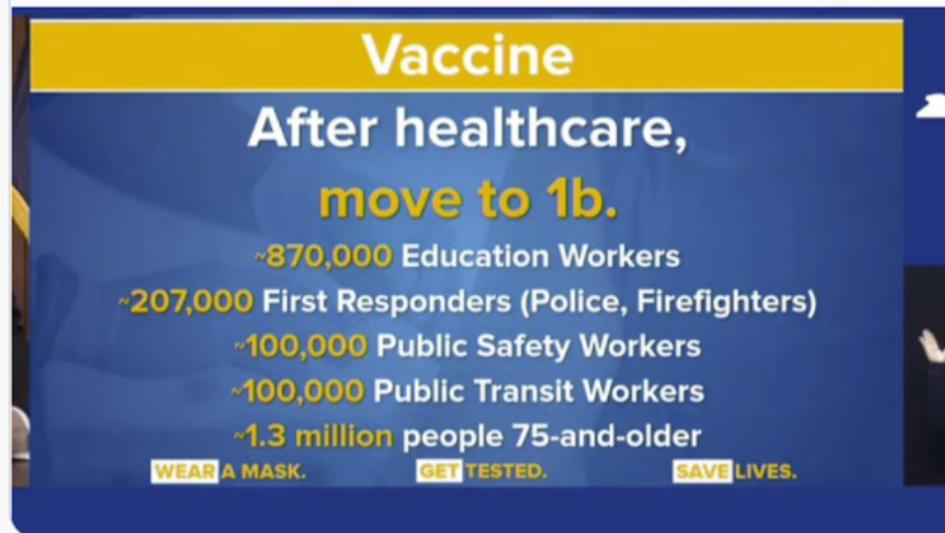
A spox for the gov yesterday confirmed to me that teachers were in 1B, but did not respond to a follow-up question about whether other in-school staff would also be included. Language below seems relatively inclusive. And some school workers are eligible NOW: nurses, PTs, OTs.



Alex Zimmerman ✅ @AGZimmerman · 22h

Cuomo says "education workers" are in line for the vaccine after healthcare folks

[Show this thread](#)



What are we going to do about prisoners? [Different people get different messages, so no one knows anything.](#)



Release Aging People in Prison Campaign
@RAPPcampaign

...

BREAKING: [@NYGovCuomo](#)'s Health Commissioner is sending mixed messages to NYers about the fate of COVID vaccines for incarcerated people. Albany journalists are reporting Zucker told Dems some in prison would be vaccinated soon, then told Republicans the opposite.

NEW YORK -- Today, during a New York State Department of Health briefing for legislators, Dr. Howard Zucker and Larry Schwarz told state legislators that incarcerated people in state prisons will be included in the next phase of vaccinations (1B) along with corrections officers. However, Senate Minority Leader Robert Ortt claimed that a different message was delivered to the Senate minority conference. In response,

Then when confronted about having sent contradictory messages to different people, they denied everything and moved on to Cincinnati:



Rob Ortt @SenatorOrtt · Jan 5

...

Why is the Legislature being given mixed messages on crucial public health information? Will inmates receive the vaccine before seniors, nursing home residents and health care workers who are still waiting? We need and deserve clear answers.



Morgan Mckay
@morganfmckay

...

UPDATE: When asked for clarification on whether inmates in NY state prisons will be receiving the COVID-19 vaccine next as part of Phase 1b, Health Department spokesperson released this statement...saying the state right now is just focused on Phase 1a

We are currently in the midst of a comprehensive approach to vaccinating all eligible populations. Corrections officers will be able to be vaccinated at community sites or other locations when they are eligible. As Governor Cuomo has repeatedly said, right now, we are focused on the population in Phase 1a.

<https://covid19vaccine.health.ny.gov/phased-distribution-vaccine#phase-1a>

So it seems like we're debating whether *prisoners* should be in the same tier, with effectively higher priority, than my very old, very high risk father who has effectively been under house arrest since March, and who would happily give up quite a lot to get it. Are they more deserving? Is he not at much higher risk? What the hell is this?

Except we're denying we're debating that because one could be blamed for either decision, and the need to decide is more than two weeks in the future, so we're not going to decide (or at least not tell people the decision) until the last possible moment.

Do you have *any* idea how enraging this is?

And of course I'm way behind all those people and who knows when I'll get a chance.

[Even DeBlasio is declaring his intent to get our collective asses in gear](#), vowing one million vaccinations in the city in January. The problem, of course, is that he's not allowed to do that yet, and who knows when he will be allowed (see: Andrew Cuomo is the Worst):

Initial estimates were that the city would have doses for more than 450,000 people available.

The current priority—"phase 1a"—remains health care workers and nursing home residents and staff, which amounts to about 1 million people in NYC.

"We want everyone in that category who is eligible to get vaccinated to actually have the ability to get vaccinated," NYC Health Commissioner Dave Chokshi said. The city wants to get the vaccine to other kinds of health care workers outside of hospitals, like home health aides, which requires amping up access points, according to the commissioner.

"For us to move quickly, as is our intent, we have to be able to expand the circle of eligibility swiftly as well so that we can match up the capacity that we have with that eligibility," Chokshi said.

The state will have eligibility screening requirements the city will follow, the commissioner added.

From NY Times via Nate Silver, so no link (Scott Alexander situation remains unresolved):

On Thursday, Mayor Bill de Blasio said the city planned to have administered doses to one million people by the end of January. He has suggested that the state is acting as a bottleneck by not authorizing the city to open up vaccinations to larger categories of people yet.

"If we're given the authorization, we can move very quickly," Mr. de Blasio said this week. "We need the state guidance in terms of the categories of people, and the more that expands, the faster we can go."

This could not be more explicit. We cannot vaccine people because there is no one we are allowed to vaccinate.

So of course [Cuomo fires back](#):



Jillian Jorgensen ✅ @Jill_Jorgensen · 23h

Well, it's 2021 and: [@NYCMayor](#) says vaccines are slow due to state rules

...

[@NYGovCuomo](#) says vaccines are slow bc leaders like [@NYCMayor](#) aren't going fast enough in their public hospitals, complete with a "MUST MANAGE" logo over his face

♪♪ same as it ever was♪♪

11

76

262



Jillian Jorgensen ✅

@Jill_Jorgensen

...

Replying to [@Jill_Jorgensen](#)

and you may ask yourself, "what is wrong with them?"
and you may ask yourself, "where is my functioning government?"

and you may tell yourself, "this is not a coherent plan"
and you may tell yourself, "this is not a way to govern"

12:00 PM · Jan 4, 2021 · Twitter Web App

Letting the days go by, letting the virus keep you home...

Cuomo's [not going to stand for this kind of undermining. Respect his authoritah:](#)



Bill Neidhardt

@BNeidhardt

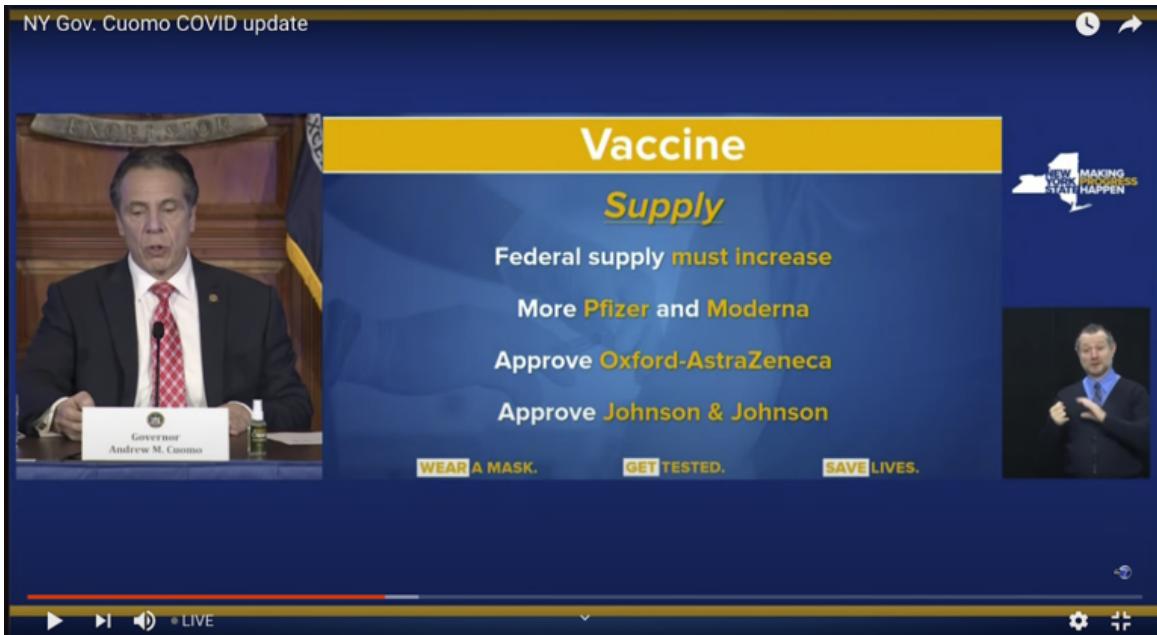
...

Today the City is vaccinating at-risk frontline workers who provide emergency medical care, including the PD medical corps.

The State has ordered the City to not vaccinate more vulnerable city workers, including DOC and officers who regularly perform CPR and administer Narcan.

4:18 PM · Jan 6, 2021 · Twitter Web App

To his credit, thanks to his motivation to shift blame to the Federal Government, Cuomo is at least yelling at those responsible for getting us more supplies to get us more supplies, and to approve more sources of supply, all of which are fine things to call for, but also cheap talk:



And DeBlasio does have that minor issue that the city that never sleeps does almost no weekend vaccinations, [making excuses seem rather weak](#)



Byrne Hobart
@ByrneHobart

...

I'm impressed that de Blasio managed to find a refrigerator that barely works on weekends and doesn't work at all during holidays. That must have taken a lot of planning.



Emma G. Fitzsimmons ✅ @emmagf · Jan 4

I asked Mayor Bill de Blasio why New York City hasn't been providing the vaccine 24/7 when he knew this was coming for months.

The mayor says there have been logistical issues with refrigeration, and the city took a cautious approach at first: "Now it's time to sprint."

Not Cuomo, but to cover bases mentioned above, here's the [quote from Newsome](#), who is trying but can't hope to compete with the champion:

“I just want to make this crystal clear: If you skip the line or you intend to skip the line, you will be sanctioned, you will lose your license,” Newsom said during a press conference. “You will not only lose your license, we will be very aggressive in terms of highlighting the reputational impacts as well.”

The first wave of coronavirus vaccines are intended for health care workers and those in high-risk congregate settings, such as skilled nursing facilities. Future phases of vaccine distribution likely will be targeted at those who work in education or child care, emergency services, food and agriculture, Newsom said.

Perhaps [he was mad about these horrible people?](#)

- At least two hospitals in Southern California have been vaccinating non-frontline workers having received 'extra' doses of the Pfizer/BioNTech vaccine
- On several occasions, word has gone out to family members of healthcare workers who might be interested in receiving a dose
- Redlands Community Hospital openly admit they did not want to see the extra vaccine go to waste and so reached out to those associated with the hospital
- A relative of a worker from Southern California Hospital say they were also 'invited' to receive the vaccine
- The hospital denies such an offer was made and instead says other frontline workers in the community were invited to be vaccinated

I'll leave this here:



Joe Bishop-Henchman

@jbhenchman

**As the Titanic was sinking,
lifeboat priority was "women
and children first." The officer
doing the starboard boats
(Murdoch) did that but
allowed men to take empty
seats left. The officer on the
port side (Lightholler)
preferred to lower boats part
empty than let men in them.**



Joe Bishop-Henchman @jbhenc... 1d

"When the most accurate estimates of survivors in each lifeboat is examined, one sees that roughly 61 per cent of the 712 survivors were in lifeboats which Murdoch directly supervised the loading of, or played a part in the loading of."



Joe Bishop-Henchman @jbhenc... 1d

James Cameron on Murdoch: "I'm not sure you'd find that same sense of responsibility and total devotion to duty today. This guy had half of his lifeboats launched before his counterpart on the port side had even launched one. That says something about character and heroism."

1 4 63 ...

Detailed story of the evacuation is [here](#).

In other New York news, I'd note that policies like this might have made sense back when New York's infection and death rates were unusually low, but make very little sense now:



City of New York @nycgov

2d

Traveling to NYC over the holiday weekend? You're required to quarantine for 14 days. This is one of the steps we're taking to fight back #COVID19 in our city. Go to NY.Gov/TravelAdvisory to learn more.



9 20 58 ...

Although I am [glad to see this](#) which seems worthwhile:



Eater NY @EaterNY 30 Dec 2020

In a minor win for NYC restaurateurs, food businesses can now use a portion of their sidewalk to accept takeout orders and sell pre-packaged goods

ny.eater.com/2020/12/30/222...



2 27

Also, [did I mention the inevitable yet?](#)



BNO Newsroom
@BNODesk

...

New York reports first case of coronavirus mutation first seen in the UK. It's a man in his 60s with no travel history - WNBC

3:50 PM · Jan 4, 2021 · TweetDeck

All I Want For Christmas is a Covid Vaccine, But They Somehow Underpriced Them So Much No One's Even Bothering To Sell Out

Israel's approach is personified [by the Pizza Guy](#):



Yaniv Erlich
@erlichya

...

Replies to [@erlichya](#)

Here is a prime example from today to Israel "organized chaos". End of the day in a vaccine center. A few doses left and will expire. Nurses go out, spot a pizza delivery guy, call him "pizza guy wanna vaccine?", jab, and another person has spike mRNA!



Nadav Eyal ✅ @Nadav_Eyal · Dec 31, 2020

נראה ביד אליו הערב: אנשי קופ"ח שנותרו עם כמה מאות פניות קוראים - "הבחן עם הפיצה, תגיע". קיבל חיסון. זו עילום.

Ours, not so much.

To be fair, it does seem [Washington, DC managed to get on that page, at least once](#), to go with the previous incident in Kentucky. This taking place in the city that isn't part of a state seems like it might not be a coincidence, although it is also where there is a lot of media.



NBCWashington ✅
@nbcwashington

...

"[The pharmacist] turned to us and was like, 'Hey, I've got two doses of the vaccine and I'm going to have to throw them away if I don't give them to somebody. We close in 10 minutes. Do you want the Moderna vaccine?'"

Then there's [the natural experiment we ran in California \(source article here\)](#).



Patrick McKenzie @patio11 · Jan 5

ooo

Not to bang on a drum, wait a second seems like this is most important drum in the world: administering vaccines is *actually not hard.*

We are *choosing* to administer them slow.

We should *stop* choosing to do this.



Anita Chabria @anitachabria · Jan 4

California Hospital lost freezer, gave out 600 vaccine shots in 2 hours -
Los Angeles Times latimes.com/california/sto...

14

162

621

↑

What happened was, they lost their freezer, so they had 830 doses and two hours to distribute them. They made the (obviously overdetermined and correct) decision to vaccinate as many as possible regardless of guidelines.

They still tried to prioritize:

Winiger got on the phone, trying to give the shots first to those on the priority lists. One local elder care facility took 40 doses for staff, and the hospital's chief medical officer drove them to the facility himself.

Also, this [just raises further questions](#):

About 200 doses belong to the county and were being stored by the hospital. Winiger said those doses were returned to the county. The county in turn gave 100 doses to the city of Ukiah, county Chief Executive Carmel J. Angelo said.

If the doses were spoiling, could they be 'returned' to the county or not? If they could, why didn't they return all the doses? If they couldn't, how did they return 200? Was there exactly that much space in their freezer? What storage facilities were available in Ukiah?

They also didn't skip over the issue of proper consent or a 'safe protocol', and prioritized on that basis:

Faced with only an hour to use the shots, sheriff's officials decided to administer them to staff and front-line personnel because they didn't think there was enough time to gain consent and organize a safe protocol for inmates. Four county medical staff began giving the shots, Bednar said.

The fire department figured out how to handle a true emergency, as one might expect:

An additional 100 doses hit the fire department about 12:15 p.m., Fire Chief Doug Hutchison said. At first, garbled information he received through phone calls left him fearing all 800 doses were coming his way, leaving him thinking, "There is no way," he said. His full-time staff of 16 had already been promised to help with other clinics.



Residents in Ukiah line up for a COVID-19 vaccine after a freezer storing it broke down. (Cici Winiger)

Hutchison headed to a city conference center, and his remaining crew "began giving shots as fast as we could sit people down and roll up their sleeve," he said. Their syringes went into the arms of police, essential city staff and firefighters — including Hutchison, who had declined earlier offerings of the vaccine to make sure his staff got it first.

"I was trying to make sure all my people got shots before I did," he said.

In the end, success!

By noon, within 15 minutes after learning of the freezer failure, shots were being administered at all four sites. Lines began to form as word spread and some staff was siphoned off for crowd control. At the site Winiger ran, about 30 people were turned away after the doses ran out. At the main site near the hospital, she estimates about 120 people left without the shot.

But by the two-hour deadline, every dose had found a patient, Winiger said.

Of course, then the conspiracy theories started flowing, because the freezer breaking let those with local power direct doses to wherever they want. I'm confident that didn't happen, but I understand the instinct to think that way.

Regardless, do [read the whole thing](#).

The conclusion is clear as day. If we cared about distribution of the vaccine, the vaccine would be distributed. Period.

The chorus rises up, and speaks as one: [Shots . In. Arms. Now.](#)

The most obvious suggestion is to [maybe use the pharmacies](#) (WSJ) that already administer the flu shot, and [let them do this one in the pharmacy as per normal](#) (MSNBC) rather than only contracting for helping with long term care facilities? Still blows my mind this one is hard. Which, in good news seen later in the week, [could be happening sooner rather than later](#), if nowhere near soon enough.

Scott Gottlieb, former head of the FDA, making multiple modest proposals, not only suggesting pharmacies administer vaccinations, but also that we administer more vaccinations:

KEY POINTS

- To speed up Covid vaccine administration, Dr. Scott Gottlieb recommended the federal government withhold fewer vaccine doses
- “Right now, every shot in an arm is a win,” the former FDA chief told CNBC on Monday.
- “Putting away 50% of all the doses, I think, is denying more people access to a vaccine,” he added.

Gottlieb said he believes states should be willing to expand the eligibility, including making the vaccine available at retail pharmacies, because it is important that high-risk Americans have access during what he called “the worst part of this epidemic right now.”

It's deeply sad that we need to use the language of 'high-risk Americans have access' to suggest using our actual existing distribution system for the purpose of distributing the things they normally distribute.

Even if we do the second shot on schedule, [surely it's madness to hold back the doses for it rather than count on future supply, right?](#)



Walid Gellad, MD MPH @walidgellad · 3h

...

The only scenario the 'get more first doses out' strategy is worse than what we're doing now is if the first shot is not effective and supply collapses.

They only used 50% efficacy for first shot(too low) and don't model potential reduction transmission.

Yes it's just modeling

1

5

Yes, if the first shot doesn't do anywhere near what we think it does, *and* we face a sudden and total supply collapse, *then* we are *slightly* more screwed with first doses first.

Options that exist in theory but somehow not in practice, [an ongoing mystery](#):



Jon Walker
@JonWalkerDC

...

The only way to balance the overwhelming need for fast vaccine distributions and make sure they go to high impact first is age bands. Easy for government/providers to find people's age, easy to check no one is "skipping," maximizes lives saved and beds free.

5:06 PM · Jan 3, 2021 · Twitter Web App



Chana
@ChanaMessinger

...

New idea: vaccinate everyone in a hospital. Staff, doctors, patients there for any reason (except covid). They're all sitting right there, hospitals have a huge % of the doses, they're more likely to be sick or have underlying health conditions, and we need to get JABS IN ARMS.

9:32 AM · Jan 1, 2021 · Twitter Web App



Chana @ChanaMessinger · Jan 1

...

If ONLY there were a bunch of UNEMPLOYED PEOPLE we could have trained MONTHS AGO to do a RELATIVELY SIMPLE TASK that will SAVE LIVES

But other communities are falling short of that rapid clip. Dr. Smith said the medical community is worried about staffing shortages when hospitals have to both administer vaccines and treat Covid-19 patients.



Chana @ChanaMessinger · Jan 1

WHAT?!?!?!

...

sick and old people just SITTING THERE are CHALLENGING what is even going on how do state and federal governments get up in the morning without initiating six slapstick routines involving banana peels???

Most vaccines administered across the country to date have been given to health care workers at hospitals and clinics, and to older adults at nursing homes. Gen. Gustave F. Perna, the logistics lead of Operation Warp Speed, on Wednesday described them as “two very difficult, challenging groups” to immunize.

Look, I totally get that nursing home residents need to get medical histories and such before you can clear them for the vaccine, but the health care workers are literally at exactly the places you give the shots and the nursing home residents are kind of there because they can't ever be anywhere else, so who exactly are the unchallenging groups? And if it's too difficult can we then open up the criteria a bit? No?

And yeah, *how about you start scrambling and thinking about that wall **that you're going to hit** BEFORE you run into the wall?*



Chana @ChanaMessinger · Jan 1

it has come time,,,to scream

...

“In the next phase,” said Dr. Jha of Brown University, “we’re going to hit the same wall, where all of a sudden we’re going to have to scramble to start figuring it out.”

Given, Dr. Jha, *that you outright said you’re going to hit that wall.*

Meanwhile, the wall we’re in now:



Christopher Hooks ✅
@cd_hooks

...

Just talked to a pharmacist in the panhandle who's currently watching 30 doses of the vaccine go bad because they can't find enough "eligible" recipients. We had a year to figure this out

1:09 AM · Dec 31, 2020 · Twitter for iPhone

16.1K Retweets 2.3K Quote Tweets 92.1K Likes

[What having a chance of being sufficiently cynical looks like:](#)



Nate Silver ✅ @NateSilver538 · 15h

...

One problem might literally be that this very good and practical solution is too simple when committees feel as though they need to design complicated solutions to prove that they've done their job and weighed all the relevant equities and so forth.



Jon Walker @JonWalkerDC · 16h

The only way to balance the overwhelming need for fast vaccine distributions and make sure they go to high impact first is age bands. Easy for government/providers to find people's age, easy to check no one is "skipping," maximizes lives saved and beds free.

43

62

576

↑

[This is beyond fully endorsed:](#)



Emily Hamilton
@ebwhamilton

...

The governors of any states where vaccines are allowed to spoil should be hauled in for a week-long congressional hearing after which they'll be paraded around the Capitol



This as well:



Andrew Rettek Retweeted



Nate Silver @NateSilver538 1d

Yeah. And the extremely byzantine vaccine priority plans developed by many states is going to make this worse. The more subcategories you create, the more things there are for people to object to, which may further erode trust in the public health system.

David Dayen @ddayen

The focus on whether every vaccine recipient is "deserving" rather than getting shots into arms as quickly as possible reminds me of trying to target "deserving" homeowners with HAMP, means-testing, all the little ways US politics stupidly pits people against one another

[Show this thread](#)

Take [the following tweet](#) far more seriously than you think:



Alex Myrrh-esianu
@ahardtospell

...

Listen, if Bill Gates was trying to use the vaccine to put computer chips in our blood, the rollout would be going a lot more smoothly.

7:11 PM · Dec 30, 2020 · Twitter Web App

1.2K Retweets 75 Quote Tweets 12.7K Likes

This is not mainly about Bill Gates being better than the government at logistics. This is because if the vaccine was being given with bad intent, then there would be no danger of [motive ambiguity](#), and those with power could safely make better choices rather than worse choices.

Also I have no idea [why Mark Primiamo is even half-joking](#):



Mark Primiano

@Doctor1Hundred

I'm only half-joking when I say they should recruit vets to help give the vaccine. On any given vaccine clinic day, I can get 60+ patients done in 2 hours, and that's patients trying to bite me.

9:31am · 3 Jan 2021 · TweetDeck

1,043 Replies **10,656** Retweets

98,798 Likes



Mark Primiano @Doctor1Hundred 21h

Back at the start of this, Illinois sent out an all hands on deck thing for medicos of any type. Not quite a draft but more or less one. We signed up. We're ready. We're waiting.

11 111 3k ...

Here's to [the courage to admit when one is both correct and isn't joking](#) and also [everybody knows this already](#):



balajis.com ✅ @balajis · 8h

If Bezos did it we'd all get the vaccine in two days.

...



Eugene Wei @eugenewei · 14h

I'm ready for the U.S. to just turn over vaccine distribution to @MrBeastYT

40

118

801

↑



balajis.com ✅

@balajis

...

Replying to @balajis

Not even kidding. Amazon has the most sophisticated logistics infra in the world.

Give them, Doordash, and Uber a few billion dollars each of the printed money, put them in touch with vaccine makers & nurses, and stop any government officials from interfering in any way.

2:22 AM · Jan 5, 2021 · Twitter for iPhone

[Texas does reasonable thing](#), and opens up Phase 1B which includes everyone over 65, which should be sufficient to find arms in which to put shots as needed:



Zoë McLaren, PhD
@ZoeMcLaren

...

In Texas everyone in Phase 1A and Phase 1B is now eligible to get the #COVID19 vaccine. 1/3
#txlege @DonnaHowardTX

dshs.texas.gov/coronavirus/im...

What's Next with the COVID-19 Vaccine in Texas

We Are Here

LIMITED SUPPLY

- * 1A: Direct Care - Hospital, Long-Term Care, EMS 9-1-1, Home Health, Outpatient, ER/Urgent Care, Pharmacies, Last Responders, School Nurses
- * 1A: Long-Term Care - Residents of Long-Term Care Facilities
- * 1B: Persons 65+ or 16+ with at least one chronic medical condition, including pregnancy

ADDITIONAL SUPPLY

- * 1C: Under consideration
- * 2: Under consideration

BROAD SUPPLY

- * 3: Under consideration

"All providers that have received COVID-19 vaccine must immediately vaccinate healthcare workers, Texans over the age of 65, and people with [medical conditions](#) that put them at a greater risk of severe disease or death from COVID-19. No vaccine should be kept in reserve."

- DSHS Commissioner John Hellerstedt, M.D.

[Meanwhile, in Israel](#), where they seem to care about vaccinating people...

Pfizer's vaccine, made with partner [BioNTech SE](#), must be administered within a five-day window after it leaves the main storage center, and six hours once out of a fridge, according to Israeli authorities, who say they are following Pfizer's rules.

To cope with that short shelf life and help authorities reach less populated and isolated areas, Israel began splitting some of Pfizer's 1,000-dose packages into smaller consignments of a few hundred each. The system, in which workers repackage the vials in workstations within massive freezers, was approved by Pfizer before being implemented, Mr. Edelstein said.

Israel also enacted a policy that allows vaccine centers facing soon-to-be wasted surplus to inoculate anyone who shows up. This has led to scenes around the country of citizens both young and middle-aged queuing up at vaccine centers, hoping to get an early shot.

Current general mood about the whole thing:



Eliezer Yudkowsky @ESYudk... 19h

I'd be agitating to move MIRI to Israel, if it was located anywhere on Earth besides god-cursed Israel.

Nick Schrock @schrockn

Replies to @webdevMason

In Cuomo's America, this nurse in Israel that grabbed a pizza boy to vaccinate when a vaccine was about to expire would be fined a million dollars.

twitter.com/erlichya/status/1345524110300013573

twitter.com/schrockn/status/1345524110300013573

7 1 51 ...

What have you to say in your defense, "ethicists"?



Daniel Eth - just Biden' my tim... 1d

US medical system be like:

**It is better for 100 vaccines to spoil
than for 1 person to cut in line and
get a vaccine early**

0 1 2 ...

OK, that's not fair. The American medical system is not Andrew Cuomo, but what would be fair?

We could... [quote the FDA's statement in full](#) so there can be no misunderstanding?

Two different mRNA vaccines have now shown remarkable effectiveness of about 95% in preventing COVID-19 disease in adults. As the first round of vaccine recipients become eligible to receive their second dose, we want to remind the public about the importance of receiving COVID-19 vaccines according to how they've been authorized by the FDA in order to safely receive the level of protection observed in the large randomized trials supporting their effectiveness.

We have been following the discussions and news reports about reducing the number of doses, extending the length of time between doses, changing the dose (half-dose), or mixing and matching vaccines in order to immunize more people against COVID-19. These are all reasonable questions to consider and evaluate in clinical trials. However, at this time, suggesting changes to the FDA-authorized dosing or schedules of these vaccines is premature and not rooted solidly in the available evidence. Without appropriate data supporting such changes in vaccine administration, we run a significant risk of placing public health at risk, undermining the historic vaccination efforts to protect the population from COVID-19.

The available data continue to support the use of two specified doses of each authorized vaccine at specified intervals. For the Pfizer-BioNTech COVID-19 vaccine, the interval is 21 days between the first and second dose. And for the Moderna COVID-19 vaccine, the interval is 28 days between the first and second dose.

What we have seen is that the data in the firms' submissions regarding the first dose is commonly being misinterpreted. In the phase 3 trials, 98% of participants in the Pfizer-BioNTech trial and 92% of participants in the Moderna trial received two doses of the vaccine at either a three- or four-week interval, respectively. Those participants who did not receive two vaccine doses at either a three- or four-week interval were generally only followed for a short period of time, such that we cannot conclude anything definitive about the depth or duration of protection after a single dose of vaccine from the single dose percentages reported by the companies.

Using a single dose regimen and/or administering less than the dose studied in the clinical trials without understanding the nature of the depth and duration of protection that it provides is concerning, as there is some indication that the depth of the immune response is associated with the duration of protection provided. If people do not truly know how protective a vaccine is, there is the potential for harm because they may assume that they are fully protected when they are not, and accordingly, alter their behavior to take unnecessary risks.

We know that some of these discussions about changing the dosing schedule or dose are based on a belief that changing the dose or dosing schedule can help get more vaccine to the public faster. However, making such changes that are not supported by adequate scientific evidence may ultimately be counterproductive to public health.

We have committed time and time again to make decisions based on data and science. Until vaccine manufacturers have data and science supporting a change, we continue to strongly recommend that health care providers follow the FDA-authorized dosing schedule for each COVID-19 vaccine.

The FDA, an agency within the U.S. Department of Health and Human Services, protects the public health by assuring the safety, effectiveness, and security of human and veterinary drugs, vaccines and other biological products for human use, and medical devices. The agency also is responsible for the safety and security of our nation's food supply, cosmetics, dietary supplements, products that give off electronic radiation, and for regulating tobacco products.

In other words, 'data and science' that do not exactly conform to standards don't exist, so until you file the proper forms representing the properly numbered trials, in triplicate, we are not going to schedule a meeting for a few weeks later to consider your proposal to prevent more deaths rather than prevent less deaths via the use of these mysterious things you call "Bayesian evidence" and "logic."

[See thread:](#)



Robert Wiblin
@robertwiblin

...

A serious reasoning error that is particularly common among educated people is to argue that if a study hasn't been done on a particular question we have 'no data', and therefore no basis on which to form beliefs or act.

This is incorrect and dangerous. 1/

1:33 PM · Jan 3, 2021 · Twitter Web App

1.3K Retweets 334 Quote Tweets 5.2K Likes

Points also awarded here:



Gordon Mohr ,,,
@gojomo

...

Replying to @ATabarrok

Despite any precedent from history, there's still 'no evidence' from gold-standard RCTs that 'thinking ahead' will work against this novel coronavirus! so it would be 'imprudent' to 'ignore the science' and risk something like 'thinking ahead'.

4:08 PM · Jan 4, 2021 · Twitter Web App

2 Retweets 16 Likes

And yes, I think that is all *entirely* fair.

FDA delenda est.

Any defense beyond that and the response to it are going to fall under "Something Is Wrong On the Internet" so feel free to skip this section, but it seems like it should be written anyway.

[This](#) seems like the line of defense for those who aren't going to fall back on the pure 'it is forbidden' and bury their heads in the sand like ostriches until someone does an RCT. In advance I admit I may be unfair here, but this seemed like a relatively thoughtful version to me, while being illustrative of the fallacies involved:



Sarah Cobey
@sarahcobey

...

Lately I've seen people say that deviations from "ethical queues" for vaccination are NBD, that we should focus on getting the vaccine into as many arms as possible as fast as possible. If you care about deaths and years of life lost, the math doesn't really back this view up.

11:06 AM · Dec 31, 2020 · Twitter Web App

42 Retweets 9 Quote Tweets 143 Likes

She then links to [this study](#) and [this other study](#), which are models that conclude that vaccinating old people before young people results in more infections and less deaths, whereas vaccinating young before old does the opposite. The second, to its credit, points out that there are other considerations than a pure count of deaths, and it's not obvious the optimization target of public policy should be purely deaths prevented – as one comment pointed out, there's a reasonable case that QALYs saved is a better health target even if you ignore economics.

Which all makes sense, and I do buy that elderly vaccinations first minimize deaths, but only if you assume that both scenarios vaccinate the same number of people.

So it doesn't address the question in a useful way. She goes on this way:



Sarah Cobey @sarahcobey · Dec 31, 2020

...

What especially troubles me is that we're seeing departures from priority groups at research-rich academic medical centers. After immunizing high-risk HCWs, they sometimes move on to loosely affiliated low-priority groups (e.g., my computational, remote, young-ish, low-risk lab).

2

7

56



Sarah Cobey @sarahcobey · Dec 31, 2020

...

I don't know if these institutions cannot return vaccines to the city/county and feel they have no way to immunize high-risk patients or even high-risk non-medical employees more systematically. This urgently needs to be worked out.

2

3

37



Sarah Cobey @sarahcobey · Dec 31, 2020

...

I've not modeled this, but my intuition is that it would indeed be better for a vaccine to sit on the freezer shelf for a week if it were then to go into the arm of a nursing home resident instead of mine.

3

2

40



Sarah Cobey @sarahcobey · Dec 31, 2020

...

Obviously, we can't afford to be too precise. Our work suggests that a robust strategy is to prioritize by age group after key occupations/settings. Once again in this pandemic, it hurts to contemplate the impact of mangled logistics, especially on lives that have been on hold.

4

4

44



Sarah Cobey @sarahcobey · Dec 31, 2020

...

I hope leaders and those in positions to influence leaders--especially at privileged institutions--can do what they can to direct vaccines to the populations in most urgent need of them, and who have suffered disproportionately.

4

4

42



The third tweet illustrates the core issue. She is thinking of the alternative as 'sits on the shelf for a week' but that all the vaccinations still happen a week later.

If we all had confidence in that, there would (I presume) not be this chorus of panic and outrage. The problem is that we are literally worried the vaccines will expire and be thrown out, and the whole timetable will be thrown off, and it will be a huge disaster on every level. Which I'm presuming she would agree, if it were to happen, would be a huge disaster on every level.

And also that, by saving the extra dose, we would give a nursing home resident a dose when they would otherwise not have gotten one. Which doesn't even parse for me at this point. When it comes to nursing homes, supply is not even an issue. When you are administering 20% of your doses after several weeks, *returning the excess supply* does actual nothing. The bottlenecks lie elsewhere.

So while I'm very confident that she's right that it's worth waiting a week if it transfers the dose to a nursing home resident that would otherwise have to wait a long time (e.g. a month or more), I don't see how that has any bearing on our current reality.

The whole thread also illustrates the general assumption, by those who think we can do central planning, that we should treat the 'baseline scenario' as being able to allocate the way the planner wants, with no waste, no delays, no additional costs, no regulatory capture, no corruption, no allocation by politics and power that does something different from their policy paper, no perverse incentives causing strange destructive behaviors, and so on, when the *actual* 'baseline scenario' is to assume you get all of that.

No, you in practice mostly can't usefully return excess doses, and that's not going to change. There are lots of logistical and storage issues. And more to the point, once you allocate doses to an institution, *what the hell did you think was going to happen?* That they were going to give it back to be reallocated no one knows when to nowhere they're associated and nowhere they'd get any credit, denying it to their employees and allies, rather than allocate the doses via politics and power to their employees, allies and associates?

That they'd spend their own resources to find a *systematic* way to *quickly* find people who fit into remarkably strict categories, to whom to give, at their own time and expense, completely free shots, taking on all the liability involved, including threats from the government for criminal prosecution if such people turned out not to fit the criteria?

And you also want, in that last tweet, to put (at least some of) the burden *on these random institutions* to then allocate the vaccine *based on who has suffered disproportionately?* While also obeying all the official restrictions or else, and also when did that become a priority?

If we lowered the thresholds at least to something reasonably common and easy to verify, such as being Over 65, we could at least give them a *chance*. Right now, I really, really [don't know what you were expecting](#).

Vaccine Allocation By Politics and Power

I [keep hearing that nurses are actually unusually likely to refuse the vaccine](#), with the usual explanation being that they know all the things that can go wrong in the medical system:



Jeremiah Get the Damn Vax

@JeremyM72014840

...

Replies to [@JeremyM72014840](#) [@MKing7403](#) and [@Noahpinion](#)

My aunt is an RN in a semi rural area and she said majority of staff refused it. Likewise LTCF staffed heavily by poor AA most likely to be anti covid vax are resorting to bribing staff to take the vax some places. It's there they just can't get staff to take it...

7:52 PM · Dec 31, 2020 · Twitter Web App

It seems that nurses' unions are taking stands to prevent the shots from being mandatory. Which takes away the whole 'people can safely use the medical system' advantage of doing health care workers first.

Luckily, nursing homes, where the most is at stake, [seem to have it less bad](#).

Nursing homes seem to have reassuringly *low* refusal rates. 90% acceptance is not so bad.



Morgan Mckay @morganfmckay · 20h

...

NEWS: NYS Health Commissioner Dr. Howard Zucker says state does not have a definitive breakdown of hospital staff that has so far refused to take COVID-19 vaccine.

But 10% of nursing home residents and 15% of nursing home staff has so far refused to take the COVID-19 vaccine.

3

12

27

↑

[Fluent Cherokee speakers eligible for the vaccine](#). If we had only known this earlier, would a bunch more people be fluent in Cherokee by now?

That one time in DC when they managed to get expiring vaccines into random people's arms was good. Any idea why the vaccine was about to expire?

[Here's a hint:](#)



Jason Crawford
@jasoncrawford

...

I literally cannot find a place to sign up for a notification when I can get the covid vaccine. I checked [@Walgreens](#), [@cvsparmacy](#), [@Aetna](#), and the CA state's covid page.

I know I'm last on the priority list, but shouldn't I be able to get an email/text when it's ready for me?

2:55 PM · Jan 4, 2021 · Twitter Web App

5 Retweets 70 Likes

[Here's another hint:](#)



The End Times
@TheAgeofShoddy

...

The simplest possible explanation for this sort of thing is that they don't really want to vaccinate people. I'm theoretically open to other possibilities, but based on the track record of the last year local governments and health bureaucracies have earned that assumption.



Spencer Brown  @itsSpencerBrown · 23h

64% of DC's appointments to get the first dose of a COVID vaccine are currently unbooked, per today's briefing by Mayor Bowser.

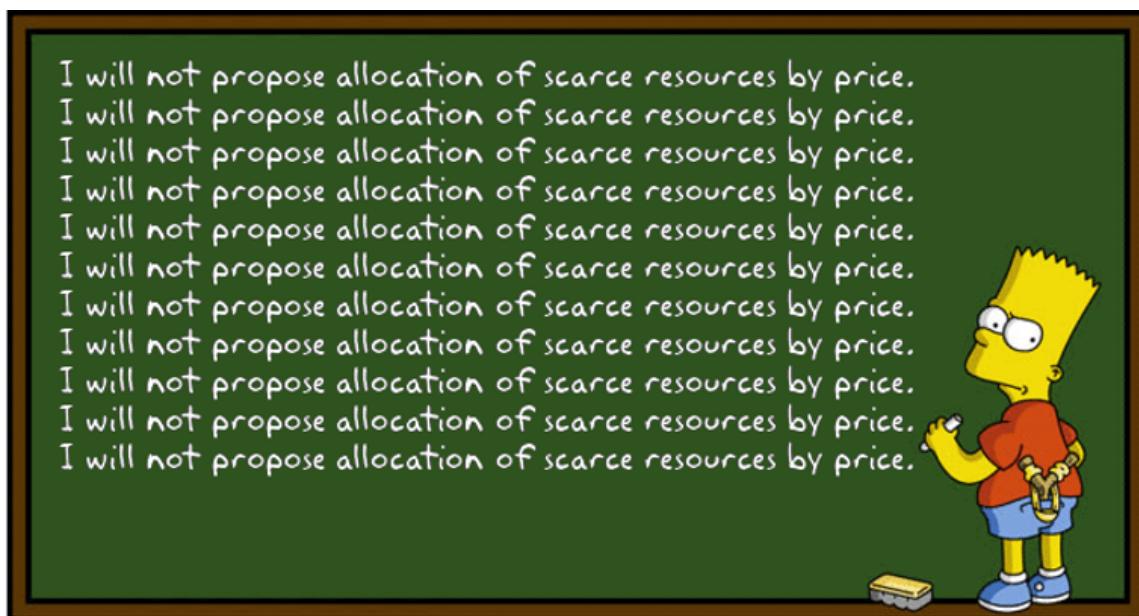
Government literally cannot even give lifesaving medicine away for free effectively.

I think that's a *bit* harsh there by Shoddy. More precise would be [they instinctively avoid being seen as taking actions which could plausibly be interpreted as motivated by getting people vaccinated](#). If they had a choice, they would totally want people to get vaccinated! But they sense a grave danger if the primacy of that motivation on their part were to become common knowledge.

Thus, they can take actions that are motivated by *avoiding blame for a lack of vaccinations within the next two weeks or so*. They can take actions to avoid blame for vaccines actively going into trash cans. What they cannot do are the things one would do if one wanted to maximize the number of needles that get put into arms.

Thus, restrictions on who can take appointments that are so harsh that *a majority of appointment times* are unclaimed, rather than open up to the next tier to ensure enough demand, or allow lower tiers to make appointments on shorter notice - e.g. if the simple solution is unthinkable, how about only those in Tier 1A (whatever that locally is) can book two days in advance, then it's opened to Tier 1B the day before, and Tier 1C same day, and you can be put on their waiting list? Or a line outside the building (at whatever tier of eligibility), and if there's an empty slot, first come first serve?

Because I've certainly learned my lesson and I'm not at all bitter about it or saying I Told You So over and over again in great detail:



Another problem is that if [something goes wrong and some doses get spoiled](#), which is something that will happen, who knows when you can get more of them?

METRO



NY nursing home says COVID-19 vaccines spoiled by defrosting issues

By Bernadette Hogan

January 5, 2021 | 6:36pm | Updated

"The pharmacist running it told me they had lost some doses of the vaccine in one of their storage facilities due to temperature control issues. They did attempt to get additional vaccine doses that couldn't get them," LaDue told The Post Tuesday.

I found this interesting in particular:

"They administered every dose they had on site which was 132 residents and 56 staff members. We were hoping to get all the residents done and a portion of the staff," he added.

They were planning to do all residents and some staff, then lost doses, so they did some residents and some staff. Sounds like some staff (and perhaps some residents) had local political power to steer allocation, and others did not.

And of course, refer to much of the section on Andrew Cuomo, as I've moved the rest of the New York examples over there.

How Bad is it Out There Right Now?

Pretty bad (this is from Southern California):



Taz Ahmed

@TazzyStar

My dad's friend died of COVID-19 today in So Cal - the Muslim graveyards are so backed up, there is a two week wait. They can only bury 4 graves a day. Muslim burials are supposed to be asap, 48 hours usually. There is not enough room & time for the burials of all the dead.

5:49 PM · Mar 21, 2021 · Twitter Web App



Taz Ahmed @TazzyStar

1d

Muslims are also supposed to partake in a last wash - ghusl - before you are buried. You are being washed/blessed for the afterlife. People that dying from COVID-19 are not getting ghusl. Ten people max at burials. All the end of life prayers and rituals are truncated.



10



169



2k

...

That was a few days ago, now [the headlines from LA are getting much worse](#). We have reached full triage mode:

Los Angeles County ambulance crews are told not to transport patients with little chance of survival

By Alexandra Meeks, Christina Maxouris and Holly Yan, CNN

① Updated 1:20 PM ET, Tue January 5, 2021

Ambulances wait for hours outside hospitals

Even when patients are lucky enough to get to a hospital, they might languish outside for hours if there's no more room.

"The Emergency Medical Services are working very hard to divert ambulances or send them to hospitals that do have potential capacity to receive those patients," said Smith, COO of Cedars-Sinai Medical Center.

"There are situations where patients are made to wait in ambulances under the care of the paramedics. We want to make sure that time is as short as possible so they can receive the necessary care."

For EMT Jimmy Webb, the wait can last several hours.



Related Article: Hospitals are feeling the impact of holiday gatherings

"We are waiting two to four hours minimum to a hospital, and now we are having to drive even further ... then wait another three hours," Webb told [CNN affiliate KCAL](#).

Local officials have urged the public not to call 911 unless "they really need to," Dr. Marc Eckstein, head of the Los Angeles Fire Department EMS bureau, told [CNN affiliate KABC](#).

"One of our biggest challenges right now is getting our ambulances out of the emergency department," he said.

Hopefully we are very near the peak of how bad things will be this wave, but that is unlikely to be true in at least some places.

[A thread on what it's like to be treated for Covid-19](#), first tweet below.



Joanna Poole
@Jopo_dr

...

Covid intensive care is: hourly arterial blood samples to check oxygen levels, lines placed into arteries to sample them, lines in the jugular vein to replace electrolytes we have driven out of the body by trying to flush excess water from the lungs, lines into the neck and groin

[Paper on deaths of despair.](#)

[Even the paper itself is not safe:](#)



Alec MacGillis



@AlecMacGillis

...

There has been such an increase in robberies and carjackings in DC that the Washington Post recently alerted some subscribers that it wouldn't start delivering papers until daylight.



| r/washingtondc

Use App



The Washington Post

November 30, 2020

Dear [REDACTED]

I am writing to let you know that, to ensure the safety of our delivery contractors, you will experience changes to the delivery of your newspaper starting immediately. In the last several weeks, our delivery contractors in your area have experienced repeated attempted robberies and carjackings while on their newspaper delivery routes.

Effective immediately, deliveries in your area will begin after sunrise each day. Waiting until daylight hours will greatly enhance the personal safety of your newspaper carrier. Our expectation is that you should receive your paper no later than 9:00 a.m. Monday-Friday and 10:00 a.m. on Saturday and Sunday. I

[A study out of the Ohio State University](#) from September (I looked for it and couldn't find a reference in past weeks, but I could have missed it) [finds that 30% of student athletes who previously had Covid-19 had heart damage linked to Covid-19](#). The whole thing is surprising

not only [because it means that athletes at the Ohio State University had hearts to begin with](#) (Roll Tide on Monday!), but also because it means that they did useful work at Ohio State, it potentially interfered with their athletic department and they still published. It all makes so little sense. My model of the world can at least take comfort in knowing that this in no way actually interfered with anyone's athletic operations.

From a tweet [on the source thread](#) that would not like to get broadcast too widely, not verified at all:

My husband is on the team developing the disability policy for this issue, from what he's told me they don't expect it to go well for people who have this for a long time. Similar to Ebola

I don't expect this to turn out to be a big deal, but I could easily be wrong and am choosing not to do the research, so if the evidence indicates otherwise, please speak up and explain.

Hospitals are not collapsed as such, [but things are very much not good](#):



COVID deaths lag case counts by 3-4 weeks

@olimay

...

Another example within a recurring theme I hear from friends in healthcare. Most are junior, so afraid to speak out due to realistic concerns of retribution from administrators. Others are worried voicing concerns publicly will undermine a united message to the public.



Emily Porter, M.D. @dremilyportermd · Jan 1

It's 2 AM. My husband is still awake and hacking. 7 d after fever started and 6 d after being diagnosed with COVID. And he's telling me no one has filled his ER shift in 4 hr so he's expected to work. CDC guidelines say a minimum of 10 days and if sx improving. Cough is not.

[Show this thread](#)

So, [I guess this might have happened](#), with 44 cases at the hospital:

The hospital is investigating whether an incident in which a staff member appeared briefly in the emergency department on Christmas Day wearing an air-powered costume with a fan may have led to air droplets being spread around the hospital.

In Other News

Off-topic but awesome: [Money Stuff is back!](#) Matt Levine returns from parental leave with the best daily newsletter I have ever known, and it is not remotely close. This is about the latest happenings of all things money, hence the name, but written in such a way that it appeals both to professional traders and to people who don't otherwise know or care about money stuff. I've been known to read the column out loud to my wife, without having read it first, and it's reliably both funny and enlightening. The perspective you get here is not the whole picture, but your picture of the world without it will be incomplete. You can fix that. Strong endorsement.

Bryan Caplan asks, have you tried... [being happy this year?](#)

Your periodic reminder that in 1947, [New York City vaccinated ~6.35 million people \(80% of their population\) for smallpox in less than a month](#). If you do not think we can do this, what changed to make it impossible?

[Your periodic reminder that the FDA's incentives are bad.](#)

What did China do to stop this virus and get back to normal? [What didn't it do?](#) (Inessential and snarky Twitter thread)

[British official report on AstraZeneca vaccine is out.](#) As we expected, less good than Pfizer or Moderna's in terms of effectiveness, far better than it needs to be to be worthwhile.

[There's an AstraZeneca manufacturing plant in Baltimore.](#) It seems we're really going to wait until April to approve their vaccine, the interpretation of which I leave to you the reader.

AstraZeneca has vowed not to make a profit on their vaccine 'until the pandemic is over.' [They have also declared the right to declare the pandemic 'over' in July 2021.](#) So due to the fear that someone, somewhere, might be earning a profit, a major corporation makes money if they sell doses after July 2021, but not if they sell them before July 2021. I wonder what they will do. ([Original Hat Tip](#))

Every few weeks, the evidence for immunity for those who have been infected gets one week longer, and by Lindy that means two more weeks of immunity, yet I *still* get people claiming "we don't know how long immunity lasts" or saying it fades after a few months, and

so on. This leads to insanity [like this](#), where in order to satisfy such folks, a star NBA player is missing games under quarantine for exposure *despite having not only already had Covid-19, but having been confirmed to have antibodies*:



Nate Silver  @NateSilver538 · 16h

...

One thing the league might think about doing is testing players for antibodies in these cases.

 **Henry Abbott**  @TrueHoop · 16h

At some point we'll have better data about protocols for a guy who tested positive last year. Can he be a spreader? I believe at this point the data says, essentially: hard to say/we don't want to take any chances.
[twitter.com/ShamsCharania/...](https://twitter.com/ShamsCharania/)

19

10

211

↑



Nate Silver  @NateSilver538

...

Replies to [@NateSilver538](#)

Ahh, that gets a bit harder to justify. I don't 100% blame the NBA because the league is catering to public opinion and there are still a lot of people who haven't kept up with the science and believe stuff like "immunity fades after 3 months!". But still—



Kostya Medvedovsky @kmedved · 16h

Replies to [@NateSilver538](#)

They are testing for antibodies. He's positive for them!
twitter.com/wojespn/status...

I don't blame the NBA either. People respond to incentives. The NBA is very well known to not be an exception to this. [Trust the process](#).

You know who has a lot of supply coming soon? [India](#):



Mike Bird
@Birdyword

...

The Serum Institute of India expects to have 100 million doses of the Oxford vaccine ready for developing countries within two weeks, eclipsing the UK's supply. Has 50 million in vials right now. The UK has 530,000 doses ready to go on Monday.

If you live in the greater Washington D.C. area, there was a super-spreader event on January 6, so stay safe. Everyone else, you too.

I'll see everyone next week. Hopefully with better news.

Simulacrum 3 As Stag-Hunt Strategy

Reminder of the rules of Stag Hunt:

- Each player chooses to hunt either Rabbit or Stag
- Players who choose Rabbit receive a small reward regardless of what everyone else chooses
- Players who choose Stag receive a large reward if-and-only-if everyone else chooses Stag. If even a single player chooses Rabbit, then all the Stag-hunters receive zero reward.

From the outside, the obvious choice is for everyone to hunt Stag. But in real-world situations, there's lots of noise and uncertainty, and not everyone sees the game the same way, so [the Schelling choice is Rabbit](#).

How does one make a Stag hunt happen, rather than a Rabbit hunt, even though the Schelling choice is Rabbit?

If one were utterly unscrupulous, one strategy would be to try to trick everyone into thinking that Stag is the obvious right choice, regardless of what everyone else is doing.

Now, tricking people is usually a risky strategy at best - it's not something we can expect to work reliably, especially if we need to trick everyone. But this is an unusual case: we're tricking people in a way which (we expect) will benefit them. Therefore, they have an incentive to play along.

So: we make our case for Stag, try to convince people it's the obviously-correct choice no matter what. And... they're not fooled. But they all *pretend* to be fooled. And they all look around at each other, see everyone else also pretending to be fooled, and deduce that everyone else will therefore choose Stag. And if everyone else is choosing Stag... well then, Stag actually is the obvious choice. Just like that, Stag becomes the new Schelling point.

We can even take it a step further.

If nobody *actually* needs to be convinced that Stag is the best choice regardless, then we don't *actually* need to try to trick them. We can just *pretend* to try to trick them. Pretend to pretend that Stag is the best choice regardless. That will give everyone else the opportunity to pretend to be fooled by this utterly transparent ploy, and once again we're off to hunt Stag.

This is [simulacrum 3](#): we're not telling the truth about reality (simulacrum 1), or pretending that reality is some other way in order to manipulate people (simulacrum 2). We're pretending to pretend that reality is some other way, so that everyone else can play along.

In The Wild

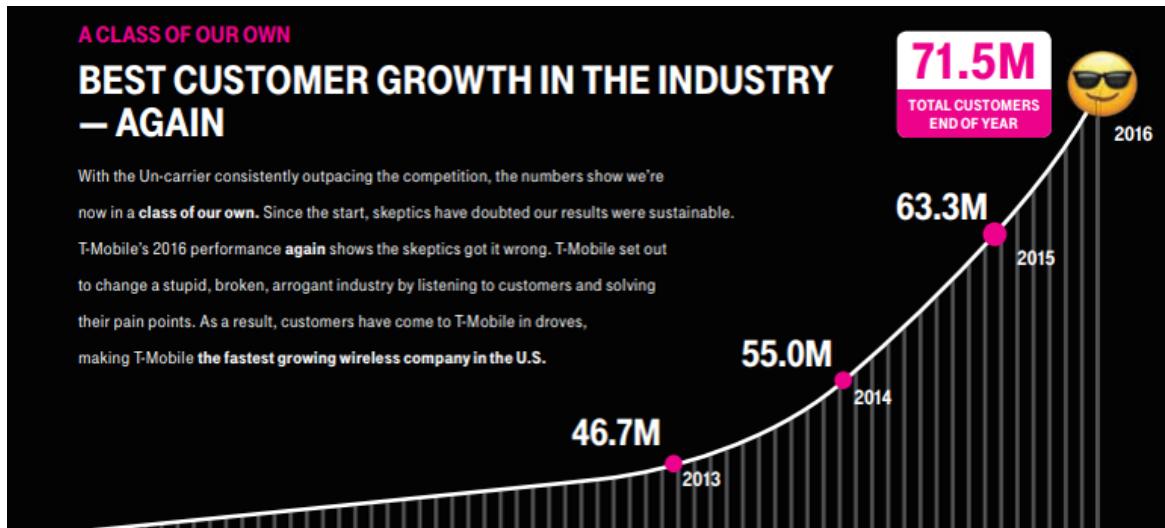
We have a model for how-to-win-at-Stag-Hunt. If it actually works, we'd expect to find it in the wild in places where economic selection pressure favors groups which can hunt Stag. More precisely: we want to look for places where the payout increases faster-than-linearly with the number of people buying in. Economics jargon: we're looking for increasing marginal returns.

Telecoms, for instance, are a textbook example. One telecom network connecting fifty cities is far more valuable than fifty networks which each only work within one city. In terms of marginal returns: the fifty-first city connected to a network contributes more value than the

first, since anyone in the first fifty cities can reach a person in the fifty-first. The bigger the network, the more valuable it is to expand it.

From an investor's standpoint, this means that a telecom investment is likely to have better returns if more people invest in it. It's like a Stag Hunt for investors: each investor wants to invest if-and-only-if enough other investors also invest. (Though note that it's more robust than a true Stag Hunt - we don't need literally every investor to invest in order to get a big payoff.)

Which brings us to this graph, from [T-mobile's 2016 annual report](#) (second page):



Fun fact: that is not a graph of those numbers. Some clever person took the numbers, and stuck them as labels *on a completely unrelated graph*. Those numbers are actually near-perfectly linear, with a tiny amount of downward curvature.

Who is this supposed to fool, and to what end?

This certainly shouldn't fool any serious investment analyst. They'll all have their own spreadsheets and graphs forecasting T-mobile's growth. Unless T-mobile's management deeply and fundamentally disbelieves the efficient markets hypothesis, this isn't going to inflate the stock price.

It could just be that T-mobile's management were themselves morons, or had probably-unrealistic models of just how moronic their investors were. Still, I'd expect competition (both market pressure and investor pressure in shareholder/board meetings) to weed out that level of stupidity.

My current best guess is that this graph is not intended to actually fool anyone - at least not anyone who cares enough to pay attention. This graph is simulacrum 3 behavior: it's pretending to pretend that growth is accelerating. Individual investors play along, pretending to be fooled so that all the investors will see them pretending to be fooled. The end result is that everyone hunts the Stag: the investors invest in T-mobile, T-mobile uses the investments to continue expansion, and increasing investment yields increasing returns because that's how telecom networks work.

... Well, that's almost my model. It still needs one final tweak.

I've worded all this as though T-mobile's managers and investors are actually thinking through this confusing recursive rabbit-hole of a strategy. But [that's not how economics usually works](#). This whole strategy works just as well when people accidentally stumble into

it. Managers see that companies grow when they “try to look good for investors”. Managers don’t need to have a whole gears-level model of *how* or *why* that works, they just see that it works and then do more of it. Likewise with the investors, though to a more limited extent.

And thus, a [maze](#) is born: there are real economic incentives for managers to pretend (or at least pretend to pretend) that the company is profitable and growing and all-around deserving of sunglasses emoji, regardless of what’s actually going on. In industries where this sort of behavior results in actually-higher profits/growth/etc, economic pressure will select for managers who play the game, whether intentionally or not. In particular, we should expect this sort of thing in industries with increasing marginal returns on investment.

Takeaway

Reality is that which remains even if you don’t believe it. Simulacrum 3 is that which remains only if enough people pretend (or at least pretend to pretend) to believe it. Sometimes that’s enough to create real value - in particular by solving Stag Hunt problems. In such situations, economic pressure will select for groups which naturally engage in simulacrum-3-style behavior.

Covid 1/21: Turning the Corner

Aside from worries over the new strains, I would be saying this was an exceptionally good week.

Both deaths and positive test percentages took a dramatic turn downwards, and likely will continue that trend for at least several weeks. Things are still quite short-term bad in many places, but things are starting to improve. Even hospitalizations are slightly down.

It is noticeably safer out there than it was a few weeks ago, and a few weeks from now will be noticeably safer than it is today.

Studies came out that confirmed that being previously infected conveys strong immunity for as long as we have been able to measure it. As usual, the findings were misrepresented, but the news is good. [I put my analysis here in a distinct post](#), so it can be linked to on its own.

We had a peaceful transition of power, which is always a historic miracle to be celebrated.

Vaccination rollout is still a disaster compared to what we would prefer, with new disasters on the horizon (with several sections devoted to all that), but we are getting increasing numbers of shots into increasing numbers of arms, and that is what matters most. In many places we have made the pivot from 'plenty of vaccine and not enough arms to put shots into' to the better problem of 'plenty of arms to put vaccine into, but not enough shots.' Then all we have to do is minimize how many shots go in the trash, including the extra shots at the bottom of the vial, and do everything we can to ramp up manufacturing capacity. Which it seems can still be meaningfully done.

The problem is that the new strains are coming.

The English strain will arrive first, within a few months. That's definitely happening, and the only question is how bad it's going to get before we can turn the tide. We are in a race against time.

The South African and Brazilian strains are not coming as fast, but are potentially even scarier. There are signs of potential escape from not only vaccination but previous infection, potentially allowing reinfection to take place. See the section on them for details, and if you can help provide better information, please do so. We need clarity on this, and we need it badly.

There are also all the *other* new strains being talked about, which are probably nothing, but there's always the chance that's not true.

But first, the good news, and it is very, very good. Let's run the numbers.

The Numbers

Predictions

Prediction last week: 14.0% positive rate on 11.7 million tests, and an average of 3,650 deaths.

Results: 11.9% positive rate on 11.3 million tests, and an average of 3,043 deaths.

Both numbers are hugely pleasant surprises, and this is the biggest directional miss I've had on deaths.

Last week we were at 3,335 deaths per day, and I figured things would keep getting worse for another week or two. Instead, things are already on their way to rapid improvement, unless there were massive shifts in when deaths were reported that made last week look worse than it was.

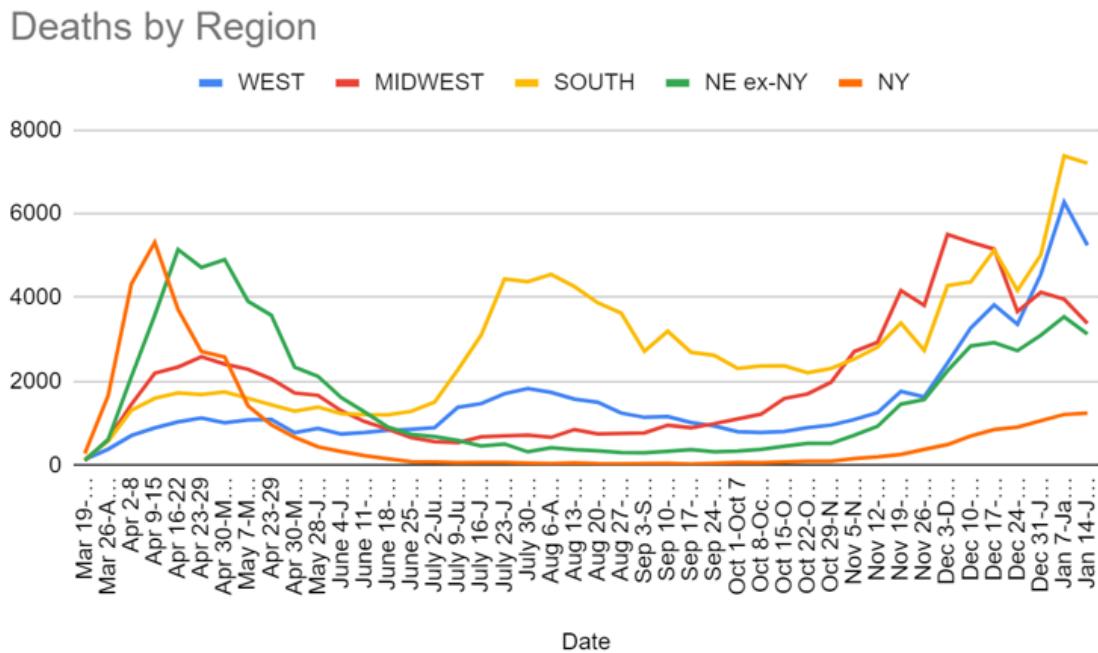
For infections, I did predict a drop (last week was 15.2%) and we got a much more dramatic drop than I expected. This was wonderful news, and it seems like this should continue.

The caveat is that Tuesday and Wednesday of this week both look suspiciously good on both stats, such that I suspect missing data. I don't know if somehow Martin Luther King Day actually mattered to reporting, or the inauguration and fears of disruptions around it were distracting, or what, but we should worry that this is getting a bit ahead of ourselves, even though test counts would indicate otherwise.

Test count predictions don't seem worth doing, so going to stop doing those.

Prediction: 10.5% positive rate and 2,900 deaths per day. I'm being conservative because I worry about the drops from this week being data artifacts, but I am confident things are improving for now. Starting next week I'll be expecting the IFR to start dropping substantially due to selective vaccinations.

Deaths

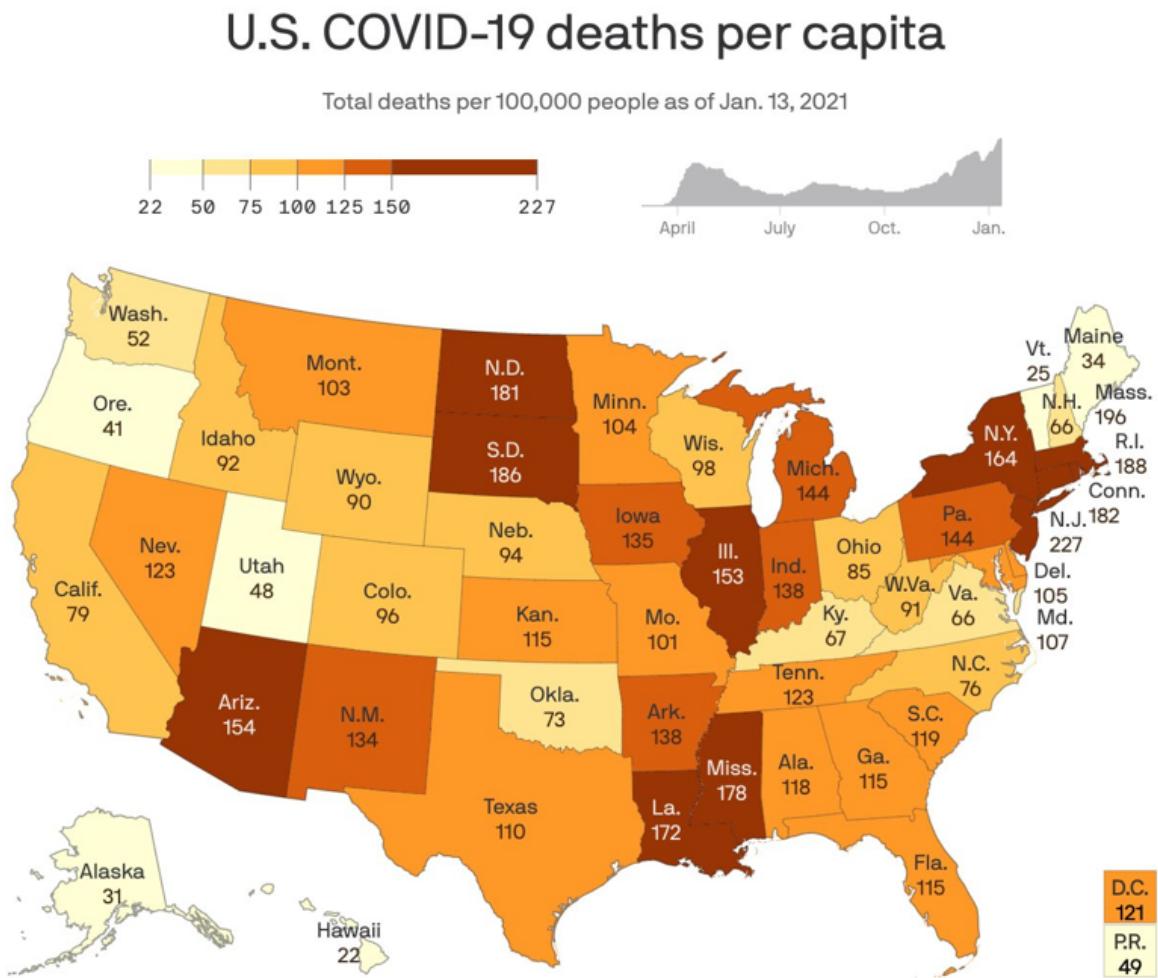


Date	WEST	MIDWEST	SOUTH	NORTHEAST
Nov 19-Nov 25	1761	4169	3396	1714
Nov 26-Dec 2	1628	3814	2742	1939
Dec 3-Dec 9	2437	5508	4286	2744
Dec 10-Dec 16	3278	5324	4376	3541
Dec 17-Dec 23	3826	5158	5131	3772
Dec 24-Dec 30	3363	3668	4171	3640

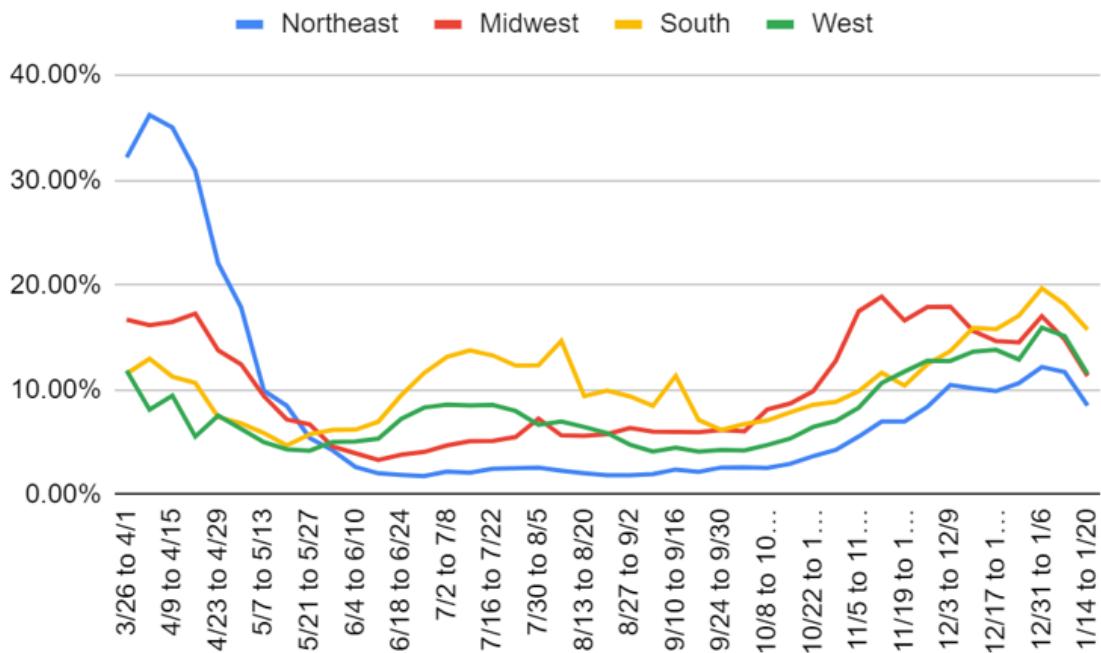
Dec 31-Jan 6	4553	4127	5019	4162
Jan 7-Jan 13	6280	3963	7383	4752
Jan 14-Jan 20	5249	3386	7207	4370

As noted above, this was expected to get much worse, and instead things started improving, although they're still in a worse spot than two weeks ago. This is very good news, and it sheds new light on what has been happening in the past few weeks. If everything we'd seen previously had been fully reflective of the situation on the ground, we would not have seen a decline in deaths this week.

This graphic of cumulative deaths comes [courtesy of Venkesh Rao on Twitter](#), seemed crisp and useful enough to include, from a few days ago:



Positive Test Percentages

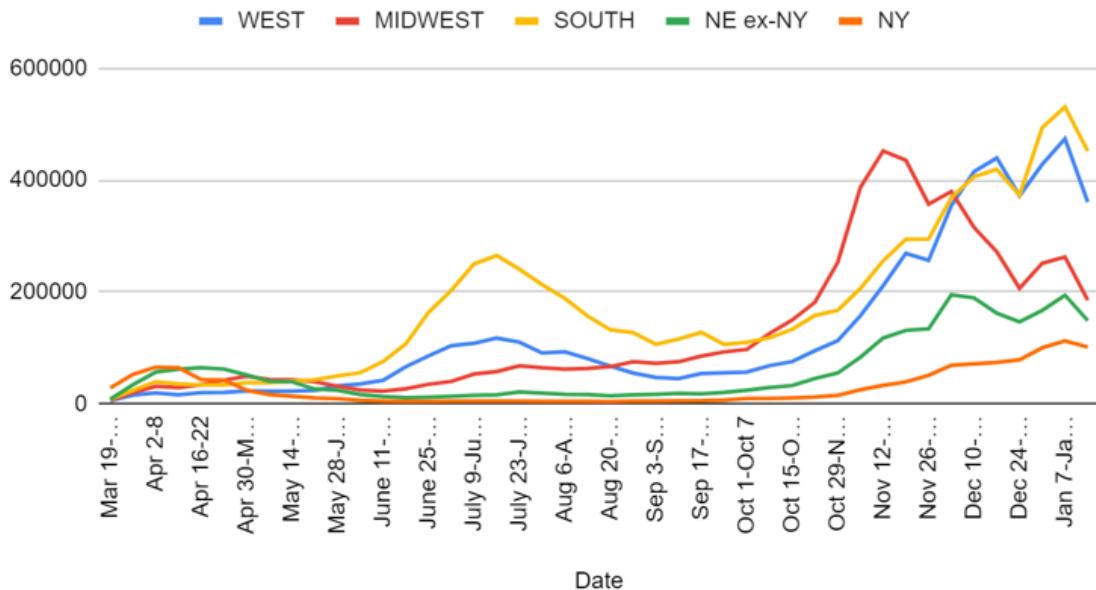


Percentages	Northeast	Midwest	South	West
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%
12/3 to 12/9	10.47%	17.94%	13.70%	12.76%
12/10 to 12/16	10.15%	15.63%	15.91%	13.65%
12/17 to 12/23	9.88%	14.65%	15.78%	13.82%
12/24 to 12/30	10.65%	14.54%	17.07%	12.90%
12/31 to 1/6	12.18%	17.03%	19.69%	15.94%
1/7 to 1/13	11.70%	14.81%	18.14%	15.12%
1/14 to 1/20	8.50%	11.32%	15.75%	11.53%

Test counts are up, positive test rates are down everywhere. Great numbers.

Positive Tests

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Dec 3-Dec 9	354,397	379,823	368,596	263,886
Dec 10-Dec 16	415,220	315,304	406,353	260,863
Dec 17-Dec 23	439,493	271,825	419,230	236,264
Dec 24-Dec 30	372,095	206,671	373,086	225,476
Dec 31-Jan 6	428,407	251,443	494,090	267,350
Jan 7-Jan 13	474,002	262,520	531,046	306,604
Jan 14-Jan 20	360,874	185,412	452,092	250,439

Good news all around, and overall test count was even up about 2%.

Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Nov 19-Nov 25	10,421,697	11.8%	1,373,751	2.9%	3.88%
Nov 26-Dec 2	9,731,804	11.8%	1,287,010	4.0%	4.23%
Dec 3-Dec 9	10,466,204	13.9%	1,411,142	4.9%	4.67%
Dec 10-Dec 16	10,695,115	13.9%	1,444,725	4.9%	5.12%
Dec 17-Dec 23	10,714,411	13.7%	1,440,770	5.1%	5.57%
Dec 24-Dec 30	9,089,799	13.8%	1,303,286	6.0%	5.95%
Dec 31-Jan 6	9,334,345	16.4%	1,365,473	7.3%	6.42%
Jan 7-Jan 13	11,084,291	15.2%	1,697,034	6.6%	6.93%
Jan 14-Jan 20	11,300,725	11.9%	1,721,440	5.9%	7.35%

In addition to the numbers listed, hospitalizations are also finally on the decline. I don't generally track hospitalizations because I worry the limiting factor is often hospital beds, but seeing a decline is definitely a very good sign.

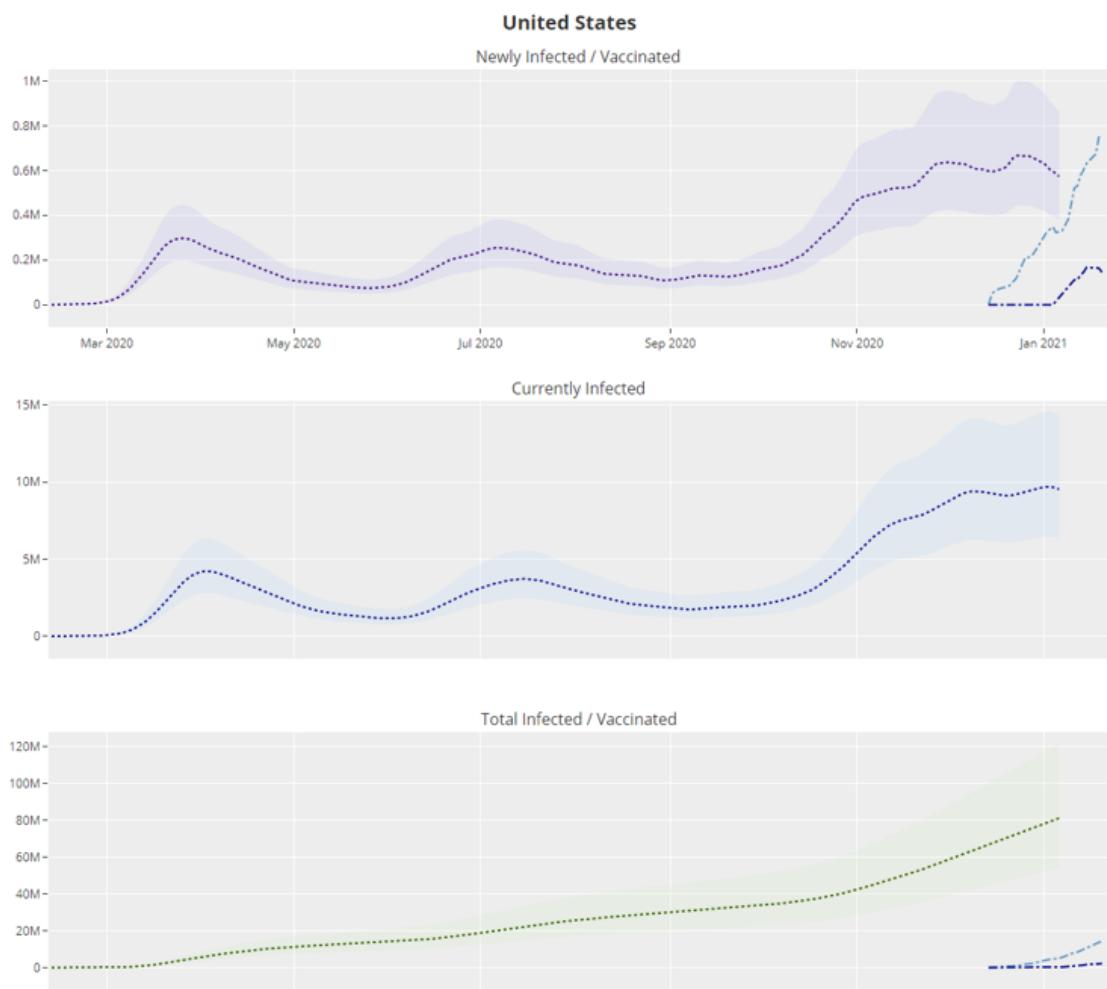
Covid Machine Learning Project

Last Updated: Thursday, January 21, 2021 (4am ET)

Newly Vaccinated (as of Jan 20): **747,000 / day** (269 / 100k)
Total Vaccinated (as of Jan 20): **4.3%** (1 in 23 | 14.4 million)

Newly Infected (as of Jan 6): **574,000 / day** (170 / 100k)
Currently Infected (as of Jan 6): **1 in 30** (2.9% | 9.5 million)
Total Infected (as of Jan 6): **24.5%** (1 in 4 | 81.2 million)

R_t (as of Jan 6): **0.92**
Adjusted Positivity Rate (as of Jan 20): **9.4%**
Infection-to-Case Ratio (as of Jan 20): **3.0** (34% detection rate)



Look at that vaccination line shoot upwards and the newly infected line start heading downwards. You love to see it.

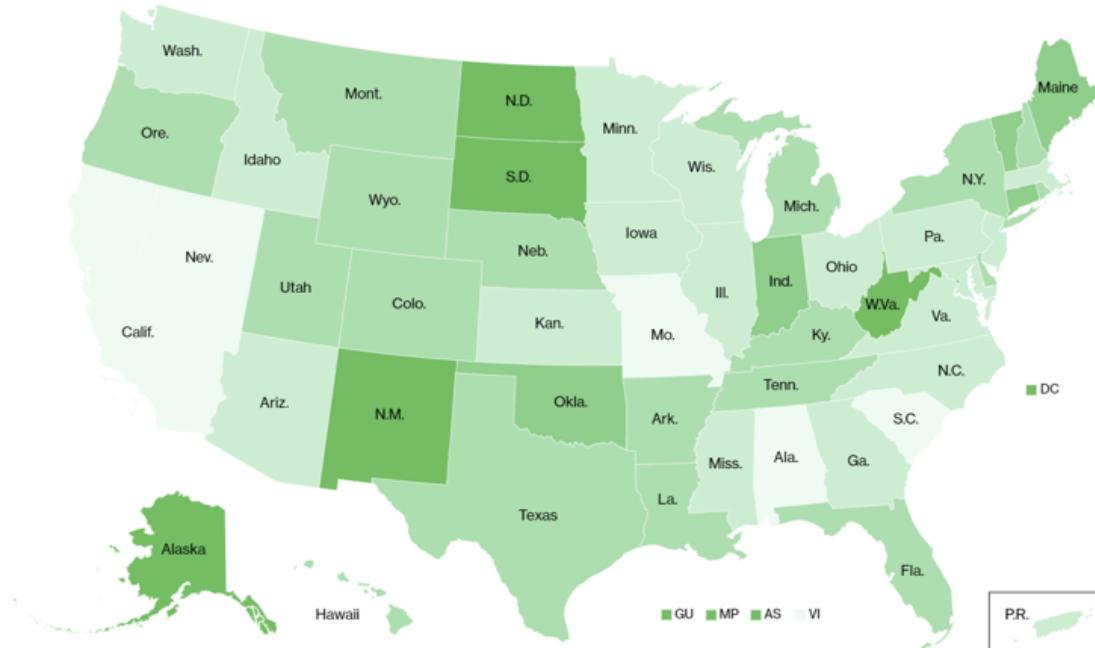
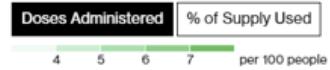
As of January 6 these projections had us at 24.5% infected, versus 23.4% a week before. This continues to be my rough lower bound for how many people have been infected. Herd immunity from infection is having a big and growing impact.

Vaccinations

Vaccinations in the U.S. began Dec. 14 with health-care workers, and so far **17.2 million shots** have been given, according to a state-by-state tally by Bloomberg and data from the Centers for Disease Control and Prevention. In the last week, an average of **912,497 doses per day** were administered.

Vaccines Across America

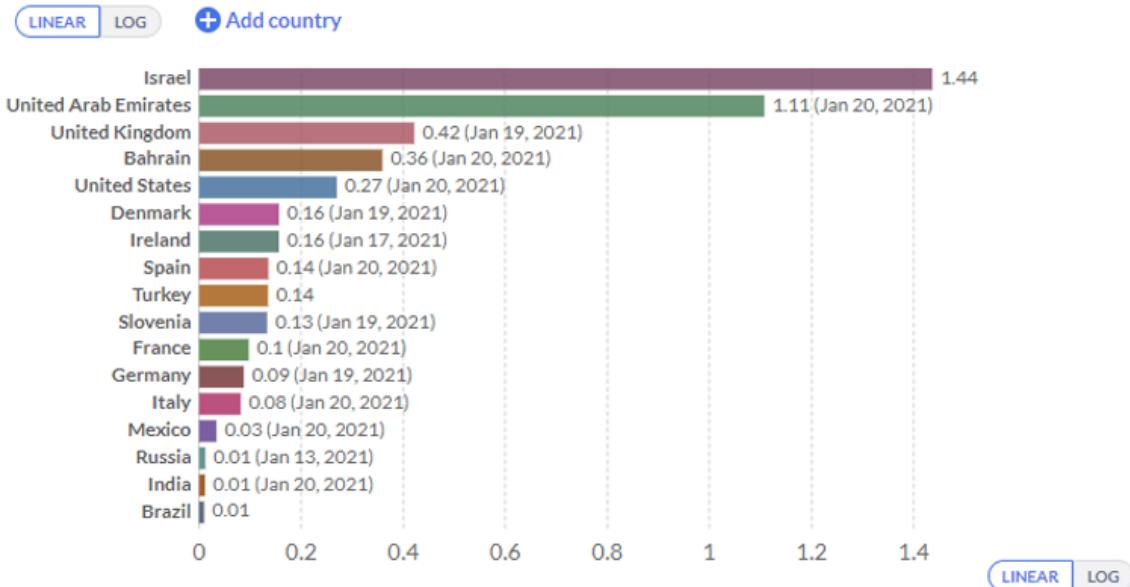
Across the U.S., 5.2 doses have been administered for every 100 people, and 48% of the shots distributed to states have been administered



Daily COVID-19 vaccine doses administered per 100 people, Jan 21, 2021

Our World
in Data

Shown is the rolling 7-day average per 100 people in the total population. This is counted as a single dose, and may not equal the total number of people vaccinated, depending on the specific dose regime (e.g. people receive multiple doses).



Relative regional progress remains unchanged. If you were behind last week, you're almost certainly even further behind now. California continues to do an unusually disgraceful job with its vaccine rollout, which will be discussed extensively later. New York, for all the complaining that gets done about it, is doing relatively fine.

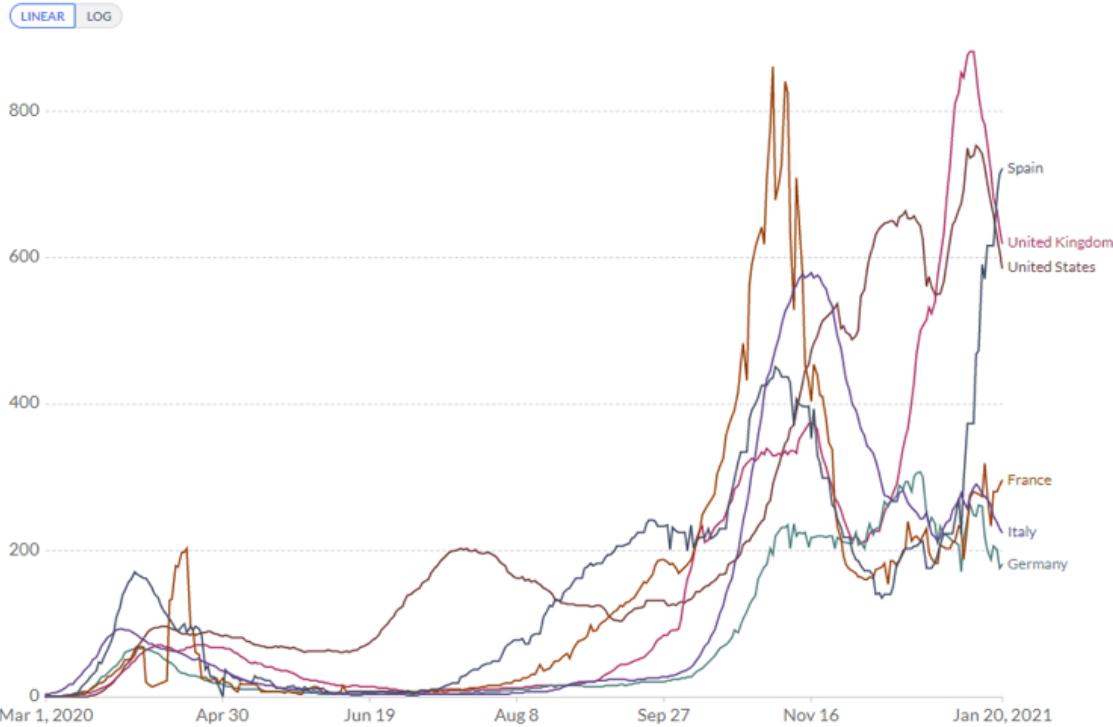
The headline number is 912k doses per day over the course of the week, the bulk of which were first doses. That's not great, and it isn't improving that quickly, but it's much less disastrous than the worst scenarios that were being pondered. It's also enough that we should start seeing the effect of those vaccinations in both infections and deaths soon if we aren't seeing it already.

Europe

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

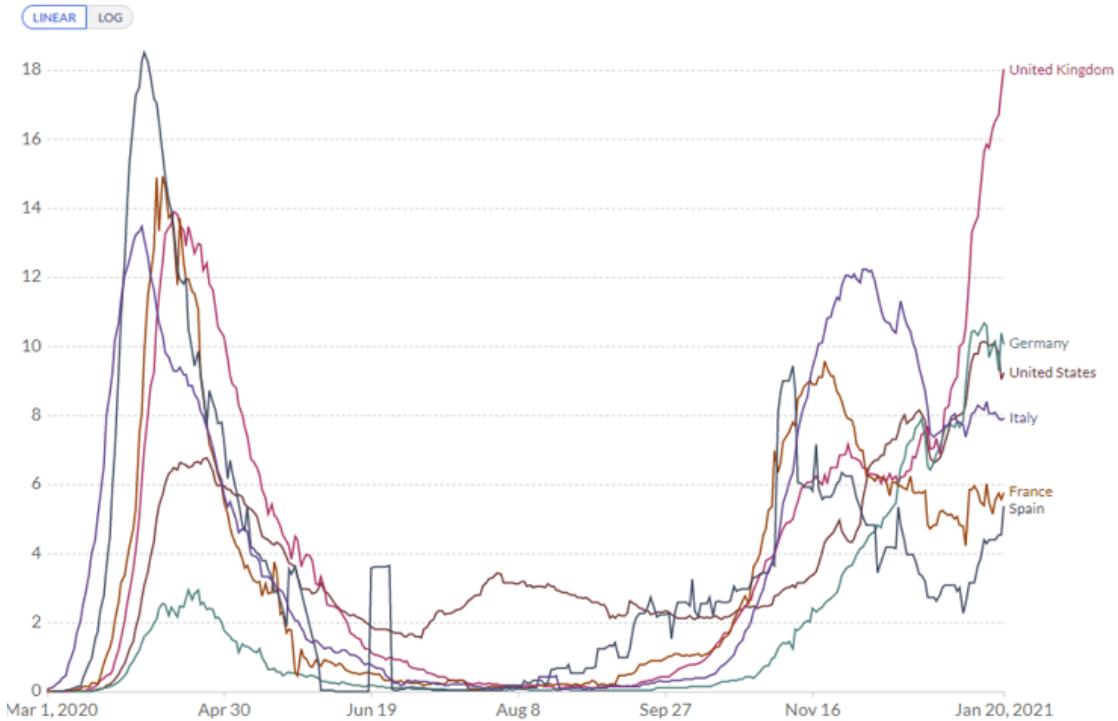
Our World
in Data



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



The first graph's story is: All hail the control system. The United Kingdom is on its way back down in infections once again, despite the domination of the new strain. The peak was January 9th. Ireland is not pictured, but it peaked on the 10th at an even higher rate and is following a similar curve.

The second graph's story is that the new strain is still killing an awful lot of people before that happens. We are currently 12 days past the peak of infections, so the line should keep going up for a few more days.

Now Spain is out of control. I don't know if that is partly because of the English strain taking over, or entirely for other reasons.

It seems clear that yes, with sufficiently strong restrictions and private reactions, the United Kingdom at least can stabilize the infection level against the new strain. I still am not sure if America could or would do the same under the same conditions. We'll find out soon, with conditions that in some ways will be substantially more favorable, with more people vaccinated and more people having already been infected.

The English Strain

[Scott Gottlieb reached the same core conclusion I did](#), at least by January 17, that the new strain will likely double every week, so we'll see a few weeks of declines and then things start getting worse again. [So did Eric Feigl-Ding](#), and many others, including the CDC itself. It seems that my core conclusions of December 24 are now rapidly becoming the official Very Serious Person perspective.

[The CDC is out with their analysis](#), accepting the basic premise of increased transmission and modeling outcomes. They are assuming a baseline of 0.5% of cases are the new strain at start of the year, which seems reasonable. [I did some toy modeling](#), and they are doing some toy modeling, except with more toys and less models.

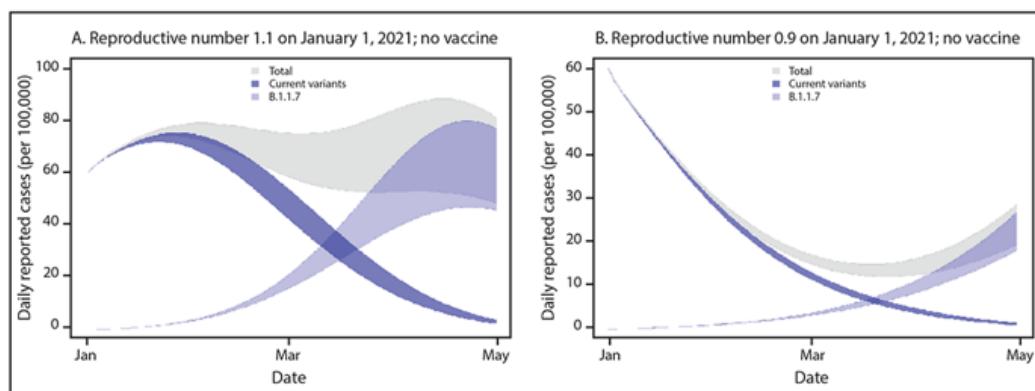
They split into the $R_0=1.1$ and $R_0=0.9$ scenarios for the current situation.

Note that they are assuming only 25% of cases are reported, so their immunity effect from infections is larger.

Without vaccinations, they

FIGURE 1. Simulated case incidence trajectories* of current SARS-CoV-2 variants and the B.1.1.7 variant,[†] assuming no community vaccination and either initial $R_t = 1.1$ (A) or initial $R_t = 0.9$ (B) for current variants – United States, January–April 2021

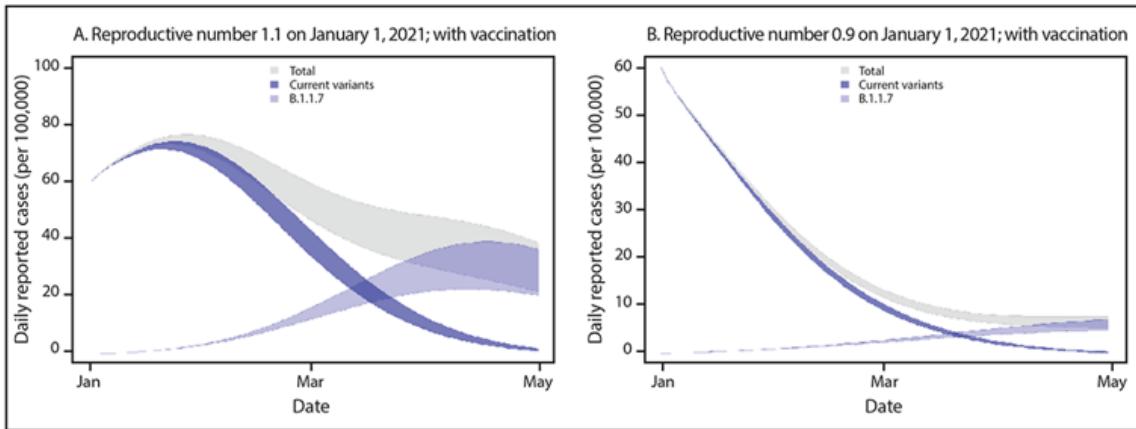
Return



Abbreviation: R_t = time-varying reproductive number.

Then here it is with vaccination of 0.15% of the population per day (e.g. shots per day equal to 0.3% of the population, with two shots per person) or about the same as my model's assumption:

FIGURE 2. Simulated case incidence trajectories* of current SARS-CoV-2 variants and the B.1.1.7 variant,[†] assuming community vaccination[§] and initial $R_t = 1.1$ (A) or initial $R_t = 0.9$ (B) for current variants — United States, January–April 2021



Abbreviation: R_t = time-varying reproductive number.

For those who criticize me for not respecting the control system, the CDC says, what's a control system and how do we talk to the people in charge?

Their recommendation, of course, is the same as it is for anything else. Universal compliance with existing policies, and more vaccinations. Thanks, CDC!

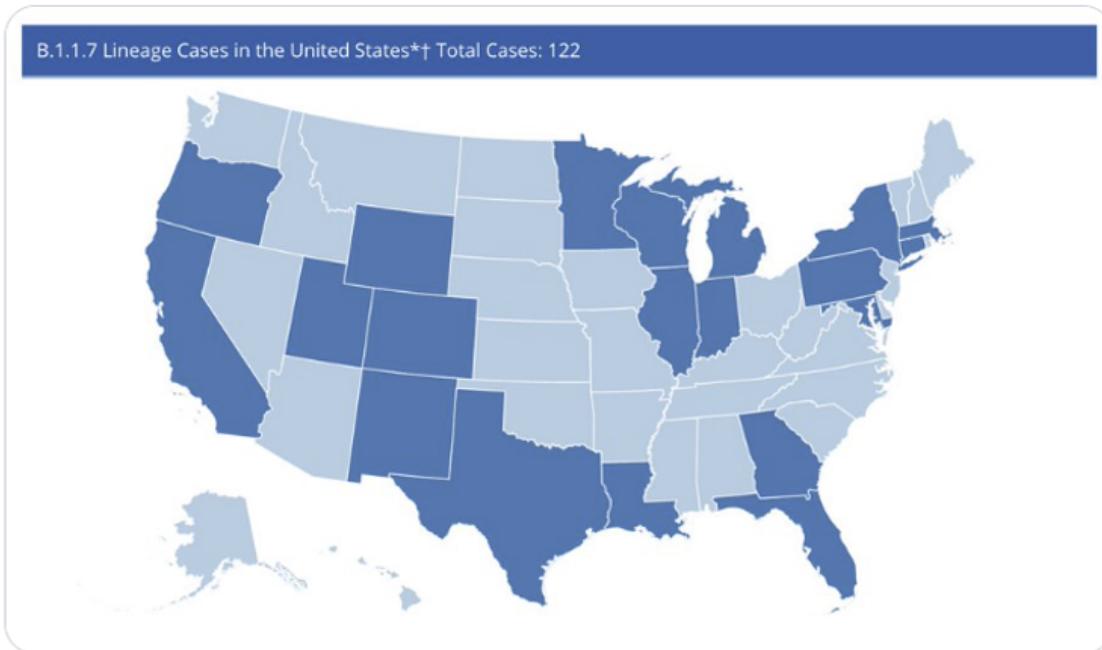
[Here's how things are progressing, right on schedule \(CDC link\)](#):



Caitlin Rivers, PhD ✅
@cmyeaton

...

CDC now reporting 122 cases of the B117 variant across 20 states, up from 88 on Friday's report. California and Florida have identified the most cases, at 40 and 46, respectively. cdc.gov/coronavirus/20...



[The vaccines are confirmed to work on the English strain](#) and I won't bother sharing further similar findings unless they put this finding into doubt.

The Other New Strains

What about all these *other* new strains, which many claim are even worse than the English strain? How likely is it that things are even worse? What do we know about these other strains?

Early in the week, we knew a lot of things that were possibly scary, but nothing definite. I know experimentation is illegal when it has to be done on people, but in this case the experiments we need can be done in a laboratory in approximately zero time for approximately zero dollars with approximately no risk to anyone - you see whether neutralizing antibodies from various sources are effective against various strains - so it's (to put it politely) rather frustrating when no one runs the tests.

The test did get done later in the week, at least for the South African strain, and the results were quite alarming. I'll get to that later in the section.

A question that isn't getting enough attention is: Why suddenly all these strains now?

I see a bunch of people saying things [like this](#), starting a high quality infodump thread:



Trevor Bedford  @trvrb · Jan 14 · ...
After ~10 months of relative quiescence we've started to see some striking evolution of SARS-CoV-2 with a repeated evolutionary pattern in the SARS-CoV-2 variants of concern emerging from the UK, South Africa and Brazil.
1/19

136 2.9K 5.7K 

And few if any of them are acting at all *suspicious* about the whole thing. Whereas my instincts whisper: [This Is Not a Coincidence Because Nothing Is Ever a Coincidence](#).

As I noted last week, the timing seems highly suspicious. There are new mutations every day. Why are there suddenly so many scary new strains?

It can't be the vaccinations, because the timing doesn't work on that. If conditions changed, it happened earlier, and it was something else.

Could it be more use of masks or social distancing somehow applying stronger selective pressure for greater infectiousness? Seems like a stretch.

This was Trevor's guess on the 14th, [from later in the thread](#):



Trevor Bedford  @trvrb · ...
Replying to @trvrb

My (highly speculative!) hypothesis is that the emergence of these variant viruses arises in cases of chronic infection during which the immune system places great pressure on the virus to escape immunity and the virus does so by getting really good at getting into cells. 11/19

12:44 PM · Jan 14, 2021 · Twitter Web App

That's plausible, but doesn't explain why the chronic infections hadn't done this earlier, and the English strain *doesn't* escape immunity in this way (and we don't know about the others) so I notice it doesn't feel like it explains things.

There are more people getting infected now than before, so it could be having more viruses around to mutate, but this seems too sudden for that alone to explain things.

There are also more people who are immune, [thus increasing selective pressure to escape from that \(Stat News](#) which seems high quality):



Eric Feigl-Ding
@DrEricDing

...

Replies to [@DrEricDing](#)

3) Why are mutations arising so fast these days??

“as more people have become protected pressure on the virus has increased. A so-so spreader might no longer be able to, but variants with mutations that help them spread can take off”.

This makes sense as an escalating factor, but once again it does not seem like it could be escalating *fast enough* to explain the sudden phase shift. There are lots of places that were previously very infected, more infected than many of the places where the new strains are now emerging.

Perhaps what changed is largely our perception of what is scary? Which raises the question of whether we're right to be terrified now, or were right before to mostly not be concerned?

Until the English strain, everyone was treating ‘there’s a new variant out there’ as nothing to be concerned about. Suddenly, every day there’s a new headline announcing another strain or travel restriction.

The new Brazilian strain is potentially terrifying. [This scared the English so badly](#) that they banned travel from not only fifteen countries in Latin America, but Portugal as well.

There are warnings that the Brazilian variant could have outright escaped not only the vaccine, but the immunity from previous infections. This would be very different from the English strain:

The NIID said the new variant belongs to the B.1.1.248 lineage of the coronavirus and has 12 mutations, including N501Y and E484K, in its spike protein – the part of the virus that is responsible for attaching and gaining entry to the body's cells.

N501Y is a mutation also found in the UK variant, called VOC-202012/01, and has been linked to increased transmissibility of Sars-CoV-2.

Research has shown that E484K could be “associated with escape from neutralising antibodies” – meaning it may be able to evade parts of the body’s natural defence memory that bestows immunity.

Ravi Gupta, professor of clinical microbiology at the University of Cambridge, said it is this specific mutation that is “the most worrying of all”.

Pfizer and German partner BioNTech said last week that their vaccine worked against the N501Y mutation found in the British and Brazilian variants.

Supporting this is that areas of Brazil that were previously very hard hit, including Manaus with 75% seroprevalence, are being hit again. Now Eric Feigl-Ding says (in a long thread of good data) [there are two Brazil variants but both can escape antibodies](#):

Eric Feigl-Ding  @DrEricDing · Jan 15

6) To be clear on nomenclature (naming), the two Brazil variants are:

👉 B.1.1.28(K417N/E484K/N501Y) ➡️ this is the one renamed P1.

👉 There is another one called B.1.1.28(E484K)

...both have the problematic E484K mutation that can escape antibodies!

10 259 692

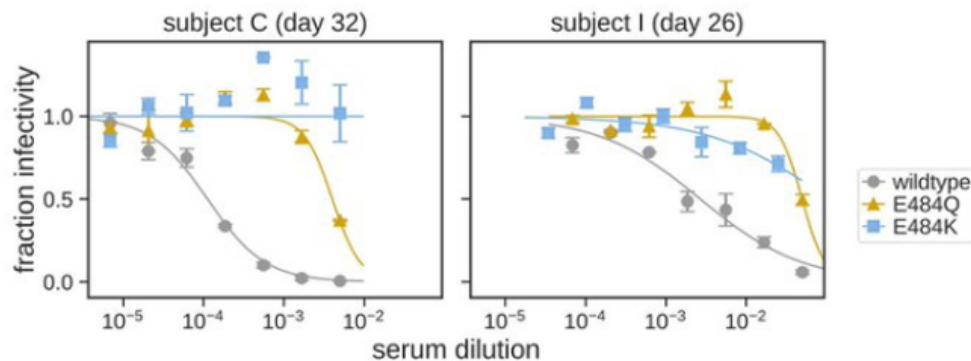
There's no weasel 'may' in that claim, although I suppose 'can' still leaves room for a 'mostly doesn't.'



Eric Feigl-Ding ✅ @DrEricDing · Jan 15

9) How much affect? the E484K shows a 10x reduction of neutralization ("neutralization" = stopping the virus) by various antibodies compared to wildtype (common #SARSCoV2) in some patients —a rather bad thing. It means the virus with E484K is worrying for "immune escape".

Mutations at E484 cause very large (>10-fold) reductions in neutralization by sera from some individuals.



Eric Feigl-Ding ✅ @DrEricDing · Jan 15

13) to be clear on vaccines, the existing vaccine will still work. Though some scientists suspect maybe a few % drop in efficacy. But not sure yet. We are awaiting studies. Researchers are backlogged with also studying the South Africa 🇿🇦 B1135 variant that also has E484K.

14

144

583

↑

Under these circumstances, halting travel as a precautionary principle seems wise. Even if the probability of full escape remains low, the consequences are beyond dire. And as noted above, we should run the tests required to know the real situation.

There's [reports of a new strain in California](#), 452R:



Soumya ✅ @skarlamangla · Jan 17

California health officials are warning that a new coronavirus variant, 452R, is being increasingly found throughout the state. It is unclear whether it is more transmissible than other variants and is different from the B.1.1.7 variant first detected in the UK.

26

496

790



Soumya ✅ @skarlamangla · Jan 17

The variant has shown up increasingly in CA since November and has been identified in several large outbreaks in Santa Clara County. "The fact that this variant was identified in several large outbreaks in our county is a red flag," said Dr. Sara Cody, the county's health officer

2

35

128



Soumya ✅ @skarlamangla · Jan 17

Santa Clara County has found that the 452R variant was present in specimens from large outbreaks where very high numbers of people exposed contracted the virus. Analysis regarding the role of this and other variants in outbreaks is ongoing, officials say.

1

23

95



Soumya ✅ @skarlamangla · Jan 17

In addition to Santa Clara, the 452R variant has been detected in Humboldt, Lake, Los Angeles, Mono, Monterey, Orange, Riverside, San Francisco, San Bernardino, San Diego and San Luis Obispo counties, officials say.

5

39

108



This feels like exactly the kind of thing that months ago would have been met with a giant shrug, a 'mutations be mutating, what you gonna do' and reminders of the importance of random factors, until more data shows up and we can run the necessary tests.

That's especially true given this:

Known as L452R, the newly announced arrival was first identified in Denmark in March. It showed up in California as early as May.

The problem is, California is rather large and has a lot of infections, [and you need to explain this](#) along with California having it especially bad right now:

Dr. Chiu said L452R grew from about 3.8% of the samples he tested in late November 2020 through early December to more than 25.2% in late December through early January 2021.

Given that the strain was identified in Denmark in March, it seems unlikely that it could be that much more infectious, or we would already know. One case in California in May, doubling every week, would have fully infected the country if the control systems hadn't kicked in at some point.

There's also the standard 'we don't know if the vaccine works on this variant' talk because this modifies the spike protein. So again, we need to run the tests, but it seems unlikely that there's a problem. It's not like California is doing much selecting for vaccine resistance.

The really scary one, at this point, is the South African strain, [because it looks a lot like it reduces neutralization capacity \(study preprint\)](#), which likely means it can reinfect people. It's worth quoting a lot of Trevor's thread:

Abstract

SARS-CoV-2 501Y.V2, a novel lineage of the coronavirus causing COVID-19, contains multiple mutations within two immunodominant domains of the spike protein. Here we show that this lineage exhibits complete escape from three classes of therapeutically relevant monoclonal antibodies. Furthermore 501Y.V2 shows substantial or complete escape from neutralizing antibodies in COVID-19 convalescent plasma. These data highlight the prospect of reinfection with antigenically distinct variants and may foreshadow reduced efficacy of current spike-based vaccines.



Trevor Bedford ✅ @trvrb · 9h

It's clear that 501Y.V2 often results in reductions of neutralization titer, quantified as "fold-reduction" where, for example, a 2-fold reduction in titer would mean that you need twice as much sera to neutralize the same amount of virus in the assay. 3/10

1

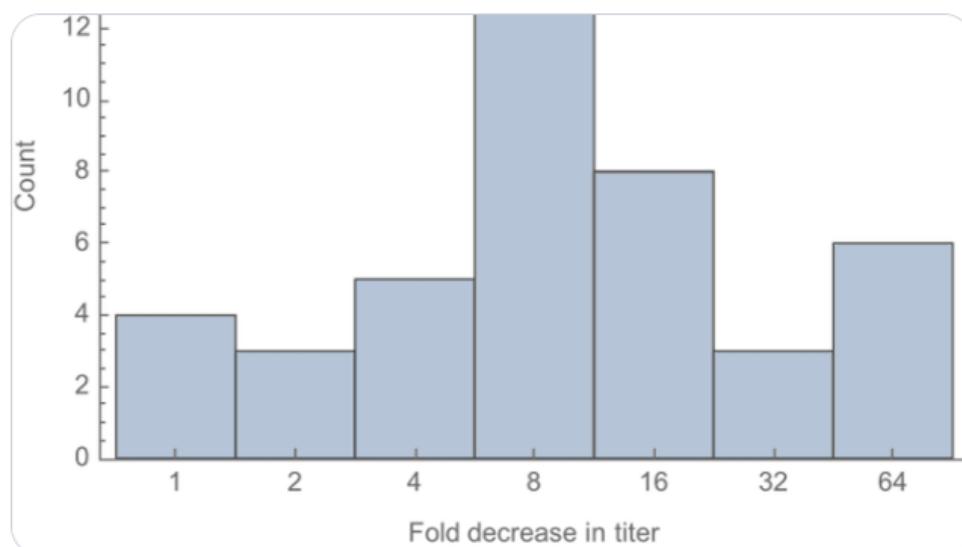
19

126



Trevor Bedford ✅ @trvrb · 9h

Here, I'm plotting distribution of fold-reduction across the 44 individuals tested. You can see there is a median 8-fold reduction in titer when comparing wildtype to 501Y.V2 virus, though some individuals show no reduction and other individuals show a 64-fold reduction. 4/10



Trevor Bedford ✅ @trvrb · 9h

To put an 8-fold drop in context, the @WHO uses an 8-fold threshold when deciding to update the seasonal influenza vaccine (note this is a different virus and neutralization results may not be directly comparable, but it at least gives a ballpark comparison). 5/10

2

33

203



Trevor Bedford ✅ @trvrb · 9h

Also note that the mRNA vaccines in particular are really good vaccines and elicit strong immune responses. A reduction in neutralization from a high starting point will have less of an impact than a reduction from a lower starting point. 6/10



Trevor Bedford ✅ @trvrb · 9h

Replying to @trvrb

We urgently need "immune correlates of protection" determined for COVID-19 vaccination. This would allow extrapolation from reductions in neutralization into expected effects on vaccine efficacy. At the moment, it's guesswork. 7/10

4

48

227



Trevor Bedford ✅ @trvrb · 9h

However, if these results are confirmed by further studies, my guess based on the seasonal influenza comparison is that we need to investigate the manufacturing timeline and regulatory steps required to update the "strain" used in the vaccine. 8/10

3

78

286



Trevor Bedford ✅ @trvrb · 9h

501Y.V2 is still largely restricted to South Africa, but it (or other antigenically drifted variants) may spread more widely in the coming months. I would be planning this potential "strain" update for fall 2021. 9/10

8

73

292



Trevor Bedford ✅ @trvrb · 9h

And all this said, I'll be getting the vaccine as soon as I'm able. We have an amazing vaccine now that works against currently circulating viruses. And if it becomes necessary, this emerging situation can be dealt with through a forthcoming vaccine update. 10/10

From what I've seen, the expectation is that the vaccine won't work as well as before, but should still work, [and could easily be updated if needed](#). This is from one of the authors:



Paul Bieniasz @PaulBieniasz · 14h

...

Replies to [@PaulBieniasz](#)

Importantly, near-identical anti-RBD antibodies are elicited in naturally infected or vaccinated individuals. Together, these data suggest that the emergence of K417N/T, E484K and N501Y mutations represent the beginnings of SARS-CoV-2 antigenic drift, as host immunity accumulates

1

7

22

↑



Paul Bieniasz @PaulBieniasz · 14h

...

The vaccines will in all likelihood still work well against the variants, but this isn't the time (in my view) to be doing population-wide experiments with vaccine regimes which delay the boost (2nd dose) that is all important for maximizing plasma neutralizing potency

5

41

93

↑

As more people are vaccinated and more people have been infected, the selection pressure for strains that escape the vaccine and/or escape prior infection intensifies, and so does the danger of leaving more people only partially protected via vaccination. With the new strains, it is becoming less clear that it would be wise to delay second doses for too long. I'd still be very strongly in favor of not holding second doses *in reserve* but am becoming more receptive to the precautionary principle, which suggests that we might not want to let people wait around for many months. Either we have the vaccine capacity required to re-vaccinate, in which case we can afford to give everyone two doses, or we don't, in which case we won't be able to re-vaccinate quickly if we need to do that.

[South Africa's CDC has issued a rather dire warning:](#)

Can you be re-infected with the new variant if you have already had COVID-19 from one of the older variants?

People who have recovered from SARS-CoV-2 infection are usually protected from being infected a second time (called re-infection). This is because they develop neutralizing antibodies that remain in their blood for at least 5-6 months, maybe longer. These antibodies bind to specific parts of the spike protein that have mutated in the new variant (K417N and E484K). We now know that these mutations have allowed the virus to become resistant to antibody neutralization.

The blood samples from half the people we tested showed that all neutralizing activity was lost. This suggests that they may no longer be protected from re-infection. In the other half, the levels of antibodies were reduced and so the risk of re-infection is not known. It is therefore important that people who have previously had COVID-19 continue to adhere to public health measures. Protecting ourselves through masks, regular washing or sanitising of hands, cleaning of surfaces, and social distancing remain the best defense against all SARS-CoV-2 viruses, including the new lineage.

That's the kind of thing a CDC would be inclined to say in terms of behavioral prescriptions, whether it was appropriate or not. What's important and terrifying is that there was so much loss of antibody effectiveness.

Of course, all of this only emphasizes how important it is now to increase capacity. If we spent a few billion to ramp up mRNA vaccine production capacity now, then by the time the South African strain becomes a problem, we'd have the ability to fix this via re-vaccination, and without effectively taking those doses away from the third world.

Also of course, we should have very strict travel restrictions around South Africa so we can slow the problem down long enough to get to that point.

But What Do We Do Now?

An interesting pivot seen this week is from 'everyone wears a mask' to 'everyone wears an effective mask.'

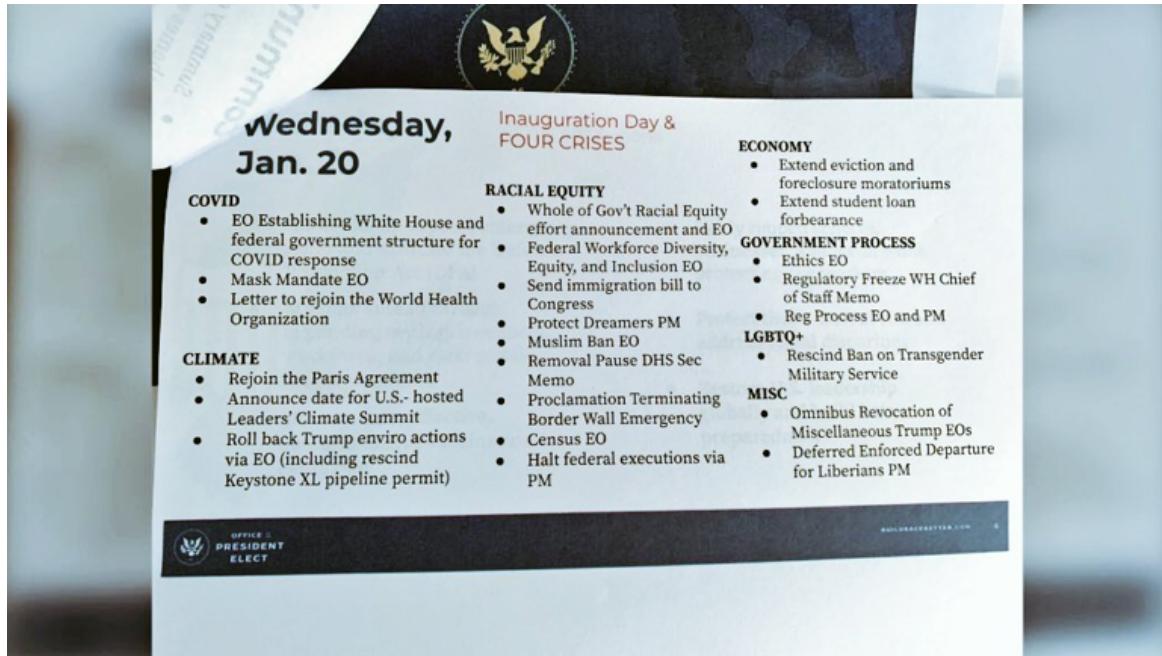
Until this year, the battle to get people to even pick up a piece of cloth was so much trouble that there was little attempt to do more than that. Periodically we'd say that a surgical mask or N95 was better than a cloth mask, but the overwhelming agreement was to emphasize 'mask at all.' If you pressure people to choose better masks, it risks making people throw up their hands and not care at all.

Now a variety of sources have decided that this won't cut it, slash certain forces are out of the picture, and we now need to step up our game and push for better masks.

I'm happy to get behind this attempt by the control system to stay on target. There is an abundance of high-quality masks available for sale on Amazon, and I was quickly able to find one that didn't feel substantially more annoying than a cloth mask. So I encourage everyone who hasn't yet done so to up their mask game.

Note that there is most definitely a [More Dakka](#) version of this where you get the fully effective \$2,000 filtration systems going, and it's overwhelmingly correct to just do that, if anyone actually does use them can you share your experiences?

What about we as in the new administration? [This seems to be a list of day one actions:](#)



So out of 23 things, 3 of them concern Covid, whereas 9 concern Racial Equity, one of which involves a fully written bill being sent to Congress. One of the 3 concerning Covid is to not leave the W.H.O. The second is a mask mandate for the places a president can issue a mask mandate, which is better than not doing it. The third is to establish a structure for future action. Arguably the two economic actions are also Covid-related.

This is better than doing actual nothing for months on end, but for now it's also remarkably similar.

You're Vaccinated, Now What?

[Congratulations?](#)



David Leonhardt @DLeonhardt · 20h

The Moderna and Pfizer vaccines are "essentially 100 percent effective against serious disease," Dr. Paul Offit, the director of the Vaccine Education Center at Children's Hospital of Philadelphia, said. "It's ridiculously encouraging."

[Israel is seeing the impact](#) and no that can't be the lockdown:



Eran Segal @segal_eran · Jan 18

Vaccine effect?

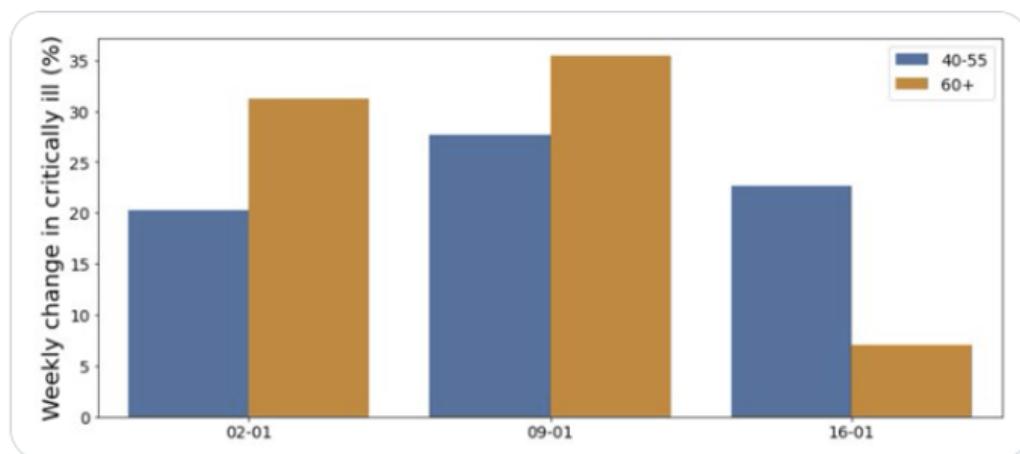
•••

~40% of all 60+ years old are now 2 weeks after 1st dose

After two weeks of ~30% increase in critically ill aged 60+, last week up only 7%

In 40-55 years old the increase was similar all 3 weeks (~22%)

Promising but too early to tell, may just be the lockdown



31

194

457

↑

Yet, as was pointed out last week, many are telling the vaccinated they still have to engage in the same behaviors as everyone else, including avoiding indoor gatherings and maintaining social distancing. These Very Serious People are looking for any way to get people to take any precaution. The Sacrifices to the Gods must continue.

To prove this, they say there is “no evidence” that vaccines prevent transmission. Then if necessary they’ll retreat to “no proof” that they prevent transmission. Then they’ll retreat to the inner motte of “no proof that they *entirely 100%* prevent transmission.”

Of course they prevent transmission. Not 100%, we don’t know the exact percentage, but a lot.

A long term worry on this is that it will make people not want the vaccine, [since what's the point if you still can't live your life](#). Then again, one could reasonably say if one were in the “lying to the American people for their own good” business, that’s a problem for Future America. Right now, there are more people who want the vaccine than there are vaccine doses. Which is true in all places that are open to those over 65. We can then turn around and tell other lies later, to solve those future problems, such people would say, and no we are not worried about our credibility, we are the authorized official sources and anyone who disagrees with us should be censored on social media.

I have [exactly PoliMath's position](#) on masks after vaccination:



PoliMath @politicalmath · 14h

...

My reaction to "we still need to wear masks after vaccination" is "why are you wearing a mask?"

If the answer is "so I don't make unvaccinated people feel like second class citizens" then fine.

But if it is "you could transmit COVID after vaccine immunity" that is nonsense

81

236

1.1K



[Here's Nate Silver](#), who also occasionally does some math and also has some understanding of public messaging:



Nate Silver ✅ @NateSilver538 · 18h

...

We need to remind people to wait for ~14 days after the 2nd dose for full protection to kick in. But it's wild to tell vaccinated people that they shouldn't hang out w/ other vaccinated people, or with people who are very unlikely to become seriously ill.

[telegraph.co.uk/news/2021/01/1...](https://www.telegraph.co.uk/news/2021/01/1...)

Millions of people in these groups have been shielding or living under solitary conditions for much of the past year and are keen to resume socialising or seeing their grandchildren again.

They are being told to continue to follow the rules on the basis that they may still be able to transmit the virus, but it is unclear what public health message will be used to stop them socialising with others who have been vaccinated or with their grandchildren, who are unlikely to become ill.

I intend to wear my mask after vaccination, if I can be vaccinated in time for that to matter, in order to reinforce mask norms. It's easy to wear a mask. There's even some tiny chance it might physically matter, and again, it's easy to do. As opposed to continuing to do costly social distancing, and yeah, no.

[This is the attempt](#) to both be honest and split the needle:



Walid Gellad, MD MPH @walidgellad · 4h

This is tough for people to understand.

...

Vaccine does not fully prevent transmission: true

Vaccine greatly reduces transmission: true

One young vaccinated person can still transmit to old person: true

Many young people vaccinated greatly reduces transmission to old people: true



Eli Klein @TheEliKlein · 11h

Telling young people to take Covid vaccines so that they don't infect old people, then claiming that those vaccines don't prevent transmission, is one of the stupidest public health strategies ever
[twitter.com/theeliklein/st...](https://twitter.com/theeliklein/status/)

This uses the weasel framing of "can" in the third claim, which is technically correct but is chosen to scare people. It is possible that this plane will crash, better drive instead.

Meanwhile, they call saying the vaccine prevents transmission "hiding the truth":



Walid Gellad, MD MPH @walidgellad · 4h

...

Replies to [@walidgellad](#)

I had such a problem with that NYT piece yesterday because it advocates hiding the truth from people in order to properly sell the vaccine.

The argument was a bad one - poorly conceived bad, and not good for public health bad.

The difference here is that this would be "hiding the truth" to say things are safe, which is not fine, as opposed to "hiding the truth" to say things are not safe, which is encouraged.

The one point of evidence potentially pointing the other way comes from Israel, where we're [getting some truly bizarre and troubling data about positive test rates for the recently vaccinated](#). I couldn't locate the original study, so I'm going by the news report, if you can link to the study please do so in the comments.

Here's my summary of key data points:

Tests up to 7 days after vaccination had a 5.4% positive rate. Vaccine shouldn't be protecting them, so consider this a baseline, out of 100,000 tested.

Tests between days 8 and 14 had an 8.3% positive rate, which is super high, higher than baseline. Perhaps those who get vaccinated go out and have parties quickly? This is out of 67,000 tested.

Tests between days 15 and 21 still had a 7.2% positive rate. Still super high. This is out of 20,000 tested.

Tests between days 22 and 28 were 2.6% positive, including some people vaccinated twice, although the second dose hadn't had much time to work. This is out of only 3,200 tested.

We don't know how they determined when and whether to test people. If testing was only done when there was a reason, these numbers don't worry me. If testing was done at random, then this is rather alarming. The declines in numbers of tests run could be because people are being vaccinated in real time, or because they were only testing people when it seemed necessary, and therefore as time went by they tested less people.

The 100,000 number is very suspiciously round, making me think that testing was randomized. Israel's general positive test rate has been rising recently up to about 7%, so that seems like a stretch – testing at random should cause a lower positive rate than that, but perhaps they stopped collecting results after 100k of them? The rise in the second week makes sense if you think those 67,000 tests weren't at random, or the timing could be weird as the situation was changing rapidly. In any case, without better data, hard to tell.

We know from elsewhere that there's a lot of protection by day 10, but the positive test rates here did not decline much until substantially after that.

Without the original source, a lot of key information is lacking, so it's hard to interpret the information we do have.

They did also show that antibody responses were robust:

Meanwhile, serological tests done on employees of Sheba Medical Center at Tel Hashomer a week after they had received the second dose of the Pfizer vaccine showed that of 102 employees tested, 100 had antibody levels between 6 to 20 times higher than they had presented a week earlier.

Ideally I'd withhold analysis until I had a better understanding here, but we don't have that luxury these days. In any case, it's data that needs to be explained.

Yes, We Can Agree Andrew Cuomo Is The Worst

In good New York State news, the state continues to open and operate additional vaccination sites. [The fifth was in SUNY Albany on the 15th](#). They'll need to close or slow down soon due to lack of supply, but that's the right problem to have.

Cuomo somehow managed to mangle the restrictions sufficiently that restaurants suing for the right to provide indoor dining [won in court, so Cuomo is now largely giving up on](#) the zone-based restrictions:



Nick Reisman 
@NickReisman

...

After a court ruling in favor of a portion of Erie County restaurants to have indoor dining, the Cuomo administration says all restaurants in Orange Zones may do so

STATEMENT FROM COUNSEL TO THE GOVERNOR KUMIKI GIBSON ON STATE SUPREME COURT DECISION IN ERIE COUNTY RELATED TO INDOOR DINING

"A court decision yesterday temporarily granted a select few restaurants located within an Orange Zone in Erie County the ability to resume indoor dining under the rules governing Yellow Zones. We are reviewing the decision. While that process is ongoing, to ensure uniformity and fairness, all restaurants operating in Orange Zones can now operate under rules governing Yellow Zones. We disagree with the court's decision and its impact on public health as Federal CDC data clearly demonstrates indoor dining increases COVID-19 spread. From the start of this pandemic, the State has acted based on facts and the advice of public health experts, and we will continue that approach."

This is part of the general sudden pivot from 'we must contain the virus' to 'we must save the economy' that is happening in many places right now. The timing seems, shall we say, suspicious, but also a lot is changing quickly.

Cuomo will pay the legal action forward [by threatening to sue the Biden Administration to get more vaccine doses:](#)



Morgan Mckay
@morganfmckay

...

Replying to [@morganfmckay](#)

NEWS: Cuomo threatens to sue the [@JoeBiden](#) Administration if New York State does not get its "fair share" of COVID relief from Congress

12:12 PM · Jan 19, 2021 · Twitter Web App

The entire NYS budget is split in two, with Cuomo demanding Washington give him money or else, and saying 'look what you'd make me do if I didn't have the money.'

[The entire system is a giant mess](#) and puts seniors in an impossible position, although to be fair I've seen reports that this is mostly true in other places as well:



JC Cassis @JCCassis · Jan 14

...

Tried to help my senior mom get a COVID vaccination appointment today in NYC. The amount of required online rigamarole is nearly impossible for most seniors to handle on their own, and there was no availability listed over the next 3 months. What a mess.



2

9



The Quest to Sell Out of Covid Vaccine

Good news, everyone. We're much closer to successfully using all our doses than previously expected, because that reserve of vaccine that Secretary Azar claimed we'd be releasing? [It never existed](#). There were no held back second doses, the government now claims, instead we had far fewer doses than we were led to believe:



Eric Feigl-Ding ✅

@DrEricDing

...

Replying to @DrEricDing

6) "Operation Warp Speed, the Trump administration's initiative to speed the development of vaccines and therapeutics, stopped stockpiling second doses of the Pfizer-BioNTech vaccine at the end of last year, instead taking second doses directly off the manufacturing line." 🔥

10:18 AM · Jan 15, 2021 · Twitter for iPhone

The whole thread is wild. Despite the administration shipping out its entire reserve, [Pfizer released a statement saying it has second doses on hand for everyone who needs them](#):

"Operation Warp Speed has asked us to start shipping second doses only recently," the spokeswoman said. "As a result, we have on hand all the second doses of the previous shipments to the US."

The U.S. Department of Health and Human Services did not respond to requests for comment.

Pfizer has shipped more than 15 million doses to destinations around the United States, primarily from its Michigan facility, and expects to be able to produce around 2 billion doses worldwide in 2021, the spokeswoman said.

The United States has been struggling to administer the shots that have been distributed, however. Only around 12 million of the more than 31 million doses that have been shipped have been administered, according to data from the U.S. Centers from Disease Control and Prevention.

If the government now claims to not have a reserve, after previously claiming it was going to release that same reserve, but Pfizer claims that instead *it* has the reserve ready to go, [what the hell is actually going on?](#)



Governor Kate Brown 
@OregonGovBrown

...

Replying to @OregonGovBrown

This is a deception on a national scale. Oregon's seniors, teachers, all of us, were depending on the promise of Oregon's share of the federal reserve of vaccines being released to us.

10:55 AM · Jan 15, 2021 · Twitter Web App

And it's not only the feds, the states and Pfizer, potentially *counties and cities* could have *their own reserves*. [New York City does!](#)



Mark D. Levine  @MarkLevineNYC · 21h

...

NYC is keeping a stock of 2nd doses in reserve. So if you've already had your 1st dose the shortage won't affect you.

But some appts for 1st doses will have to be cancelled unless the feds come through with more supply soon. Outrageous that they've put us on this position.

It isn't clear to me whether there was never any reserve of second doses on a mass scale, or if there was a reserve but it's already gone, or there were *two* reserves of second doses sitting around idle and now one of them has been deployed, or even possibly if there were *three or more distinct* reserves of second doses because yes we really are that dumb, and it seems states are talking about distinct distributions of "their" first and second doses and took delivery in two sections, [or something?](#) Then combine that with city or county reserves.



Dan Clark @DanClarkReports · Jan 15

...

New: Overall, 74% of the first doses of vaccines have been used, Cuomo says.

731,285 first doses, and 96,430 second doses.

2

2

4

↑

I don't *think* there were an average of two distinct reserves let alone three or more, but it's so confusing I can't rule anything out.

None of the potential answers cover us in glory.

The quest of then selling out what *has* been distributed goes better in some places than others.

[Here's one theory of what went wrong. \(Link to WSJ\)](#)



Scott Lincicome ✅ @scottlincico... 4m

"Trump administration officials said states have struggled to increase vaccinations partly because some adhered too strictly to federal guidelines on who gets shots."

wsj.com/articles/biden...

0 1 2 4 000

[New York City is now crossing over into the camp that has successfully sold out](#) (while of course holding onto a complete reserve of second doses for everyone who got a first dose):



Emma G. Fitzsimmons ✅ @emmagf · 41m

Mayor Bill de Blasio says New York City is going to run out of vaccine doses by Thursday and will have to cancel many appointments after that. The city isn't getting enough new vaccine shipments.

0 36

1 265

0 464

0

THE RECOUNT **NYC**

DELIVERING VACCINES

NYC HAS:

- Administered more than 220K vaccines **last week alone**
- Last week a New Yorker was vaccinated **every 3 seconds**
- 455,737 doses have been administered in total.

NYC NEEDS:

- **More vaccine doses**
- We have 32K first doses left
- We will hit 0 by Friday
- Without federal support, we'll have to begin closing vaccine sites as of Thursday

0:28 | 7.4K views

This in fact happened, and Thursday and Friday first dose appointments were cancelled en masse.

How are some places doing the rollout much faster than others? Here's a CNN article about that, [suggesting what matters is basic logistics and planning in advance](#), and an emphasis on speed. If you focus on allocation to where vaccine can be used, it gets used. Makes sense to me. I'd also add that such techniques require de-emphasizing prioritization, and not threatening people with huge penalties for giving the wrong person a vaccine shot.

If you don't want to succeed, there are always plausible ways to not succeed. For example,

California has [decided to not administer what seems to be hundreds of thousands of doses in a giant Moderna shipment](#), while they 'investigate possible severe allergic reactions,' all of which occurred at only one location, and while as far as I can see none of the other states that got the rest of that shipment (almost a million doses have been given out) either are halting use or reporting any concerns.

Offit, a top US vaccine expert who is a member of the FDA's vaccine advisory committee, said California's decision to hold back doses of vaccine carries its own risks, especially since allergic reactions can be monitored and treated and, in this case, they occurred at only one location.

"The thing about anaphylaxis is although it is frightening to watch, it's easily identified, it's quickly identified, and it's easily treated with epinephrine," Offit said. "I don't see how taking this off the market is a conservative thing to do or exercising an abundance of caution. I

This seems like the latest variation on 'make vulnerable elderly people sit together indoors in close quarters for observation after getting vaccinated, to monitor for extremely rare reactions, thus exposing them to infection right before they become immune.'

The new administration looks to be [moving ahead as quickly as possible to distribute via pharmacies, which seems ideal, and also to use FEMA and the National Guard for distribution](#), which [doesn't seem like it should be necessary](#) but given how things are going, sure, why not try throwing everything at the wall and seeing what sticks.

The math on selling out via pharmacies on their own seems rather strong:

WOONSOCKET, R.I. — The following is attributed to Karen S. Lynch, currently Executive Vice President, CVS Health and President, Aetna, who will become the company's next President and CEO on February 1:

"We agree with President-elect Biden that pharmacies will play a critical role in the next phase of the COVID-19 vaccine rollout and appreciate his leadership in the pandemic response. CVS Health has more than 90,000 trained health care professionals standing by, with the capacity to administer approximately one million shots per day through our 10,000 CVS Pharmacy locations across the country once the federal program is fully activated. This will build on our success in providing vaccines to one of the nation's most vulnerable populations, having administered more than one million shots at long-term care facilities to date.

"We also agree that despite the growing availability of vaccines, wearing a mask, maintaining safe distances and avoiding crowds remain the best ways to slow the spread. These guidelines are based on sound science and will be critical for months to come."

If CVS can do a million shots a day, *that alone* is the entire goal of Biden's 100 million shots in 100 days.

CVS has less than 10,000 pharmacies in the United States, [out of a total of about 88,000 pharmacies](#). That seems eminently doable, with the limiting factor being supply.

[Here's a thread on all the people who could put shots in arms now, if we had both shots and arms but needed professionals to bridge the gap.](#) 752k practicing physicians, 3.1mm registered nurses, 125k physician's assistants, 265k paramedics and EMTs, 322k pharmacists, 422k pharmacy technicians. Professionals simply are not the limiting factor. Full stop.

How is the experience trying to book an appointment for your elderly parents? [It could be compared to trying to get concert tickets](#). We've gone over problems in New York, and it seems similar issues exist everywhere. Information is in different places, confusing and contradictory. Everything is booked, no one has supply, the people who want it make tons of calls and try lots of methods. That link has information by state, including links to everyone's websites and phone numbers. Hopefully that can help.

Vaccine Allocation By Politics and Power

[Patrick McKenzie sums up](#) what happens when you go around threatening anyone who disrupts the properly ethical priority order with personal ruin, as New York and California have done:



Patrick McKenzie @patio11 · Jan 17

...

Replying to [@patio11](#)

If you want to cause heavily regulated professionals to be far more conservative with their interpretation of your guidance than you believe your guidance actually calls for, that is an excellent way to signal your command intent.



2



10



106



When you emphasize how bad it is to 'jump the line' [you also get stories like this](#):



Brad Shapiro @btshapir · 19h

Story. Name changed to protect the (mostly) innocent.

...

My friend Todd is over 65 and lives in CA. The other day, Governor Newsom goes on national TV and says those over 65 are now eligible for the vaccine. Todd goes online, and makes an appointment at a mass vaccination site.

79

796

1.4K



Brad Shapiro @btshapir · 19h

Upon arrival, Todd checks in and is asked for his health care worker ID. He's not a health care worker. He says "The Governor got on TV and said >65 was eligible. And the website allowed me to make an appointment"

...

Check-in worker says "NO" and puts a large X on his windshield

2

32

398



Brad Shapiro @btshapir · 19h

Funny enough, there's no place to turn around the car, so Todd has to go through the vaccination line, but is not allowed a vaccine. As he goes through the line, he keeps arguing. Shows the article on his phone. Re-iterates that he was allowed to make an appointment.

...

1

24

413



Brad Shapiro @btshapir · 19h

Director of the vaccination site eventually agrees with him. They have the dose ready. The governor told them >65 was eligible! They gave him an appointment. None of this is his fault.

...

2

29

483



Brad Shapiro @btshapir · 19h

Check-in worker loses her mind. Says if he gets a vaccine, she's calling the police. She won't be silent while someone tries to "jump the line"

...

Knowing if the police come, vaccines will be disrupted for the rest of the day, the director relents. Todd leaves unvaccinated.

36

121

629



Brad Shapiro @btshapir · 19h

Unclear if that dose was administered at all or trashed.

...

Moral of the story: SIGH. We are self owning in so many ways...

32

55

1K



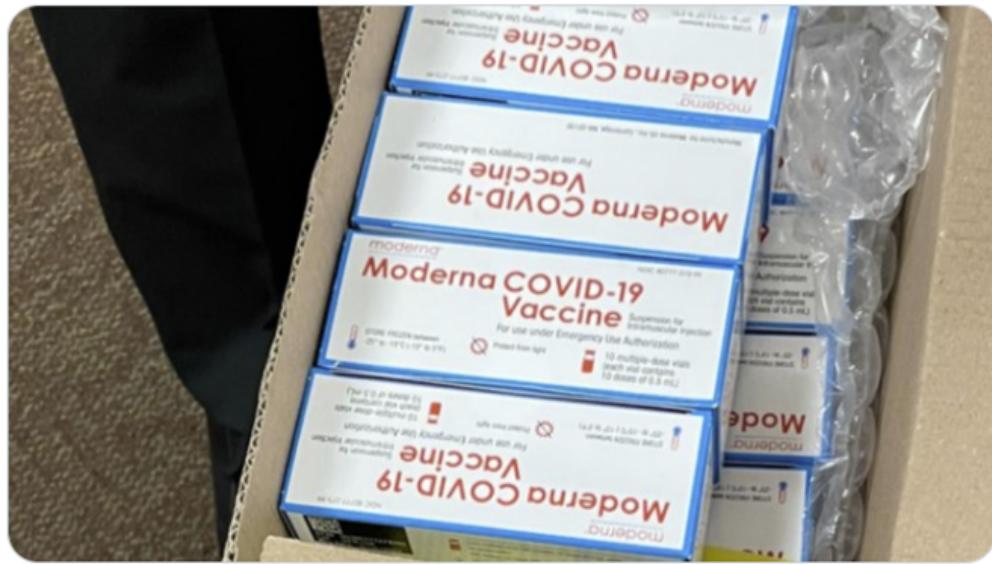
Or like this:



doug hirsch @dougjoe · Jan 14

...

Here's a picture from 3 days ago when 1000 doses of Moderna vaccine were delivered to a major pharmacy near my house (I was there; the nice pharmacist opened for me). 3+ days later, all 1000 doses are still in their fridge because [@MayorOfLA](#) chooses to hold them hostage. Awful.



31

104

347

↑

[Or this, from of all places TMZ:](#)

The clinic -- the Men's Health Foundation in Inglewood -- contacted people who were not on the priority list but desperately wanted the vaccine, and these folks **got the remaining doses**. Incredibly, that runs afoul of the County Health Dept's guidelines, which say ONLY people on the priority list should be vaccinated, EVEN IF THE VACCINES WOULD OTHERWISE END UP IN THE TRASH.

We worked this story for 4 days, and went back and forth with the Health Department, which never gave us an answer to our question -- ISN'T IT BETTER TO VACCINATE NON-PRIORITY RESIDENTS THAN THROW THE PRECIOUS VACCINES IN THE GARBAGE? We never got an official answer to that question, but sources who were on a call with the Health Dept. tell us they decided they would not take punitive action against clinics that replicated what the Men's Health Foundation did. But, that's a terrible solution.

Here's the problem ... the policy still says non-priority people cannot get the vaccine now, so clinics are essentially being told it's still wrong to vaccinate non-priority residents, even if the vaccines would otherwise go to waste. What's even worse, the County Health Dept. has NEVER EVEN STATED PUBLICLY THEY WOULD NOT TAKE ACTION AGAINST THESE CLINICS, so the fear is that vaccines are getting thrown in the garbage.

Supervisor Hahn is outraged, and wants the Health Dept. to get with the program and clarify this insane regulation ... IMMEDIATELY.

That's also how you [get outcomes like this](#) via the LA Times:

By JACK DOLAN | STAFF WRITER

JAN. 15, 2021 | 9 AM UPDATED 6:20 PM

As public health officials scramble to clear a backlog of unused COVID-19 vaccine by opening the process to anyone 65 or older, new data show they failed to quickly deliver shots to the vast majority of California's most vulnerable residents, who were supposed to be the priority.

As of Sunday, only about 5% of long-term care facility residents in the statewide vaccination program — including people in skilled nursing homes and assisted living centers — had been vaccinated, according to California Department of Public Health data obtained by The Times.

To summarize, emphasis on prioritization has led to large amounts of vaccine sitting around unused because people are waiting for 'authorization' to use it, to people blocking valid appointments at vaccination sites, and also to only 5% of the actual most vulnerable people, the group that is 1% of the population and over a third of the deaths, getting their shots a month into the campaign.

Meanwhile, in Florida, [they're requiring government ID and proof of residency](#) to get vaccinated, to avoid accidentally giving doses to undocumented immigrants, or to people who noticed Florida was doing a decent distribution job and came down to get vaccinated. [Nebraska](#) is attempting to exclude immigrants as well. This will doubtless trip up numerous people, especially poor people, who lack or forget or are afraid of using proper documentation:

 [seminolecountyfl.gov](#)

- Proof of employment is required for non-hospital healthcare workers, including but not limited to a paystub from healthcare employer, employee ID badge or other form of verification.
- All patients should bring a government-issued ID and proof of Florida residency.

[Meanwhile,in Wales](#) (Twitter HT), they are delaying vaccinations so the curve of vaccination times is smoother *for each individual vaccine type*, [no really they are literally doing that, what more do I have to say:](#)

4d ago About 300,000 doses of the vaccine have been delivered to Wales, the first minister, **Mark Drakeford**, has confirmed. Drakeford said the figure “in very broad terms” was made up of 50,000 doses of the Oxford/AstraZeneca vaccine and 250,000 of the Pfizer/BioNTech vaccine. Drakeford said:

¶ *We will be using all the Oxford vaccine that we get as we get it, the Pfizer vaccine has to last us until into the first week of February.*

So we have to provide it on a week-by-week basis. What you can't do is to try and stand up a system which uses all the vaccine you've got in week one and then have nothing to offer for the next four weeks.

We won't get another delivery of the Pfizer vaccine until the very end of January or maybe the beginning of February, so that 250,000 doses has got to last us six weeks. That's why you haven't seen it all used in week one, because we've got to space it out over the weeks that it's got to cover.

We are expecting a significant upswing in the Oxford vaccine coming to Wales next week and we will use all of that because it is a much easier vaccine to use, it can be used in GP practices and so on.

We will continue to use the Pfizer vaccine in a way that will mean that we will use it all before we get the next delivery.

Prioritization by Lack of Virtue

What do politics and power reward and punish, in the end?

At some point, the system stops pretending it is rewarding virtue and punishing lack of virtue.

Then, at some point, the system stops pretending it is not punishing virtue, and starts punishing virtue and rewarding lack of virtue.

The official CDC recommended guidelines suggest prioritizing those with various ‘chronic conditions’ and include **giving priority to smokers**.

This is being followed in at least Alabama, Nevada, New Jersey, Mississippi [and Washington D.C.](#)

In other words: If you, on a regular basis, pay for and then consume poison, then that puts you at higher risk, so we will prioritize that you get life-changing and life-saving medicine before others who do not on a regular basis consume poison.

[Every year, the poison in question kills more people, and costs more years of life, than Covid-19 was responsible for in 2020](#). It is highly plausible that, should this guideline be followed, smoking would gain status, people would have a new excuse for their smoking or not quitting, and this act alone could result in sufficiently more smoking to be a bigger health cost than the entire Covid-19 pandemic.

I expect that, for the rest of time, anyone who wants to justify smoking, or not having a healthy weight, or any other issue they don’t want to deal with, will often pull out “hey, at least it’ll get me priority health care!”

In addition, did you know you can *just lie about this*? It's not as if they check in any way whatsoever. So...

And the extensive list of chronic conditions — which includes habitual smoking and diseases like cancer, heart failure, diabetes and Down syndrome — raises the question of whether residents eager to be vaccinated will be honest about whether they smoke, for example, or have a body mass index over 25.

In addition, there's the question of whether you are sufficiently shameless to use the fact that you smoke to step in line ahead of an elderly person who is at actual risk in a way that has nothing to do with their life choices. So in that sense they are prioritizing the *selfish and shameless*.

Most of all, **they are prioritizing liars.**

You don't even have to say what you're lying about! In DC you can simply say you have *one* of the conditions, never mind which one, and get vaccinated at age 17:

Shah said that to protect those people's privacy, they will simply be asked a yes-or-no question: Do you have one of the chronic conditions on this list? Those who say yes, live in the District and are at least 16 years old will get a vaccine, no further proof required.

Here's the actual prioritization scheme they're about to have in Washington, DC, then:

Would you like a vaccine? If so, check this box.

At that point, what are the ethics of checking that box? Should this kind of rule be respected?

Do you think people will respect such a rule? What will that do to their respect for such rules in general?

Once you add obesity as a chronic condition, everybody knows that the dice are loaded, the system's sole purpose is to punish the honest and honorable, [and we'd wish there was no prioritization at all](#):



Nate Silver 🌐

@NateSilver538

So, >50% of DC residents would qualify based on the weight criteria alone. Add in criteria based on other conditions like smoking (how is that defined or verified; do ex-smokers count?) and there's basically no prioritization at all.



Nate Silver ✅

@NateSilver538

8m

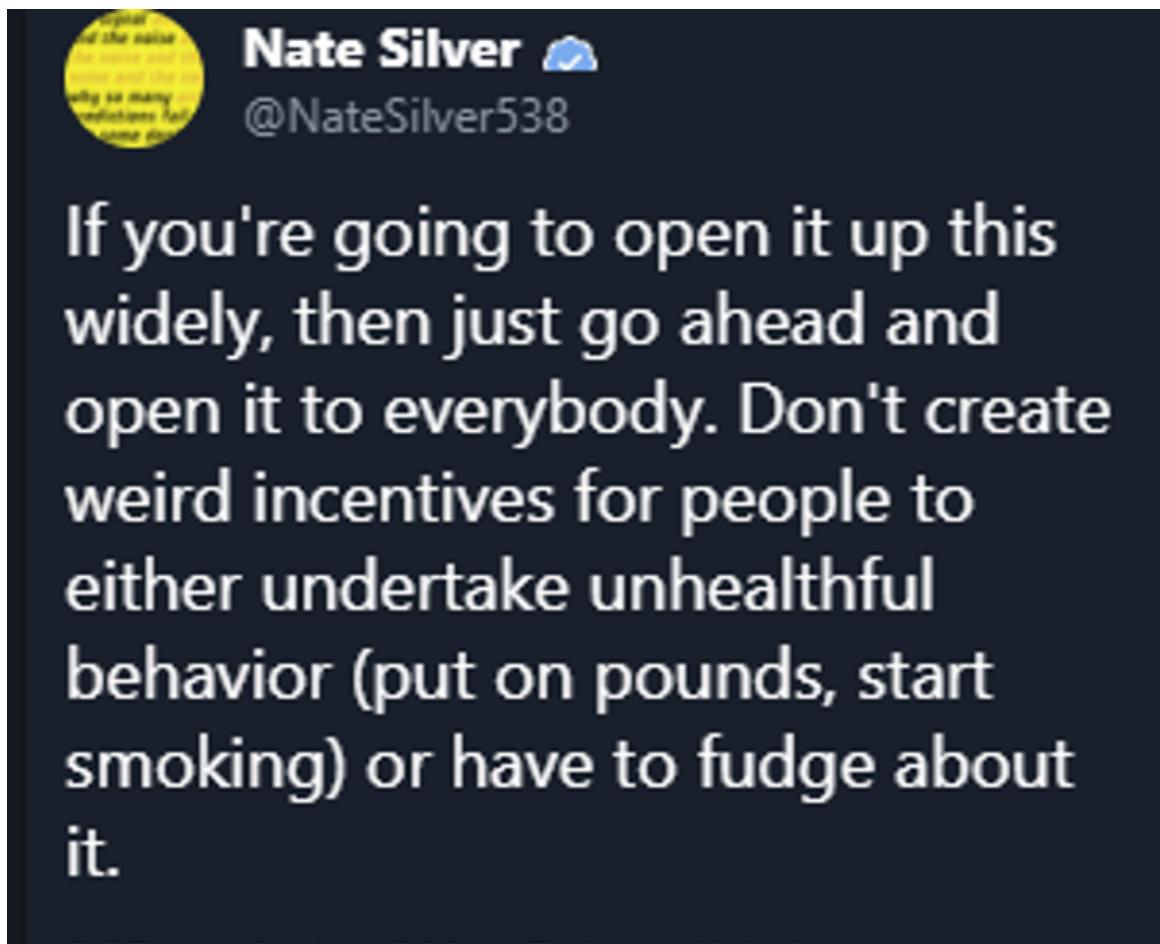
The average US male has a body mass index (BMI) of 29 and eligibility in DC starts at 25! There is no scientific basis for this at all.

Note that starting at 25 not only includes the majority of people, thus making sure that the elderly can't get vaccinated any time soon, it also doesn't make physical sense at all even if you buy the supposed premise of people being at higher risk:

"We don't see anything in our data — and this is a very large data set that I'm talking about — I don't see an association below 30," Tartof said. "And for death, we don't see a statistically significant association below 40."

This is the ultimate result of allocation by politics and power. Those who learn to work the system, to invest their resources in such games, to be comfortable using special rules and appropriating from others, get the scarce resources.

Those who play by ‘the rules’ and do ‘what is fair’ are left out in the cold. If you did what the Responsible Authority Figures said to do, you’re now behind *most* other people and will have to spend additional months of your life hiding at home while those who smoke or are overweight or just decided to lie about it frolic around town like it is nothing.



If you're going to open it up this widely, then just go ahead and open it to everybody. Don't create weird incentives for people to either undertake unhealthful behavior (put on pounds, start smoking) or have to fudge about it.

So let's be clear. If you don't want to have priority, you can just... not have priority. Allocate by willingness to make phone calls or stand in lines or reload web pages. Find out who is willing to destroy more real resources.

You could also allocate by willingness to pay more real resources rather than destroy more real resources, but whenever I talk about the only known good way to allocate scarce resources people get into demon threads complaining about how that is Just Awful, so once again I'm not going to suggest that.

I strongly suspect (and hope) that there will be a lot of vaccination sites that are told that this is the priority list, but if you call them and say you're eligible because you are a smoker or have a BMI of 27, suddenly there won't be any appointments available and you'll be put on a waiting list and never called back.

Luckily, it seems the majority of states realize what these CDC guidelines imply, and are mostly disregarding them.

I considered not writing this section to avoid highlighting the issue, because highlighting the issue risks accelerating the negative consequences involved, and it didn't seem like anyone was noticing this. Then Nate mentioned it, and I wrote the section.

On reflection, I shouldn't have hesitated.

This is not a small effect. This could easily, where adopted, delay an honest and honorable person's vaccine access by *several months*.

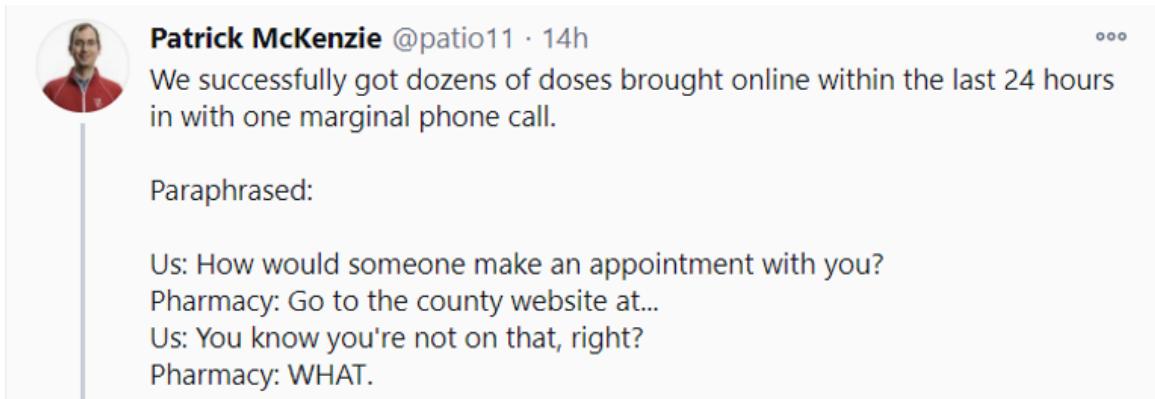
Not mentioning destructive behaviors because the noticing of such behaviors creates destruction is a horrible, horrible incentive that leads to harmful crimes being continuously covered up and rewarded. If one sees something, one must say something. If the law is unjust, one should not keep quiet about that out of fear that more others will then notice the law is unjust and might take advantage of it.

The silver lining of such policies is they absolutely create enough eligible arms in which to put all the shots. This is systematic injustice for injustice's sake, but at least it does get shots into arms.

Useful Resources

[VaccinateCA](#) is a project that calls hospitals and pharmacies in California daily, and checks which are currently administering vaccines. I heard about it from several sources, originating with [Patrick McKenzie](#).

[Here's how necessary that project is](#) from another angle:



Patrick McKenzie @patio11 · 14h
We successfully got dozens of doses brought online within the last 24 hours in with one marginal phone call.

Paraphrased:

Us: How would someone make an appointment with you?
Pharmacy: Go to the county website at...
Us: You know you're not on that, right?
Pharmacy: WHAT.



Patrick McKenzie @patio11 · 14h
Us: Yep. We're quite sure.

Pharmacist called county to complain. Someone pushed some buttons. They started getting patients scheduled.

[Here's a similar project in Massachusetts](#).

[Here's a similar project in Texas](#).

[Here's the start of something similar for New York City](#).

If you'd like to direct few-questions-asked funding to a similar operation to VaccinateCA in another state, I know someone looking to do that, and I'm happy to direct you to that person

if you contact me via email, Twitter DM or LessWrong PM.

Or, if you know about existing similar places for other states, share in the comments, and I'll include in future updates.

[This CNN article linked above](#) has some useful phone numbers and websites to try.

Note that different locations have decided to use different standards for who they will vaccinate. Some are allowing anyone 65+, others are only allowing 75+. Of interest to many readers, Alameda County's three cities are all (as of writing this section on Tuesday) only doing 75+. Other areas are place after place with no supply.

To get an appointment in New York State from the state's facilities (as opposed to other places, or using NYC's system) you are officially [asked to start here](#). It looks like some upstate places have appointments available. It's up to you to decide how far you're willing to travel. My answer would be quite far.

If you're looking in NYC you could try starting [here](#) or look [here](#) but I expect best answers to change. There is also now the NYC vaccine list above.

How Bad is it Out There Right Now?

It's so bad that the states are starting to turn to actual logistics experts. No, not Amazon. [Starbucks!](#)



Evan Rosenfeld @Evan_Rosenfeld

...

Starbucks assigns 11 employees with expertise in labor & deployment, operations and research & development to work full-time on Covid-19 vaccine distribution, with a goal of helping WA distribute 45k doses a day, company says. nbcnews.com/news/us-news/w... via [@tylerkingkade](#) [@stephgosk](#)

To be clear, I do not say this to mock. This is a *very very good* development. Let the experts do what they do best.

Meanwhile, in Los Angeles, we have moved on from leaving people to die without transporting them to hospitals, [to then having to temporarily suspend air-quality regulations in order to cremate them when they die](#):



So many people have died in Los Angeles County that officials have temporarily suspended air-quality regulations that limit the number of cremations. Health officials and the L.A. County coroner requested the change because the current death rate is “more than double that of pre-pandemic years, leading to hospitals, funeral homes and crematoriums exceeding capacity, without the ability to process the backlog,” the South Coast Air Quality Management District said Sunday.

You Should Know This Already

There are extra vaccine doses in the vials, [but you can only fully extract them with a low dead space syringe](#) and we are not reliably using such syringes, wasting a substantial percentage of all potential vaccine doses. This could plausibly be a much bigger loss than throwing unused doses away at day's end.

Once again, [do not throw doses in the garbage](#). In an important sense this is the most important thing to care about, for most people, on the margin. Of course, the hospital gets attacked for breaking with 'priority' and also roasted alive for wasting doses, meaning that they keep everything quiet and destroy all records of what happened. The key is to [choose the side to be on in the dark](#).

[Matt Yglesias makes the case for vaccine challenge trials](#). He makes a strong and clear case, which is admittedly easier when something is overwhelmingly obviously correct. In any case, additional voices on this are always welcome.

Mentioned from another source above, but reminder: [Israeli study on Pfizer vaccine sees 100 of 102 develop significant antibodies](#), editor says participants likely won't spread virus further.

Periodic reminder: Pay for something and you get more of it, [well, maybe](#), but yeah, you do: [Utilization of the United Kingdom's Covid subsidies by region correlates to cases of Covid \(pdf\)](#).



Kirsten Bibbins-Domingo @KBibbinsDomingo · Jan 14

...

UK subsidized 50% food & drinks in restaurants from August 3-31, 2020

areas with higher uptake saw higher increases in new COVID19 infection clusters

program may have accounted for 8-17 percent of all new local infection clusters during that time period

(Not the biggest concern these days, but can we stop using the word 'may' and then providing a range? Studies show I may have between 2 and 17 burgers for lunch next Tuesday.)

[The program details were even worse than you know](#):

Participating establishments may include:

- restaurants, cafés, bars or pubs
- work and school canteens
- food halls

All diners in a group of any size can use the discount.

Can I use Eat Out to Help Out on takeaways?

No, the discount is only available if you actually dine out at a restaurant, so even if you go there to pick the food up you will have to pay full price.

This is, shall we say, most definitely not how any of this works ([via MR](#)) and explaining why would only insult your intelligence:

According to [the Guardian](#) the First Minister of Wales explained their policy of doling out the Pfizer vaccine evenly over the next six weeks:

the Pfizer vaccine has to last us until into the first week of February.

...We won't get another delivery of the Pfizer vaccine until the very end of January or maybe the beginning of February, so that 250,000 doses has got to last us six weeks.

That's why you haven't seen it all used in week one, because we've got to space it out over the weeks that it's got to cover.

[Your periodic entirely correct rant](#) that we should consider allocating scarce resources by price rather than by politics and power, and letting people do the things they want to do to stop the pandemic because that would actually work, from The Grumpy Economist John Cochrane.

Europe [has been informed that it will get fewer vaccine doses from Pfizer](#) than expected for a while, so that they can upgrade their factory and produce more doses in the future. That's an excellent reason to temporarily produce fewer doses, given you are in a world where there's no better way to increase production capacity than you already implemented months ago for trivial amounts of money. It seems that Europe negotiated a lower price point in exchange for going to the back of the line, [so now they're going to the back of the line](#).



Bruno Maçães

...

Replying to [@MacaesBruno](#)

Someone will have to explain to me why EU negotiated a reduced price for the jab, moving to the end of the line for delivery. Surely half a year of 5% fall in GDP seems more expensive than 20 euro price per unit (adding some 20 million euros to total)

The Very Serious People will always criticize anyone who does not defer to the Very Serious People, even when they are obviously wrong, as we are reminded this week by two old Guardian links from Marginal Revolution.

First, the Very Serious People declared that since Brexit was in defiance of them and their dictates, that bad things must follow, so they declared [pulling out of the European Medicines Agency would slow down the UK's vaccine rollout](#). Because, somehow, the union that spends

all day telling people what they cannot do and how exactly they must do everything that is still done, and being unable to make any decisions, would obviously get the vaccine first. By implication it is unforgivable therefore to leave since that will deny your people the vaccine. Instead you should participate in the EU's plan to... negotiate a lower price in exchange for going to the back of the line.

Then in July, when the United Kingdom decided not to take part in the European Union's plan of paying less for vaccines in exchange for going to the back of the line while putting lots of regulatory hurdles in place, [that was called 'unforgivable'](#) because it would 'set the UK up as a competitor' and because the UK might decide to secure more doses rather than less doses:

“We urge the UK government to follow the EU’s lead and only secure vaccine doses for those who need it most (healthcare workers, over-65s and other vulnerable groups).”

What is morally unforgivable to the Very Serious People? Not going along with their schemes (the title of the Guardian article *actually calls it a “scheme” by name*) and deciding instead to attempt to create better outcomes rather than worse outcomes, [instead of their explicit calls to aim for worse outcomes rather than better outcomes](#). No wonder, for example, that *every single super-rich person was terrified to be seen actually helping*. I know it’s easy to not see such statements but perhaps consider that they could be literally true?

Your periodic reminder that [people are crazy and the world is mad](#) and none of the rules make about children make any sense:



PoliMath @politicalmath · 23h

The hospital: You are late for your well-child check. Don't worry about us, we're super clean and safe.

...

Also the hospital: You may neither leave your other children in the car nor bring them into the hospital. If you are a single parent, they must stay home by themselves.

15

48

262



PoliMath @politicalmath · 23h

This is like when I went to the Cheesecake Factory the other day and I couldn't get a cheesecake b/c they wouldn't let me bring all my children inside with me b/c they have a 3 person limit at the counter b/c everyone has lost their fucking minds.

10

12

212



In Other News

With the new administration, [the CDC will now review all of its guidance on everything](#):

Statement from Rochelle P. Walensky, MD, MPH, Director, Centers for Disease Control and Prevention

It is truly a privilege to join the world's premier public health agency. For 75 years, CDC has carried out a mission to protect America's safety, health, and security at home and abroad.

I am proud to join this agency, and I recognize the seriousness of the moment. The toll that the COVID-19 pandemic has had on America is truly heartbreaking — for the loss of our loved ones and our beloved ways of life. At Massachusetts General Hospital, I saw firsthand the many difficulties this pandemic brings to our frontline workers and first responders, hospitals and public health systems, communities, and loved ones.

Better, healthier days lie ahead. But to get there, COVID-19 testing, surveillance, and vaccination must accelerate rapidly. We must also confront the longstanding public health challenges of social and racial injustice and inequity that have demanded action for far too long. And we must make up for potentially lost ground in areas like suicide, substance use disorder and overdose, chronic diseases, and global health initiatives.

America and the world are counting on CDC's science and leadership. Just as it has since the beginning of the pandemic, CDC will continue to focus on what is known — and what more can be learned — about the virus to guide America. As part of that promise, CDC's Principal Deputy Director Anne Schuchat will begin leading a comprehensive review of all existing guidance related to COVID-19. Wherever needed, this guidance will be updated so that people can make decisions and take action based upon the best available evidence.

There are two recent studies out about immunity to Covid coming from past infection. My analysis of those studies [is available in its own post](#) rather than as part of this post, so it is easy to link to.

Israel had already secured vaccines in part by promising to provide good data in return. [Now it seems they've struck another data-for-vaccine deal](#). For everyone who says there are no more doses to be had, it's gotta be odd that more doses keep being had.

Meanwhile the W.H.O. thinks that countries and companies should stop making deals entirely, so they can direct all the vaccine shots wherever they think is best. [Anyone who disagrees with this, they declare, is deeply unethical](#). How dare people with money pay for things to be created, and then take delivery of those things! The horror. Yes, that logic has other implications. Remember to be consistent.

[Could it be happening? Please?](#)



NBC News Business @NBCNewsBusiness · 18h

...

NEW: Amazon has extended an offer to President Joe Biden to assist with the national distribution of Covid-19 vaccines, a move that could expedite the federal effort to combat the pandemic.

More from [@DylanByers](#).

[If the attack on the plan is 'how dare this not have happened sooner'](#) then that's perfect, let's do it now and yell at each other about how awful and political the timing was, come on, everyone, we can do this:



Brendan Nyhan ✅ @BrendanNy... now
Democratic president o'clock

Bobby Lewis @revrlewis

Fox & Friends is table-slammimg
Outraged that 4,000 Americans are
dying of COVID-19 every day but
Amazon didn't "step up" earlier to
help distribute vaccines. Quite the
complaint coming from former de
facto presidential advisers who
cheered Trump through pandemic
mismanagement.



1 2 5 ...

[You know who isn't wasting doses or time? The Department of Corrections!](#)



[**mitrebox**](#)

@mitrebox

Replying to @politicalmath

If you're looking for interesting points, the bloomberg dashboard has federal bureau of prisons at 105% of allocated vaccinations.

[12:33pm · 20 Jan 2021](#) · Twitter for Android



PoliMath

@politicalmath

lol, so does the CDC

Apparently the Bureau of Prisons has received 16,750 doses, but has administered 17,628 doses

(This is not to make fun of the CDC, it's to say that data gathering and reporting is often riddled with these discrepancies)

PoliMath assumes this is a data error, but my presumption is that this is no error. There are extra doses in each vial, so it's perfectly reasonable to get a few percent more shots in than there were doses allocated to you. That should be the standard by which one is judged.

Via MR, [this long detailed post](#) goes over the mRNA vaccine supply chain. Most of the steps, while non-trivial in an important sense, seem straightforward to scale as far as we'd need to scale them, including making the mRNA itself. It's known tech.

The limiting factor seems, according to this article, to be Lipid Nanoparticle (LNP) production. I don't know anything about that process beyond what is seen here, so I don't know how much that could be scaled or at what cost. There weren't any indications we were punching anywhere near the limits of what could be done.

[Studies suggest](#) saliva tests are as accurate as swab tests while being cheaper and easier to use ([synthetic review 1](#), [review 2](#)).

[It is almost certainly safe to be vaccinated while breastfeeding.](#)

[Marginal Revolution](#) links to a Reason interview of Alex Tabarrok on First Doses First.

[Claim that NSAIDs dampen immune response to Covid in mice.](#)

[This seems like a good method](#) of explaining how to stay safe:



Michael Story 
@MWStory

...

The best way I've found to communicate risks with older relatives (especially if they've been listening to that "covid secure" meme) is get them to imagine everyone has a cig on the go and try not to breathe in the smoke

For Those Who Actively Want to Give Me Money

An increasing number of people have asked about giving me money, to show appreciation for these posts and the work required to create them. You really *really* don't have to do this! I don't need the money! I don't do this for money, I have a day job and I don't need to worry about money any time soon.

But if you choose to contribute, I believe this would be motivating rather than demotivating, and you have my thanks.

If you wish to do this on a small scale, [I have set up a Patreon for the blog](#) as a means to do that. There won't be any rewards beyond things like 'I am happy and motivated, and I respond more to your comments.' There won't be any locked posts.

If you want to give enough that the fees involved in Patreon are worth avoiding, you can PM me on LessWrong or DM me on Twitter, or email me, and I'll provide details for PayPal or the relevant crypto address.

Once again, *please do not consider yourself under any obligation whatsoever to do this.* It brings me joy that others are finding these updates useful, and ideally spreading the word about them and putting the information and ideas into practice, and that we are building better models of the world together. That's what is important.

Until next week.

Pseudorandomness contest: prizes, results, and analysis

This is a linkpost for <https://ericneyman.wordpress.com/2021/01/15/pseudorandomness-contest-results-and-analysis/>

(Previously in this series: [Round 1](#), [Round 2](#))

In December I ran a pseudorandomness contest. Here's how it worked:

- In [Round 1](#), participants were invited to submit 150-bit strings of their own devising. They had 10 minutes to write down their string while using nothing but their own minds. I received 62 submissions.
- I then used a computer to generate 62 random 150-bit strings, and put all 124 strings in a random order. In [Round 2](#), participants had to figure out which strings were human-generated (I'm going to call these strings **fake** from now on) and which were "truly" random (I'm going to call these **real**). In particular, I asked for *probabilities* that each string was real, so participants could express their confidence rather than guessing "real" or "fake" for each string. I received 27 submissions for Round 2.

This post is long because there are lots of fascinating things to talk about. So, feel free to skip around to whichever sections you find most interesting; I've done my best to give descriptive labels. But first:

Prizes

Round 1

Thank you to the 62 of you who submitted strings in Round 1! Your strings were scored by the *average probability of being real* assigned by Round 2 participants, weighted by their Round 2 score. (Entries with negative Round 2 scores received no weight). The top three scores in Round 1 were:

1. **Jenny Kaufmann**, with a score of **69.4%**. That is, even though Jenny's string was fake, Round 2 participants on average gave her string a 69.4% chance of being real. For winning Round 1, Jenny was given the opportunity to allocate **\$50** to charity, which she chose to give to the **GiveWell Maximum Impact Fund**.
2. **Reed Jacobs**, with a score of **68.8%**. Reed allocated **\$25** to **Canada/USA Mathcamp**.
3. **Eric Fletcher**, with a score of **68.6%**. Eric allocated **\$25** to the **Poor People's Campaign**.

Congratulations to Jenny, Reed, and Eric!

Round 2

A big thanks to the 27 of you (well, 28 — 26 plus a team of two) who submitted Round 2 entries. I estimate that the average participant put in a few hours of work, and that some put in more than 10. Entries were graded using a quadratic scoring rule (see [here](#) for details).

When describing Round 2, I did a back-of-the-envelope estimate that a score of 15 on this round would be good. I was really impressed by the top two scores:

1. **Scy Yoon** and **William Ehlhardt**, who were the only team, received a score of **28.5**, honestly higher than I thought possible. They allocated **\$150** to the **GiveWell Maximum Impact Fund**.
2. **Ben Edelman** received a score of **25.8**. He allocated **\$75** to the **Humane League**.

Three other participants received a score of over 15:

1. **simon** received a score of **21.0**. He allocated **\$25** to the **Machine Intelligence Research Institute**.
2. **Adam Hesterberg** received a score of **19.5**. He allocated **\$25** to the **Sierra Club Beyond Coal campaign**.
3. **Viktor Bowallius** received a score of **17.3**. He allocated **\$25** to the **EA Long Term Future Fund**.

Congratulations to Scy, William, Ben, simon, Adam, and Viktor!

All right, let's take a look at what people did and how well it worked!

Round 1 analysis

Summary statistics

Recall that the score of a Round 1 entry is a weighted average of the probabilities assigned by Round 2 participants to the entry being real (i.e. truly random). The average score was **39.4%** (this is well below 50%, as expected). The median score was **45.7%**. Here's the full distribution:

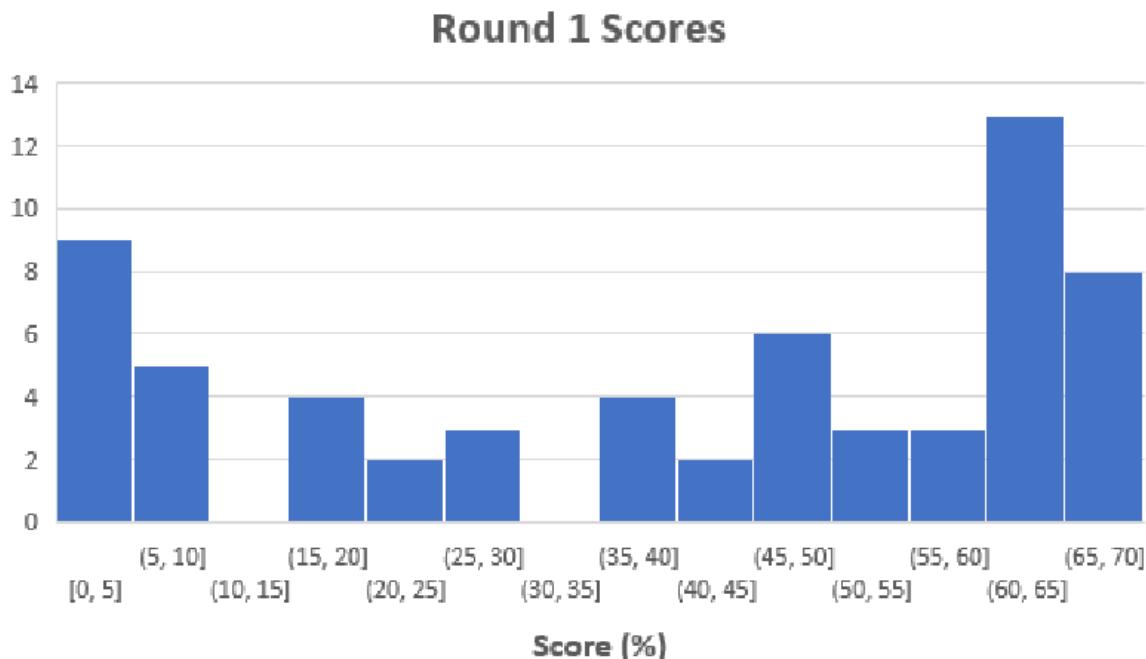


Figure 1: Histogram of Round 1 scores

Interesting: the distribution is bimodal! Some people basically succeeded at fooling Round 2 participants, and most of the rest came up with strings that were pretty detectable as fakes.

Methods

I asked participants to describe the method they used to generate their string. Of the 58 participants who told me what they did with enough clarity that I could categorize their strategy:

- 14 participants used a **memory-based** strategy. That is, they used a poem or text they had memorized, or digits of numerical constants, or their friends' birthdays, to generate their random numbers. The average score for strings in this category was **47.3%** (above average). Jenny, our Round 1 winner, used this strategy: she recited the poem "Renascence" and took the parity of the number of letters in the last word in each line.
- 19 participants used their **brain** as a built-in source of randomness. This meant things like just using their intuition to come up with 0s and 1s, or coming up with random words and taking the parity of each word's length. The average score for these strings was **41.7%** (about average).
- 14 participants used their **motor functions** as a source of randomness. This included things like "mashing the keyboard" as their free will dictated. The average score for these strings was **16.8%** (way below average).
- 11 participants used some sort of **mix** of these strategies. For instance, maybe they generated some bits from a poem and some bits using their intuition. The average score for these strings was **62.0%** (substantially above average).

Although I didn't participate, the strategy I thought to use was a mix of memory-based and brain-based. Specifically, I would have taken the digits of Pi as one source of randomness, used my brain to come up with random-ish bits, and then taken the bitwise XOR. (I'm not totally sure I would have been able to come up with 150 digits in 10 minutes with that method, though.)

Common pitfalls

By far the most common mistake made in Round 1 was not having sufficiently long (or sufficiently many) **runs**, i.e. consecutive zeros or consecutive ones. There is only a 0.5% chance that a random 150-bit string will have no run of length 5 (i.e. 5 zeros or 5 ones in a row); in comparison, 10 of the 62 strings submitted in Round 1 had no length-5 run. Many strings with insufficient runs were generated by either intuition or motor functions, though a few were not. An instructive example:

I took the opening lyrics from Hamilton, which I have in my head, and followed them letter by letter. Each letter later in the alphabet than the previous letter got a 0, while each letter earlier in the alphabet than the previous letter got a 1.

This generated the following string (Round 2 ID #80):

```
001010100101001100110110100110110100110001101110101100010001  
001101001111011001010101011001010101101010110110101100110101  
011101
```

The fact that this string has no runs of length more than 4 makes a lot of sense! After all, suppose you got four 0's in a row. That means you had a sequence of five letters, each later than the previous. The last of those letters is likely pretty late in the alphabet, which means that there's a much greater than 50% chance that the next letter will come earlier in the alphabet (so you'll write down 1).

The moral here is that there are many ways to mess up besides having a faulty intuitive grasp on randomness: you could also take a pretty good random number generator, but mess up when deciding how to convert it to zeros and ones. One other such pitfall was having a conversion method that resulted in an unbalanced string. One example:

I divided the alphabet into two groups of letters, then recited a poem I know by heart picking the bits based on which group each word's first letter was in.

I'm not sure how this contestant divided up the alphabet, but their string ended up with only 57 ones. The probability of a random string being that skewed (in either direction) is just 0.3%.

Some contestants who didn't get caught for insufficient runs or imbalanced strings got caught with more sophisticated tests (see the discussion of Round 2 below).

Interesting sources of randomness

Here are some Round 1 techniques I found interesting (whether or not they did well):

I tilted my keyboard so zero was "right and down" and 1 was "up and left" and tried to play sections of DDR songs from memory. I let "jumps" across 1 and 0 hit at the same time and left the order to randomness.

Jon

I moved around from state to state on a mental map of the United States. Move from state 1 to state 2. Convert the first letter of each state to a number 1-26. Write a 1 if state 2's number is in the thirteen above state 1's, wrapping around from 26 to 1, write a 0 otherwise. Repeat 150 times. My map contained mistakes, and I took multiple different routes around it.

Multicore

I listed the birthdays of people by how well I know that person, and then took the month (if between 1 and 8) and the day (if between 1 and 16).

Anonymous

A nearly perfect source of randomness, if you know enough birthdays! (This person didn't; they said that they then used song lyrics.)

I used the letters of a [poem] that we had to memorize in elementary school. I went through all the letters in my head. For every two consecutive letters x,y such that x precedes n in the alphabet and y does not, I wrote down 0. If y precedes n in the alphabet and x does not, I wrote down 1. Otherwise, I skipped the letters.

Anonymous

This seems to me like a great way to extract uniformly random bits from a text (though I'm not sure why they didn't just put down 0 or 1 based on which of x and y came first).

Round 2 analysis

In Round 2, I gave participants the 62 strings generated by humans in part 1 ("fake" strings) alongside 62 strings randomly generated by a computer ("real" string), all in a random order,

and asked them to assign each string a probability of being real. You can see all 124 strings [here](#).

There's a lot to say about Round 2. Let's dive in!

Fake strings were systematically noticed

Figure 1 showed the distribution of Round 1 scores: the average probability assigned by Round 2 participants to each fake string being real (weighted by Round 2 score). What if we similarly assigned scores to *real* strings?



Figure 2: histogram of probability assigned by a consensus of Round 2 participants to strings being real. Real and fake strings correspond to the orange and blue bars, respectively.

As you can see, real strings got systematically higher scores than fake strings, by a lot. Only 5 of 62 real strings were judged to be more likely to be fake than real, compared to a majority of fake strings. That's pretty cool (even if expected).¹

What, more concretely, helped people tell real and fake strings apart? Contestants used many methods, but the vast majority of the mileage came from looking at runs (consecutive 0s or 1s): as I mentioned, many fake strings failed to have sufficiently long runs (though in some cases people overcompensated and made their runs too long).

But “looking at runs” alone isn’t enough to get a good score. Let’s say you find a string that has one run of length 5 and no runs of length 6. That’s mildly suspicious, but how do you turn that into a *probability*? Converting a qualitative amount of “suspicion” into a probability was perhaps the most important part of the contest to get right. To understand why that was so important, let’s take a look at the scoring system used for Round 2.

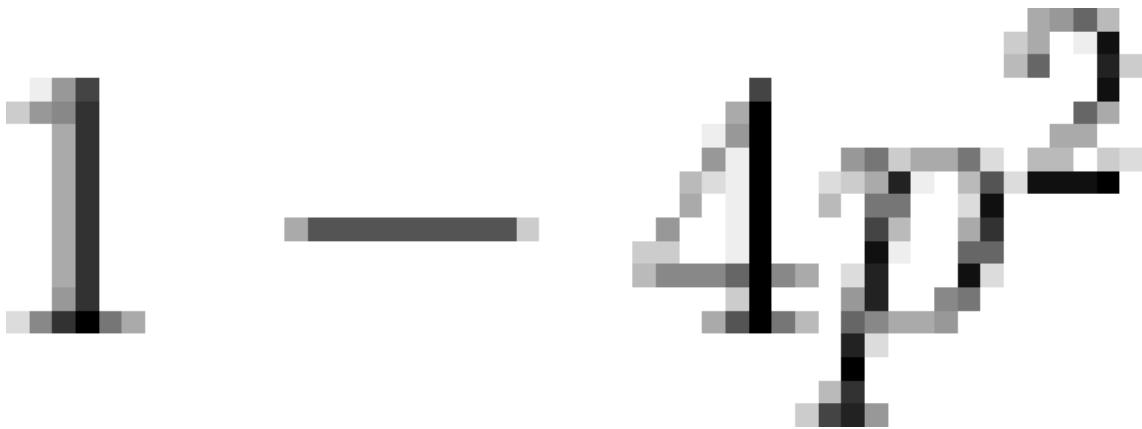
Scoring system

Slightly paraphrasing/rearranging my [Round 2](#) post:

*Let’s say you submit a probability p for a string. If the string is **real**, your score will be*

$$1 - 4(1 - p)^2$$

*If it is **fake**, your score will be*



*So you get a score of 1 for a string if you're completely right and -3 if you're completely wrong. Your total score will be the sum of your scores for all the strings. If you're going for maximizing your expected score, this scoring system [incentivizes you to be honest](#). Note that **your score will be 0 if you say 50% for every string**.*

(For those of you familiar with scoring rules, this just means I was using Brier's quadratic scoring rule, scaling it so that a totally uninformative forecast would result in 0 points.)

I also didn't catch that string #106 (which was fake) had a 9 in it; I ended up asking everyone to put down 0% for string #106. This means that everyone had the opportunity to say 0% for #106 and 50% for everything else, for a score of 1.

The other benchmark I set was 15, which was my estimate for what a "good" score would constitute. As I mentioned in the prizes section, five of the 27 Round 2 contestants cleared this bar.

Before going on to the next section, you might want to make a guess about the distribution of scores!

Summary statistics

Here's the distribution of Round 2 scores.



Figure 3: Histogram of Round 2 scores. Red = below 0, green = above 15.

Scores in the red part were **negative**: participants with these scores would have been better off saying 50% for everything. Scores in the green were the ones that met my "good score" benchmark.

The average score was -5.5. The median was -1.4. **A majority of scores (14 of 27) were negative.** This means that most participants would have gotten a better score if they had said 50% for every string!

The reason for this is **overconfidence** on the part of Round 2 participants. Proper scoring rules punish overconfidence pretty harshly. For example, if a contestant managed to classify 70% of strings correctly (better than almost anyone actually did) but was 90% confident of their classification for each string, their score would have been 0.² Let's take a closer look.

Basically everyone was overconfident

One nice thing about proper scoring rules is that *you can use people's answers to infer what score they expected to get*. For example, let's say someone submitted 70% for one of the strings. That means they think there's a 70% chance of the string being real, in which case they'll get a score of

$$1 - 4(0.3)^2 = 0.64$$

and a 30% chance of it being fake, in which case they'll get a score of

$$1 - 4(0.7)^2 = -0.96$$

That means their *expected score* for that string is

$$70\% \cdot 0.64 + 30\% \cdot -0.96 = 0.16$$

You can calculate an expected score for every string and add those up to find the total score that the participant expected.

If the previous paragraph didn't make sense, here's a simplification: you can tell what score someone expected to get based on how confident their answers were (how close to 0% or 100%). The more confident someone's answers, the higher the score they expected to get.

Of course, someone's *true* score may differ a lot from their expected score if they aren't calibrated. If someone is overconfident, their actual score will be below the score they expected to get. If they're underconfident, their actual score will be above their expected score. So we can figure out whether someone was over- or underconfident by comparing what score they expected to get to their actual score. Here's a plot of that.

Round 2 scores vs. expected scores

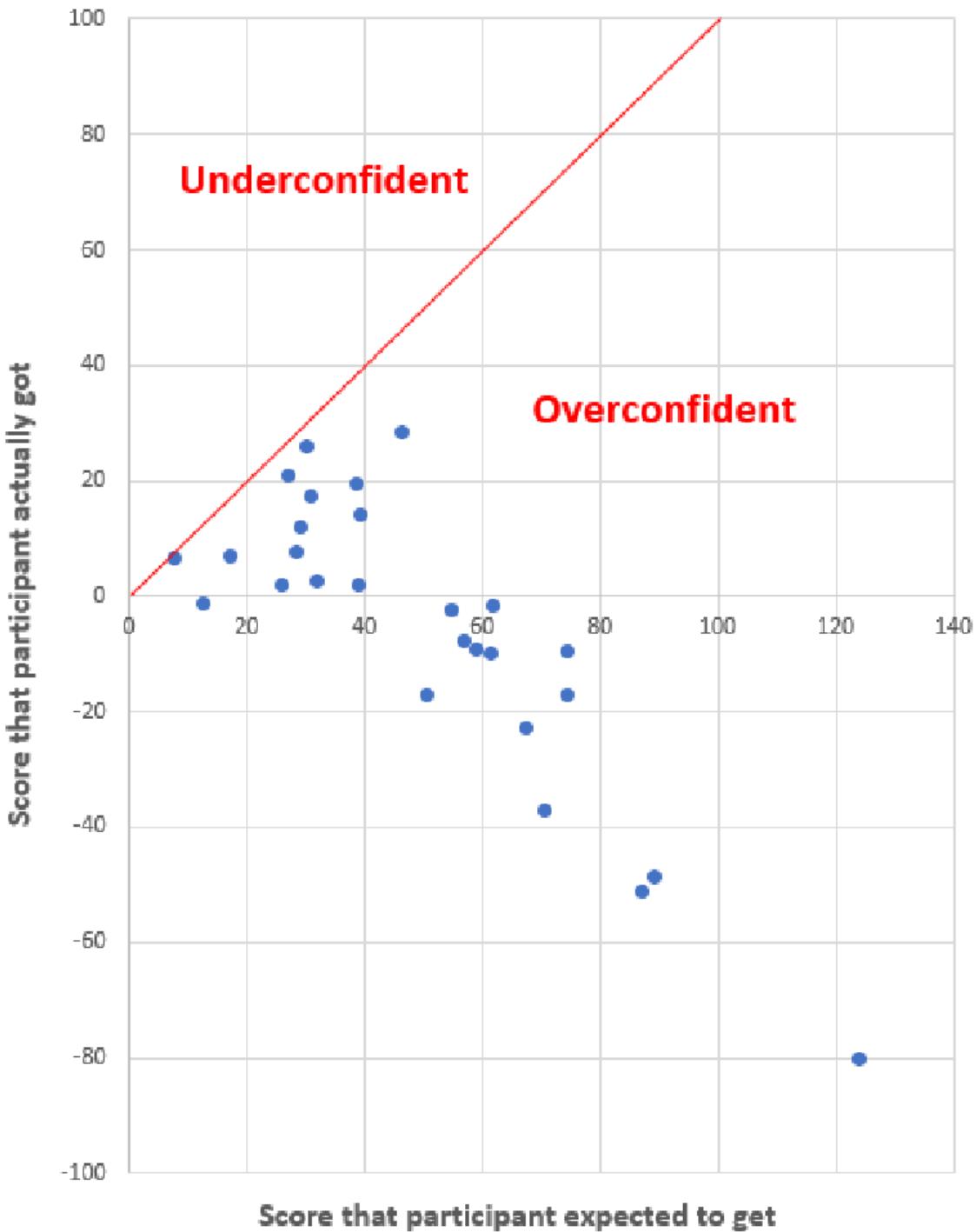


Figure 4: overconfidence in Round 2. Distance from the red line roughly corresponds to amount of overconfidence.

Every single point was below the red line, meaning that every single participant was overconfident (though three were close enough to the line that I'd say they were basically calibrated).

In fact, a nearly perfect predictor of whether someone would get a positive score or a negative score was whether they *thought* they were going to get a score above 50 or below 50.

Everyone who thought they were going to get a score above 50 got a negative score. With one exception, everyone who thought they were going to get a score below 50 got a positive score. This is pretty remarkable!

One concrete way to measure overconfidence is “how much squeezing toward 50% would the participant have benefitted from”. For example, squeezing predictions halfway to 50% would mean that a 20% prediction would be treated as 35% and a 90% prediction as 70%. We can ask: **What percent squeezing would optimize a participant’s score?**³

For a perfectly calibrated forecaster, the optimal amount of squeezing would be 0%. But everyone was overconfident, so everyone would have benefited from some amount of squeezing. How much squeezing?



Figure 5: histogram of optimal amounts of squeezing probabilities submitted in Round 2 toward 50%

The range of optimal squeezing amounts went from 7% (achieved by Ben, who got 2nd place) to 82%, with an average of 47%. So on average, it would have been best for participants to squeeze their probabilities about halfway toward 50% before submitting. (For comparison, the winning team would have benefitted from a 19% squeeze.)

The next chart shows participant scores as a function of how overconfident they were.



Figure 6: Round 2 score versus amount of overconfidence (measured by optimal amount of squeezing toward 50%)

This plot shows that being calibrated was crucial to getting a good score. But calibration wasn’t everything: Ben submitted the most calibrated entry (the red point) but got second place, while the winning submission (green), submitted by Scy and William, was only the fourth most calibrated.

So what did the winning team do so well?

Before we answer this, it’s instructive to consider the second most calibrated entry, which is the yellow point in the previous chart. This entry got 10th place out of 27 — better than average, but not by much.

What happened? Out of 124 probabilities, this contestant submitted 50% 114 times, 10% 6 times, and 0% 4 times. (They went 4/4 on the 0% entries and 5/6 on the 10% ones, though the one miss was bad luck — I’ll touch on that later.) This contestant did fine, but didn’t do great because of poor **classification**, i.e. an inability to systematically sniff out fake strings.

One rudimentary measure of classification is what fraction of strings were “called” correctly. That is, one could look at how well contestants would have done if I had simply awarded each contestant a point for every real string to which they assigned a probability greater than 50% and every fake string to which they assigned a probability less than 50% (and half a point if they said 50%). Here’s a plot of contestant scores vs. the fraction of strings they called correctly.



Figure 7: Round 2 score (y-axis) versus percentage of strings classified correctly (x-axis)

By this metric, Scy and William were the best at classification by a substantial margin. The 2nd and 3rd best entries by classification got 3rd and 4th place, respectively, in Round 2, behind Scy/William and Ben.

The plot below is meant to illustrate in more detail how exactly Scy and William beat out Ben despite being less calibrated.



Figure 8: probability assigned to each bit string being real by Ben (most calibrated; 2nd place on the x-axis, versus by Scy and William (1st place; 4th most calibrated) on the y-axis. Blue points are real strings; orange points are fake strings.

Here, every point corresponds to one of the 124 strings. Fake strings are orange; real ones are blue. The x-value is the probability Ben (the most calibrated, 2nd place contestant) assigned to the string being real; the y-value is the probability assigned by Scy and William (1st place, 4th most calibrated).

Here's the same chart again, this time with emphasis added to two parts of the chart.



Figure 9: same as Figure 8, but with rectangles indicating strings Scy and William were confident were fake but Ben was not (green) and strings Ben was confident was fake but Scy and William were not (purple)

The green rectangle contains all the strings that Scy and William said were very likely fake but Ben did not. There are 13 such strings, 11 of which were in fact fake.

The purple rectangle is all the strings that Ben said was very likely fake but Scy and William did not. There are 3 such strings (you only see one dot because they're on top of each other). One of those was Ben's string. (The other two had long substrings of the digits of Pi mod 2, which Ben looked for but Scy and William didn't.)

So basically, even though Scy and William were less calibrated than Ben (three strings they assigned probability 0 to were real), they discerned fake strings better. And ultimately this made them win, earning 28.5 points to Ben's 25.8.

This is in part why I used the quadratic rule instead of the log rule for scoring: the log rule punishes miscalibration at the extremes really harshly, and I didn't want someone who was slightly less calibrated but much more discerning to lose. Had I used the log rule (and rounded 0% probabilities of real strings to 1%), Ben would have narrowly won.

Calibration vs. classification

Did more calibrated contestants also classify more strings correctly?



Figure 10: optimal amount of squeezing toward 50% on the x-axis (note that the axis is reversed so that "more calibrated" is toward the right) versus percentage of strings classified correctly on the y-axis

The answer is yes, somewhat. The r^2 of the relationship here is 23%, meaning that contestants' calibration explained 23% of the variance in how well they "called" strings. This makes sense: the more effort you put (and the more experience you have with this, etc.), the better you'll do on both fronts.

(By the way, it's pretty nice how the top three finishers formed the calibration-classification Pareto frontier!)

If you're interested in reading more about the calibration-classification dichotomy, I also have a [more detailed post](#) about it.

Okay but what strategy did the winners use?

I'll quote Scy and William's description of what they did verbatim (modulo small stylistic edits), alongside clarifying comments.

We assigned 0 or close to for ~20 numbers that looked "obviously fake", then added another ~10 to the 0 list with:

- *Binomial CDF (too big a skew got a low number)* (I think this means if the string had too many 0s or 1s -Eric)
- *Plotting the mean and standard deviation of contiguous bits with a bit string (i.e. the lengths of runs -Eric) and superimposing it on a plot where 360 pseudorandom bitstrings of my own devising (i.e. computer-generated random strings -Eric) were plotted, and eliminating outliers*
- *Count the distribution of length-2 bitstrings within each string and figure out the difference between the most and least common (should be fairly low for random strings - the difference topped out at ~30 for most runs for ~124 randomly generated strings, so we gave a low probability to anything > 30)*
- *gzip'ed them and flagged one that was unusually compressible*

For the remaining, we couldn't figure out a satisfactory way to formalize the probability that they were fake. We ended up assigning 0.67 (62 / 62 + un-eliminated human generated strings) to them by default, and using numbers-out-of-ass to dock or bump entries visually, based on how far they were from the center of mass of our pseudorandom bitstrings plotted on the mean + standard deviation 2D distribution.

Scy and William kindly shared with me their plot of the mean and standard deviation of run lengths:

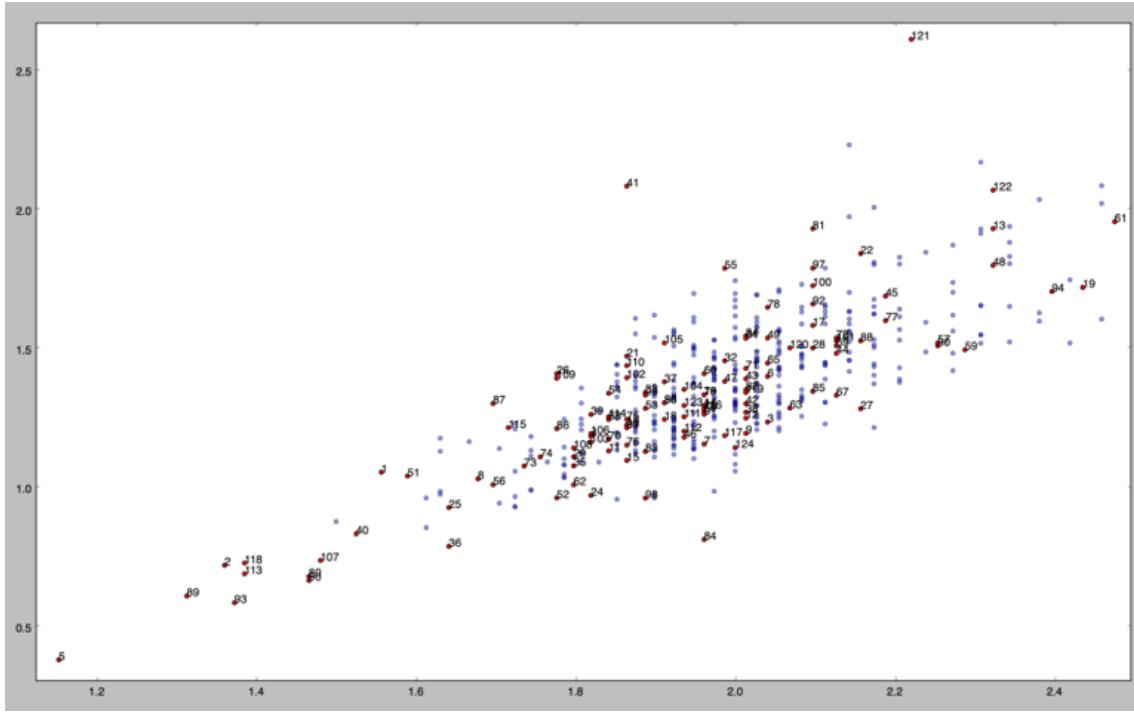


Figure 11: the average length of runs of the same digit consecutively on the x-axis, and the standard deviation of these values on the y-axis. Red points are the 124 given strings (with ID numbers written). Blue points are 360 computer-generated random strings. [See here](#) for a larger version.

([Click here](#) for a larger version of the plot.) In this plot, red points are the 124 strings given to participants (along with their [ID numbers](#)), while blue points are computer-generated random strings. The x-axis is the average length of a run; the y-axis is the standard deviation of run lengths.

The idea is pretty simple: blue (truly random) points formed a pretty narrow cloud on this plot. A red point that was far away from this cloud was therefore likely to be fake.

Many people looked at average run lengths, but not many looked at the *standard deviation*. This helped them catch a few fake strings that the consensus failed to catch, including strings 27 and 124 (whose standard deviation was too small), and 81 and 105 (too large). I haven't done a deep dive, but it looks to me like this strategy helped set Scy and William's entry apart.

What other strategies did well?

I'll again quote the other top contenders' strategies verbatim. First, Ben (2nd place, most calibrated):

*On a hunch, I looked at the number of times a 1 followed a 1 or a 0 followed a 0 in each string ("runs"). This empirical distribution had **way** higher variance and somewhat lower mean than would happen by chance. So I used Bayes' rule to update my prior probabilities, with likelihoods coming from a smoothed version of the observed runs distribution (a sketchy technique, but perhaps not too sketchy). This was my main source of information.*

Separately, I searched for various substrings within the strings: long stretches of 0s (and 1s), parities of digits of pi and e forwards and backwards, alternating 1s and 0s, alternating

pairs of 0s and 1s, and a few others. A few strings had unrealistically long substrings of these forms, so I gave them 0 probability of being random.

I then applied normalization to ensure the sum of predictions was 124.

I also found that the variance of the empirical distribution of Hamming weights was quite high (but not nearly as high as the runs distribution variance). ("Hamming weight" means "number of 1s" -Eric) This seemed to be explained mostly by the tails, so I applied some semi-ad-hoc scaling to my predictions for strings with particularly high or low Hamming weight, and applied normalization again.

These methods are very solid, but nothing here is particularly *special* in terms of doing things no one else did. My best guess, then, is that Ben had great execution. That is, his Bayes updates (which many people tried to do with varying degrees of success) were roughly of the correct size. This theory is supported by the fact that Ben was well-calibrated.

There is something I found really interesting about Ben's submission, which is that his probabilities went as high as 88%. Ben's was an outlier among the best entries in this sense: most of the entries that did best had a maximum probability in the 60-80 percent range. This confused me: could you really be 88% sure that a string was random, when you know you're assigning something more like 60-65% to a *typical* random string? Can a string look *really random* (rather than merely random)? This sort of seems like an oxymoron: random strings don't really stand out in any way sort of by definition.

In response, Ben told me:

It's because those strings had run counts that were uncommon among (a smoothed histogram of) the given strings but not too uncommon among the random strings.

But the frequency of particular run counts should be at most twice as rare in the 124 Round 2 strings and in a sample of random strings, since half of Round 2 strings are random. So is there overfitting going on? Ben's submission was well-calibrated, so perhaps not. I remain confused about this.

Next we have simon (3rd place):

I used LibreOffice calc - I checked for extreme values of: (1) total number of 1's (2) average value of first XOR derivative (don't know if there's a usual name for it) (I think "XOR derivative" refers to the 149-bit substring where the k-th bit indicates whether the k-th bit and the (k + 1)-th bit of the original string were the same or different. So this is measuring the number of runs. -Eric) (3) average value of second XOR derivative and (4) average XOR-correlation between bits and the previous 4 bits (not sure what this means - Eric). In each case, I checked to see if the tails were fatter than expected for random chance and penalized strings in the tails to the extent that was the case (and only the amount of tail that was in fact fatter than expected). Then I multiplied the resulting probabilities obtained for each statistic together and rescaled for the final answer... Only a minority of the strings were sufficient outliers in any test to receive a penalty, which is why most strings have the same 64 percent (rounded from 63.6 percent) estimate.

I'm curious how much, if any, of simon's success came from (3) and (4). I'm not sure what (4) refers to, but (3) was not commonly done.

EDIT: See [here](#) for simon's explanation of the XOR-correlation, as well as an answer to my curiosity. The answer is: they both helped a decent amount!

Next we have Adam (4th place).

I checked each of the following: length of runs of consecutive 1s, same for 0s, squared distance of the distribution of substrings of length 6 from the uniform distribution of substrings of length 6, same for 5, same for 1 (didn't bother with other numbers),

maximum length of a substring shared with another answer, and maximum-length shared substring with any mapping of digits 0-9 to 0-1 applied to digits of pi. I then multiplied a running odds ratio by the odds that an answer is that extreme relative to the expected most extreme value for that statistic among 124 random strings.

I'm curious about the particular choices of 5 and 6. Would Adam have done better by incorporating evidence from every length?

Finally, Viktor (5th place):

I made a simple program that counted the sequences of consecutive numbers. For example, the sequence 0010111001100 Would generate:

Zeros: 7. Sequences with 3 consecutive numbers: 1. Sequences with 2 consecutive numbers: 4. Sequences with 1 consecutive number: 2.

For example if an entry had more than 35 sequences with just 1 consecutive number (meaning alone), it was very likely humanly generated. If the number of zeros was under 60 or over 90, that was also an indication the sequence was fake.

Then I used my gut feeling to give odds to "odd seeming" entries that might have had abnormally many sequences with 5 consecutive numbers or something. If something was just a little odd, I would just give it slightly lower probability of being truly random. I also made a blind guess that entries with 8 or 9 as its longest sequence was slightly more likely to be truly randomly generated. I still don't know if that assumption was correct.

I entered all my guesses into an Excel spreadsheet, and there I saw my average guess was like 35% chance of being truly random, so I added a 1 too all numbers, and multiplied with something like 1.4, so the average would be close to 50%.

I think I might have made a mistake by multiplying by a constant, because if I had two numbers, let's say 40 and 50, if I multiply by 1.4, the difference between them increases 16% (not sure where the 16% is coming from -Eric), which was probably a bigger difference than what I originally might have believed.

It's pretty interesting that this did well, given the reliance on gut feeling and the crude 1.4x scaling at the end, but it seems to have worked out! I wonder if a good gut feeling outperforms all but the best automated methods because automated methods are pretty easy to mess up.

Other interesting strategies

Most people did some variation of "I looked at runs". Relatively few people did anything that surprised me. My favorite of these non-orthodox methods:

Another thing that I tried was to plot the strings with "turtle graphics", so you can visualize each string "fingerprint". For this to be done, at first I turned the turtle 90° degrees left if the digit is a 0 and to the right if it is an 1 and advanced one step per digit. Strings with a lot of 0/1 groups tend to curl and pack, while strings without those groups tend to go away. (Strings 5 and 66 are nice to be see because they are clearly different from each other.)

Jorge Tornero

Cool! I bet this visualization makes real and fake strings look pretty different.

Another interesting method:

I had some linalg code lying around, so I just formed small matrices over F_2 (the field of 2 elements -Eric) out of the sequences, and calculated their ranks to tease out linear-ish dependencies.

Fishlips

Interesting, I wonder how much you can get out of just looking at linear dependencies like this.

Finally:

I calculated the relative frequency of various substrings of equal length — e.g. “0101” vs “1111” — up to a length of 10, and compared the squared error of the competition strings to that of a population of 1k random strings.

Measure

Interesting: Measure looked at relative frequencies of *all strings* of a given length, rather than only looking at runs. I wonder how much that helped.

Final comments

Did participants who did well in Round 2 do well in Round 1?

Interestingly, there was no correlation between Round 1 and Round 2 scores (though as a group, people who *participated* in Round 2 did quite well in Round 1).



Figure 12: Scores in Rounds 1 and 2 of the 23 contestants who participated in both rounds

A bit of bad luck for Round 2 participants

[Figure 2](#) showed that two of the random strings looked pretty fake. These were #121 and #122.

```
#121:10111111010101100110001010111111111001111111111110000001000110100  
110110001011100101011110111011110110000000001101101001011101010111111111  
10101
```

```
#122:10101100111000111010010110111011011000111101001100111111111110111000  
1111111011011100011101111111110111001010011110110001100111011101001110  
00110
```

What made them look fake is that they had, respectively, 56 and 53 zeros — much less than half of 150 (3.1 and 3.6 standard deviations too few zeros, respectively). The probability of having as unbalanced a string as #121 in a sample of 62 random strings is 12% (not too unusual). The probability of having as unbalanced a string as #122 is just 2%.

Given that balance between zeros and ones is one of the most basic things to look at, Round 2 participants got *unlucky* in that they were fed a string that looked much faker than the fakest real string should look — in fact, arguably two such strings.

Contestants *justifiably* gave probabilities near 0 for both strings. This cost participants a few points. This also means that the previously discussed overconfidence is *slightly* an artifact of this bad luck. If not for this, three participants (the top three by calibration) would have been essentially correctly calibrated; everyone else would have still been overconfident, though.

Raw data

- The 124 strings (62 fake and 62 real) are available [here](#). The answer key is on the second sheet.
- [Here](#) is a (mostly anonymized) spreadsheet containing all Round 2 entries, as well as some summary statistics for each participant. I think the data is fascinating (as you can probably tell from the length of this post), and I'd be thrilled if people played around with the data and found other interesting things!

1. This is admittedly a little sketchy, since I'm determining the weights based on ability to discern fake strings. However, I think there are enough total strings that this causes only a small amount of bias.
2. If I had chosen the log scoring rule instead, I would have been punishing overconfidence even more harshly!
3. Interestingly, there is a nice pictorial representation of the optimal amount of squeezing: in Figure 4 (repeated below), it is the distance from the point to the red line, divided by the point's x-value, and then divided by the square root of 2. Alternatively (using the red point below as an example) it is the ratio of the lengths of the green segments.



Figure A: the optimal amount of squeezing for the submission represented by the red point is equal to the ratio of the lengths of the top and bottom green segments.

Cryonics signup guide #1: Overview

This is the introduction to a sequence on signing up for cryonics. In the coming posts I will lay out what you need to do, concretely and in detail. **This sequence is intended for people who already think signing up for cryonics is a good idea** but are putting it off because they're not sure what they actually need to do next. I am *not* going to address the question of how likely cryonics is to work – that's been [covered extensively elsewhere](#).

If you have no idea what cryonics is, or if you want a thorough refresher, I recommend WaitButWhy's [Why Cryonics Makes Sense](#).

Biases

This sequence is US-focused, since I went through the process in the US. It's also somewhat Alcor-biased, since I chose Alcor quite early on in the process. However, I've collaborated with both non-US cryonicists and people signed up with the Cryonics Institute, so I'm confident there will be useful information no matter where you are or which institution you choose to keep you in a vat of liquid nitrogen.

Epistemic status

I approached all questions in good faith and have documented my reasoning to the best of my ability, but I don't have a high level of confidence in my conclusions. Commenter Josh Jacobson is signed up with the Cryonics Institute and [had a very different experience](#) than the one outlined in this sequence, and I don't think I have any special knowledge or abilities that he doesn't. My recollections of the research that led to these posts has also faded with time.

Caveats

This sequence was researched and written in late 2020, and just two years later, it seems that the landscape has already changed significantly. For example, Alcor has changed their membership options, their fee structure, and their payment options, and they've also introduced an online signup flow that I have no experience with. As such, please be aware that some of the logistical advice in this sequence may be outdated. I have tried to update the sequence where possible, but I'm not going to go through and overhaul it.

Acknowledgements

Thanks to Connor Flexman, Daniel Filan, Gordon Worley, Mati Roy, Seraphina Nix, and nameless others for letting me ask them endless questions. Thanks also to Eli Tyre and [Oge Nnadi](#) for their previous writeups on this topic, from which I borrowed liberally.

Summary of the process

The first thing most people probably want to know is: What do I do now? It turns out to be really hard to figure this out, and I think unnecessarily so – the information is out

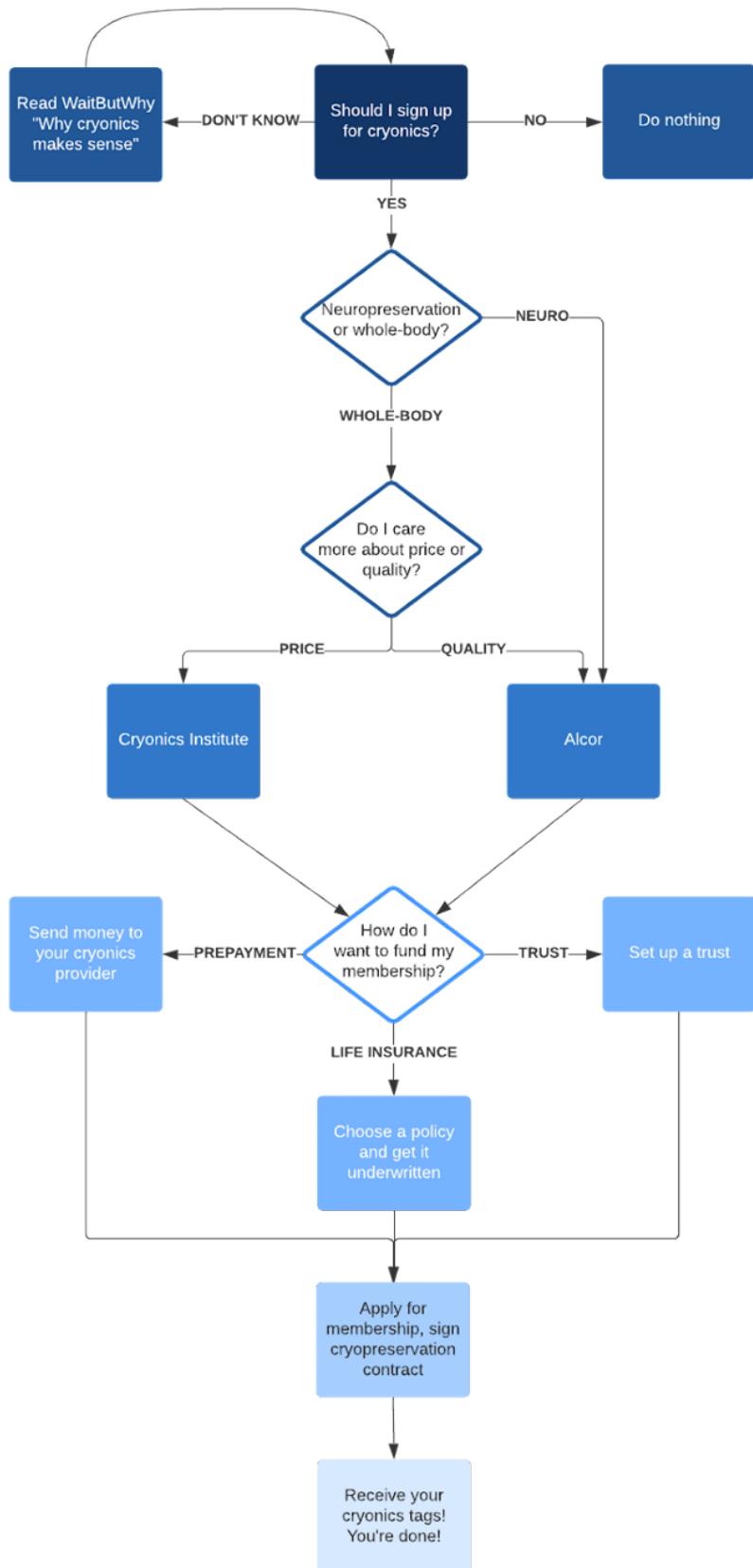
there, but it's not all written down clearly in one place. This sequence is my attempt to rectify that.

Basic process overview

Here is a basic overview of the cryonics signup process from start to finish:

1. Preliminary decisions
 1. Neurocryopreservation vs whole-body cryopreservation
 2. Cryonics Institute vs Alcor
2. Contact an agent to get life insurance
3. Fill out and submit cryonics membership application
4. Sign cryopreservation contract
5. Optional additional paperwork
6. Keep your policy and membership up-to-date forever
7. Be cryopreserved upon your legal death

For those who want to get oriented visually, here's a flowchart covering the basics:



Sequence outline

And here is the outline of this sequence:

1. Introduction (*you are here!*)
2. [Neurocryopreservation vs whole-body cryopreservation](#)
3. [Cryonics Institute vs Alcor](#)
4. [Intro to life insurance for cryonics](#)
 1. [Types of life insurance](#)
 2. [Cryonics-friendly life insurance carriers](#)
 3. [Cryonics-friendly life insurance agents](#)
 4. [The insurance underwriting process](#)
5. [Making it official](#)
6. [Optional additional steps](#)
7. Actually putting someone in cryostasis (possibly forthcoming late 2022)
8. [Appendices](#)

You may notice similarities to the process overview above, with the main difference being an outsize focus on paperwork, and particularly life insurance. This is because life insurance is a cesspool of bureaucratic bloat, and I wanted to lay things out really clearly so that you can navigate it without crying as much as I did. Once you've secured your funding method (whether that's life insurance or something else), the rest of the process is very straightforward!

I think the preliminary decisions – on whole-body vs brain and which provider to use – merit a fair amount of consideration as well. I've already made my decisions there, but you may have different cruxes than I do; the questions raised can get pretty philosophical.

What I chose

If you just want to offload all of the complex decision-making to me (the person who spent several months writing this sequence but has no other relevant qualifications), I chose Alcor neuropreservation, which I funded by a \$200,000 indexed universal life insurance policy from Kansas City Life, with help from the insurance agent David Donato. I made these choices as a fairly well-off 25-year-old female US citizen with no major health problems and no history of drug use. If you are in a substantially different situation but still want to defer to my judgment, send me a DM and I can help you figure out what's right for you.

Should I sign up?

Even though this sequence assumes you think cryonics is a good idea in the abstract, you might be wondering if you, personally, should actually sign up for it, and if so when. Below I'll discuss a couple factors that might help you make that decision.

Costs

Monetary cost

Cryonics is not just for rich people. It does cost money, but it's really not exorbitant, especially if you're young and healthy. There's a wide range of possible costs (corresponding to different choices of cryonics provider or life insurance policy type) that bottom out around \$25 a month. I personally (25-year-old female, who did not make decisions primarily based on price) pay about \$240/month.

For most people, I think this cost is probably worth a small (but not infinitesimal) chance at immortality. Sure, if it's a choice between paying life insurance premiums and having enough to eat, feed yourself first. But if you're at all financially secure, and you think cryonics is a good idea but just haven't gotten around to signing up, I don't think you should view the cost as a major deterrent.

Time cost

Signing up for cryonics takes a fair amount of time, even if you come in understanding exactly what to do, and also offload the paperwork to an insurance agent. So if you're busy with something very urgent – like, if you've been spending all your mental energy this year advising national governments on their pandemic response measures – then now is indeed probably not the best time to sign up. But if it's been like five years and you *always* feel too busy to get around to it, I recommend just biting the bullet and doing it now.

If you're not price-sensitive, you could pay someone to do nearly all of the work for you, but you'd still have to hire that person, provide them with your personal information for filling out forms, look over their decisions, sign papers, and potentially submit to a medical exam. My guess is it'd be hard to get it below ~10 hours total.

If you're not willing or able to pay someone to go through the signup process for you, expect more like ~40 hours. That's a significant chunk of time, but not an unmanageable one.

Attention cost

The signup process just takes a while, even if you do everything right ([Oge reports](#) 11 weeks between seeking out a life insurance policy and receiving his medallions), and so there's a sustained back-of-the-mind attention cost until it's over. Being signed up for cryonics also requires a bit of ongoing maintenance (something I'll cover in a later post), but not much more than, say, taking out a rental insurance policy does.

Now vs later

I know someone who signed up with Alcor in their twenties, and then the next year was diagnosed with potentially fatal cancer. If they had waited any longer, they would have been uninsurable, or in the best case, their life insurance premiums would have been crazy, unaffordably high. As it turned out, they remain insured, grandfathered in at their previous price. Sure this is just an anecdote, but it really drives home for me that, while you may be at an age when it's statistically highly *unlikely* that you'll die, it's never *impossible*.

All that's to say: If you think it's a good idea, do it now; don't put it off. If you're uncertain whether it's a good idea, find the root of your uncertainty and make a real decision, rather than just indefinitely [driving at half-speed](#).

But I'm not in the US!

Not a problem! You can still sign up with Alcor or CI, and fund your membership using life insurance, so nearly everything in this sequence will still apply to you.

If you're looking into the signup process and are not in the US (or need to work with anyone outside of the US), I strongly recommend finding cryonicists in the relevant country; they'll be able to help you with bureaucratic specifics more than I can. Here are some links I found (disclaimer that I'm not endorsing any of these and they might not even still exist):

- [Australasia](#)
- [Belgium](#)
- [Finland](#)
- Germany: [Cryonics Germany](#), [German Society for Applied Biostasis](#)
- [Greece](#)
- [Italy](#)
- [Netherlands](#)
- Portugal: [Alcor Portugal](#), [Cryonics Portugal](#)
- [Québec](#)
- [Southern hemisphere](#)
- [Switzerland](#)
- [UK](#)

Likely-outdated email contact info for additional groups available [here](#).

What's the lowest-effort thing I can do right now?

If you don't expect yourself to go through the full process right away for whatever reason, but you want to increase your chances of cryopreservation in the event of your death, you should **sign a [Declaration of Intent to Be Cryopreserved](#)** (form [here](#)).

This constitutes informed consent, making it *much* more likely that it will be legally possible to preserve you in case of an emergency. I expect this to take less than 30 minutes in total.

(Note: I previously recommended that people also become an [Alcor Associate Member](#), but as of September 2022 Alcor is no longer accepting new associate members.)

That's it for now! Stay tuned for many more posts that are very technical and much longer than this, and please comment if you have any questions!

Fourth Wave Covid Toy Modeling

Epistemic Status: Highly speculative. I threw this together quickly, and wrote this to document how I went about it. This is an attempt to create a first toy model, so others can error correct and improve, and upon which less-toy models can hopefully be built. You can see the spreadsheet with my work [here](#). Please take this and run with it, and please don't take this as more than trying stuff to see what's wrong with it.

No one seems to be creating models of various scenarios in a way that feels remotely realistic, or even in a way that feels super simplified but that can be used as intuition pumps or baselines.

This post aims to fix that, or at least provide a first step.

At this point, we mostly know we're f***ed, and that the new strain is at least ~40% more infectious, probably 50%+ more infectious, perhaps as high as 65%-70%.

What happens now?



Nate Silver @NateSilver538

...

It really does seem like it's going to be a race between how fast the new variant spreads and how quickly we can get people vaccinated. I suppose I like the variant's odds, in the short run. But I don't think people should be fatalistic; speeding up vaccines could help a lot.

4:03 PM · Jan 4, 2021 · Twitter Web App

Speeding up vaccinations seems to be the opposite of what we are doing, but seeing exactly what it would take to get us out of this mess seems like a worthwhile exercise. Time to start building toy models.

Completely Naive Model

To start off, I did the quickest thing at every step to see what happened.

Note that we're measuring and predicting *actual cases* not *positive tests*. The observed and official numbers will always be several times lower than the real ones.

This model assumes:

1. Covid Machine Learning's estimate of old infections is correct.
2. Covid Machine Learning's estimates of current infections are roughly correct, about 600,000 per day.
3. Based on sequencing results, we estimate 0.06% of new cases for cycle ending December 19.

4. We assume every 'infection cycle' is five days, so if you are infected on day 0, you infect others on day 5.
5. The initial strain had an effective $R_0=1$ on December 19, when things began.
6. The new strain is 50% more infectious than the old one.
7. Immunity is total if you have been infected.
8. Vaccinations don't exist.
9. Control systems don't exist, people don't adjust behavior at all.
10. Heterogeneity doesn't exist, people are all the same. Straight SIR.

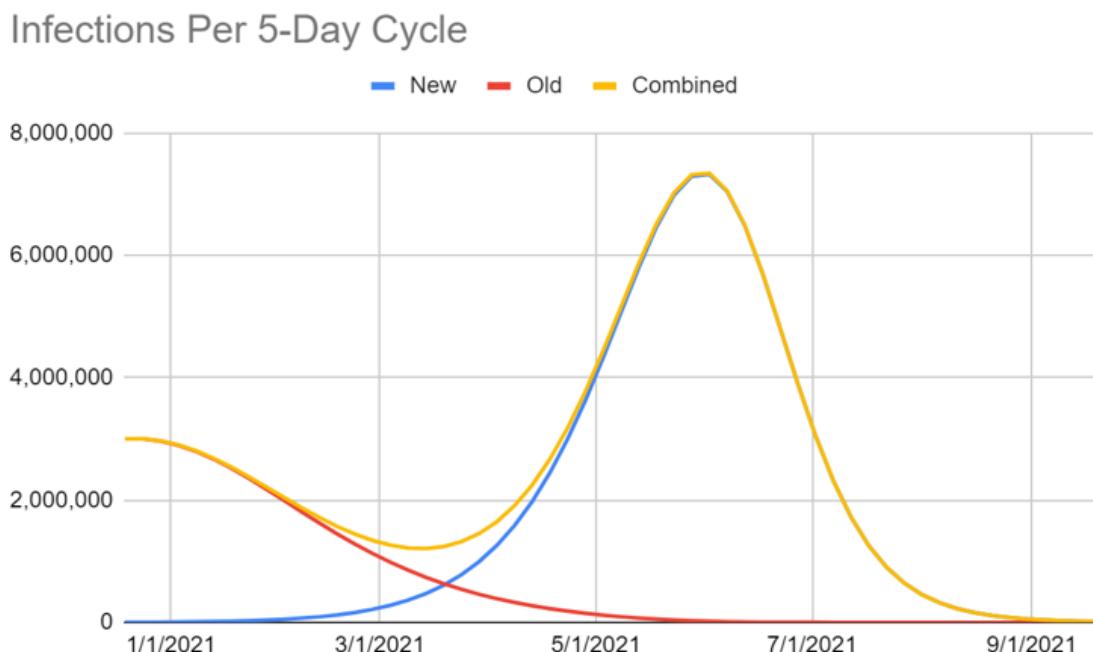
That's *deeply silly*, of course, but it's not *more deeply silly* than models that have been offered and taken seriously in the past, despite being obvious nonsense. Assuming away vaccinations is the most obviously nonsensical part of that, but control systems and heterogeneity point in the other direction in many places, and our vaccination progress has been pathetic, so it makes sense to start off extra naive.

I do believe approximately in assumptions 1, 2, 3, 4, 5 and 7, so we have three things we'll need to fix: Heterogeneity, control systems and vaccinations. And we'll need to vary the infectiousness level of the new strain, considering at least the 50% and 65% cases.

We'll also want to add deaths, especially since the control system likely depends on that number.

First let's check out what you get without fixing any of the three issues.

You get this:



This is rather disastrous.

The safest day between now and the endgame is around Pi Day, March 14, at roughly 40% of current levels.

The most dangerous day is around June 1, at around 244% of current infection levels. If you are not infected by then, you'd be 380% more at risk than today, because of increased

immunity in others. In addition, the hospital system would presumably be under at least extreme strain.

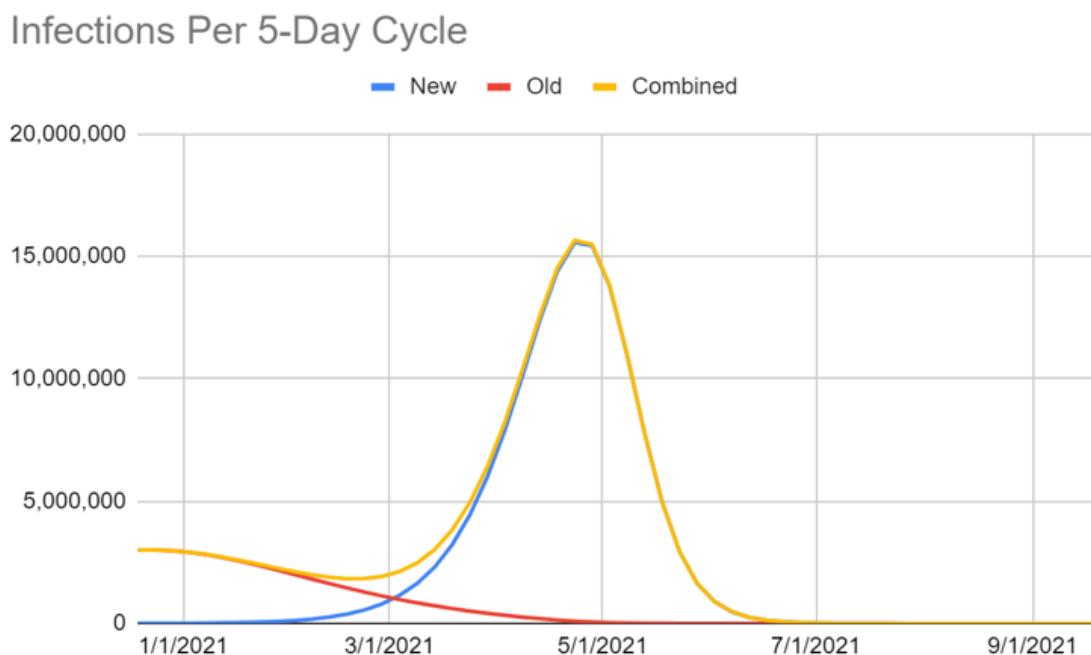
The number of infections reaches current levels twice. On the way back up, it happens on about April 20, and on the way back down, on about July 1. Note that July 1 is still much more dangerous than today for those still vulnerable.

We get to 10% of current levels around August 8, 10% of that around September 1.

The new strain is 1% of infections by January 20, 10% by February 20, 50% by March 20.

Approximately 63% of the population ends up infected.

We can contrast that with a 65% more infectious strain:



That is a true disaster. This is not the same scale as the previous graph. Check the Y-axis.

The safest day is now February 17 or so, most dangerous is April 24 and it's really bad with over *three million infected* on that day alone. Hospitals would be facing five times the current case levels, and would be fully in triage mode at best.

We get back to current levels by around May 21, 10% of that by June 10, 10% of that around July 1.

The new strain is 1% of infections by January 15, 10% by February 9, 50% by March 2.

Approximately 75% of the population ends up infected.

We are not currently approximating deaths. If we did, since we are assuming zero vaccinations and heterogeneity, we'd choose a fixed IFR and then adjust for the collapse of the hospital system. It would not be pretty.

That's all an intuition pump of why there's a *very large* difference between 50% and 65% more infectious. It compounds quickly, and the crisis hits much faster and much harder. The

second scenario's peak crisis has twice as many simultaneous cases.

Adding Vaccination

The most obvious missing knob is vaccinations.

We presumably add some number of people who are vaccinated, and thus effectively immune, on some accelerating function from week to week.

Given how things are going, it seems safe to treat being vaccinated and previously infected as mostly uncorrelated.

As of January 5, about three weeks after approval, 1.4% of the population (4.73 million doses) has been confirmed as vaccinated. That's a really pathetic pace, and didn't even require second doses. It seems safe to assume we'll do better than *that* going forward, but how much better? That's much harder to say. The limiting factor could quickly shift to the limited vaccine supply soon, so one approach is to assume that we administer most of the doses we have purchased. Biden has vowed 100 million shots in 100 days, so my 'baseline scenario' will assume we can do 250,000 vaccinations a day from December 14 until January 20, then 500,000 vaccinations a day starting January 21, and they are (for now) chosen at random.

Once someone is vaccinated, how immune are they and how fast? I will assume for now that everyone gets shot number two at the three week mark. My understanding is that you're 80% immune by about day 10, and that with the booster we will say that jumps to 95% immune on day 30. We'll say that you get no effect until day 10, people will be careful and be dealing with side effects at home, but also will have had to take some physical infection risk to get the shot, so let's say those effects cancel.

To make my life easier without changing the math much, we'll say that you get full 95% protection on day 15, and before that you have nothing, and simulate that by counting vaccinations as if they happen on day+15. It's basically the same thing, and avoids having to keep track of who catches the virus before the vaccine can work.

There's the model where such folk are X% immune to each potential infection, and there's the model where such people are X% likely to be immune period and (100-X)% likely to be fully vulnerable. I'm going to model the second version, so all we have to do is multiply our vaccination numbers by 95%.

So each cycle, there will be an additional 2.5 million people vaccinated, which we will add to either the "Vaccinated and Previously Infected" or "Vaccinated and Never Infected," and both categories will be fully safe.

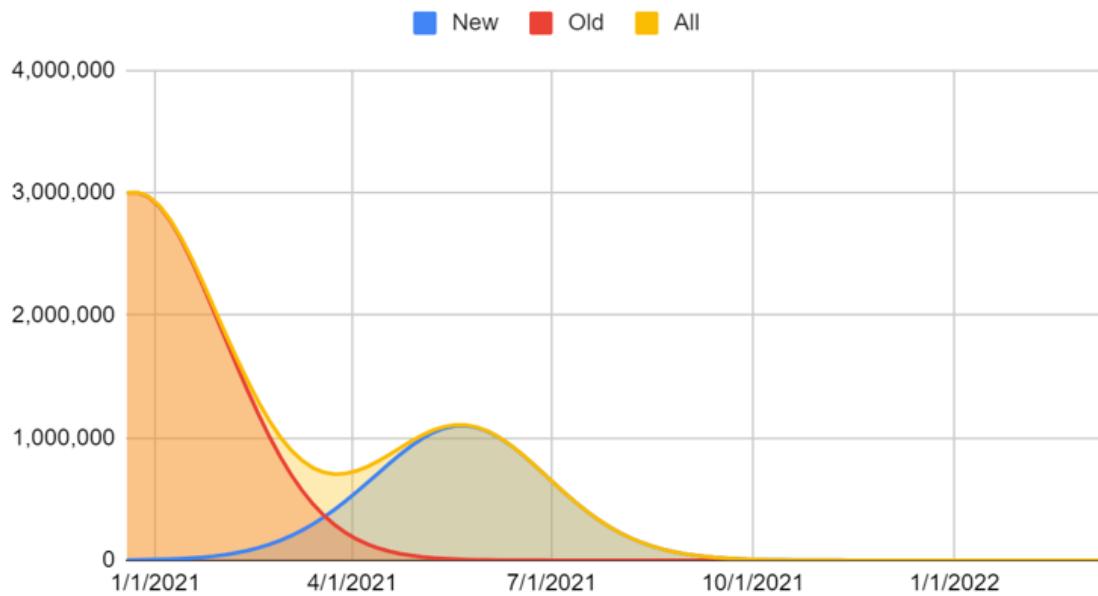
Let's see what that does.

Excellent news! It's a much better picture if we're looking at only 50% more infectiousness, despite being a rather pathetic pace of vaccinations.

For a 50% more infectious new virus, the worst remaining day is already behind us.

It takes over from the old strain at exactly the same pace (of course), but the peak is no longer terrifying, because there's enough immunity to turn the tide quickly.

Infections Per 5-Day Cycle By Strain and Date

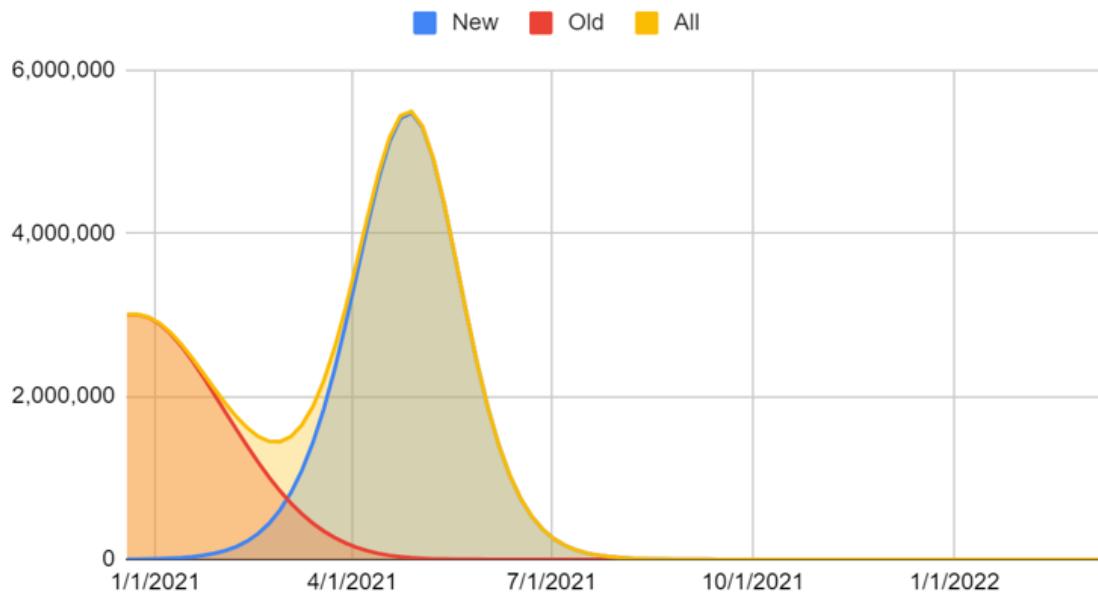


There is still a secondary peak later on, effectively prolonging our agony for those not yet vaccinated from May until September. That sucks, but it's not a crisis situation.

We end up with 36.7% infected, versus 63% infected without the vaccine. Big improvement.

However, if you change from 50% more infectious to 65% more, we get a very different picture.

Infections Per 5-Day Cycle By Strain and Date



Now we have a terrifying peak at the end of April.

We end up with 51.5% infected. Much better than 74%, and a peak of 5,000,000 per cycle we still see what looks like hospitals very overwhelmed in May and June.

This still is a more reassuring picture than I had going on. At 50% additional infectiousness rather than 65%, we have a very good sporting chance to have the vaccine arrive almost exactly on time, and keep the final wave in check.

This is strong motivation to push hard on the vaccine side of the race. If we can accelerate faster than this, it will help a lot. Conversely, much slower than this, and things get much worse.

Now we need to add control systems. This will potentially help in some places, but also hurt us in others.

Adding Control Systems

It is known that the pandemic has involved powerful control systems of various sorts.

As people see infection rates, hospitalizations and deaths go up, they take less risk. As they see such things go down, they take more risks. Governments do the same, opening and closing various businesses, schools and other locations and activities to keep things in balance. For our purposes here, these forces all look similar. The question is how they would react in these new circumstances.

My model has the following gears:

1. People are slowly getting what might be termed pandemic fatigue. Over time, they are less willing to put their lives on hold, and thus do riskier things, which has been enough to prevent us from winning via herd immunity so far. Magnitude of this is hard to say.
2. People react to some combination of infections, deaths, hospitalizations, and the reactions of governments and other authorities. It's not clear what is the central mix of these elements. Hospitalizations and deaths seem to play a large role, thus introducing lag. It's not clear what would happen if deaths and infections diverged, such as if nursing home residents were vaccinated and thus protected, but it could get weird.
3. People react to levels, not rates of change. Things rapidly getting worse or better doesn't much matter to them. Mainly they care about how dangerous things seem to be. This is a lot of the evidence for pandemic fatigue – if we are stabilized at relatively high levels now versus before, despite substantial immunity and better knowledge, there must be a factor pushing in the other direction.
4. Because people's observations lag at least a week behind (for infections) and as much as a month or more behind (for deaths), and people then take time to adjust especially for the official systems, the control system will be slow to react to changes. This is much of what causes waves, peaks and valleys.
5. In a full 'back to normal' scenario we'd be looking at R_0 of somewhere between 2.5 and 4.5 before any immunity effects. In the past I've used 4, but often seen smaller numbers. Which means that given R_0 is currently close to 1 with 20% infected (e.g. before immunity $R_0 \sim 1.25$) people really, really aren't taking this all that seriously in aggregate.
6. The alternative hypothesis to #1 and #5 is that the virus has already mutated at least once to be more infectious, maybe several times, and even the old strain has baseline R_0 higher than 4.

For now, we are assuming vaccinations are random, so they won't impact the IFR. While that remains true, we can approximate the combination of all these factors by an average of the

infection counts of the past several weeks, with infections lagging one cycle, deaths lagging four or five, and some reaction time, so let's say we take the average of the last six cycles.

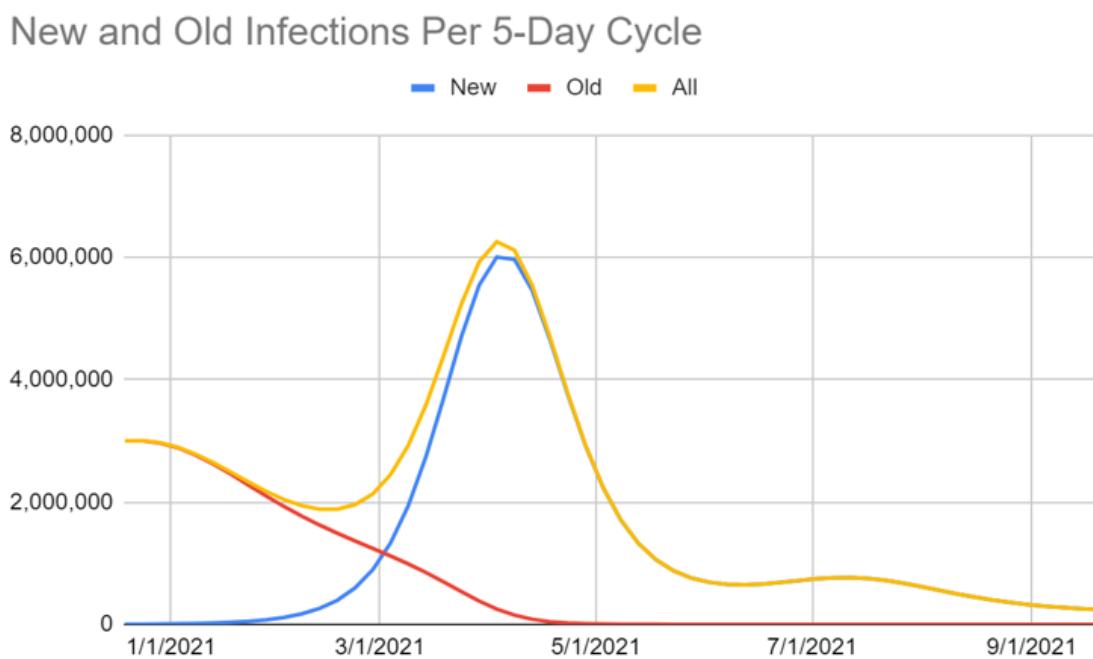
To start, let's also ignore phase shifts like overloading hospitals, and ignore fatigue on the hopes that vaccines coming soon will cancel it out, although there's an argument that in practice some people do the opposite.

As a first guess then, let's take that average of the previous six cycles, the ratio of that to current levels, then raise that to some exponent. As sanity checks, I ask myself what I think people would be capable of doing from here if things got how much worse, and when they'd mostly return to normal, and I noted that I'd be highly surprised if such forces were sufficient to cause an additional peak after the new strain started declining again, with vaccinations continuing and a lot of immunity.

This settled me on an exponent of 0.25, which is the limit of where a final peak does not emerge with 65% additional infectiousness. You can see it almost happening on the chart but it stalls out.

Now what happens?

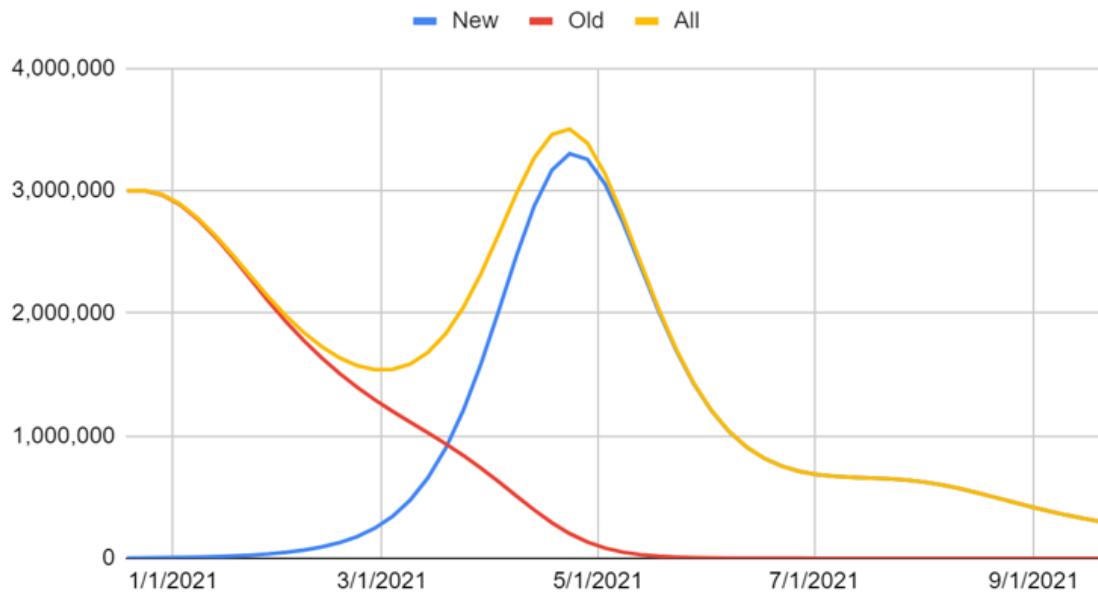
With 65% more infectiousness:



That's a higher peak than before. That makes sense, because lag is your enemy here. By the time people realize things are getting bad again, they've let the situation get further out of control, and it happens about a month faster, in early April instead of early May. We also see an extended die-out period afterwards, which seems realistic. We end with 54% infected, slightly higher than without the control system.

With 50% more infectiousness instead:

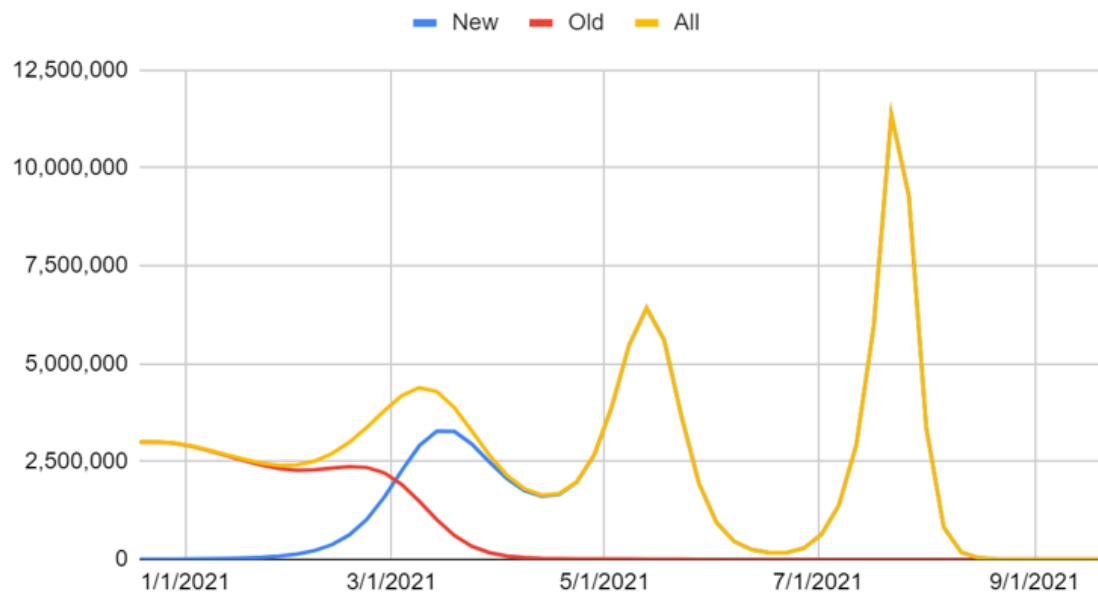
New and Old Infections Per 5-Day Cycle



Instead of the new peak being below the old peak, now it's slightly above it, although it does not last long which gives less time for disastrous scenarios to accumulate. Final infected percent is around 49%, again somewhat higher.

Weird things happen with overeager control systems. I don't expect this at all, but if you set the exponent fully to 1 and allow full pre-pandemic behavior to emerge, there's even a 'false dawn' scenario, where it looks like it's over, people go fully back to normal, and then it isn't over, and there's a final even bigger crisis, and *then it happens again even bigger*:

New and Old Infections Per 5-Day Cycle



That's because even with a lot of immunity, by assumption the new virus spreads *really, really fast* when everyone ignores it. Reducing to an old base R0 of 3 makes both later peaks stop at 1.2mm cases a day.

That doesn't mean I think such things are likely, but it is worth noting that we could be this stupid. I can't rule it out.

I also added a knob to make current R0 only move a percentage towards the target R0 to avoid dramatic shifts like the ones in that graph, if one wants to do that. It's not clear the knob does much extra work.

The big tricky thing is that the control system largely depends on hospitalizations and deaths, and those depend on who gets sick. If we vaccinate the vulnerable first, hospitalizations and deaths will begin to lag behind case counts. That's great in terms of patient outcomes, but has a dangerous side effect.

What does that do to the control system?

Adding Heterogeneity

There are two major heterogeneity effects I think require incorporation before the model will seem complete: Selective vaccination, and selective infection.

Selective infection, as I see it, is mostly about different levels of risk taking, and also different levels of vulnerability, which presumably is correlated with superspreading. If you take twice as much risk, you are twice as likely to be infected each day. Depending on the distribution of risk taking, this can mean either not much, or a hell of a lot. If the superspreaders slash super-risk-takers are super out of control, they can get taken out quickly, and a little immunity can go a *long* way.

This seems very plausible to me. If anything, I continue not to believe that immunity isn't doing a lot more than it is. Some people wear masks and hide, others disdain masks and go to crowded bars. Some get to work at home, some have to be essential frontline workers. This doesn't seem like it should be complicated.

Me and many of those I know, both family and friends, are doing effective prevention, well above 90% below pre-pandemic activities, and were generally doing less of the risky things to begin with. Yet if people overall took more than 75% less risk than they started with, we would not have a pandemic before the new strain arrives.

It seems likely to me this effect is roughly *fractal*. If we say that one third of the people take two thirds of the risk (which seems crazy low) then I'd suspect that one third of that top third takes 4/9ths of the total risk, and so on, in both directions. Then each person is likely already infected proportional to how much risk they take.

We then also need to consider that those taking more risk are probably less likely to accept a vaccination, although some policies put them at the front of the line. No idea how that works out.

It all gets complicated, but we can *probably* treat this for now as one knob, by saying the first 50% of those who get infected will have taken on X% of the combined risk, and treating vaccinations as still roughly random with respect to risk taking (see the other adjustment for death rate concerns) since we have forces pulling in both directions, at least for now. Previously we've set this X to 50, which is too high, and thus we have been underestimating immunity from infections (assuming that reinfection remains super rare).

Then we have heterogeneity from vaccination. One aspect of this as noted before is how much risk such people take. For now, I'm willing to not worry about that, or set a knob and

default it to no adjustment.

The more interesting question is what happens in terms of vaccinating the vulnerable. Right now, we are vaccinating on two fronts.

Nursing home residents, who are 1% of the population and 40%, are in the first wave of vaccinations everywhere. At some point, all areas add the elderly, but most are waiting.

Simultaneously, health care workers start the other angle of vaccination attack, followed by other essential workers, politically powerful or influential people, and those "most deserving" of allocation via politics and power. There are some old people in this group, but overall they are frontline and so tend to be younger than average, and less vulnerable than average, at least for those over 18.

Because risk grows so dramatically with age, doing some very old people early is more important than worrying about the other group being unusually young. Thus, we should expect to continue to lower the IFR further as more elderly get vaccinated.

On the flip side of that, IFR will go up if hospitals are overwhelmed. We don't see evidence of that now, so it seems like hospitals can mostly handle current levels, but we are seeing definite signs of strain. At a minimum, patients with other conditions are suffering. So the behavioral effects should come into play more rapidly anywhere above current levels.

So presumably, we should add a column for IFR baseline, and one for IFR effective given infection levels and who has been vaccinated at a given stage, and adjust people's reactions based partly on that, with a knob for how much weight they give hospitalizations/deaths versus infections.

To be conservative, I'll assume an IFR of 0.45% to start, to make us line up with recent death numbers.

How much can we cut deaths? Let's mostly focus on age:

Age Alone

To adjust from a baseline we need a baseline. For New York the relative risks look like this:

	%Of Deaths	Population	Relative Risk
0-4	0.02%	7%	0.2%
5-14	0.04%	14%	0.3%
15-29	0.3%	22%	2%
30-39	1%	16%	9%
40-49	4%	14%	26%
50-59	10%	9%	116%
60-69	20%	7%	287%
70-79	27%	5%	525%
80-89	25%	2%	1066%
90+	12%	1%? (cuts off at 85)	1500% or so?

Nursing homes have about a third of all Covid deaths, presumably almost all from people who are 70+, so we'll want to note that we're going there first, then assume we have two tracks, one for politically approved adults who have random risk, and one for old people in descending order.

Right now, the vast majority of doses in most places are going to the politically approved – there are many times more prioritized workers than there are residents of nursing homes. In the second phase, it plausibly reverses, as we hit a wall of who is plausibly better than the elderly. Things have been so crazy and random it's hard to know.

About 6% of the population are health care workers, depending on your definition. If you expand that to other 'essential' workers you can get more or less as many extra political choices as you want. We'll make a knob, but let's assume for now that the first stage does all health care workers and those 70+, combined 14% of the population. Nursing homes going very early means that the early doses should be at least that effective. Then we have an even split between random people and those 60-69 or so, who are 20% of deaths (and the majority of *remaining* deaths) while being 7% of the population.

So the first 14% of the population protects against roughly 67% (two thirds) of deaths, then the second 14% protects against another 20%, which is about two-thirds of the remainder once again. Hospitalizations should be somewhat less extreme, but still not that dissimilar. After that, it's likely that they open things up to everyone, but still exclude kids, so there's a quarter of the population left out that basically never dies, and we sample from the rest.

Then we have to decide how much such factors matter to people, versus observing their own risk or case numbers, when deciding what to do. And also we have to consider that if people you know are vaccinated, you might then not care as much because you can't infect them, and you sense less risk.

As a default let's do an even split. Half of consideration is deaths. The other half is case counts, unadjusted for immunity because people don't seem to make that adjustment.

Thus, at 14% vaccinated, *relative* death rates will be down by 62%, which reduces *perception* of risk in this model by 31%. At 28% vaccinated, relative death rates are down by 75%, reducing perception of risk by 37%. Then we don't improve further.

In the interests of simplicity, let's call that a decrease in risk perception of 33% and deaths by 66%, phased in linearly over the first 15% of the population vaccinated. Close enough.

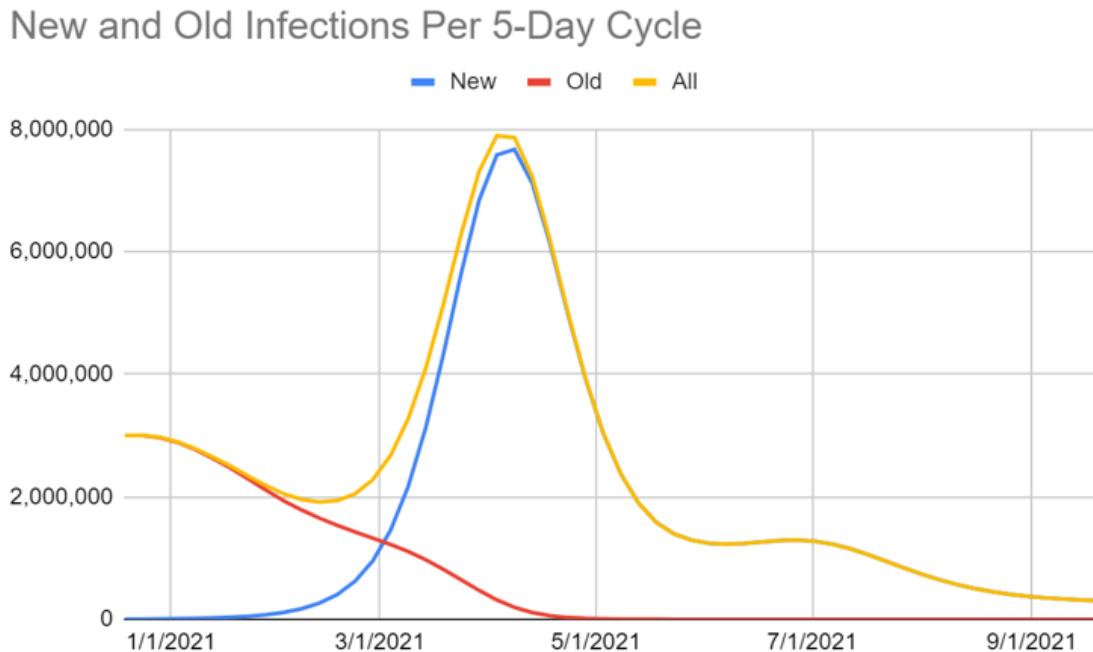
Of course, it takes time to kick in, so we'll give it 8 cycles (3 for the vaccine to work, 5 for people to actually die). Again, that's an approximation but should be fine, as the effect comes in continuously anyway.

We're not considering kids here at all, who are almost never hospitalized but also relatively rarely tested, I'm not sure how that adjustment works.

Before setting that up, the obvious prediction is that this will create much higher peaks in terms of number of infections under our control system assumptions, but still greatly reduce deaths.

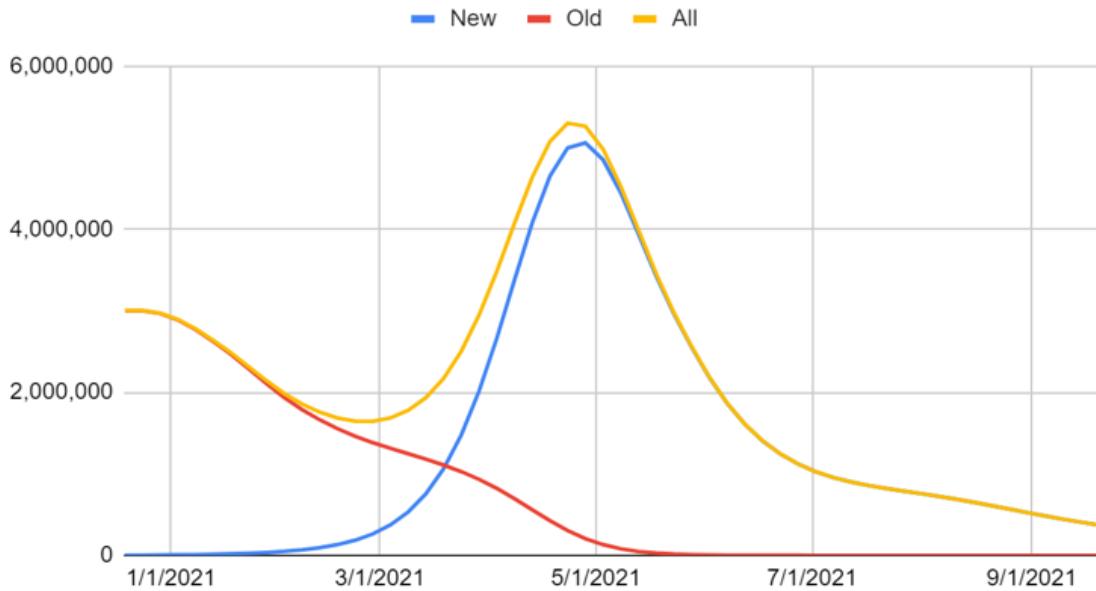
Here's what we get at max effect, when we have the control system only look at deaths, and with zero control system memory (so it doesn't look at how safe we were being last week):

65% more infectious:



50% more infectious:

New and Old Infections Per 5-Day Cycle



Those are substantially higher peaks than without the modification. 63% and 56% of people get infected respectively in these new scenarios by the end, again substantial boosts.

If the control system is 50% deaths/hospitalizations and 50% infections, you get an answer halfway between the two extreme scenarios, as you would expect.

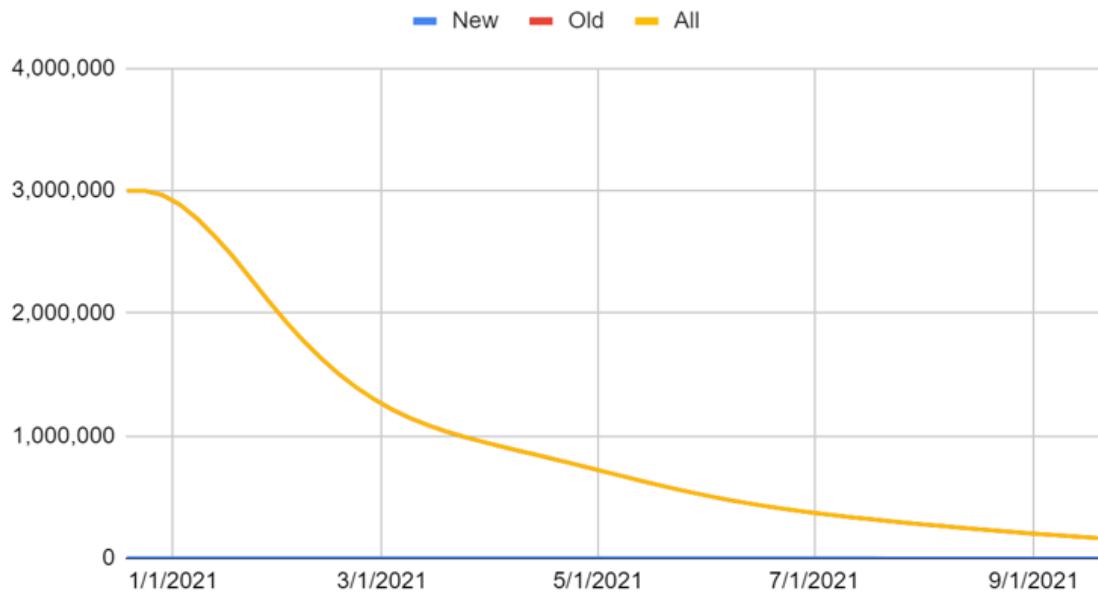
While substantial, those are much smaller bad effects than the positive effect on saving lives.

(Reminder: So far there have been about 350k dead, and the next three weeks are already baked in.)

There's a clear strong benefit to vaccinating old people early in terms of deaths. With the more infectious variant under plausible numbers, we end up with 663k dead, but with random vaccinations we would have had 844k dead.

Note that if we take the new strain out entirely, our model predicts a slow but steady improvement from here, with a combined 525k dead:

New and Old Infections Per 5-Day Cycle



I think I'll stop there, but give everyone a chance to look at the spreadsheet, spot errors, and make copies to try for yourself. [You can use this link.](#)

No doubt there remain errors of all kinds, both spreadsheet errors and conceptual mistakes, and things that weren't incorporated, including things I know about (e.g. I'm intentionally not using the fractal nature of risk taking, which is hard to estimate but could be a very large advantage).

This is not meant to be an answer. This is meant to get juices flowing, and at least ask some of the questions and throw out concrete possible futures. Let's work to improve it, and/or inspire other attempts.

Covid 1/14: To Launch a Thousand Shipments

Pardon me while I make my way to the rooftops.

So I'm sure it's not that simple especially because of regulatory issues, but... [did you hear the one where humanity could have produced enough mRNA vaccine for the entire world by early this year, and could still decide to do it by the end of this year, but decided we would rather save between four and twelve billion dollars?](#)

If not, there's a section on that.

Meanwhile, we also can't figure out how to put the vaccine doses we already have into people's arms in any reasonable fashion. New policies are helping with that, and we are seeing signs that things are accelerating, but wow is this a huge disaster.

We took some steps this week towards sane policy. Everyone over 65 is eligible in most places due to new CDC guidance. All doses we have will be distributed rather than reserved. Distribution will be based on the speed at which existing doses are put into arms. Mass vaccination centers are coming online. One can be hopeful for the path of future policy. At a minimum, it's a start.

In the near term, deaths are way up, the majority of which is almost certainly a real increase, and they probably won't peak for another week or two. Positive test rates are down a little, but that's explained by the thankfully rising test counts, so we have not yet meaningfully turned the corner after Christmas.

We likely won't have that much time to do so before things get worse again. The English strain is definitely coming. It looks like it won't be as bad as it looked like it might be, so we'll have some more time and the end result won't be as severe, and it looks like England has been able to use sufficiently extreme measures to contain it in the last few days, but it's definitely here, definitely spreading and definitely going to make things a lot worse.

Let's run the numbers.

The Numbers

Predictions

Prediction last week: 17.0% positive rate on 9.5 million tests, and an average of 2,800 deaths. The holidays are over, there will be some fallout, with things getting slightly worse, but with the main boost in deaths from Christmas mostly coming later.

Results: 15.2% positive rate on 11 million tests, and an average of 3,325 deaths.

This guess turned out to be backwards. We got a *lot* more rise in the death rate than I guessed. The positive test rate improved as we cleared past the holidays and got more testing online, but reported deaths increased much faster than I expected despite there not being enough time for people to have died from infections over Christmas, a lot of which was presumably our reporting getting back online and catching up.

I treat this as likely a single conceptual error, where I underestimated and misunderstood the warping effects of *reporting and measurement* from the holiday. We saw a real effect from Thanksgiving, and I expected that again, but this was larger, and I didn't properly adjust for

that. I think that's where most of the mistake lies, rather than a misunderstanding of the physical situation on the ground in terms of infections or deaths.

The new strain is having a small effect on infections, but not enough to notice on a graph or chart. That will come later.

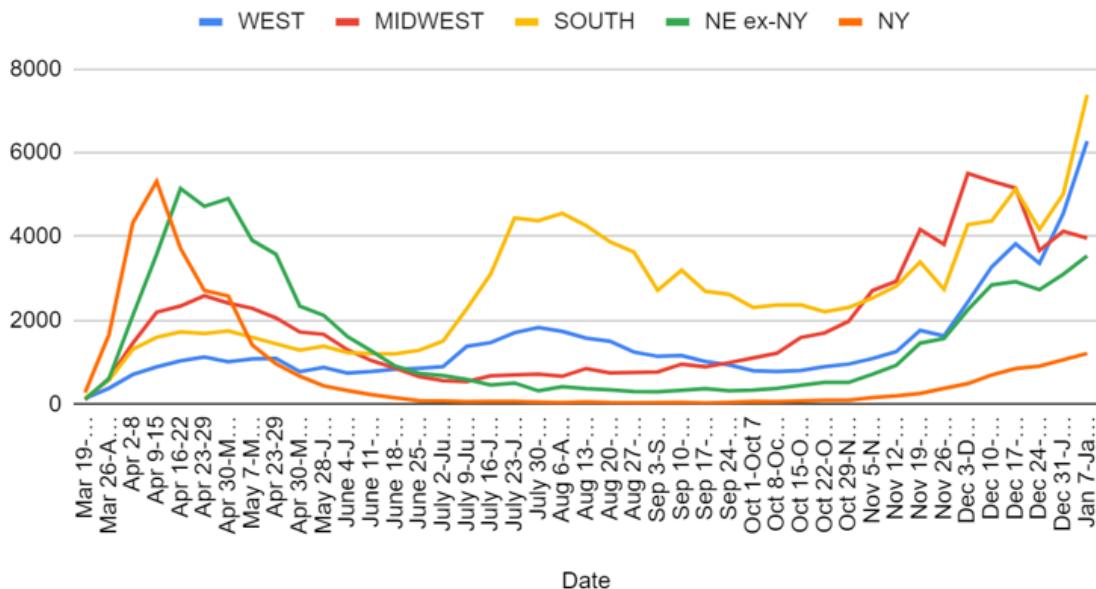
Prediction: 14.0% positive rate on 11.7 million tests, and an average of 3,650 deaths.

We're already almost three weeks past Christmas, so I don't expect the death rate to rise that much more, but I also didn't expect this much this week. My guess here attempts to split the difference.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 5-Nov 11	1089	2712	2535	870	7206
Nov 12-Nov 18	1255	2934	2818	1127	8134
Nov 19-Nov 25	1761	4169	3396	1714	11040
Nov 26-Dec 2	1628	3814	2742	1939	10123
Dec 3-Dec 9	2437	5508	4286	2744	14975
Dec 10-Dec 16	3278	5324	4376	3541	16519
Dec 17-Dec 23	3826	5158	5131	3772	17887
Dec 24-Dec 30	3363	3668	4171	3640	14842
Dec 31-Jan 6	4553	4127	5019	4162	17861
Jan 7-Jan 13	6280	3963	7383	4752	22378

Deaths by Region



That's a disaster, far worse than I expected. Christmas lags in reporting can explain some of it, but with rises this big, that is little comfort. Things are very, very bad out there. We do

know the Midwest has peaked, but the other regions are definitely not there yet, and this is far above previous peaks.

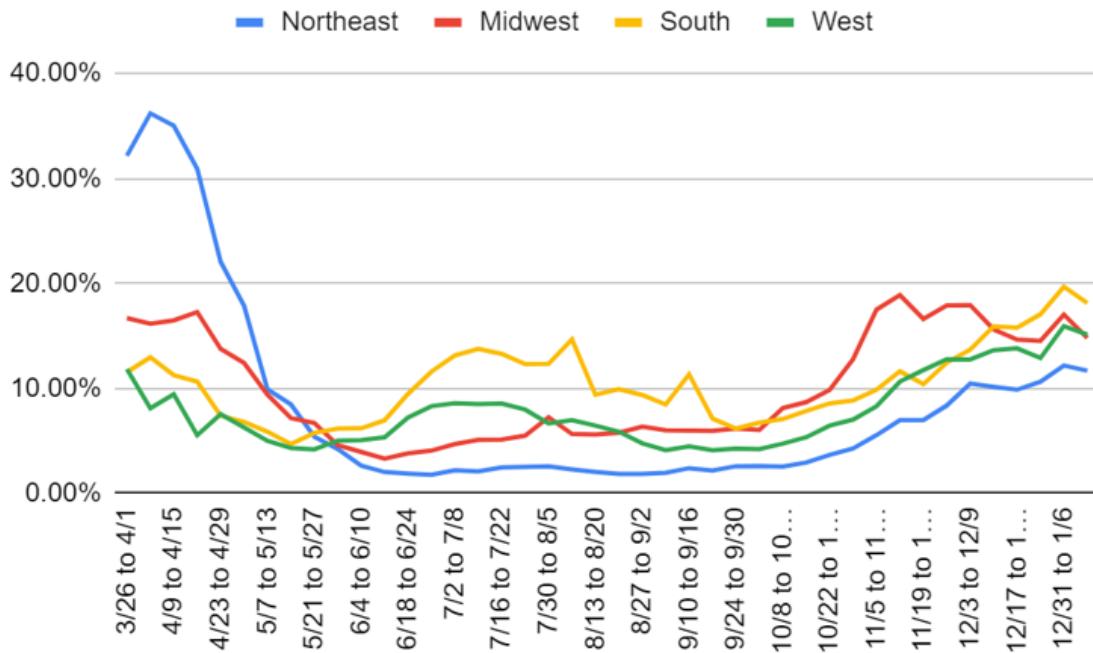
Two things I track on my spreadsheet but don't post are the 14-day and 21-day lagged rolling 7-day average CFRs. Those numbers had been dropping steadily over time, rose a bit and peaked on 12/22, then dropped once again. They are now back at their 12/22 peak, presumably because of cases we missed two or three weeks ago due to lack of testing.

There's at least some worry that the true IFR has also risen more than I realized, due to hospitals being overwhelmed in many areas.

I still think this wave of deaths will peak within a few weeks, and will then start declining until the new English strain shows up in force, as we feel the full Christmas effect in terms of both deaths and reporting, but I'm far from certain, and it's not likely to be that steep a slope down even before the next crisis starts.

Positive Test Percentages

	Northeast	Midwest	South	West
11/12 to 11/18	6.99%	18.90%	11.64%	10.66%
11/19 to 11/25	7.00%	16.62%	10.41%	11.75%
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%
12/3 to 12/9	10.47%	17.94%	13.70%	12.76%
12/10 to 12/16	10.15%	15.63%	15.91%	13.65%
12/17 to 12/23	9.88%	14.65%	15.78%	13.82%
12/24 to 12/30	10.65%	14.54%	17.07%	12.90%
12/31 to 1/6	12.18%	17.03%	19.69%	15.94%
1/7 to 1/13	11.70%	14.81%	18.14%	15.12%

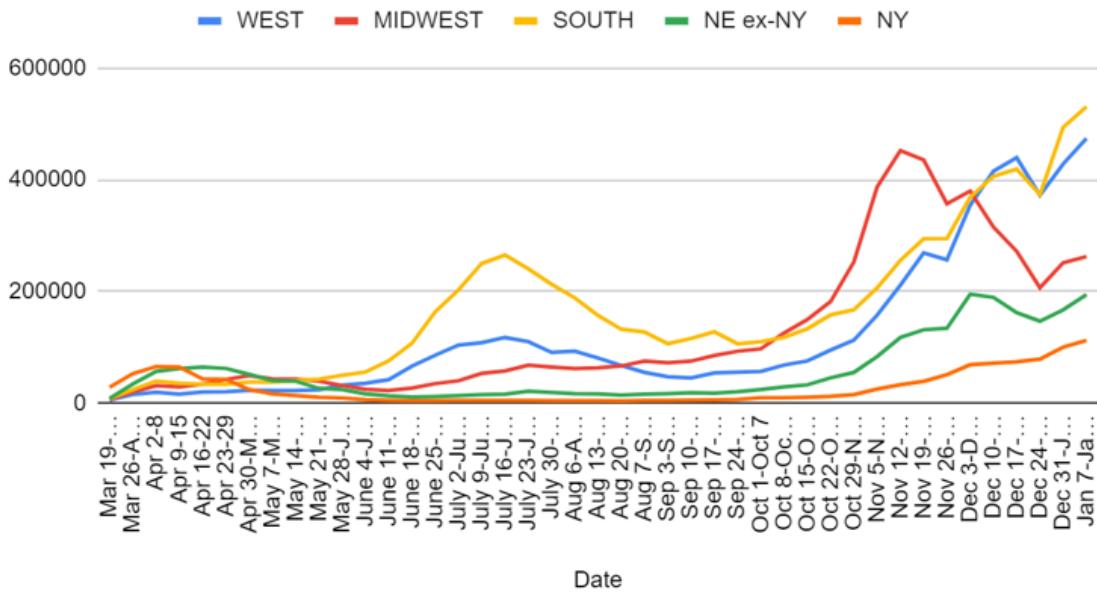


Encouraging numbers, but likely mostly due to increased testing.

Positive Tests

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Nov 26-Dec 2	256,629	357,102	294,734	185,087
Dec 3-Dec 9	354,397	379,823	368,596	263,886
Dec 10-Dec 16	415,220	315,304	406,353	260,863
Dec 17-Dec 23	439,493	271,825	419,230	236,264
Dec 24-Dec 30	372,095	206,671	373,086	225,476
Dec 31-Jan 6	428,407	251,443	494,090	267,350
Jan 7-Jan 13	474,002	262,520	531,046	306,604

Positive Tests by Region



These increases are due to more testing rather than higher test percentages, so no reason to be alarmed, but also no reason to feel especially good about the situation either.

Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Nov 5-Nov 11	8,290,417	10.8%	1,059,559	2.4%	3.16%
Nov 12-Nov 18	9,040,426	12.4%	1,155,670	2.9%	3.50%
Nov 19-Nov 25	10,419,059	11.8%	1,373,751	2.9%	3.88%
Nov 26-Dec 2	9,747,026	11.8%	1,287,010	4.0%	4.23%
Dec 3-Dec 9	10,465,254	13.9%	1,411,142	4.9%	4.67%
Dec 10-Dec 16	10,701,134	13.9%	1,444,725	4.9%	5.12%
Dec 17-Dec 23	10,716,189	13.7%	1,440,770	5.1%	5.57%

Dec 24-Dec 30	9,082,257	13.9%	1,303,286	6.0%	5.95%
Dec 31-Jan 6	9,333,470	16.4%	1,365,473	7.3%	6.42%
Jan 7-Jan 13	11,054,685	15.2%	1,697,034	6.6%	6.93%

Testing is back, and we've finally broken 11 million tests in one week. Hopefully this represents a return to growing test counts, and we'll see continued increases. It does mean that the drop to 15.2% positive tests is not as encouraging as it otherwise looks, as the rise in tests mostly explains it. We can't be confident that things are improving much yet. It does mean we don't have to be worried about the rise in positive test counts we saw this week.

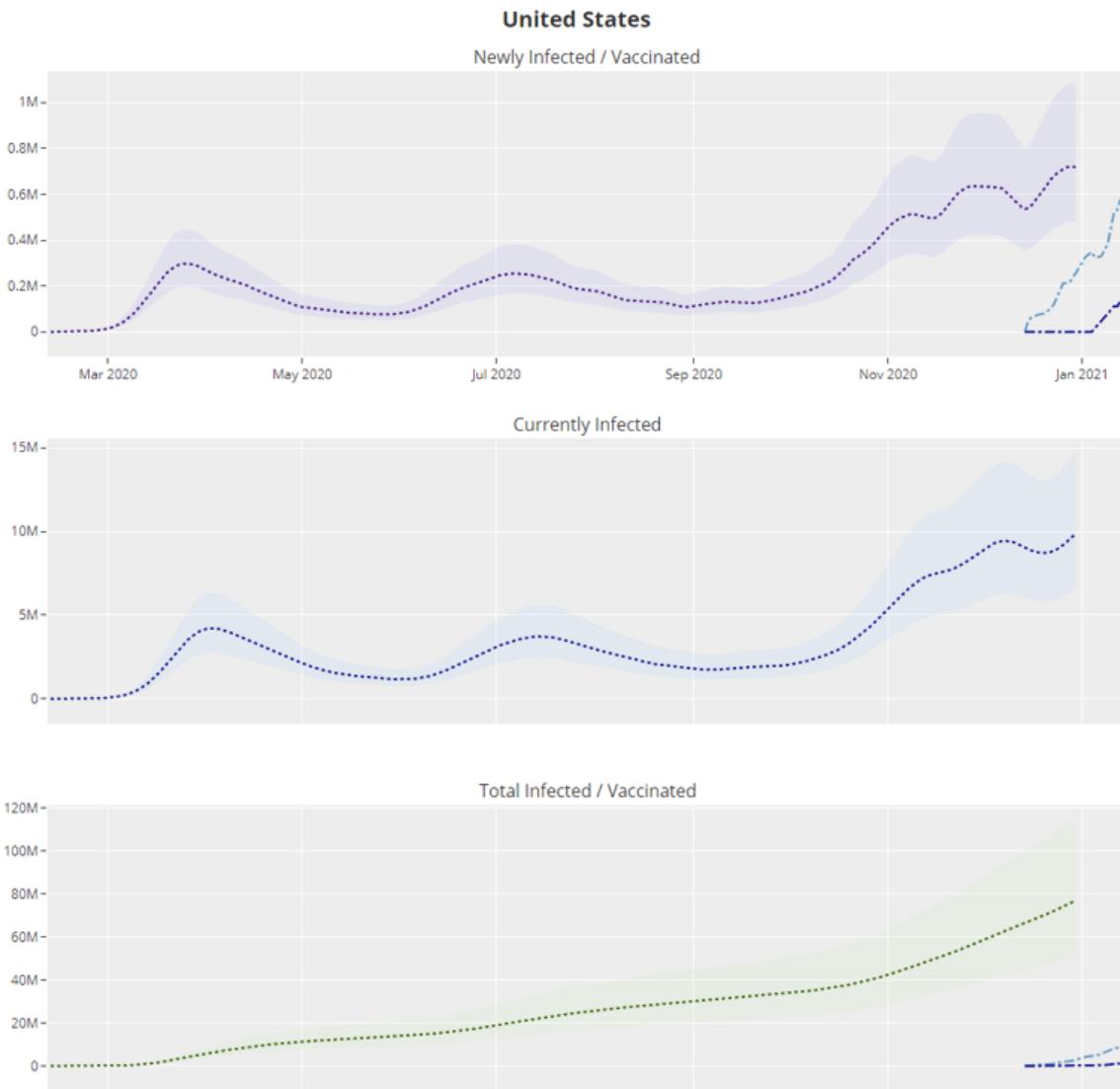
Covid Machine Learning Projections

Now with vaccinations! That's what the extra two lines are on the top graph, with the light blue being first dose and the dark blue being second dose.

Newly Vaccinated (as of Jan 13): **577,000 / day** (214 / 100k)
 Total Vaccinated (as of Jan 13): **2.7%** (1 in 36 | 9.1 million)

Newly Infected (as of Dec 30): **718,000 / day** (220 / 100k)
 Currently Infected (as of Dec 30): **1 in 30** (3.0% | 9.9 million)
 Total Infected (as of Dec 30): **23.2%** (1 in 4 | 76.9 million)

Rt (as of Dec 30): **1.05**
 Adjusted Positivity Rate (as of Jan 13): **11.8%**
 Infection-to-Case Ratio (as of Jan 13): **3.0** (34% detection rate)



The projections here seem to not think much of Christmas, which is odd. They had total infections as of December 23 at 21.7% last week, which the model still believes, and it thinks as of December 30 the number is now 23.2% infected. As usual, I consider these lower bounds, and that means herd immunity is making rapid progress on two (alas, mostly uncorrelated) fronts.

For vaccinations, the model is smoothing out weekends, which it probably *shouldn't* do, as the day-of-the-week cycle of activity is very much real, but it does make the graph easier to

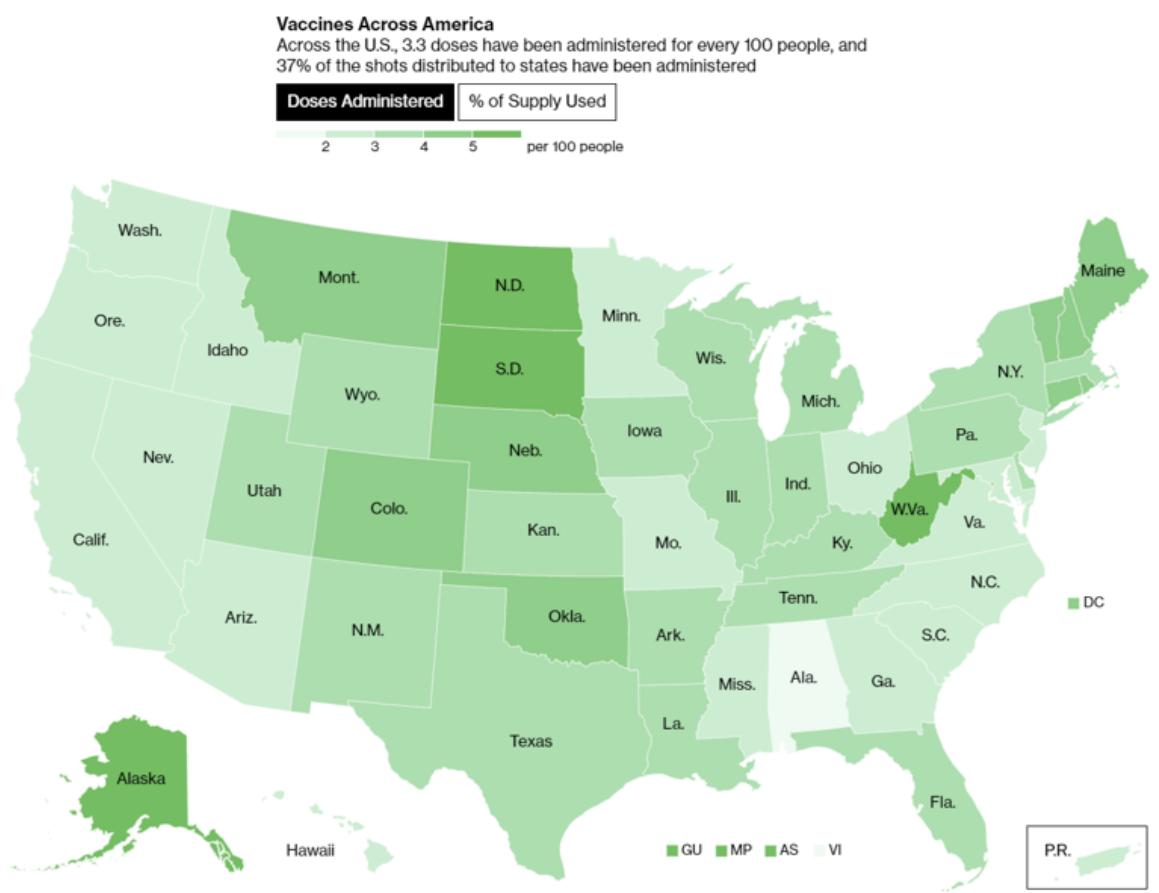
read.

The point where vaccinations exceed new infections looks like it will arrive sometime near the end of January. That doesn't mean we're 'winning' in any real sense, but it is at least something.

Vaccinations

[This twitter thread](#) can be useful, because it provides the daily delta in number of doses administered in an easy to view format.

To avoid getting too depressed by the overall look, remember that the colors indicate *relative* vaccinations, because the color scheme adjusts each week.

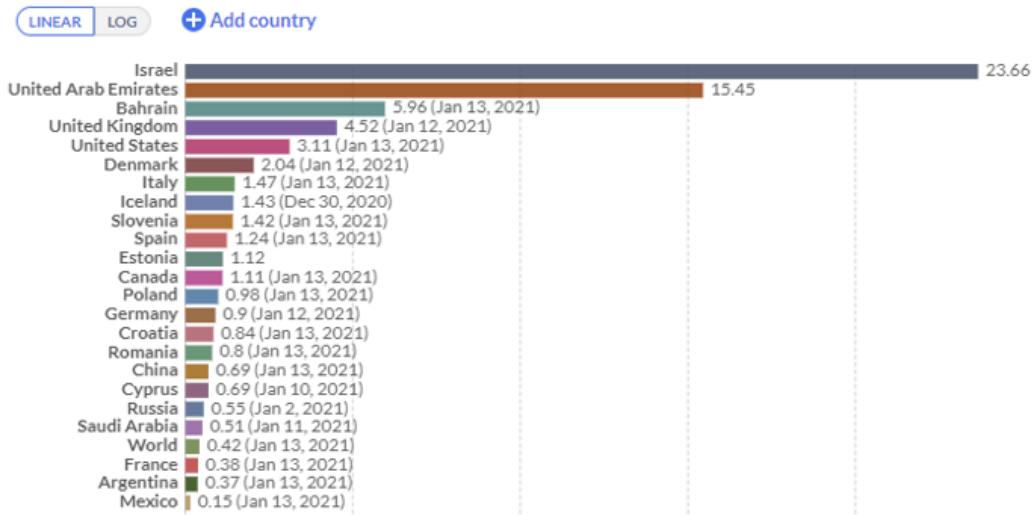


That doesn't mean this isn't a horrible, no good, very bad performance. Only 37% of all distributed doses have been given, with some of them ending up in the trash by choice. We need to do much better.

The good news is that things *are* accelerating, and so far decisions made by the incoming administration are highly encouraging and should speed things along further. As I've noted before, doing things fast early on is crucial since gains compound, but the real game is ensuring that as we scale up supply we are ready for that, and it all gets shot into arms, and we don't let supply go to waste.

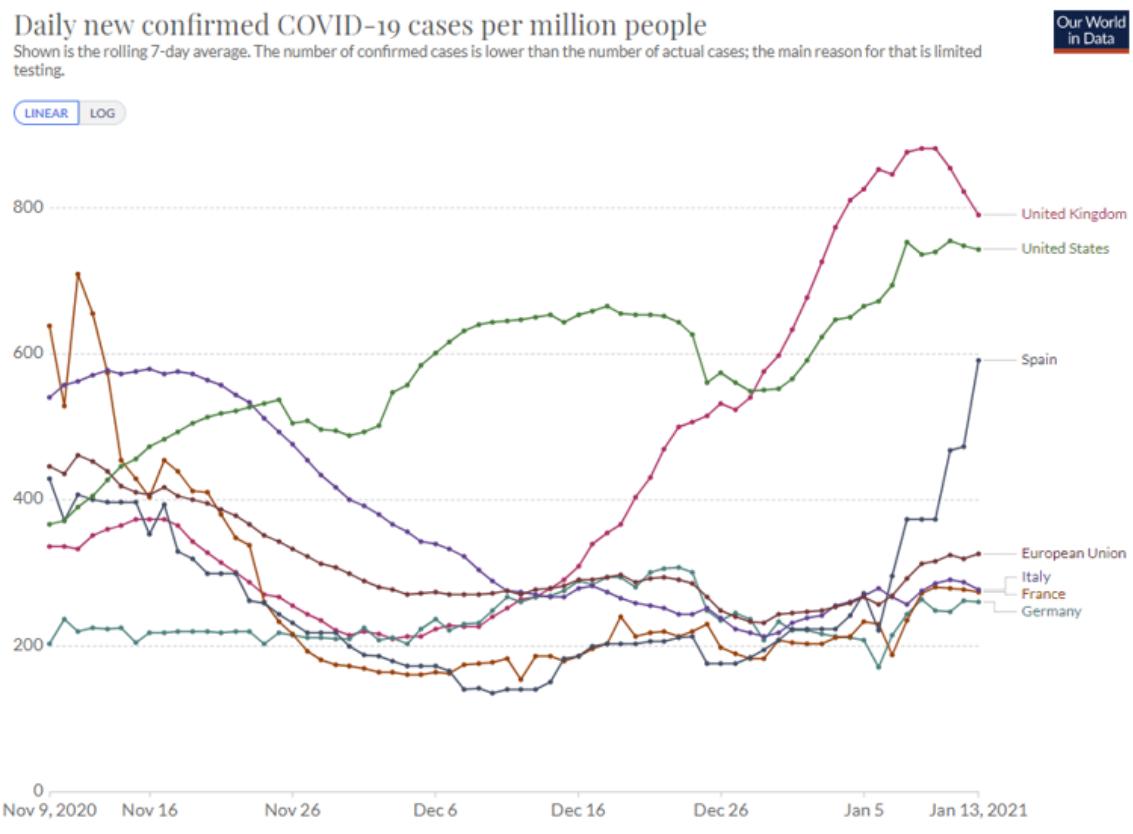
COVID-19 vaccination doses administered per 100 people, Jan 14, 2021
 Total number of vaccination doses administered per 100 people in the total population. This is counted as a single dose, and may not equal the total number of people vaccinated, depending on the specific dose regime (e.g. people receive multiple doses).

Our World
in Data



Europe

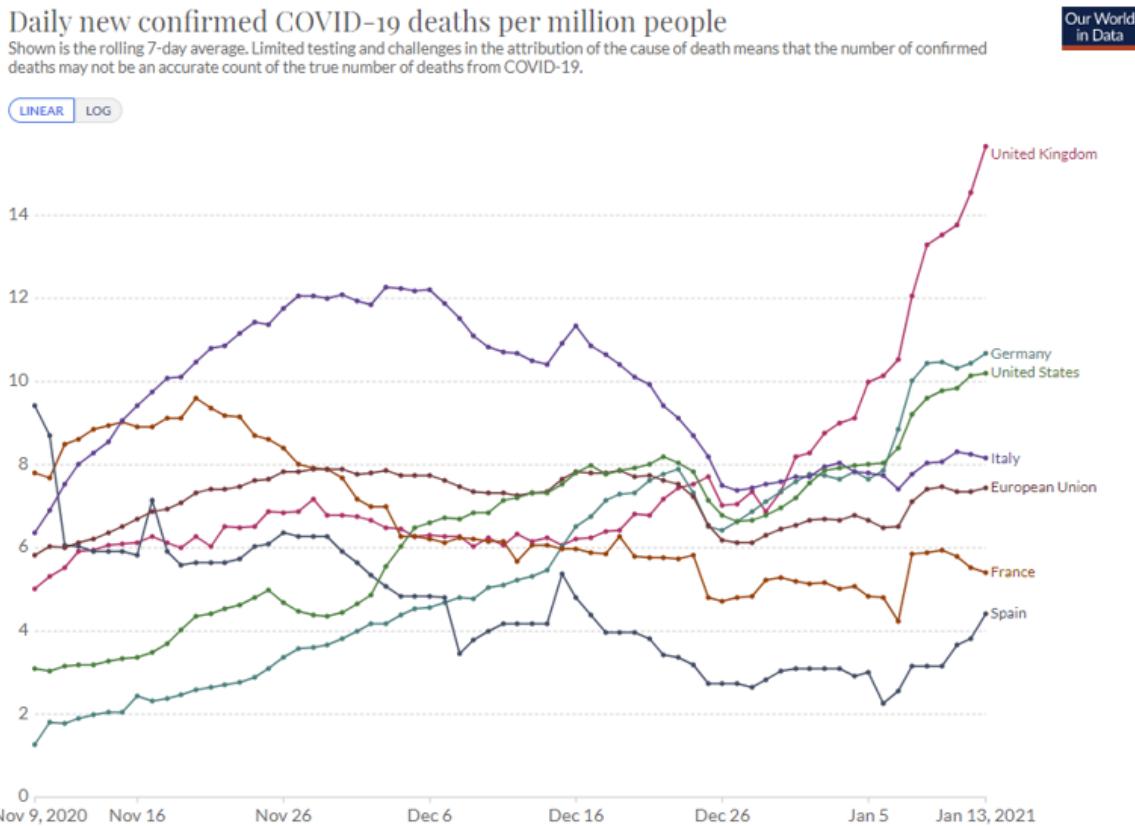
In the interest of consistency I'm not going to alter who is on the graph, but see the next section for a discussion of Ireland.



Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 14 January, 13:02 (London time)

CC BY

From what I see it looks like the Spanish hockey stick here is not due to increased testing. I also don't know any sequencing data from Spain, so we don't know if the new strain is involved here, but this rise is too fast for that to even be a full explanation, either there or in Ireland.



The most interesting part of all these graphs are those last few days from the United Kingdom. Positive test counts are declining again, as is the positive test percentage.

Did the control system manage to pull it off once again?

Even if that did happen, things got *quite bad* before that happened. The death rate in the UK is now 50% higher per capita than it is in the United States, and they have several weeks to go before that is likely to stabilize, so it's likely going to go at least 25% higher than that.

Despite that, this seems like it happened, and it happened fast, and it shows that the control system does have enough ammo left to stabilize matters if those involved care enough to do so. I'm highly skeptical the United States could match this performance, even if the United Kingdom does sustain things from here, especially before things get much worse than they are right now in America.

Still, one must acknowledge that this happened, and without much vaccine help, and we should update our expectations accordingly.

Also worth noting is that despite their massive lead in vaccinations, Israel's short term situation in terms of cases is quite bad. They test a lot, so their positive rate is still only a little over 6%, and their death rates remain well behind ours, but it's clear that the vaccines have not let them turn the corner on infections yet despite being over 20% complete. I do expect them to turn the death rate corner soon, since so many of their vulnerable are now protected, but it seems infections are continuing to rise.

This seems like more evidence that people have the perverse response of *increasing* risk taken when they are about to get the vaccine, due to some sort of risk budget or perception of risk, and a sense that ‘it’s over,’ rather than seeing it as the time to be extra careful. That’s not a good sign.

The English Strain: Are We F***ed? Is it Over?

We’re fucked. It’s over.

That still leaves plenty of room for our decisions to matter, and determine how bad things get before they get better. The strain is not growing as rapidly as we feared. This is going to suck, but my attempts to model things have me more optimistic that we can vaccinate enough people to muddle through without things getting *too* bad.

But the central assumption seems clear. The strain is here, and it’s going to take over by March.

Ireland is being overwhelmed by the new strain, [and it’s about as bad a graph as we’ve seen \(link to Reuters\)](#):



Kai Kupferschmidt ✅ @kakape · Jan 11

According to @Reuters, new variant #B117 now accounts for 45% of #sarscov2 sequences in Ireland.

I'd like to see an official statement to be sure this comes from a random sample.

But if true, percentage of #B117 in last 4 weeks has been:

9%

13%

25%

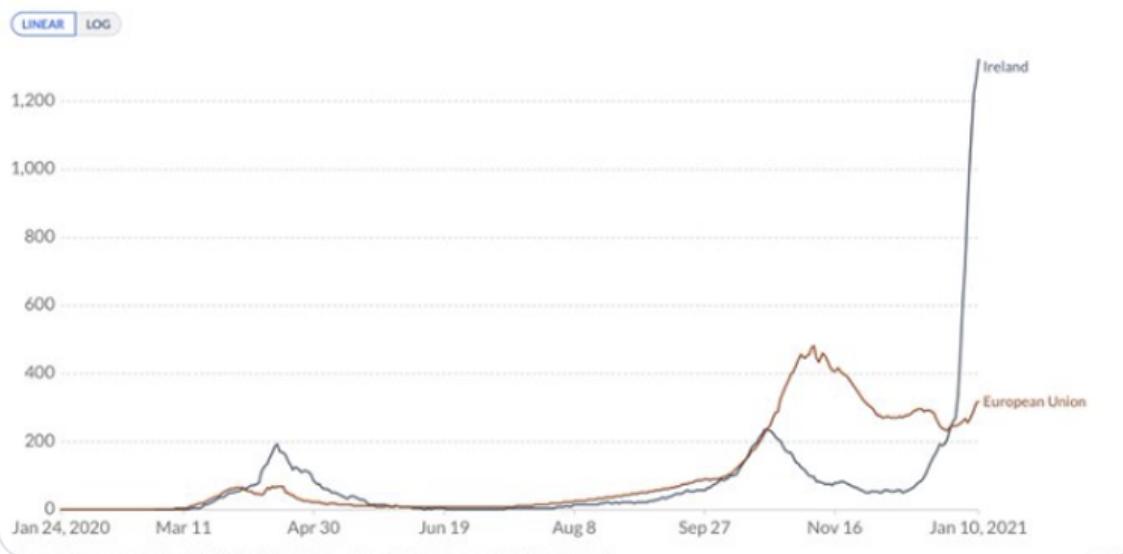
45%

reuters.com/article/us-hea...

Daily new confirmed COVID-19 cases per million people

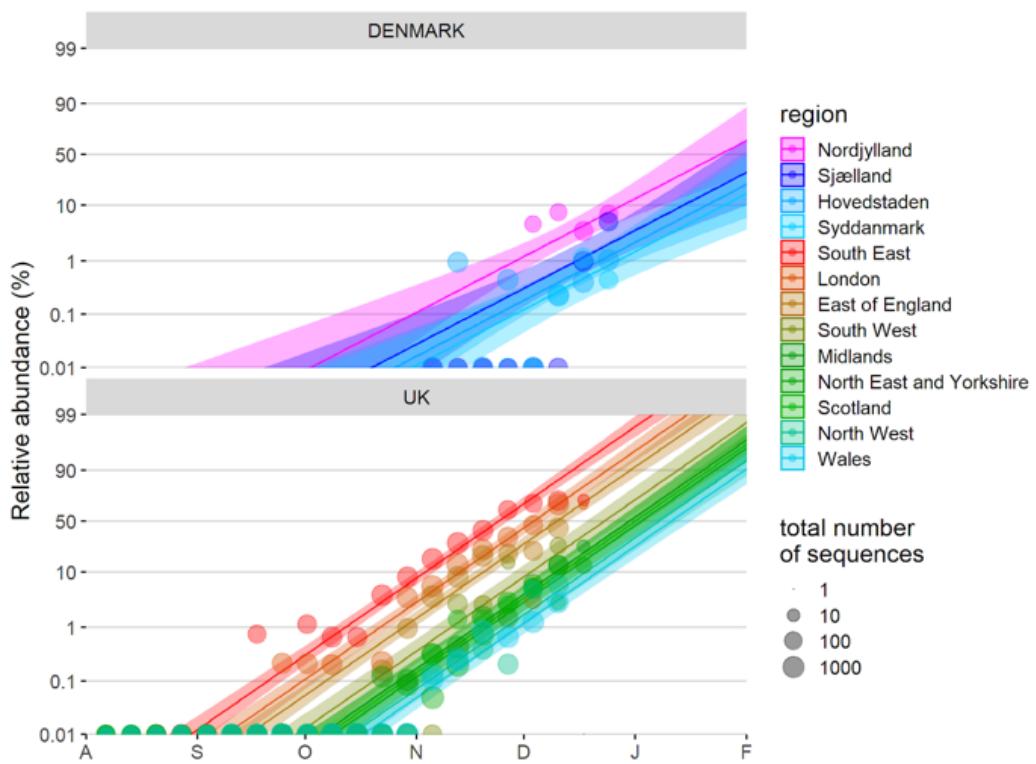
Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World
In Data



[Here's the data from Denmark](#), suggesting 59% additional infectiousness.

GROWTH OF VOC STRAIN 20B.501Y.V1 IN DENMARK AND THE UK



@TWenseleers
data COG-UK

[Here's another data source](#) that gives these results:



Caitlin Rivers, PhD Retweeted



Kai Kupferschmidt @kakape

5h

New data shows #B117 still
expanding in Denmark:

Last weeks of 2020:

0.4%

0.8%

2.0%

2.4%

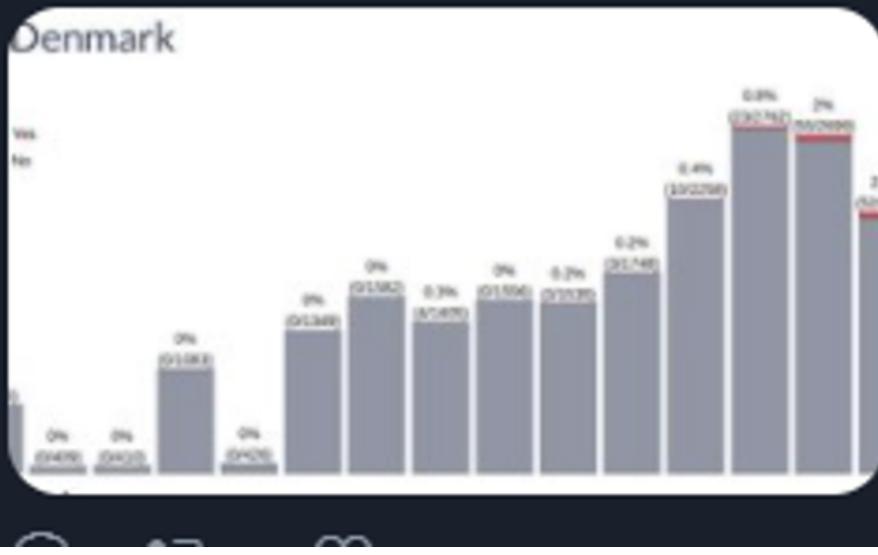
First week of 2021 (initial data):

3.6%

(Full report:

covid19genomics.dk/statistics, h/t

@MadsAlbertsen85)



The good news is that this is noticeably less than one doubling per week.

There seems little remaining doubt over the infectiousness or growing presence of the new strain. The only question is what we do about it, and how bad things are going to get.

So, how far along is this? [Here's a guess by Helix posted on January 11](#):

What's new?

- Helix has identified 74 cases of B.1.1.7 (up from the 51 reported in our January 6 update).
- B.1.1.7 appears to currently represent 0.27% of all positive tests and 43% of SGTF samples, up from 0.17% and 34% last week, respectively.
- B.1.1.7 has been found in 3 additional states (TX, MN, IN), adding to the 4 previously reported (CA, FL, PA, GA)
- B.1.1.7 has been found in 4 of the 5 states where we have statistical power to detect it (CA, FL, PA, IN), as well as 3 additional states where we are underpowered (GA, TX, MN).
- We've added a breakdown of statistics for the three states in which we have sequenced the highest numbers of samples
- Despite having the power to detect it, we have yet to identify B.1.1.7 in MA
- 43% of total SGTF sequenced samples are identified as the B.1 variant, which was initially associated with the spread of COVID-19 in Italy during the early months of 2020 but has not been found to be more contagious
- Presence of the B.1 variant varies significantly by state, however, representing 80% of SGTF samples in MA, 56% of samples in FL, and 2% of samples in CA.

This is not a complete listing of sequencing tests, only listing the places they have the statistical power to look. New York last week identified four cases, three of which are linked and one that was distinct from the first three.

Also note the sample sizes here for SGTF-positive samples are not large until recently.

Longitudinal trends in SGTF and B.1.1.7 data

National data ▼

	Dec 13 - 19	Dec 20 - 26	Dec 27 - Jan 2
Percent of positives that are SGTF	0.5%	0.5%	0.6%
Percent of SGTF sequenced samples that are B.1.1.7	10.0%	34.2%	43.2%
Percent of positive tests that are B.1.1.7 (extrapolated)	0.1%	0.2%	0.3%

Clade	Dec 13 - 19	Dec 20 - 26	Dec 27 - Jan 2
B.1	7 (70%)	14 (36.8%)	60 (43.2%)
B.1.1.7	1 (10%)	13 (34.2%)	60 (43.2%)
B.1.2	0	6 (15.8%)	0
B.1.262	0	3 (7.9%)	13 (9.4%)
B.1.315	1 (10%)	0	1 (0.7%)
Other	1 (10%)	2 (5.3%)	5 (3.6%)

The data for that final week seems reasonably robust, so I'm down with saying that tests completed that week are 0.3% positive for the new strain. The previous weeks seem to have much less robust sizes, and the Dec 13-19 one is literally 1 out of 10, so extrapolation from that would be unwise.

Taken at face value, thus, this is all good news. We only have 50% week over week growth in cases detected, rather than a doubling each week. Even boosting that to factor in the first week of data, that's a much slower acceleration than we see in Ireland. If we do have almost twice as much time and also the endgame isn't as bad as feared, then that's definitely a 'we can muddle through this' scenario.

In terms of where we are along the curve, 0.3% on tests done at the end of the year (which are presumably infections from a week before that) is modestly ahead of my previous estimates when I created my toy model, corresponding to a 'starting strain size multiplier' of about 4, and putting us about two weeks further along.

The South African Strain: Are We Even More Fucked? Is It Even More Over?

As far as I can tell, we continue to await the experimental data that will tell us for sure.

For now, what we do know is that [the mutation common to both new strains does not interfere with the mRNA vaccines](#).

What we still do not know is whether the other changes in the South African variant do interfere with the mRNA vaccines, or to what extent this might compromise vaccine effectiveness. We can find this out easily if we use post-vaccination serum and see if it works against the new variant, but so far no one has done this, so we don't know.

My father's current best guess is that the vaccines will work for some people, but not others: "E484K is a radical mutation. E (glutamic acid) is negatively charged at physiological pH and K (lysine) is positively charged. Antibody to the epitope that includes 484 is not likely to work against the variant, but there are other epitopes that are targets for neutralizing antibody. That's why convalescent serum from some people, but not others, neutralize the variant. So my guess is that the vaccine will protect some people, but not others."

I also have not seen any reports that this strain is being detected much in other places. That can be little comfort in an exponential growth situation, but at a minimum this problem is not as imminent as the English strain. Hopefully soon we will know more about the situation.

The “Columbus” Strain

When I posted about the English strain, the standard counter-argument was that mutations happen all the time, and new variations end up dominating all the time, and mostly it doesn't mean much. With that in mind, [we now have a third strain emerging, this one seemingly homegrown, that carries the 501Y mutation](#) that is the presumed primary cause of the increased infectiousness.

It's going about how you'd expect if that were the case: "This strain quickly became the dominant [coronavirus](#) variant in Columbus, Ohio, over a three-week period from late December 2020 to early January, according to the researchers, who hope to post their findings soon on the pre-print database bioRxiv."

There are no signs this version is functionally different from the English strain in a way we'd be worried about, so presumably all this does is speed up the timeline. The mutation is already dominant in a mid-sized city.

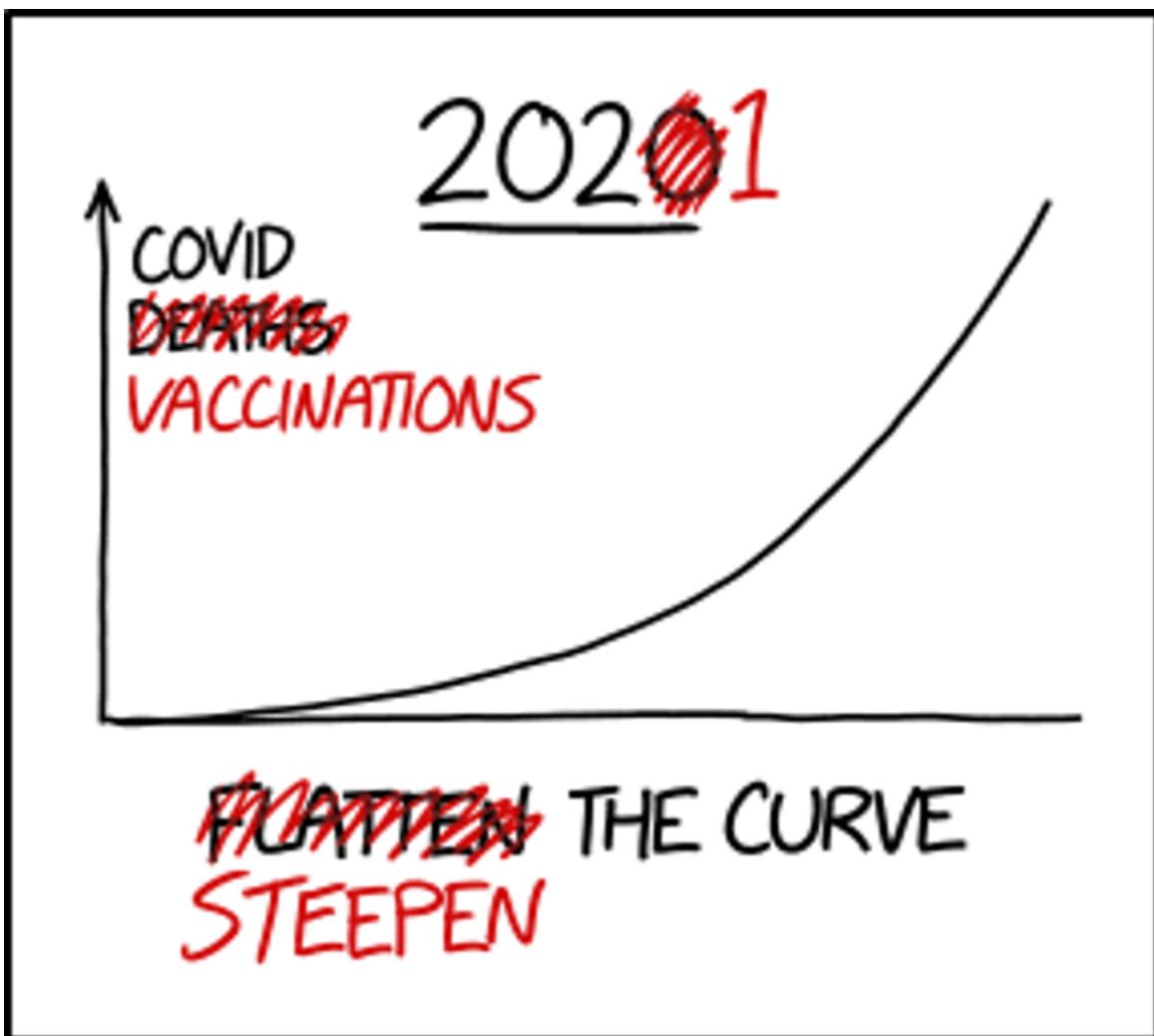
What jumps out is that this is now *three* new versions with this mutation, emerging now.

Twice can be coincidence. [Three times is enemy action](#). So, presumably, either there's a common origin for these strains in some way that seems implausible, or this mutation is suddenly very likely to emerge.

I don't know yet what to make of this development. The evolution is sufficiently fast and the timing sufficiently wrong that the vaccines can't have anything to do with it, especially since it looks like they still work. It could be some combination of many more cases plus luck, which is *kind of* a coincidence, yet I want to be suspicious of such explanations.

But What Do We Do Now?

[We should do this:](#)



[And think like this:](#)



Leah Libresco Sargent

@LeahLibresco

...

Replying to @LeahLibresco

The more vulnerable the population, the higher proportion of fraud you may need to be willing to tolerate to make sure the people least able to navigate your roadblocks get through.

9:43 AM · Jan 11, 2021 · Twitter for iPhone

As opposed to, [well, deciding to instead do the opposite:](#)



Kamala Harris @KamalaHarris · 15h

...

The first 100 days of the Biden-Harris administration will focus on getting control of this pandemic—ensuring vaccines are distributed equitably and free for all.

5.2K

12.5K

149K



Two Dose, One Dose, Who Knows, You Knows

One piece of great news this past week was the Biden administration [announcing they were shifting to at least the weak form of First Doses First, and it's happening](#). We will release all our vaccine doses now, and count on future supply to provide any second doses. That's not a full 'delay second dose to get more first doses' policy, but at least it is a 'a first dose now is better than delaying everything by weeks to be certain the second shot will always be available if there are supply disruptions' policy.

That both has a big direct impact, and also gives hope that we might see additional helpful policies from the Biden administration.

[As Tyler Cowen points out](#), the most striking thing is that once the new policy was announced, everyone 'fell in line' with it and neither of us has seen a single objection, because now the new policy is the default.

He also points out that there was never an expected value case that could be made in defense of the policy of holding back doses. The simple argument for First Doses First is not only compelling, it is overwhelming. If you can make two people 80% protected, or make one person 95% protected, and you want to turn the tide of a pandemic and protect people, that's that.

[Last week I was linked to the only plausible objection one could raise](#), which is a claim that the 80% number is not accurate, so that should be addressed:



Alyssa Vance

A response I got from the relevant FDA official:

"Dear Ms. Vance,

Pfizer and Moderna submitted EUA requests with data supporting two dose regimens for their COVID-19 vaccines. The data in the firms' submissions regarding a single dose are commonly being misinterpreted. Over 90% of participants in both of the trials received two doses of the vaccine at either a three- or four-week interval, respectively.

For example, the 94.5% efficacy as the primary endpoint for the Moderna vaccine is based on all COVID-19 cases diagnosed 14 days or more following the second vaccine dose compared with placebo (this is when the protocol specified evaluation of the primary endpoint). The 80% "single dose" efficacy is based on COVID-19 cases diagnosed any time after the first vaccination compared with placebo and includes the 92% of individuals who received the second vaccine dose. Thus, the 80% figure represents protection against COVID-19 during a significant interval of time when the overwhelming majority of individuals in the trial had already received their second vaccination.

It is therefore erroneous to conclude anything definitive about the depth or duration of protection after a single dose of vaccine from these single dose percentages reported by the companies. Though it is quite a reasonable question to study a single dose regimen in future clinical trials, we simply don't currently have these data.

Indeed, using a single dose regimen without understanding the nature of the depth and duration of protection that it provides is concerning, as there is some indication that the depth of the immune response is associated with the duration of protection provided. If people do not truly know how protective a vaccine is, there is the potential for harm because they may assume that they are fully protected when they are not and alter their behavior to take unnecessary risks.

FDA will continue to work with vaccine manufacturers to advance development and production to help the greatest number of people possible.

Peter Marks, M.D., Ph.D.

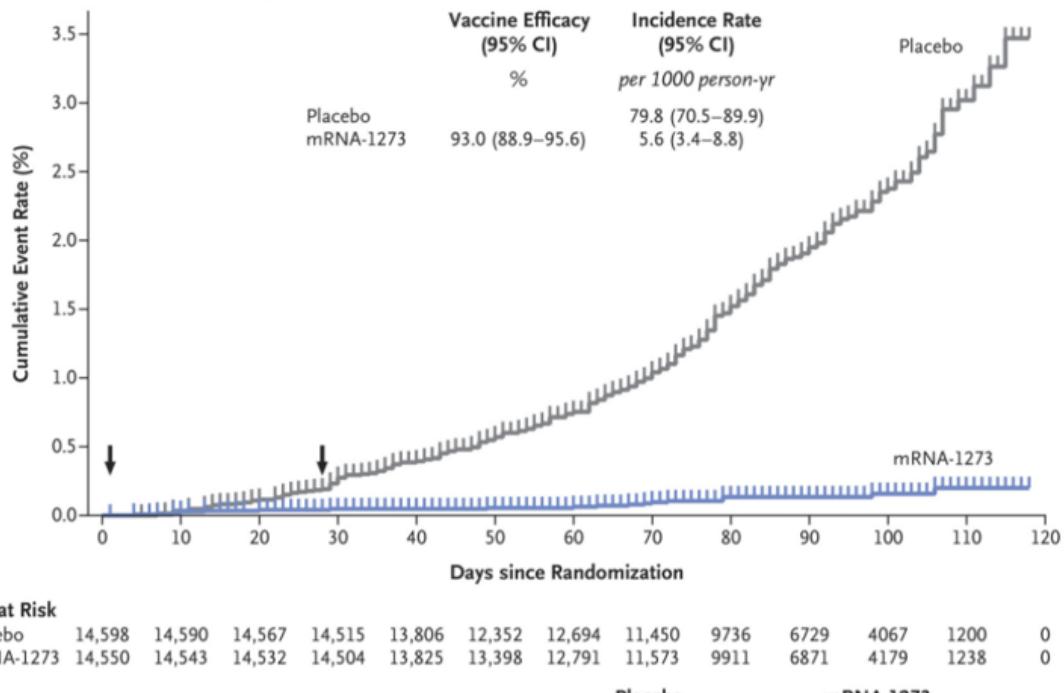
Director

Center for Biologics Evaluation and Research, FDA"

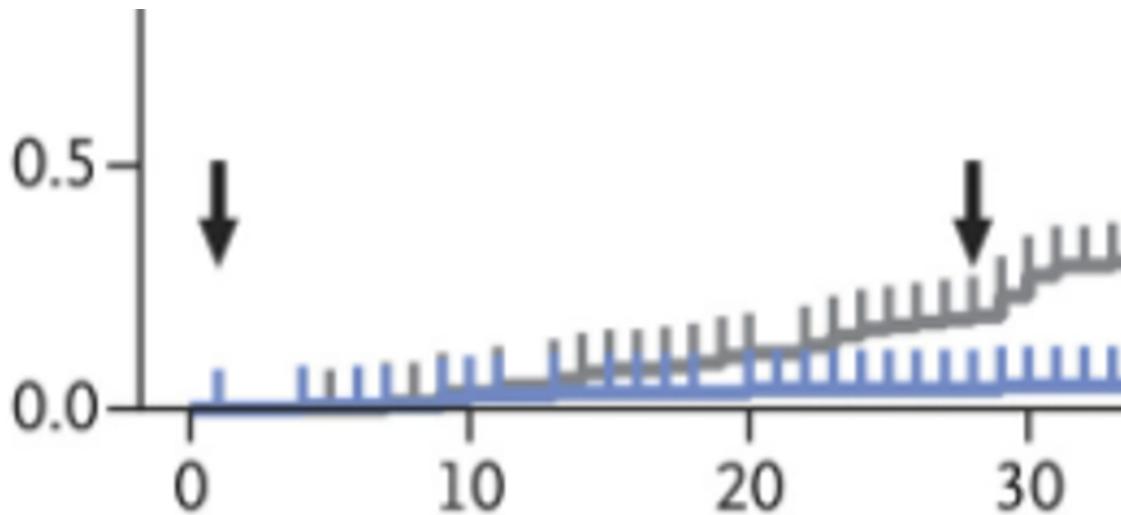
40

Is that so, Dr. Peter Marks of the FDA? Excuse me while I pull out the "[hot damn, look at this chart](#)" response [from the data](#):

B Modified Intention-to-Treat Analysis



Let's zoom in a bit:



So... does that look like it's less than 80% effective between day 10 and day 28?

Or in chart form:

Covid-19 Onset	Placebo (N=14,598)	mRNA-1273 (N=14,550)
Randomization to 14 days after dose 1	11	5
14 Days after dose 1 to dose 2	35	2
Dose 2 to 14 days after dose 2	19	0
Starting 14 days after dose 2	204	12
Total (any time after randomization)	269	19

If we assume that infections after day 10 show up after day 14, we could reasonably look at the “14 Days after dose 1 to dose 2” part of the chart as the relevant one, and the score is... 35-2. That’s actually way *more* than 80% effective.

The question of *duration* of that protection is less obvious, but based on experience from prior vaccines, and the protection gained via infection, it seems highly unlikely that we will see big drop-offs within the first few months. There is a reason booster shots are often given a year or more after the first dose. If we do see such drop-offs, as Marginal Revolution points out, we could easily change our policy, and then administer the second dose at that time.

I’d also point out that Dr. Marks says that duration and depth of protection are potentially correlated, which seems very reasonable. But that means that if depth is strong, as it seems here, then we should expect duration to also be strong.

Yes, We Can Agree Andrew Cuomo Is The Worst

Last week, Andrew Cuomo was the worst because (among other things) he took the following steps, which are what you’d do if you wanted doses to end up in the trash:

1. Severely restricted who could get the vaccine.
2. Including not allowing vaccinations for 75+ or any other age band.
3. Gave contradictory and confusing answers about who was or would be eligible for vaccination.
4. Authorized vaccinations for different groups with no warning, such that county officials (let alone citizens) had no idea what was happening.
5. Spent his bandwidth threatening anyone who skipped the queue or let someone skip with severe penalties.
6. He carried those threats out in at least one case, [confiscating vaccine doses from New Rochelle](#) that tried to vaccinate its first responders, and fining the town, per DeBlasio.
7. He then threatened those who didn’t use their full vaccine allocations quickly.
8. Was actively ignoring [pleas from NYC Mayor DeBlasio](#) that he wanted to vaccinate and didn’t have anyone he was allowed to vaccinate.
9. I didn’t notice in time for last week’s update, but it also became increasingly clear that the #YouHadOneJob of vaccinations, residents of nursing homes, was going, shall we say, [Not Great, Bob](#):



Mary Whitney @WordyMary · Jan 7

Replies to @emmagf and @maggieNYT

My 89 year old aunt is in an assisted living facility in Brooklyn. We just learned that they won’t start being vaccinated until Feb 15 to April. There are 12 covid cases in the facility NOW. How is this acceptable?

1

2

10

↑



Sarah Rose  @thesarahrose · Jan 7

...

Replying to @emmagf

My mom is 1A in assisted living in midtown where residents died of covid last spring — they have no idea when they're getting the allocation

 1



 3



At this point, of course, it is ultimately their fault, they are in New York City, so all they have to do is fill out the proper online forms:



Scott M. Stringer  @NYCComptroller · Jan 10

...

This very minute, there are more than 200 vaccination slots available on TUESDAY on the [@nycHealthy](#) website.

I am concerned this signals twin failures of outreach and technology by the City.

Scott M. Stringer  @NYCComptroller · Jan 10

...

The [@nycHealthy](#) site for signing up for a COVID vaccination is complex, burdensome, and buggy.

It will present an obstacle for too many people—particularly seniors—trying to sign up. This is a major problem.

 22

 184

 885





Scott M. Stringer  @NYCComptroller · Jan 10

...

But the problem is fixable, and it should be fixed immediately.

[@NYCHealthSystem](#) and [@nycHealthy](#) have two different sites.

The H+H site is more user-friendly, but both are buggy. This is from the H+H site.

Oops!

Sorry. Your request could not be carried out because of an error.

Please log out and sign in again.

 LOG OUT



Scott M. Stringer
@NYCComptroller

...

Replying to @NYCComptroller

The [@nycHealthy](#) site has a multi-step verification process just to set up an account, and then a six-step process to set up an appointment.

Along the way, there are as many as 51 questions or fields, in addition to uploading images of your insurance card.

9:31 PM · Jan 10, 2021 · Twitter for iPhone

They are lucky they have a website at all. Here in Orange County, no officials had any idea what was going on for days when we asked, and there was no website on which to book. Phone calls were the only option. Lots of phone calls, and waiting lists.

Many of which turn out to be the wrong places. [Here's Erie County](#) getting not all that many calls with no ability to book appointments that way:



Mark Poloncarz

@markpoloncarz

Please do NOT call our Health Department to schedule an appointment for a vaccine. We are NOT setting up appointments by phone. Per our IT Department, yesterday our main @ECDOH phone number received more than 18,500 calls. It normally would receive less than 250 per day.

9:42am · 12 Jan 2021 · Twitter Web App

[Here's another experience:](#)



Kelsey Piper @KelseyTuoc · 22h

After reading this I tried to help my NY grandparents, age 87+83, book vaccine appointments. They eventually got one, 90 miles from where they live, after hours of effort from four people two of whom have law degrees. If they could no longer safely drive, it'd have been hopeless.

...

The thing is, yes, Andrew Cuomo is absolutely The Worst, but New York *isn't actually doing so badly in terms of shots in arms*. We are actually doing slightly *above the mean* on that metric.

We could, for example, [be Virginia \(MR\)](#). That link walks through exactly how trivial the logistical issues with vaccine *distribution* (as opposed to manufacturing) would be if anyone cared about solving them.

Which means that other states are doing things that are, on average, *even worse than this*, in terms of getting shots into arms. [As PoliMath notes](#), New York is full of people looking for

bad news. I'd add it also has a governor not only causing much of that news, but striving to make us as aware of much of that news as possible. Think about what kind of disasters you would find elsewhere with the same attention.

Meanwhile, back in New York, the situation is fluid, and in important ways it is vastly improving.

Cuomo did the most important single thing he had to do prior to being told to by the CDC, which was to open up things to Group 1B, which includes everyone over the age of 75. Then things were extended to those over the age of 65.

There are now more than enough arms in which to inject shots, and the most important arms are among those in which shots can be injected. And [there's more clarity on who exactly is eligible in addition to the elderly. And he skipped the Bills game.](#) That's all great. Credit where credit is due.

The flip side of that is that there are lots of *other* arms that are also eligible, which are more likely to have success navigating the systems and getting an appointment they can use. So we should expect a lot more stories about the elderly being unable to get appointments, while young people working in administration get slots instead. As you do. But it's still a vast improvement.

Also that it seems we've decided that [a lot of the state's focus should be on the "climate crisis"](#) where we will somehow lead the way. Aren't you, I dunno, kind of busy, sir?

It also seems he's going after a record that was previously (I can only presume and won't check) was held until now by former Arkansas Governor Bill Clinton, and which we thought would never be broken:



Andrew Cuomo 

@NYGovCuomo

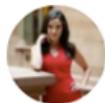
**Will be delivering Day Three of my
2021 State of the State address at
11:30am ET.**

**Join me and watch live:
governor.ny.gov**

#SOTS2021

On a personal level, I can report great success. As a health care worker, my wife called infinite places, got exactly one of them to call her back, and got her shot over the weekend. Armed (partly by me) with the information on how to book, the proper urgency and reasonable skill at using the internet, my elderly parents, both over 75, got their first shots on Monday morning in New York City. This is a great relief all around, even though I will likely be waiting many months for my turn.

Meanwhile, Cuomo has noticed that testing is a thing we could do a lot more often, but rather than use it to solve a pandemic, [he's going to use it to reopen entertainment venues and office buildings:](#)



Morgan Mckay
@morganfmckay

...

Replies to @morganfmckay

"Testing is the key to opening the economy," Cuomo says. Specifically ramping up rapid testing.

Points to the Bills game- took 5 minutes per car for everyone to get a rapid test
Says it seems it was a success and could be used for theaters and others events



Morgan Mckay @morganfmckay · 23h

...

Replies to @morganfmckay

"Cities are by definition centers of energy, entertainment and cuisine. Without that cities lose much of their appeal."

"We must bring culture and arts back to life." Questions what is NYC without Broadway

1

7

6

↑



Morgan Mckay @morganfmckay · 23h

...

Cuomo vows to open hundreds of rapid testing sites so people can return to office buildings

3

10

10

↑



Morgan Mckay @morganfmckay · 23h

...

Cuomo announces "New York Arts Revival" which will aim at bringing arts back to cities across the state starting February 4th. These will be outdoor pop-up events headlined by Amy Schumer and Chris Rock

~

~

~

~

I suppose there's only room for one 'actually stop pandemic' thing, and the vaccine already called dibs. We have to respect dibs. We're not barbarians.

If you are in New York, [this is the official place to start arranging for an appointment](#).

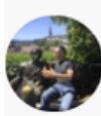
[Here's an NYC-specific Twitter thread with a few additional links](#) and thoughts. The author of that thread [has offered to help anyone in NYC who needs assistance navigating the system](#)

[and booking an appointment.](#)

All I Want At This Point is to Sell Out of Covid Vaccine

There are three potential bottlenecks for Covid vaccine distribution. You need shots to put into arms, you need professionals to put those shots into people's arms, and you need arms in which to put the shots.

[Good news \(WaPo\)!](#)



Walid Gellad, MD MPH @walidgellad · 3h

For clarity, the Biden team agrees on need to expand population to over 65.

...

The disagreement is expanding to under 65 with comorbidity, which is a big population and can add to confusion.

They have a point, given limited supply,

[The guideline to extend to over 65 has been issued by the CDC.](#)

We fixed the first bottleneck, at least where the CDC's guidelines are followed. The elderly are eligible now. Wonderful! Time to make an appointment. [There's a problem:](#)



Chealsye Bowley @chealsye · Jan 11

...

I had an over 70 patron come to the library. Her DR gave her a slip with a very long URL listed to register for a COVID vaccine. DR said "the library will help you."

She's never used a computer.

An email AND phone number that supports texts was required to make an appointment.

14

127

274



Chealsye Bowley @chealsye · Jan 11

...

I understand mistakes happen and a lot of people make assumptions about what access to technology people have (of any age and any income). But when your audience is just healthcare workers AND 70+ PEOPLE... how do you assume *everyone* has an email and a phone that texts?

2

17

146



Chealsye Bowley @chealsye · Jan 11

...

A non-internet option for scheduling an appointment should be available. Or at least the online form have a checkbox for "I have a home phone, but no email - please call me."

We're also the only library fully open in the surrounding counties. DRs should just schedule the appt.

2

12

144



Chealsye Bowley @chealsye · Jan 11

...

People eligible for the vaccine w/o a computer will make a trip to the library (likely their lib isn't open, so they'll travel to mine) putting themselves + library workers at risk.

Current solution: we'll help you make an email and you'll have to come to the library to check it

A lot of elderly people are not going to be able to get through this process, and a lot of others are going to get infected on their multiple trips to the library.

How is all that going to go?

We can start with how it's gone so far:

The US COVID-19 Vaccine Shortfall

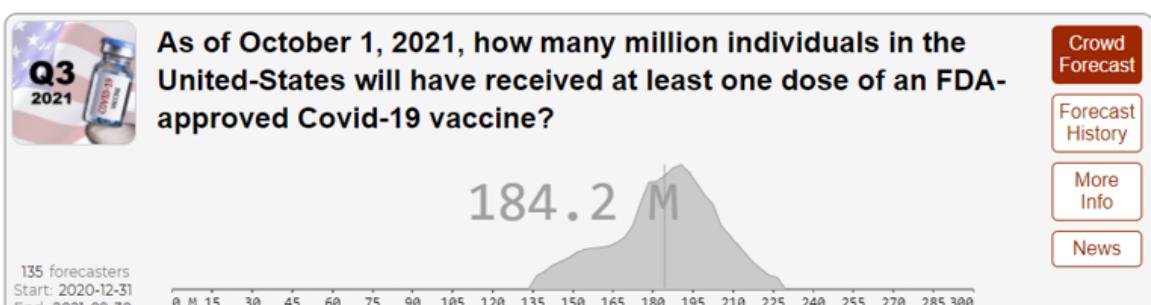
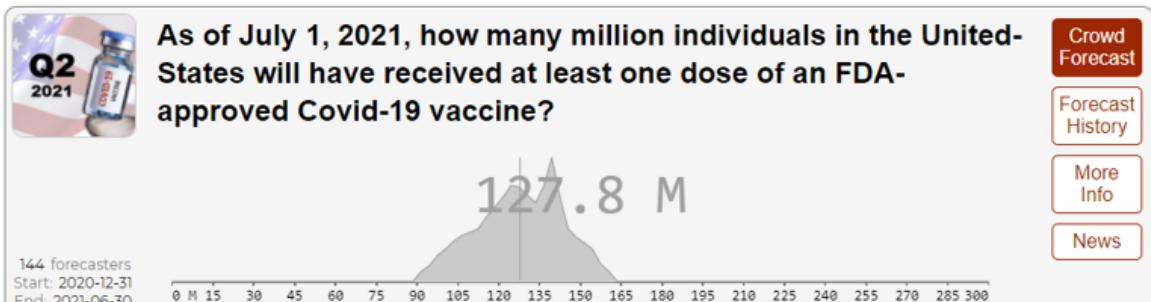
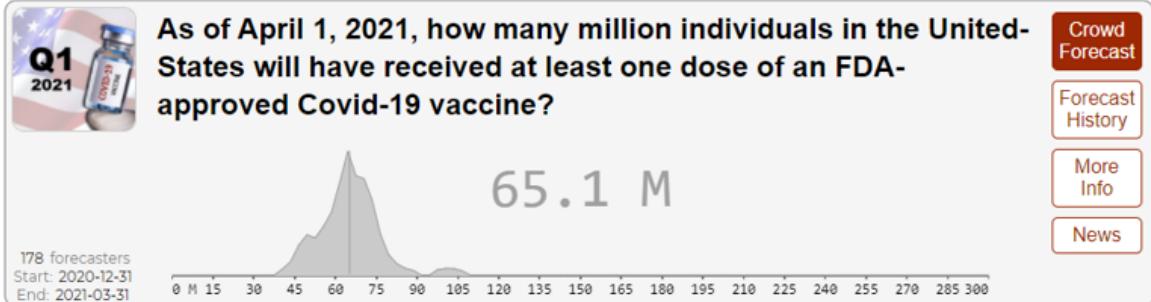
Operation Warp Speed has shipped more than 25 million vaccine doses, but less than 9 million people have received shots.



**CDC statistics do not report how many people have received second doses.*

Chart: BuzzFeed News / Dan Vergano • Source: [CDC](#)

[Let's check the predictions over at Hypermind](#), where the numbers haven't moved much this past week:



This is more optimistic than [my naive Covid modeling](#). I had about 38 million vaccinated by April 1, 81 million by July 1, 131 million by October 1, so they're also predicting roughly linear progress but have it happening a little over 50% faster.

I hope these predictions are right, or better yet pessimistic. The difference in those guesses is a *really big deal*. If we take my model's guess and boost vaccinations an additional 50%, and the additional infectiousness of the new strain is only 50%, we plausibly (mostly) stop the fourth wave, despite the new strain currently being ahead of my previous guess for that. I've added knobs in the spreadsheet to vary both numbers.

Under the optimistic scenario, this increased vaccination pace would be the difference between ending with 40% or 50% of the population having had Covid-19 when it's all over, with about 24% having already been infected. So this would prevent over a third of all future infections. Speed matters more than anything.

There are 209 million Americans over the age of 18. Given how many don't want to be vaccinated, these predictions above imply that by April 1 you can likely get the vaccine as long as you care about finding it, and by July 1 you can get the vaccine provided you want it at all, as there would be enough vaccinations to cover every adult who wants one at all.

What about [the Good Judgment Project](#)?

When will enough doses of FDA-approved COVID-19 vaccine(s) to inoculate 100 million people be distributed in the United States?	Today's Forecast	1-week Forecast	1-week Change
Today's Forecast:			
A Before 1 February 2021	1%	+1	
B Between 1 February 2021 and 31 March 2021	7%	-7	
C Between 1 April 2021 and 31 May 2021	80%	+10	
D Between 1 June 2021 and 31 July 2021	11%	-4	
E Not before 1 August 2021	1%	0	
Forecast History			

That's distributed doses rather than vaccinations, but by then I presume those numbers will be similar. This market seems overconfident in the timing, even if the median of early May is reasonable. In general, I find Good Judgment prediction markets to be overconfident.

It is worth noting they were deeply overly optimistic about how much distribution we would get in December and January, as were essentially everyone as far as I could tell, and it seems to have taken them far too long to realize their mistake.

Meanwhile, [proper epistemic procedures are in place at MIRI](#):

Eliezer Yudkowsky  @ESYudkowsky · 22h
I have bet \$100 with epistemic peer Nate Soares that we won't be vaccinated (by mRNA, by a US-government-approved process) by end of July. (I don't think I know better than Nate, but bets are important for recording first-order disputes.) [@AnnaWSalamon](#) at 2/3 for vaccination.

14 2 101 

Eliezer Yudkowsky  @ESYudkowsky · 22h
("We" defined as Nate+myself.)

1 2 13 

I like Nate's side of this bet and think Anna's 2/3 estimate is a little bit low, given how much I expect Eliezer and Nate to make an effort. They should usually be (in effect) very near the front of the no-priority line.

Vaccine Allocation By Politics and Power

Headlines that tell the story and don't require a click: [Who can get the COVID vaccine in Florida? Hint: It helps if you have donated to a hospital.](#)

There's an extra advantage to 24-hour vaccinations, as they give an efficient way for people who aren't allowed to pay money to bid on getting the vaccine. Who cares about it most? [Offer some highly inconvenient appointments and see who shows up:](#)



Rachel Holdsworth

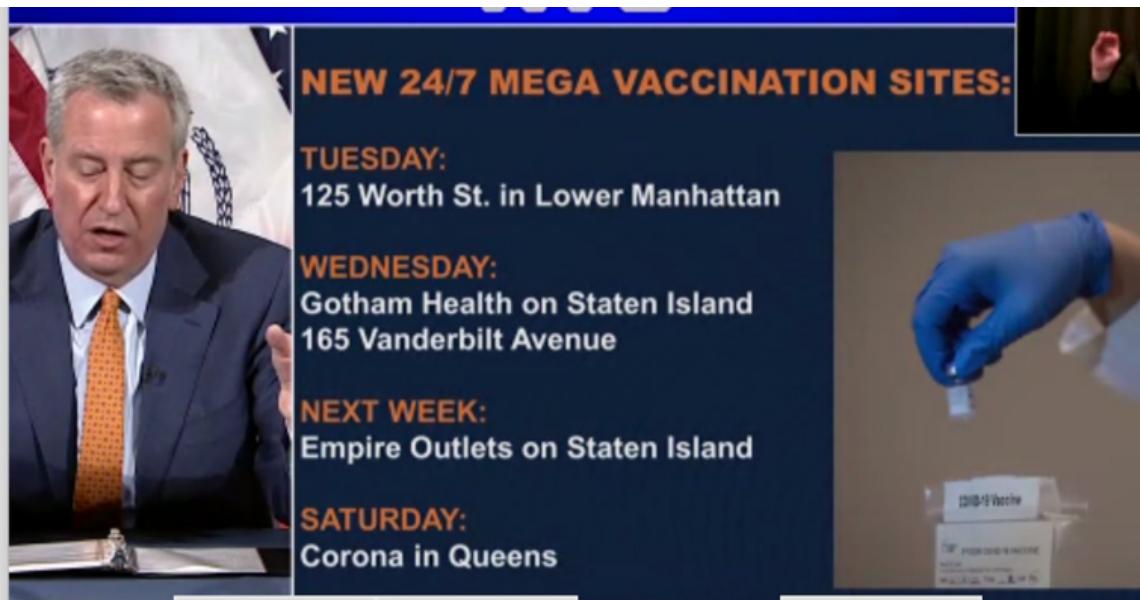
@rmholdsworth

...

Speak for yourself, [@guardian](#); if someone offers me a vaccine slot at 4am I'll be there so fast my shoes will have scorch marks

Nadhim Zahawi, the vaccines minister, also floated the possibility of a 24-hour service in an interview this morning. (See [10.08am](#).) Given what is likely to be the very limited enthusiasm for getting vaccinated at, say, 4am, it may be that Starmer and Zahawi were being more aspirational rather than literal with this proposal.

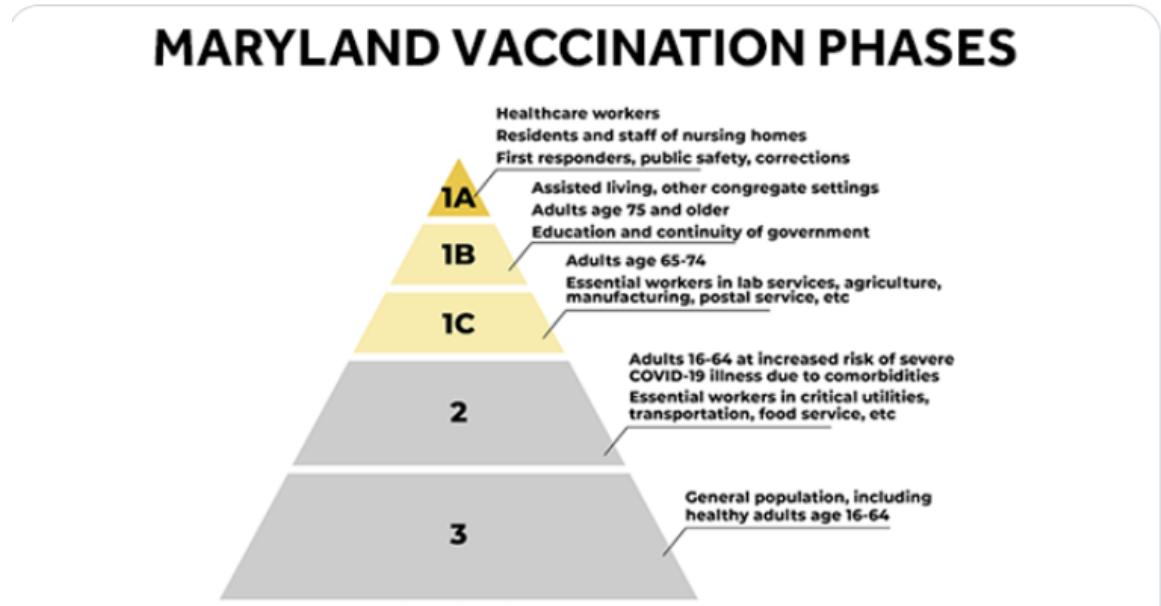
Believe me, I would happily show up any time, day or night. [We're about to find out](#) if I'm alone in that, which I am very confident that I am not:



The list is growing. We also have [the Washington Heights Armory](#).

Not listed is the best mega vaccination site, Citi Field, home of the newly owned-by-someone-who-cares-about-winning New York Mets, [which will start on January 25](#).

[Maryland's headline choices](#), prior to the new CDC guidance, where teachers were to go before those aged 65-74, and 8% of the population is in front of those over 75:



Or to be as clear as possible on what was going on as of January 11, prior to the new guidance:



Walid Gellad, MD MPH @walidgellad · Jan 11

...

Replies to [@walidgellad](#)

Just to make this really concrete for people.

A 74 year old person in Maryland with obesity, diabetes, and heart failure is of lower priority for the covid vaccine than a 28 year old perfectly healthy kindergarten teacher.

4

5

10

[Meanwhile, the latest from Los Angeles \(2nd link from post\)](#), where we keep getting signs that things are being handled especially destructively:



Ben Casnocha



@bencasnocha

LA County plans to wait to vaccinate seniors until the 500k remaining healthcare workers get it:
latimes.com/california/sto...

But huge % of healthcare workers decline the vaccine. In Fresno County, 50%+ of healthcare workers opt out:
gwwire.com/2021/01/05/ove... cc:
@eladgil @tylervcowen



California allows everyone 65 and older to get COVID-19 vaccine
latimes.com

[One way to look at it:](#)



David Chapman

@Meaningness

52% of Americans were vaccinated for flu last year, mostly in 3-4 months. This involved no sense of urgency and did not require extraordinary efforts.

The current extraordinary efforts are halving the speed from what we'd get with business as usual.
[cdc.gov/flu/fluview...](http://cdc.gov/flu/fluview/)

Do Not Throw Away Vaccine Doses: Somehow a Section Title

I am mostly doing a solid job of not getting angry about how things are going. I make an exception for when *people literally throw out doses of vaccines*.

[Welcome to Cuomo's New York:](#)

The nurse at the clinic called her supervisor at home, asking what to do with the remainder. The supervising nurse then called her contact at New York City's health department for guidance, Dr. Calman said. She was told to find people who fit the eligibility criteria and was encouraged to contact a nearby nursing home, an urgent care center and a women's shelter.

The nurse at the clinic set out on foot. She was turned away at a few places, including a nursing home and fire station, although she finally found one eligible health care worker willing to be vaccinated, Dr. Calman said.

He said the nurse eventually threw out the remaining doses after the health department told the clinic that it could vaccinate only members of eligible groups.

Fortunately, also in Cuomo's New York, some people realize [it's not the incentives, it's you and decide not to do that](#). Obviously vaccinating a 26-year-old healthy reporter is not ideal, but it's infinitely better than throwing the dose out.

In the former control group, [Sweden is throwing out the extra vaccine doses in Pfizer vials](#), because we might dislike the FDA but at least we don't have to deal with the European Medicines Agency, who are totally Delenda Est Club members:



Farmor @Farmor45593288 · Jan 10

...

Replies to @DavidSteadson

The instructions for use say that you should throw away anything that is over the 5 dose recommended by Pfizer. Then EMA had to approve that you can use 6 doses/bottle. It came this Friday. Swedish Medicines Agency must also approve the same. Bureaucracy takes time, costs lives.

1

3

8

↑

But good news, [they could soon give the go ahead](#) to stop throwing away vaccine doses.

England by contrast has moved on to [doing distribution via pharmacies](#). Brexit has had some very high costs and was not pitched to the people remotely honestly, but getting away from European Union regulatory bodies while one has the chance should not be underestimated.

[And here in America:](#)



Ashish K. Jha, MD, MPH @ashishkjha · Jan 10

...

Replies to [@ashishkjha](#)

They were looking for unvaccinated employees

Most employees there had been vaccinated. Rest were unwilling

Found several EMTs & patients who were excited to be vaccinated

But hospital policy was clear: non-employees aren't eligible

My friend, ER Doc, incensed, intervened



Ashish K. Jha, MD, MPH @ashishkjha · Jan 10

...

He tried to persuade vaccine team but they wouldn't over-ride hospital policy.

He called ER leadership. They wouldn't over-ride

Next, hospital leadership. They initially said no, claiming state mandate

He is persuasive and persistent....so they eventually relented



Ashish K. Jha, MD, MPH @ashishkjha · Jan 10

...

But by then, vaccinators had left

He tracked them down. Their shift was over and per protocol, they had discarded the doses

Hearing more stories like this

No idea if these are one-offs or systemic

We aren't publicly reporting detailed data on state of vaccinations

Also, there are some things you should *definitely not* do if doses are about to expire, because they're technically dangerous, and against the rules, and set bad examples, and reasons, and they're so bad I'm not even going to say what they are, but in any case please *definitely do not* do them.

Buy More Vaccine? You Can Do That?

I'm primarily *not* being sarcastic here, I'm actually still boggled can someone please explain [what the hell is up with this:](#)

MICHIGAN

Whitmer asks feds if Michigan can buy 100,000 vaccine doses directly from Pfizer

I ask, because if Michigan is asking permission that implies that with permission they would be able to buy doses (from the factory, which is in Michigan) *which implies there are doses still for sale that no one has bought.*

Which seems utterly insane, but so does the idea that the Feds would be giving Michigan part of their allocation in this spot, which is the only other explanation I can come up with? But that does not seem like something the Feds would ever agree to do.

So I notice I am extremely confused, and hope someone can help me out on this one.

Also, it certainly *does seem like* the main bottleneck for solving *the entire global pandemic* much faster was indeed *very small amounts of money* combined I presume with regulatory barriers:

[If true, and we'd been willing to spend approximately zero additional dollars, we'd already have all the doses we need.](#) If true, we could still do it and have the rest of the world covered within the year.

Not doing this last year was... [I mean I don't even have the words:](#)



Max Roser @MaxCRoser

18h

This study estimates the cost to produce the missing vaccines to protect *the entire world* from COVID.

[mdpi.com/2076-393X/9/1/...](https://mdpi.com/2076-393X/9/1/)

Facilities to produce 16 billion doses of a Moderna type vaccines would only cost around \$4 billion, ~\$2 a jab.

Easily the best deal of the decade.

🕒 34 🔍 565 ❤️ 1k ⚡



David Manheim

@davidmanheim

Replies to @MaxCRoser @DrEricDing

And we could have done this 6 months ago for all 3 leading vaccines, at a cost of \$12 billion, despite being unsure of which would work, and had enough doses already.

David Manheim @davidmanheim

A Simple Proposal for Jumpstarting Vaccine Production

In about 12 months, the world will need to start producing massive quantities of a COVID-19 vaccine, but we don't have enough vaccine production facilities to quickly produce the billions of doses we need globally. (1/9)

Show this thread

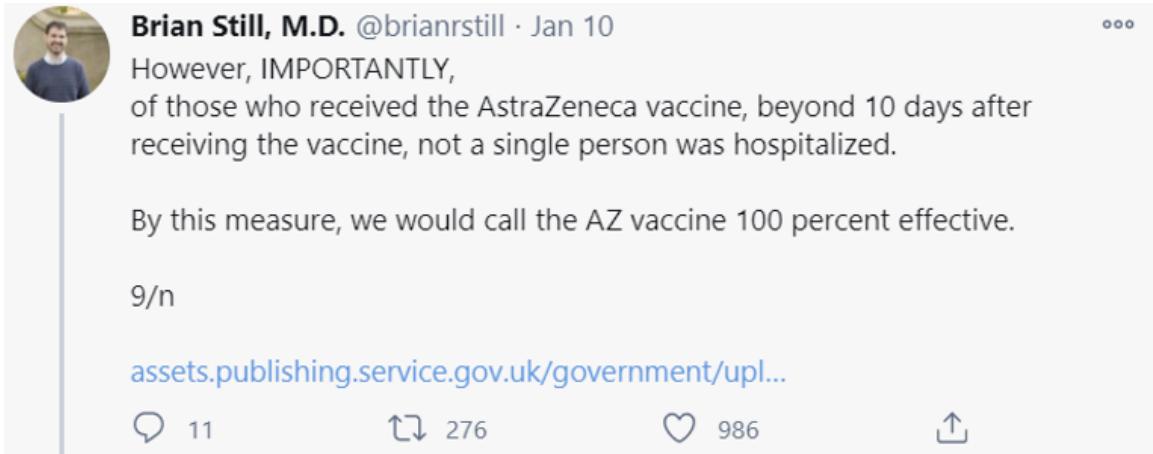
If this is all remotely true, then this was separately within the budget of dozens of countries and at least a hundred private individuals, many of whom are primarily philanthropists. None

of them made a serious attempt to do this or even consider doing this (at least that we know about), is a very important fact to fit into your model of the world.

So: [Ask yourself why.](#)

Approve Safe and Effective Vaccines: Somehow Also a Section Title

Your weekly reminder from Marginal Revolution that we should approve the AstraZeneca vaccine. [No, really, we should approve it](#) and I would happily accept it myself ([link from screenshot](#)):



Brian Still, M.D. @brianrstill · Jan 10

However, IMPORTANTLY,
of those who received the AstraZeneca vaccine, beyond 10 days after
receiving the vaccine, not a single person was hospitalized.

By this measure, we would call the AZ vaccine 100 percent effective.

9/n

assets.publishing.service.gov.uk/government/uploads...

11 276 986

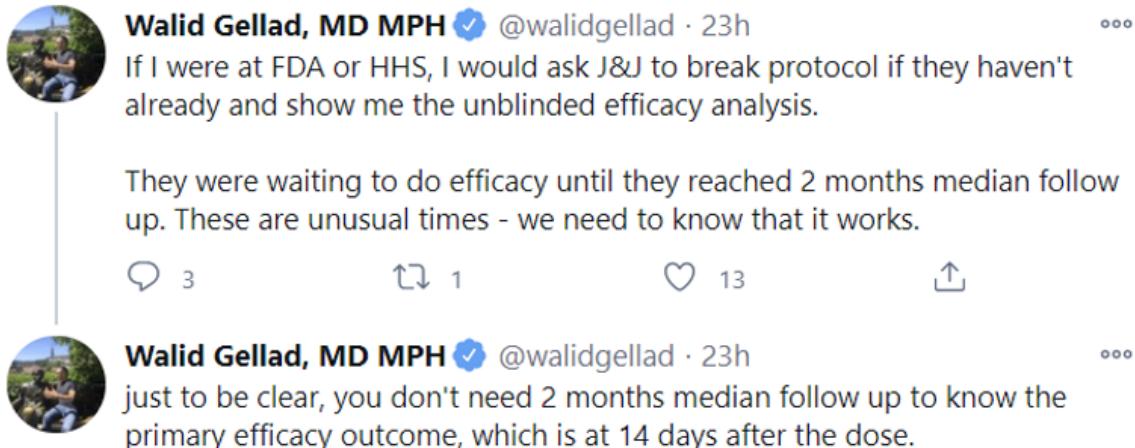
Ensuring this happened would be my first order of business if I was the new President.

The second order of business would be approving Johnson & Johnson's vaccine.

That's right, the best news of the week is that this section title can be officially plural now, because [we can welcome Johnson & Johnson's one shot vaccine to the list](#). So we're now waiting on multiple approvals.

Technically, we don't know that it works. What we do know is it generates a robust antibody response. We won't know the phase three results until later in this month, but if anyone thinks those results won't be good enough to justify approving the vaccine, I will happily book your wager.

[We have the information](#) we actually need:



Walid Gellad, MD MPH @walidgellad · 23h

If I were at FDA or HHS, I would ask J&J to break protocol if they haven't already and show me the unblinded efficacy analysis.

They were waiting to do efficacy until they reached 2 months median follow up. These are unusual times - we need to know that it works.

3 1 13

Walid Gellad, MD MPH @walidgellad · 23h

just to be clear, you don't need 2 months median follow up to know the primary efficacy outcome, which is at 14 days after the dose.

We won't know exactly how well it works yet, but we know it's safe and we know it's a lot better than nothing, so all waiting does is kill people and prolong the pandemic.

Alas, we also got very bad logistical news from Johnson & Johnson. They will likely only be able to deliver a few million doses (only one needed per person) by the end of February, if results prove good and their application for emergency use authorization is approved. All help is welcome, and we should take it now and give them every incentive to move faster, but we were hoping for a lot more.

How Bad is it Out There Right Now?

It's not great.

I was asked a few weeks back about the dangers of getting infected while getting your vaccine. This is a real worry, and is also the ultimate way to waste a vaccine dose and expose exactly the people we most want to protect. It was suggested that if things get bad enough during the fourth wave, that getting vaccinated might be so risky as to not be worthwhile.

I responded that while such infections will happen, it seemed super unlikely that getting the vaccine wouldn't be worthwhile. Then [you read descriptions like this one](#) from England, [via Stuart Ritchie](#), which I recommend reading in full, and it makes you wonder. There's no question that what they are doing there is going to get a lot of people infected.

You Should Know This Already

We worked for weeks on a detailed set of prioritizations and procedures and regulations and approvals, [and Israel just... gave out the vaccine to everyone as quickly as possible](#).

We sat around worrying about 'wasting money' buying too many doses, and now we don't have enough, [and Israel just... paid two to three times the "market price" to make sure it had enough doses](#). They also committed to sharing data to help make the deals go through. The rest of the thread has lots more info on how Israel pulled this off. And again, it's all... they went ahead and did the thing. There was no [shutting up and doing the impossible](#), both because nothing involved was remotely impossible, and also because what would be truly impossible is getting Israelis to shut up.

[And now:](#)



Approve AstraZeneca ✅ @GarettJones · Jan 11

They did it:

...

Israel has now prevented the majority of all possible future Covid fatalities.

70% of the elderly in Israel are now vaccinated.

Just wonderful, a model for other nations.

Israel has already given first doses to some 1.7 million people out of a population of 9.29 million, by far the highest vaccination rate in the world. The Health Ministry has prioritized at-risk groups and people over 60, with more than 70 percent of Israelis in that age group having now received the first shot.

Alex Tabarrok reminds us that [we should be doing preliminary testing now on vaccines for potential future pandemics](#).

[John Cochrane righteously rants about our failure to approve AstraZeneca](#). He's right but no need to click.

[Alex also reminds us \(in a repost\) that experiments make people uneasy and suggests reasons](#). My understanding continues to be that the experiment result is largely about [The Copenhagen Interpretation of Ethics](#) and [Asymmetric Justice](#), as one who experiments is now responsible for all potential harms and all potential lost benefits, as well as all potential or realized inequalities, and for trading off any and all sacred values, while getting none of the credit for anything since the choices were randomized.

Do not click on this because it is a case of Something is Wrong on the Internet, but I needed an example of the sort of insane scaremongering guidance [that is being given to many people that they should remain paranoid even after being fully vaccinated](#). And thus say things like this:

If your grandkids live in the area, you could definitely safely see them outside, 6 feet apart. If you want to see them indoors, there is going to be some level of risk. That risk will be much lower than if you were not vaccinated, but the risk is still going to be there to you. And you could still be a risk to the unvaccinated members of your family, as you could be an asymptomatic carrier who transmits to them.

If you really want to spend time with the grandkids indoors, the safest way to do this is still for everyone to quarantine for at least 10 days and lower their risk during these 10 days.

Quarantining for seven days and a negative test is an option too, but everyone also has to do the quarantine — a negative test alone is not enough.

This is obvious nonsense. If it is reasonable for you to consider doing full quarantines after getting fully vaccinated, I can only deduce that you are either immune deficient or have a dead man's trigger that launches nuclear missiles.

At some point we have to call such sentiment what it is, which is that it is opposed to and incompatible with life. One should also look at all similar recommendations made in other

circumstances *in light of its inability to adjust to circumstances*. You should be skeptical of cries of “Wolf!” from people who keep saying “Wolf!” no matter how obviously there is no wolf.

You Should Know This Already [in the sense that this screenshot has a 1 where there should have been a 0](#):



BNO Newsroom @BNODesk

...

WHO's team of experts will be allowed to travel to Wuhan on Thursday to investigate the origins of the novel coronavirus, China says

4:01 AM · Jan 12, 2021 · TweetDeck

[Reminder that this Covid tracker exists for NYC.](#)

[Fans were not socially distancing after the National Championship game.](#)

[Offered without comment:](#)



PoliMath

@politicalmath

I'm enjoying taking walks in my local park but the signs crack me up

on their list of rules is "follow current CDC recommendations" which is a hilarious admission that they have no idea what those are or if they've changed but whatever the current ones are, you follow them

12:10pm · 13 Jan 2021 · Twitter Web App

In Other News

From the CNBC Johnson & Johnson article linked above, we also see this tidbit worth noticing and applauding, because incentives matter, and they matter more than you think even after knowing this fact:

The data comes as U.S. officials complain that the pace of vaccinations has been too slow as the supply of vaccine doses exceeds demand. The Centers for Disease Control and Prevention expanded [Covid vaccine eligibility](#) guidelines Tuesday to include people 65 and older as well as people with preexisting conditions. The government is also changing the way it allocates Covid vaccine doses, now basing it on how quickly states can administer shots and the size of their elderly population.

[Wired covers the micro-covid project at length.](#)

[The University of Wisconsin used rapid testing and managed to catch 80 percent of cases with symptoms and 41 percent of asymptomatic cases.](#) The article chooses a somewhat different framing.

A question that needs to be asked, [if you have good additional info not in the thread please share:](#)

Eliezer Yudkowsky  @ESYudkowsky · 5h
Is there a good central summary of the allegations I'm hearing that mRNA-vaccine papers had trouble getting published? Because if true, that plus the replication crisis - good unpubishable, bad published - seems to indicate that the journal-review system should just be burned.
8 replies 2 retweets 88 likes

Eliezer Yudkowsky  @ESYudkowsky · 5h
Considering selection biases and forum shopping, good papers being unpubishable is a much worse sign than bad papers being publishable. Bad papers can seek out bad journals or editors having bad days. Good papers being unpubishable means there are no good journals!
4 replies 4 retweets 57 likes

Another source of [weekly Covid updates](#) is Tom Frieden. I'd skip it, but figured I'd note it exists. There's a lot to like here, but it didn't teach me anything new, and it reflects the moralizing explanation of Covid spread. There's some 'luck' from super spreaders but otherwise it's all about whether you did enough sacrificing, or you didn't do it right or didn't do it for long enough, and the goal of writing such updates is to scare people into bigger sacrifices.

I have no idea why Dr. Moncef Slaoui, the head of Operation Warp Speed, [was asked to resign and transition things over to someone else](#). Seems like if someone does their one job

this effectively you'd want to keep them around.

After being confined in a room for hours with a number of Republican congressmen on Wednesday, January 6, you know, *because of reasons*, [Congresswoman Jayapal tested positive for Covid-19](#), and is now in quarantine. [She reports](#): "The duration in the room was multiple hours and several Republicans not only cruelly refused to wear a mask but mocked colleagues and staff who offered them one."

Youyang Gu of the Covid Machine Learning projections [notes that there is essentially no correlation](#) among the US states between current infection rates and estimates of existing immunity. One could interpret this as places that are more vulnerable having had more infections, and thus now having more immunity but doing so from a baseline of greater vulnerability. Certainly the correlation between state precautions in the past and state precautions now is positive. If one sets 'effect of accumulated additional immunity' equal to 'effect of the conditions that caused those infections' then that gives us a model we can test.

Never let a crisis go to waste, medical innovation edition, [as mRNA vaccine tech is applied to multiple sclerosis](#), here's the [paper in Science](#) (login required). [Next up: Nipah, HIV and seasonal flu.](#)

It's still early, but still. Note that lack of speed doesn't kill *as effectively* here, but lack of speed still kills here, too.

Also, off topic, to end on a positive note, if anyone wants to never feel cold again, I'm here to report back: [The coat is here, it is real, and it is spectacular.](#)

Tentative covid surface risk estimates

My household previously made some highly uncertain estimates of the covid risk from bringing random objects that other people have recently been touching into our home, for instance salads and groceries an endless stream of Amazon packages. The official guidance is very vague, e.g. "[...not thought to be the main way the virus spreads](#)". Our bad estimates were fairly low, so we decided to basically ignore it in our covid risk accounting, except for habitually taking some reasonable precautions.

Then the covid rates here increased by a factor of ten, so we decided it would be good to look at it again.

So today I tried to estimate this from [this paper](#) (HT Ben Weinstein-Raun and Catherine Olsson) in which a group of researchers swabbed various door handles and trash cans and crosswalk buttons and the like in a small Massachusetts city and measured the covid RNA detectable on them. They also used the amounts they measured to estimate the infectiousness if someone else were to touch the surface and then touch their face.

[Here](#) I offer you a draft of an elaborate Guesstimate spreadsheet on the topic, in case you are interested in such things. Hopefully someone is, and will tell me ways that it is wrong, then later I might offer you something reliable. At present, it is probably pretty unreliable, and should only be used insofar as you would otherwise use something even more unreliable. Some non-exhaustive evidence of its unreliability:

- I haven't actually read much of the paper
- The answers in the spreadsheet have changed substantially while I have felt about as confident in them as I do now
- There are numerous places where it seems dubious to me, or where I made up numbers
- I try not to be the sort of person who only shares things if they are of high quality, even if the stakes are high
- This calculation is ignoring the efforts everyone is making to be safe, so you might underestimate the risks if surfaces look low risk in this study because supermarket employees are actually constantly wiping them down, for instance. So it should probably be interpreted more like 'if you take levels of caution similar to those taken with the study surfaces...'.

Interesting tentative conclusions so far, rely upon at own risk:

- a person with covid touching something then you touching it then touching your face is worth extremely roughly 13 [microcovids](#) (uCov) (with 90% confidence of 0.7 to 110 according to my estimate, but I wouldn't trust that)
- thus such a touch-touch-face sequence with a random person in San Francisco (where I live) at the moment is ~0.7 uCov (give or take an order of magnitude)
- adding further wild guesses about how many touches are involved in acquiring groceries, I get that a 30 item San Francisco grocery trip is worth about 5 uCov (give or take an order of magnitude)
- that would mean about two cases from groceries in San Francisco per week, (give or take an order of magnitude) which doesn't sound crazy to me. (You might think it does sound crazy, because surely we would notice that by now, but I'm pretty uncertain about our collective ability to observe and infer things.)

The basic reasoning is this:

- If an infected person touches a surface, it looks like it has about a 13% chance of becoming detectably infectious by this method (based on 36 samples from grocery store surfaces which are estimated by me to receive very roughly 2 covid-infected touches per day yielding 4 positive samples, along with some complication with distributions that make the arithmetic strange.)
- Average risk from touching one of these positive-testing surfaces is very roughly 100 microcovids (taking an estimate half way between the average grocery surface infectiousness and the average surface infectiousness, according to the paper)
- So if a person with covid touches a surface, then you touch it and then touch your face, this gives us about 13% of 100 microcovids = 13 microcovids (.0013% chance of covid)

I welcome all kinds of comments, e.g. even ‘I want to know why this cell says 27, but I can’t be bothered reading all these things’.

Catching the Spark

Multiple readers have noted that this is very similar to a method called “Thinking At the Edge”, taught by Eugene Gendlin. I don’t know much about Thinking At the Edge, but what little I do know suggests that he and I have developed his focusing work in similar directions, perhaps for similar reasons. If you like this essay, you might want to check out [Gendlin’s introduction to Thinking At the Edge](#).

There are a lot of different ways to dance.

In one kind of dancing, you move your body around spontaneously in response to music. That’s a fun dance that even infants can do. But some dances have a more specific goal than “fun”. Partner dances enable conversation through touch and motion rather than symbolic language. Folk dances encourage participation in the soul of a community. Ballet pairs choreography with music to tell a story. If you’ve ever tried any of these specialized dances, you know that they take deliberate effort to learn.

Similarly, there are a lot of different ways to engage with the world in a spirit of curiosity.

“Naturalism” (an allusion to [19th century naturalists](#)) is the name I use for a specific way of engaging curiously with the world. It’s a method of inquiry that uses patient observation and original seeing to build models that were previously unthinkable. It takes longer to learn than the spontaneous curiosity of pure exploration, but this specialized method helps you to make deliberate progress in areas where your existing concepts are leading you astray.

It takes much more than one essay to talk about naturalism as a whole, so *this essay* is just about the very first piece: catching the spark.

Imagine two pieces of flint colliding with each other and sending sparks flying. By default, the sparks will go dark and fall to the ground. But if you catch one in a bundle of dried grass, and then blow on it gently, you can make a flame.

Curiosity is like that. It's being sparked all the time, but most of the sparks flare out and disappear.

(If you watch your experience closely for just thirty seconds, with attention to teeny tiny searching sensations, [you can probably feel it happening \[1\]](#).)

This essay is about my strategy for turning some of those sparks into steady flames. When skillfully tended, a steady flame of curiosity can sustain a whole line of research, even when neither you nor anyone else knows how to think about the problems involved.

It took me a long time to come up with this method. I meandered a lot, and I am not done meandering. What I will share with you is where my meanderings have taken me so far. I hope you will make a better version that is closer to the place where you want to be.

So, how does a naturalist study begin?

I think there are several good answers to this. But like a hiker finding north, it usually starts with some kind of orientation.

What follows is an orientation procedure. It is the first lesson in my course on naturalism, adapted to text. The procedure is meant to prepare you to drive with curiosity toward a more

intimate relationship with a region of territory where crucial data are likely to live. (More on that later.)

I think my way of orienting is especially good for people who have some idea of what they're interested in studying, but not *much* of one ("Something about forecasting, maybe?"); and for people who know what bug they want to solve, but aren't relating to it with curiosity ("I just want to *make the problem go away!*").

This might not be quite the right way to help you catch the spark. When I guide someone through this one-on-one, we work together to design *their* orientation process. As you read, I recommend paying attention to your frustrations and desires, and holding a question like, "What might work better for *me*?"

This orientation procedure has three parts: articulating stories, squinting at stories, and choosing your quest. In each part, I'll say a little about the point of it, show you one way it can go with a real-life example I work through on the spot, describe the instructions I followed to do it, then share another example before I move to the next part. There will be some discussion and philosophizing, and then, at the end, I'll include an appendix with a couple more real-life examples.

If you want to follow along, this is the time to [think of a spark you would like to catch \[2\]](#).

The Orientation Procedure

Part 1: Articulating Stories

You'll start by finding a handle for a potent [felt sense](#) that's backed by curiosity. That handle will eventually become a torch as you carry your flame into deliberate exploration.

Example 1: Geometry

While designing a new cover for my greenhouse the other day, I used the Pythagorean theorem to figure out the dimensions of the roof. Something happened while I was doing those calculations. There was a moment of confusion, puzzling, and curiosity. A spark.

Let's see if I can catch it.

(Context: I know almost zero geometry.)

Does it involve "calculation"? No, that doesn't seem right.

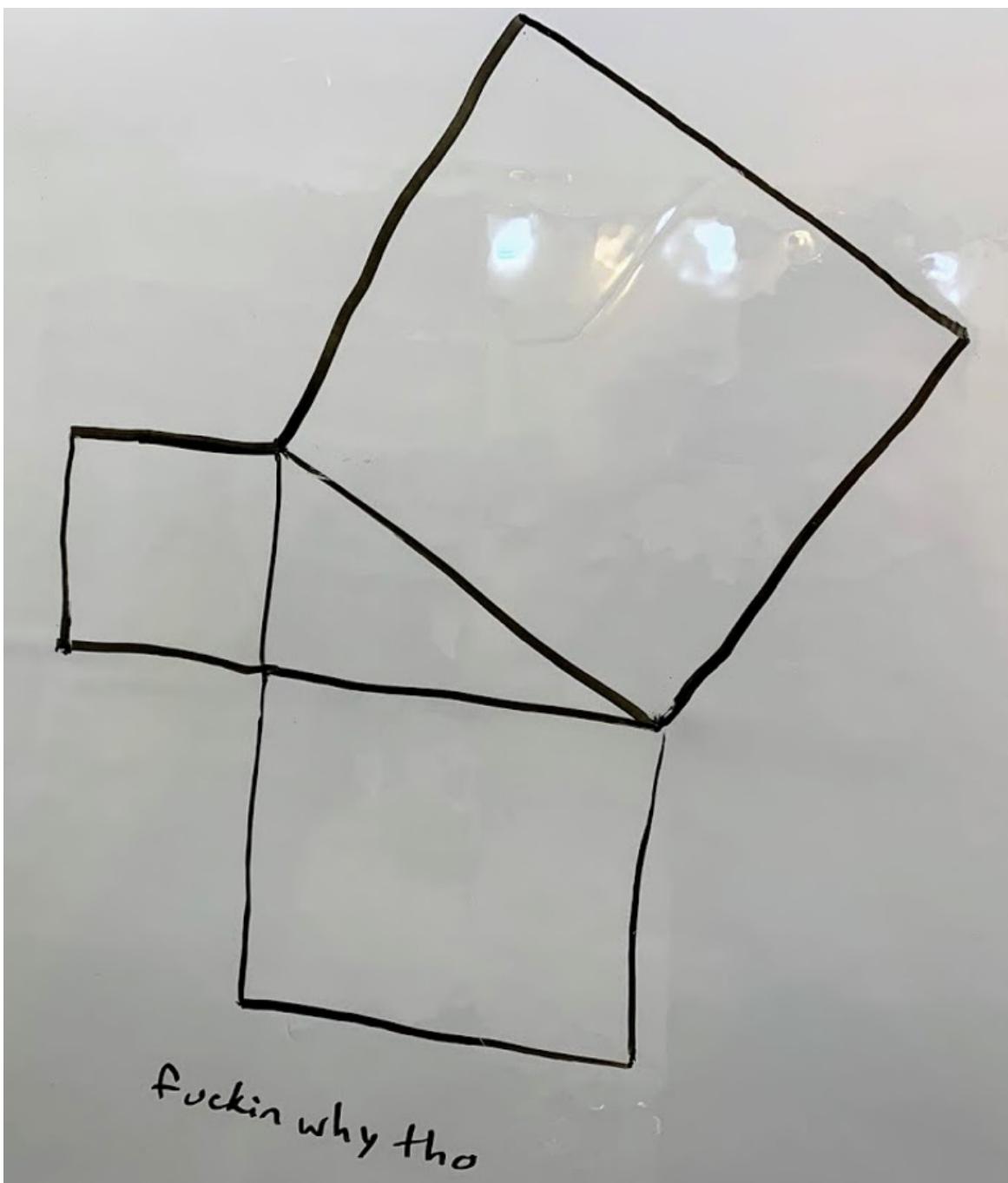
Something about "measure" or "physical dimensions"? Not quite, but "physical dimensions" is moving in the right direction.

The picture I drew on my whiteboard of the triangle representing the roof, that feels important. I remember tilting my head and puzzling at it. I thought things like, "Why does this work?" and "Why does it have to be this way?".

"Inevitability." That one clicks.

"The inevitability of physical relationships." Yes.

"Discovery." "Proof." "Comprehension." "The real world." All of these things.



Let's try a sentence, now that the short phrases are clicking: I think I'm interested in geometry because... I want to discover the inevitability of physical relationships?

Almost. I would like... I want a different relationship with the physical world than the one I have. I want to have a kind of relationship where I believe deep down not just that it's governed by laws, but that I personally can discover those laws and prove the law-ness of them. And I think facility with (Euclidean?) geometry could shift me toward that relationship.

Yeah, that feels right.

My story (a summary of all of that): I want mastery over the relative arrangements of physical properties.

So I started with a vague spark of curiosity, and now I have a clear articulation I can work with, like a flame on a single tiny twig.

Now that you've watched me demonstrate with a specific example, I'll go back through and explain what I was doing in more general terms.

Instructions for Articulating Stories

1. **Find a felt sense of the general topic you're interested in.** It doesn't matter if you don't know what it's called, or how to think about it. Just reach out with whatever part of you feels like reaching, until you recognize a sensation like, "Ah, yeah, something I care about is going on over there".
2. **Offer some words and images to your felt sense:** words like "learning", or "overwhelm", or "that thing that happens to people when they spend too much time reading analytic philosophy". You're not necessarily looking for a single word or phrase that clicks. You're just locating relevant concepts and finding plausible handles for them.
3. **Weave some of those handles into a statement about your topic.** I'll call this your "story". You'll treat your story as (roughly) a hypothesis.

As an alternative: Your story does not actually have to be made of words. It could be a series of images, or maybe even a numerical model of some kind.

Or it could be highly poetic, an ungrammatical string of metaphors.

What matters is that you experience it as asserting something about how the world is, and that it be made of parts you can consider independently.

4. **Tinker with your story statement** until it has the following properties:
 1. It is propositional (something that could be true or false).
 2. It is a short, big-picture summary of what you sort-of-probably believe about the thing you're interested in.
 3. Most importantly, it captures the core of the felt sense you've been talking with.

You've probably articulated your story correctly and completely if

- each part of it emotionally impacts you at least little as you read it, and
- when you ask the felt sense, "Is anything missing?" you get a "no" or a sense of satisfaction.

(Throughout this essay, I'll present instructions as well-ordered lists. In practice, the orientation process tends to be a lot more fluid than that. If you find yourself jumping back and forth between steps, taking things out of order, or inserting new steps I failed to mention, that's totally fine. In fact it's exactly what I hope you'll do. These numbered lists are here for ease of reading.)

Example 2: Courage

(I worked through the geometry example in real time, but for this one I'm using a months-old memory. It's a true story, but don't trust the details.)

A while back, I was watching [this video](#) of ten-year-old Alex Shumaker playing the drums. And I was feeling some things about it.

The things I was feeling were happy and sad at the same time. My chest was warm and light, buzzing with electricity, but also constricted. My throat tightened, my eyes teared up. I felt admiration, love, and excitement. I also felt envy, regret, and longing. There was something *important* there. I wanted to know what, and why, and how it works. I wanted to catch the spark.

I was sad because I wanted so badly to be like *that*, and knew that I wasn't. And I was happy because *he was* like that, and maybe I could be too.

Like what, exactly? What's the "that" Alex is like?

In the video, he's playing "Don't Stop Believing" by Journey. It takes a while for the drums to come in, so at first he's sort of just sitting there. If I were in that position, I'd be frozen, awkwardly gripping my drumsticks and sweating as everyone watched.

Only he's *not* just sitting there. He's moving around, singing along, and twirling his drumsticks. He hasn't even started and he's already so *alive*.

"Alive" fits, the click of a focusing handle. But that's not all of it. As I watched, I had a battle going on in my head.

(*Before the second verse, he spins his drumstick six times, then hits the [high hats \[3\]](#) twice.*)

Part of me said, "This is the way to be. Alive and free and unafraid."

(*Finally the drum part comes in, and Alex finds a rhythm on the snares.*)

Another part of me said, "He's like that because he's a child, and children don't know about danger yet. Anyone can be fearless when they don't see any reason to be afraid."

(*On the word "night", he tilts his head back to sing with his mouth wide open, puts one hand over his heart, and continues drumming with the other hand.*)

"But the dangers that make me afraid are real, and I am not blind to them. I can't be like Alex without somehow becoming ignorant again."

(*Then he jumps to come down on two high hats at once.*)

But the part of me that said "this is the way to be" would not give up. I was just so sure of it, so sure that there must be some way not to give up on ourselves just because we've grown up and learned more about the world.

I kept thinking about the video for days after, and each time, it got a little louder: "Maybe it is possible."

Maybe it is possible to be alive, even though.

Maybe it is possible to move without restraint, when you know that the danger is real.

Maybe when there is danger on all sides, and you have to act decisively while accepting risk, it is necessary to be like Alex, even while you are afraid.

For a week, these concepts swirled around in my head, chasing each other, weighing, asking questions: "Freedom." "Maybe it is possible." "Fear." "Maybe it is necessary." "Alex playing the drums." "Freedom in the face of fear." "The person I want to be." "Courage." "How?"

So after a few days, I found this story: "To be myself, in the world I find myself in, I will have to learn courage."

Part 2: Squinting At Stories

Once you've articulated a story, it's time to transfer the flame from the tiny twig onto a torch that is more durable and mobile. You'll do this by squinting at it. Uncovering its hidden assumptions. Inviting it to tell you way more about itself than you originally wanted to know.

This gentle interrogation has two purposes.

The first purpose, accomplished by listing assumptions implicit in your story, is to de-anchor yourself from your most familiar way of thinking about the topic.

Why do this? [Eliezer wrote](#), “Curiosity requires both that you be ignorant, and that you desire to relinquish your ignorance.” But also, curiosity is most effective when you *recognize* your ignorance.

I believe that typos exist in this essay. I also want to correct them. But in an ordinary, habitual frame of mind, it's difficult for me to recognize the typos when I see them.

Why are they difficult for me to recognize? Because I know what I meant to say, and I've rehearsed these sentences many times. The typos are invisible to me as long as I am hallucinating my expectations onto the page. If I *really* want to give myself an opportunity to notice where the symbols on the page differ from my expectations, I will have to do something a little unusual (such as reading the essay backward).

Concepts we have rehearsed many times tend to feel both unitary and inevitable. A year or two after learning the alphabet song and singing it over and over, I thought that one of the letters was called “elemeno”. When I got more serious about learning to read, I asked my teacher, “Which one is elemeno?” and I could not make any sense of her response. It took me a while to figure out that elemeno is in fact *four completely different letters*.

A couch that gets stuck no matter how you try to shove it through your doorway may pass through easily once you've separated the backrest from the seat and arms.

The story you articulated in Part 1 is a product of your habitual patterns of thought and observation. It's the way the world seems to you now, given all the grooves you've worn into your perceptual systems. But the real world is far more detailed.

By listing assumptions implicit in your story, you find the pieces your story is made of, and you begin to take it apart. The more pieces you break off, the more opportunities you create to consider the components independently, or in novel arrangements. You may begin to recognize your ignorance in places you never knew existed.

This approach (and the rest of naturalism) is especially useful when you'll need a lot of original seeing to get anywhere worth going. But no matter what you're hoping to learn, the efficacy of your curiosity will increase with even just a little bit of original seeing.

The second purpose of the gentle interrogation, accomplished by brainstorming questions, is to educate yourself about *where you'll need to look* if you want to *shift* your understanding toward a more accurate representation of the world.

Unlike assumptions, questions point outside of themselves. They are like charters for nautical expeditions.

By asking questions *inspired by* assumptions, you can follow those questions out into the world, where you'll encounter data related to your assumptions—and encounters with data related to your assumptions are often opportunities to change your mind.

Example 1: Geometry

My story: I want mastery over the relative arrangements of physical properties.

First I will list the assumptions implicit in this story, or in the parts of me that produced the story.

Assumptions:

- I mean something by "mastery".
- Physical properties are arranged.
- I care about relative arrangements, as opposed to absolute arrangements.
- The thing I want is a kind of mastery.
- There's a thing I want mastery over.
- I'm interested in physical properties, as opposed to something else.
- I'm interested in physical properties.
- There are "physical properties".
- Physical properties have arrangements.
- Physical properties have relative arrangements.
- What I am doing at mastery over the thing is "wanting".
- The thing I am wanting mastery over is "the relative arrangements of physical properties".
- Mastery is something you can have "over" things.
- It's not just physical properties I want mastery over, but their relative arrangements in particular.
- There are things with physical properties.
- There is a space in which physical properties are arranged.
- "Physical" is a thing.
- "Relative" is a thing.
- "Properties" is a thing.
- There's a kind of relationship I can have with something that I call "mastery over".
- The space in which physical properties are arranged is a kind of thing I can relate to in a mastery-over sort of way.
- The difference between what I'm interested in and "physics" has to do with a focus on "relative arrangement".
- "The relative arrangements of physical properties" is a thing.
- "Geometry" has something to do with the relative arrangements of physical properties.

(This kind of list can go on indefinitely, but for me, five to ten minutes of listing is usually enough.)

Next I will tilt my head at these assumptions, wondering to myself "What is this? What's going on here? Is this right? Is it really so simple? Could I be confused somehow?".

I won't write about my experience of tilting my head at every single one of these. That would take too long, and I usually skip over some anyway (prioritizing by aliveness). I'll just pick a couple to show you.

"I mean something by 'mastery'."

- Hm. (For the record, I literally tilted my head just now, not on purpose.) If this is true, I wonder how clear I am on what it is I mean. If it's false, then what am I doing with this "mastery" thing, other than meaning something by it? I do suspect it's possible to accidentally pepper your thoughts and sentences with "meaningless" words, words that serve some sort of auxiliary purpose. I don't *think* that's going on with "mastery"? But I am a little suspicious of this word. I feel as though I'm packing an awful lot into it, and it seems like the stuff is sort of swirling or tangled up. What is the stuff, and how easily does it untangle?

"Physical properties have arrangements."

- "Have"? Really, they "have" arrangements? Where do they keep them? I wonder what "properties" is doing that isn't covered by "physical". Arrangements like the things you do with funerals and wills and stuff after someone dies? Arrangements like the thing you're threatening to change when you tell someone you'll rearrange their face? Arrangement seems to have something to do with order. But is it about lower entropy, or is it more a description of the relationships among parts of a system, whatever its entropy? Is that why I talk about "relative" arrangements? But then why do I bother specifying "relative", if it's part of what I mean by "arrangement" already? What kind of "system"? Is a triangle a system? Multiple parts of me are trying to call bullshit on "physical". One part says that nothing isn't physical. Another part says that since mathematical objects like circles obviously aren't physical, it isn't physical properties I'm interested in at all. I also call bullshit on knowing what mathematical objects are. Maybe I should have said "properties of physical objects" instead of "physical properties". Anyway, whatever these things are, I seem to think that they're arranged, and it's interesting to think of non-physical things as having arrangements. I don't know what "physical" means. Maybe I should try to make one list of definitely physical things, and another list of definitely not physical things, and then see how confused I feel afterward.

And so forth.

I don't tend to think in words, and for me it doesn't take nearly as long to tilt my head at one of these as it takes to read all of that (let alone to write it). But what I do with my mind at least has the flavor of those paragraphs.

When I look back over my list of assumptions, after tilting my head at them, I notice that my interest snags most on

- "The thing I want is a kind of mastery."
- "There is a space in which physical properties are arranged." (When I got to this one, I said "ohHO!" out loud.)
- "'Geometry' has something to do with the relative arrangements of physical properties."

When I say that my interest snags on these assumptions, I don't necessarily mean that I'm more skeptical of these than of the others (though I may be). It's more like, "There is definitely something interesting going on here, and I want to know what it is."

Besides "snag", another way to describe it is "drawn into". I am *drawn into* these assumptions by my interest, my curiosity, the center of my quizzical squinting. I'm eager to engage with them.

Finally, I make a list of questions. I focus especially on those snaggiest of assumptions for input.

Usually I don't write anything down during the head-tilting phase, but you can see in my accounts of head-tilting that this attitude generates many questions. The purpose of the list is to articulate and capture them. I'll just put a few here.

- Are there other kinds of mastery?
- What do I mean by mastery?
- Do I want a thing?
- What do I want?
- How can I tell when I've mastered something?
- Have I ever mastered something?
- If I wanted to master something, how might I go about it?
- Is "the relative arrangements of physical properties" a thing?

- What kind of a thing is it?
- Can physical properties be arranged non-relatively? Absolutely? Are there absolute arrangements?
- What sort of properties am I talking about here?
- What exactly does any of this have to do with triangles, quadrilaterals, conic sections, coordinate planes, and so forth?
- And what does it have to do with Euclidean axioms and postulates?
- What is the relationship between Euclidean axioms and the relative arrangements of physical properties?
- What is "proof", and why does that feel so closely tied to "mastery"?
- Why do I want to prove things?
- What's so juicy about the Pythagorean theorem in particular?
- What is this elegance business I like so much, and why does geometry seem like a good place to look for it?
- What is geometry?
- How is geometry different from topology?
- I've read that topology is a generalization of geometry; if so, does that mean a geometry is a topology with certain properties? Which properties are they?
- What would happen if I tried to understand Euclidean geometry very backward, like maybe from a perspective of topology and model theory?
- What would happen if I tried to understand geometry very forward? What would that mean? Would it mean studying the tree outside my window? Or building a shed? Or starting with algebra?

As you can probably see, this list could get very long, and could roam around a lot. But it's also going in a direction, and it sort of has a destination that it's searching for.

I don't know how apparent this is from the outside, but to me, the questions change as the list goes on. They become more spring-boardy. I started with questions like, "What do I mean by 'mastery?'?", which had a little bit of a diligent feeling behind it. Gradually, I narrowed in on what I'm most curious about. As I did so, avenues of investigation presented themselves ("What's so juicy about the Pythagorean theorem in particular?"). Subsequent questions tended to have a little bit more of whatever propels me to stick my hands in the dirt and figure out what is *up* with the world ("What would happen if I built a shed?")—and specifically that part of the world picked out by the felt sense in Part 1.

So I ask a lot of questions, without much of a filter, but I also feel for what matters as I go.

Instructions for Squinting At Stories

1. **Make a list of assumptions** you think are implicit (or explicit) in your story.

As you do, spend a little time on each substantive word (or symbol) in your story statement.

Also, bother to write down the blindingly obvious. Include stuff that's so apparent, you tend to experience it as invisible. You might by default be inclined to restrict yourself to assumptions that seem shaky. Don't.

(If you find that this, or any other part of Squinting At Stories, is taking too long, skip down to the appendix for troubleshooting.)

2. **Look over your list of assumptions, and [tilt your head \[4\]](#) at some of the items**, like a puppy who's just heard a kazoo for the first time. For each one, try on an attitude of, "What is this? What's going on here? Is this right? Is it really so simple? Could I be confused somehow?"

Pay attention to which ones your interest snags on. Some of them will catch your attention more easily than others. Notice which assumptions stand out to you.

3. Brainstorm questions around your topic.

How many questions to brainstorm depends on a few factors. The main ones are how important the topic is to you, how much time you have to spare, and how easily the questions are coming. Usually I decide whether to set a five minute timer or a twenty-five minute timer to start, and commit to brainstorming for the duration. If I get to the end and want to keep going, I set another timer. Otherwise, I move on.

You've probably squinted at your story correctly if

- you have a sense that, in several respects, your topic is rich in fascinating detail, and
- you're eager to uncover its secrets.

Example 2: Courage

(I don't have notes from when I did this while first studying courage, and it happened in a slightly different form. I've done my best to remember my old perspective and recreate the thoughts, but I suspect there's a lot more of my current perspective in here than my old one.)

My story: To be myself, in the world I find myself in, I will have to learn courage.

Assumptions

- I want to be myself.
- I am something.
- I mean something by "be myself".
- I find myself in a particular kind of world.
- The world seems at odds with "being myself".
- I mean something by "the world".
- I "find" myself in the world.
- I am in the world.
- Courage is learnable.
- There's no other way to be myself in the world than to learn courage.
- Courage somehow makes being myself compatible with being in the world.
- The world and I are incompatible while I am not courageous.
- I do not currently know courage.
- I mean something by "learn courage".
- Courage is a thing.
- There is some kind of necessity in my future learning of courage.
- It is possible to be other than myself.
- If I'm living in the world and I am not courageous, then I am being something other than myself.
- It is possible to not be courageous.
- It is possible not to have learned courage.

Some thoughts I have while tilting my head at a few of these assumptions:

- "Courage is learnable.": Learnable by whom? What makes something learnable or not learnable? Is there such a thing as "not learnable"? What exactly am I saying is learnable? What do I mean by "courage"? Are there some parts of it that can be learned, and some that can't? Does it break down into parts? Is it a coherent thing that easily comes apart, or more of a swirling cloud that will dissipate on closer examination? If it's learnable, what category of knowledge will be gained? Is it a skill? A capacity? A model? A perspective? Something else? Have people learned courage in

the past? How have they done it? I have a hunch that this is a kind of thing people emphasize in *childhood* education, part of “being brought up right” or something; what does that mean, if anything, about what it takes to learn it?

- “I ‘find’ myself in the world.” I couldn’t even get all the way through typing this one in the original list without tilting my head so much that I put “find” in quotation marks. What a strange sentence this is. I find myself in the world. When a person says that they find themselves in this or that situation, they’re emphasizing the accident of it, their having arrived in that situation without having chosen deliberately to get there. The “find” in my sentence here has that flavor. It’s as though I’m trying to deny responsibility preemptively. As though I’m saying, “Hey, it’s not *my* fault I’m here.” I wonder why that matters. Why is that an important part of how things are shaped in my head? And what do I mean by “the world”? Do I mean it in the modal semantics sense of “actual world”? I think not. But what *do* I mean? What aspects of the world do I care about here? What state of affairs do I have in mind when I say “the world”, while thinking about courage? And what is it I am finding, exactly? Is there something I’m picking out about *me* in particular, or am I talking about any given human? I think it’s a little of both. I wonder what about me in particular is relevant here.
- “Courage is a thing.” Is it? Suppose it’s not a thing. Then what would be true? Some people, myself included, would be confused. Specifically, they’d be confused about something to do with virtue, fear, action, responsibility, conscience, danger, risk, safety, strength, or weakness. Probably they’d be wrong about how those things relate to each other, or to the world, or they’d be wrong about the internal structure of some of those things. Maybe they’d be conflating a lot of things that really need to be kept separate to understand the world accurately. Or maybe they’d be making some nonsensical distinctions. But even if everyone is confused enough to believe in courage, the way that some people claim others are confused enough to believe in free will, the concept is still doing something. What might it be doing? How does it influence behavior? How does it influence people’s conceptions of themselves and their communities? Anyway, if courage *is* a thing, then what kind of a thing is it? And how could I know?

Looking back through the list of assumptions, I notice that my attention snags on

- It is possible to be other than myself.
- Courage is a thing.
- I find myself in a particular kind of world.

Questions

- What is courage?
- Is courage a thing?
- How can you tell when someone is being courageous?
- Are there things that look similar to courage but are actually different?
- What kind of a thing is courage?
- What can be courageous?
- Is there a difference between being courageous and behaving courageously?
- Is courage a property of actions?
- What does it mean to “learn” courage?
- Can I “know” courage and choose not to be courageous?
- Is every action either courageous or cowardly?
- Is cowardice the opposite of courage?
- What is cowardice?
- What does courage have to do with fear?
- What does it mean to be myself?
- If I suppose that “myself” is a thing, and that it’s made of stuff I already think exists and am not even opposed to, what kind of a thing is it?
- What do people mean when they say “just be yourself” or “I was not myself [when I yelled at you while drunk]”?

- Where does courage come from?
- Why might courage matter?
- If a world is such that I can't be myself without learning courage, what might be the properties of that world?
- Why do I think that maybe I can't be myself in this world without courage?
- In what respect do I "find" myself in this world? What other ways can one be in a world?
- Do I perhaps mean one very specific part of the world? Society, for example?
- What am I scared of?
- What is fear?
- How can I tell when I'm afraid?
- What happens when I'm afraid?

Part 3: Choosing Your Quest

After articulating your story and then squinting at it, your spark has (hopefully) become a steady flame atop a torch.

Next, you will set out to illuminate some part of the world you think could teach you something.

You'll do this by finding a conceptual crux.

I don't think I mean "crux" in *quite the usual CFAR sense*, but I mean something very close. Rather than "a proposition that some other particular belief rests on", what I'm talking about is "something near the foundation of your whole way of thinking about things" (where "things", in this case, is everything to do with your topic).

Naturalism is not very interested in improving specific beliefs. If grokking whatever it is you want to grok were as simple as flipping the truth values you assign to particular propositions, or even adjusting the distribution of probability mass, naturalism probably wouldn't be the most efficient method for you. That would mean you're merely incorrect, rather than deeply confused or lost.

Naturalism is more interested in improving the concepts that your beliefs are made of. It helps you recognize when the world is shaped in a fundamentally different way than you originally thought—like when elemeno turns out to be L, M, N, and O.

Your quest will be a question with the potential to do *that*.

And once you have such a conceptually crucial question, you'll set out to answer it, packing your other questions like provisions for your journey.

Example 1: Geometry

My story: I want mastery over the relative arrangements of physical properties.

I've taken a moment to get back in touch with the felt sense behind my story. I can tell I'm in touch with it because several of the words are hitting me in the chest: "mastery", "relative arrangements", and "physical properties" all go "Pow pow pow! Yeah, those things!".

So now I will look over my list of questions.

The first one that jumps out to me a little is, "What do I mean by mastery?". It jumps out at me because when I imagine seeking the answer to that question, the person I am afterward is in a better position to understand the thing I care about. They're closer than me to the region of territory picked out by the felt sense I am holding. "Yes, crucial data over there!" is the name of the feeling, for me.

But it's not a very strong feeling, for this particular question.

"What kind of thing is [the relative arrangements of physical properties]?" That one jumps out a lot.

When I offer this question to the felt sense behind my story, the connection is like a gigantic tome falling open. I have a hunch that whatever is over there, it will not only result in me gaining crucial information, but will also lead me to ask many new questions that I can't even conceive of yet, much better ones than I've asked so far.

I get about the same thing with, "How is geometry different from topology?"

The second and third of these both seem like fine quest objects, but I suspect that I'll get further faster with the second. "What kind of thing is [the relative arrangements of physical properties]?" is the sort of question I expect any high school geometry textbook could help me with, for example. It also seems like looking at a tree all by myself could be productive.

By contrast, I expect the third one ("How is geometry different from topology?") will require reading stuff written for people with very different math backgrounds than mine.

So I have chosen the first quest of my geometry adventure:

My quest: What kind of thing is the 'relative arrangement of physical properties'?

Instructions for Choosing Your Quest

1. Set aside all the words from part 2.
2. Look back at your story from part 1, and re-connect with the felt sense that generated it.
3. Hold onto that sense, and look over your list of questions.
4. Imagine trying to investigate each question. Notice when you expect that pursuing an answer to one of them could change the way you think about or relate to your topic—when it could *change your story* about what's going on. (Those questions are conceptually crucial.)
5. Consider each crucial question, noticing what you would most enjoy attempting to answer, and what could get you traction quickly.
6. Choose your favorite question as your quest.

As an alternative: If you have multiple excellent questions and don't know which path to take, you could plan to spend one short block of time investigating the first, and another short block of time investigating the second. This might give you more clarity on how to proceed.

You've probably chosen a good quest if,

- when you hold the question up to the felt sense of your story, it's clear that the question *matters*,
- you're at least a little excited, and
- part of you is already making plans to investigate.

Example 2: Courage

My story: If I want to be myself, in the world I find myself in, I'll have to learn courage.

It's remarkable how much fear is in the felt sense of this. And resolve. "If I want to be myself" is somewhere between an invitation and a challenge. "In the world I find myself in" comes with this image of swirling white electric chaos, with my body in the middle. "I'll have to learn courage" is like pressure and heat in my chest, and an image of a rhinoceros lowering its head toward the ground and stamping the dust as it prepares to charge. Overall, the story comes from a feeling of opposition to limpness, plus concern for not just myself but also the society around me and where we're headed.

As I look back over my list of questions, the ones that feel most likely to quickly change how I relate to the topic are

- What kind of a thing is courage?
- What happens when I'm afraid?
- What does it mean to be myself?

Of these, I think I'd like to start with, "What happens when I'm afraid?" I like the groundedness of it. It's just so obvious where it's suggesting I look.

Plus, I'd be shocked if I didn't learn something crucial about courage while improving my understanding of fear.

My quest: "What happens when I'm afraid?"

Reflection and Conclusion

Why this particular way?

The orientation procedure is meant to prepare you to drive with curiosity toward a more intimate relationship with a region of territory where crucial data are likely to live.

Which is a bit of a mouthful. Let's break it down.

Why use "curiosity" to drive?

One way to categorize people is to ask what usually drives them to learn.

Many things can be learned out of duty, out of ambition, or out of a desire for the problem to be solved. But naturalism is geared for situations so novel, fraught, or hopeless that your whole way of thinking about them is totally inadequate. In those situations, there's no way around it: You need rationality.

And there's a reason curiosity is the first virtue or rationality.

With duty, there's a particular thing you have a duty to do, and your thoughts can hang themselves on that frame. If you can't conceptualize your duty, though, you won't know what is right or wrong, so there's no way to steer.

With ambition, your frame is made of the abilities, skills, or resources you want to command. But if you're too lost to know which resource you're after, ambition won't help much either.

With goal-oriented problem solving, your frame is a pre-defined problem. But if you used broken concepts to define that problem, you won't find traction, because no ground exists where you'll try to put your wheels.

(I conjecture that this is usually what's happening when people try and fail to solve the same problem over and over again.)

It seems to me that all three of those engines (duty, ambition, and problem solving) share a sort of inward gaze.

Duty is about *you* being a good person. When you use it to drive, there's a constant dialog between your behaviors and what you think is right.

Ambition is about *you* becoming powerful. When you use it to drive, there's a constant dialog between the resources you have and the resources you want.

Goal-oriented problem solving is about what shape *you* want the world to take. When you use it to drive, there's a constant dialog between the state of the world and the one you envision for it.

You might argue that curiosity is about what *you* want to know, but I think I would disagree. That perspective on curiosity is problem solving, ambition, or duty in disguise.

Pure curiosity is about your reflection of the world. Although it tends to take the world piece by piece, it does not fundamentally rest on a desire to know a particular thing for a particular practical purpose. It does not feel like, "If only I knew *this*, then I could accomplish *that*." Instead, it feels like, "I must fill this hole in my knowledge, or I will be forever incomplete."

A person driven primarily by curiosity is so in love with the world that they want to become a perfect reflection of it. And to do that, you put your perceptions of the world in dialog, not with yourself, but with your other perceptions of the world.

So when your concepts are broken, inadequate, or nonexistent, curiosity has a special advantage. It doesn't rely so much on your concepts to judge.

Why drive toward "a more intimate relationship with a region of territory"? What do I even mean by that?

When I was six or seven years old, my dad wouldn't let me open my Christmas presents early. But I wanted desperately to know what was inside the opaque boxes. So I got Dad's blowgun, snuck the presents into the basement, and shot the darts through the cardboard.

In some places, the darts penetrated almost all the way through. In other places, they got stuck on the surface of whatever was in there. The idea was to infer the object from its outline.

This method... needed some improvement, and did not actually cause me to know what I was getting for Christmas. Still, I was literally piercing through a region of territory, over and over again, to learn whatever I could about it through direct observation.

That is the kind of intimacy I'm talking about.

I read a story once about a boy who wanted to master the wind. He went to an old wizard whom he'd heard could direct the wind, and asked to be his student. The boy was ready to do anything: he'd read every book in the library; he'd study late, night after night; he'd listen and research and write essays until he knew every rune, magic word, and alchemical recipe in existence, if that was what it took.

But the wizard refused to lecture. He offered no textbook.

Instead, he made the boy stand naked on the roof of a tall building during a storm. That is the kind of intimacy I'm talking about.

When Darwin went to the Galapagos Islands, he collected many finches, and he measured each of their beaks. That is the kind of intimacy I'm talking about.

It's easier to find an answer when you know what the question is. But what about when you don't? What if you *know* that you're as yet unable to formulate the question whose answer you need? How can you make progress anyway?

What you can do is cut out as many intermediaries as possible.

Do not ask an encyclopedia to tell you *about* the world. Do not think harder and longer, waiting for something to emerge from the implications of your existing beliefs. Instead, seek out opportunities to encounter crucial data. Directly expose your own sensorium to reality. Become *intimate* with the territory.

When successful, this orientation procedure draws out your own felt senses and hooks them up to your own curiosity. By this method, it leaves you undisguised and naked, with a longing for the world to impact you. It prepares you to learn without mediation, which is essential when your intermediary concepts will lead you astray.

Why go "where crucial data are likely to live"?

If you want to learn about birds, digging in the dirt [won't \[5\]](#) help.

To get anything in particular out of directly exposing your own sensorium to reality, you have to be strategic about which bits of reality it's exposed to. You're bound to learn *something* while in contact with reality, but what you learn will depend on where you are. You can see and process thousands of things without ever getting close to anything that might answer your questions.

If you want to draw a map of the coastline, it's necessary but not sufficient to have your eyes open. You also have to go to the sea.

What comes next?

Once you've stopped superimposing so many preconceptions onto your observations—once you've lowered your broken map—you can turn your curiosity toward the mountains and shadows that currently surround you. You can look at them through a longing to understand something, which is how you will deliberately move toward a destination.

That's what it's like in *my* mind, anyway, at the end of this orientation procedure.

But you're not me. You'll probably have to do something at least a little different to get all the way there. The goal is to look directly at the world with genuine curiosity and awareness of what you care about. How you do it is up to you, but that's what's needed to move on.

The rest of naturalism is about figuring out how to go where crucial data live when the map you're starting with is all wrong.

In the next essay (unless my plans change before then), I'll show how I begin to seek out experiences that are rich in crucial data.

Appendix

A Note On Duration

For both of the below examples, I followed my instructions from this essay. I also time-boxed each step to five minutes (five minutes for story articulation, five minutes for assumptions, five minutes for head-tilting, five minutes for questions, five minutes for quest choosing.). Only "assumptions" and "questions" took the full five minutes (and I could easily have kept going).

If it's taking you a lot longer than this, then either you feel good about how long it's taking and you should carry on, or you feel bad about how long it's taking and you should probably change something.

If you want to change something, maybe try one of these:

- Do a time-boxed version and run with whatever results from that, even if it's not perfect.
- Warm up with a brainstorming exercise: Write down as many animals as you can in one minute. Make a few notes about what strategies you used for coming up with animals, and what other strategies you could have used but didn't. Then spend another minute naming animals. When you're done, come back to the "assumptions" and "questions"

sections with the same mental posture you used in your second minute of animal naming.

- Skip "Squinting At Stories" entirely, and try the whole thing again after you've been questing for a week.

Example 3

Story: *When I'm trying really hard, I'm afraid that if I let go, things will fall apart.*

Assumptions

- Sometimes I try really hard.
- Sometimes I try.
- There's a difference between trying a little and trying "really hard".
- Sometimes I try only a little.
- At some point I've experienced a thing in my mind and/or body that I'm now calling "trying really hard".
- When that "trying really hard" thing happens, there's something I'm afraid of.
- The thing I'm afraid of while trying really hard is what will happen if I "let go".
- If I let go, things will fall apart.
- There's a thing I can do called "letting go".
- Things can "fall apart".
- There's a particular thing I mean by "fall apart".
- Sometimes I'm afraid.
- Sometimes I'm afraid of something that will happen if I act in a certain way.
- I don't want things to fall apart.
- I want something besides "things falling apart".
- I think something bad will happen if I let go.
- I think something bad will happen if things fall apart.
- I think something will happen if I let go.
- It's important to me that I prevent whatever it is I think will happen if I let go.
- It's important to me that I prevent things from falling apart.
- The fear of letting go happens specifically while I am trying really hard.

Questions

- What does it feel like to "try really hard"?
- What does it feel like to "try"?
- Are there multiple kinds of trying?
- Are there kinds of trying that do not fall on a linear scale from "only trying a little" to "trying really hard"?
- Do the different kinds of trying feel different?
- Do the different kinds of trying have different purposes?
- When I try only a little, is there fear?
- When I try a lot, is there fear?
- Is there a difference between trying "a lot" and trying "really hard"?
- What is "trying"?
- When I'm trying really hard, what is the fear like?
- What is fear usually like?
- What is fear trying to do, in general?
- What is fear trying to do while I'm trying really hard?
- When I'm afraid of something while trying really hard, am I imagining a particular state of the world, or is something else going on?
- What would it mean for "things" to "fall apart"?
- Which "things" do I not want to fall apart?
- Have I ever experienced "things falling apart"?

- If things fell apart, how would I know?
- Can one individual thing fall apart, or is it some kind of systemic property?
- What is the bad thing I think will happen if things fall apart?
- Can I prevent things from falling apart?
- Is it sometimes true that if I let go, things will fall apart?
- What might I do *after* things fall apart, if that happens? How bad is it exactly?
- Does trying really hard prevent things from falling apart?
- What does "trying really hard" actually cause?
- Can I choose whether to try really hard or to let go?
- Is there anything it's almost impossible to do at the same time as "trying really hard"?

Quest: *What happens when I'm trying really hard?*

Example 4

Story: *Seedlings are tricky.*

Assumptions

- Seedlings are trickier than adult plants.
- There's something I mean here by "tricky".
- The trickiness is a property of the seedlings.
- All seedlings are tricky.
- There's such a thing as a "seedling".
- There aren't any seedlings that aren't tricky.
- There are several different ways in which seedlings are tricky.
- Lots of different things can go wrong with seedlings.
- When I want one plant, I should sew ten seeds.
- I don't have much control over how many of the seeds I sew grow up into adult plants.
- It's very hard to help new gardeners grow plants from seed without them becoming discouraged.
- Mold is an especially tricky thing about seedlings.
- It's not that I'm dumb about seedlings, it's that the seedling part of the universe is mysterious.

Questions

- What causes damping-off mold?
- Are there kinds of mold that grow on the surface but don't affect the roots?
- Is there some clever way to keep seedlings moist enough to grow, but not so soggy that mold grows, without manually checking the moisture level every few hours?
- What causes seeds not to germinate?
- How many of the seedlings that never break the surface of the soil do in fact germinate?
- What happens when you spray a lot of copper fungicide on a seedling?
- Are some plants more resistant to damping off mold than others?
- What causes those plants to be more resistant, if so?
- What role does nutrient balance play in the germination of seedlings?
- Is it perhaps better to let seeds germinate in a damp paper towel before transplanting them to soil?
- Is there something about growing in tiny containers that makes things harder, such that more seedlings would prosper if sewn directly into large containers or into the ground?
- How much does the actual germination rate of my seeds tend to differ from the germination rates claimed by professional horticulturists?
- What factors affect seedling health that I've completely failed to consider?
- Could I grow damping-off mold on purpose without even planting any seeds?
- How does industrial agriculture treat seedling care?

- How do horticulturists working with very rare and expensive plants handle seedling care?

Quest: *Can I find a plant that's very difficult to cultivate from seed and raise it to maturity?*

Coda: Empiricism, Objectivity, and the Soul of Science

(You're still here? It's over! Go home.)

(Ok, well, if you insist...)

In earlier drafts of this essay, readers asked me, "How do I know what kind of spark I'm supposed to catch?"

This is a straightforward question to which I unfortunately do not have a straightforward answer. The closest I can come to a short answer is, "Try catching any spark whatsoever, and see how it goes," which I recognize might not be satisfying.

The realer answer is, "Wrong question." I think I can explain why it's a wrong question, and I think that in the process of explaining, I can probably give you the thing you were groping toward by asking it. You'll likely get a much better sense of where all this "naturalism" stuff is coming from along the way, too.

It's going to be a long and convoluted process, but I'll give it a shot.

I said that "Catching the Spark" is a text adaptation of the first class in my naturalism program. But that's not *quite* true.

In fact, there's a zeroth-class beforehand that's a combination of "interview" and "going over the syllabus". For people who either decide to continue with the program or are on the fence, I give an assignment to complete before the first proper lesson: "Spend two weeks paying attention to patterns in what you're actually drawn to in the moment. As you do, hold the question, 'What feels important or exciting but also daunting/confusing/impossible?'"

In the first lesson, sparks are drawn from memories of those experiences.

Ultimately, I'm working toward a general approach to applied rationality research and collaboration, so I've tended to encourage people to pursue rationality-related topics. Plus, naturalism is designed to be a method for researching such topics, which may succeed where other approaches are likely to fail.

But I expect this particular procedure will work well for any object of curiosity. When I imagine taking a shortcut like, "Look, if it would show up on your bugs list at a CFAR workshop, it's fine," I feel... disappointed. That scope is so narrow. I really want to find out what happens when lots of people try this in lots of different domains with lots of different goals.

I am interested in combating a kind of listlessness or impotence I perceive among the modern intelligencia, rationalists included.

I have a perception that many of the people around me think that they are *not allowed* to come to conclusions based on their own observations. And as a result, they do not bother to observe carefully or deliberately. Since they have so often rehearsed the assertion that

subjective experience is untrustworthy and anecdote doesn't count, when they want to know something that isn't readily googleable, it does not occur to them to personally check. (That's my story so far, anyway.)

I have a guess about how this has happened.

In the early 20th century, there was a lot of excitement about the scientific study of phenomenology. But the methodology turned out to be really tricky to reconcile with what the larger global scientific project was trying to do. The pendulum swung very hard away from phenomenology, reaching its anti-subjectivity peak in the 60s with behaviorism, to the extent that the very existence of subjective experience was frequently denied, even by psychologists and philosophers.

Today, it's common for people to reject out of hand any empirical observation that is not obtained through double blind randomized controlled trials, peer reviewed, and published in a reputable scientific journal. Many people think that's what science *is*. After all, you can't just believe everything written by some crackpot blogger tinkering in his basement lab.

And, in one sense, they're right. There are good reasons that what we tend to consider "scientific discoveries" are usually the result of a whole bunch of coordination, orchestration, and bureaucracy. Establishing anything with certainty is very hard, and as readers of a rationality blog know better than most, humans are incredibly proficient at duping ourselves into confirming whatever it is we wanted to believe all along.

But in the early history of modern science, circa the founding of the Royal Society of London and a couple hundred years afterward, people didn't really know that yet. And science looked very different.

John Aubrey spent a bunch of time walking around the Avebury henge monument recording details of the structures and their surroundings. In 1663, he presented his findings to the Royal Society, thereby creating the modern field of archeology.

In 1665, Antoni van Leeuwenhoek and Robert Hooke shared their drawings and descriptions of things they observed while looking at a whole bunch of stuff under microscopes, thereby creating the modern field of microbiology.

In 1752, Benjamin Franklin described how he captured electrical charge by attaching a kite string to a Leyden jar and flying the kite near storm clouds, which strongly suggested the electrical nature of lightning.

In the late 1700s and early 1800s, John Audubon attempted to make a complete pictorial record of all the bird species of North America, discovered twenty five new species in the process, and dramatically advanced the field of ornithology.

In 1830, Charles Darwin set out to study geology by observing and documenting volcanic activity in the Galapagos Islands, became very curious about differences among local finches along the way, and ended up completely revolutionizing natural history itself.

I think that as a species, we sacrificed something vital in our wholesale rejection of scientific phenomenology. We lost the feeling that when you aren't sure what to believe, and there's no reputable publication around to tell you, sometimes you can,

personally,

go check.

The world discussed in all those Google Scholar search results is in fact the very same one you yourself are sitting in right now. When you wonder how the world is, it is often possible to just... look.

But what does that have to do with phenomenology? Isn't phenomenology about what happens inside your own mind? What does it have to do with understanding the mind-independent natural world?

This is a lot like asking what eyes have to do with literature.

In a sense, eyes have nothing to do with literature. Literature is stories and essays and poetry. It involves things like plot tension, and narrative structure, and linguistic rhythm. It exists in the realm of language, not the realm of optics or physiology.

But if you're severely far-sighted and you lose your glasses, it will suddenly become immediately obvious that eyes have *everything* to do with literature. Without eyes (or ears, or fingertips, depending on the format), you have no access to literature. Written language is universally mediated by vision.

Similarly, everything you can possibly interact with—not just cognitive phenomena like confusion and defensiveness, but also the fraying threads of a kite string as they stand on end with static charge, and the liquid crystal display screen on your calculator as you double check the statistical claims of an epidemiological study—presents itself to you as immediate subjective experience.

So even if you can coordinate all the lab subjects, and technicians, and grant committees, and publications, to minimize observer bias as you try to learn whatever it is you want to know,

it still seems

rather irresponsible, to me,

to completely neglect the practice of phenomenology, all together.

It *also* seems irresponsible to routinely learn through direct observation without getting very good at taking your subjective experiences as object, examining them in detail, and taking as little as possible for granted.

One of my favorite things about the rationality community is its recognition of Bayesian evidence.

Unlike most other modern cultures, it understands that probabilities exist independently of authoritarian decree, that it's possible to update incrementally on many pieces of relatively weak evidence. I like this for several reasons, but one is that it means a rationalist (theoretically, at least) has a lot more personal power to learn things on their own, without the support or hindrance of large institutions. It makes them more like Ben Franklin and Robert Hooke, and less like Lab Technician Number 73.

Don't get me wrong, I think it's good to be Lab Technician Number 73. It took many such people to build the vaccines for covid-19. We all understand the value of that work.

But I happen to really like Ben Franklins, and I think we need more of them than we have—especially in the field of applied rationality, which is as much the Wild West as volcanic geology was in Darwin's day.

I think that more rationalists could be more like Ben Franklin. But as a culture, we're so stuck between our fetish for empiricism and our fetish for objectivity that we don't even notice the way in which empiricism is *opposed* to objectivity.

So when we feel a spark of curiosity, and consider setting out to personally investigate, we find that our shoelaces are hopelessly knotted together.

I would like to un-knot our shoelaces.

You ask me what kind of spark you're supposed to catch? I ask you who is doing the supposing, because it sure isn't me.

The appropriate topic for your study as a naturalist is between you and God. Hook wondered what fleas look like up close. I've wondered what courage is for and how it works. My dad raises salamanders and studies their genetics. As long as you're willing to look at the real world with your own eyes, curiosity can fuel your project, provided you're bold enough to catch the spark and take it somewhere worth going.

I now do this kind of work independently. If you like it and want there to be more, you can support me through [Patreon](#).

Matt Levine on "Fraud is no fun without friends."

This is a linkpost for <https://www.bloomberg.com/opinion/articles/2021-01-13/fraud-is-no-fun-without-friends>

Someone recently told me "Matt Levine's Finance Newsletter is really good", and then I [signed up](#). Most post so far are about... well, I don't actually know the jargon to say what they're about, but "good typical econblogger stuff" I guess.

This is an excerpt from a particular newsletter that seemed to tie into other LessWrong interests.

The link is paywalled. If you sign up for the newsletter you'll get future essays via email for free. I'm not entirely sure about whether it seemed reasonable to quote this substantively, but FWIW I endorse signing up for future newsletters.

The way a lot of financial crime works is by slow acculturation. You show up at work on your first day, bright-eyed and idealistic, and meet your new colleagues. They seem like a great bunch of people, they're so smart and know so much and seem to be having so much fun. They go out for beers after work a lot, and sometimes they let you tag along and listen to their hilarious jokes and war stories.

During the day, they teach you how to trade Treasury futures, and it is all so exciting and high-stakes and important. You shadow one experienced trader and quickly find yourself imitating his mannerisms, looking up to him, hoping to be like him one day. "Here is where I put in some fake orders to spoof the price higher," he says; ["a little razzle dazzle to juke the algos."](#) "Isn't that, uh, illegal?" you ask timidly. "Hahahaha illegal!" he replies ambiguously. You do not press the matter. Three months later you are bragging in the desk's electronic chat room about your own big spoofing victories. As you type "lol i just spoofed em so good hope i dont go to jail" into the chat window, you feel a rush of pride; now you really fit in, you are one of them. You go out for beers that evening and you are the center of attention; everyone congratulates you and celebrates your achievements. It is a great day. Six months later you are arrested.

Now imagine the same story except that you show up at work your first day on Zoom, and your colleagues seem kinda nice but talking to them is awkward and disjointed, and you have no idea what they do after work because nobody leaves their house, but you have a Zoom happy hour once and that's pretty awful. And there is an electronic chat room, sure, and your colleagues make jokes in the chat, but you don't get a lot of them because they reference stuff that happened in the office, in person, before you arrived. You learn to trade Treasury futures by reading some training materials. "I just put in some fake orders to spoof the price higher," says one experienced trader in the chat one day. You frown and reference the training materials, which say "spoofing is super duper illegal and should be reported to compliance immediately." You shrug and send the chat transcript to compliance. Your colleague gets fired and prosecuted. He may or may not feel a sense of personal betrayal that you turned him in, but you'll never know or care.

The SEC [knows what I'm talking about](#):

- > The work-from-home phenomenon has triggered a fresh frustration for U.S. corporations: Americans are blowing the whistle on their employers like never before.
- > The proof is in the data, with the U.S. Securities and Exchange Commission receiving 6,900 tips alleging white-collar malfeasance in the fiscal year that ended Sept. 30, a 31% jump from the previous 12-month record. Officials at the agency, which pays whistle-blowers for information that leads to successful investigations, say the surge really started gaining traction in March when Covid-19 forced millions to relocate to their sofas from office cubicles.
- > The isolation that comes with being separated from a communal workplace has made many employees question how dedicated they are to their employers, according to lawyers for whistle-blowers and academics. What's more, people feel emboldened to speak out when managers and co-workers aren't peering over their shoulders.
- > "You're not being observed at the photocopy machine when you're working from home," said Jordan Thomas, a former SEC official who helped set up the agency's whistle-blower program a decade ago. "It's never been easier to record a meeting when you can do it from your dining room table," added Thomas, who now represents tipsters as an attorney at Labaton Sucharow in Washington.

> Adam Waytz, a psychologist and professor at Northwestern University's Kellogg School of Management, agrees.

> "When you feel disconnected from work, you feel more comfortable speaking up," said Waytz, who has studied the motivations of whistle-blowers.

Also I would guess that somewhat more financial crime is now coordinated via email and electronic chat than it was a year ago, when you could just turn to the person sitting next to you and talk live about your crimes. And if you're going to blow the whistle to the SEC, it helps to have electronic chats to forward to them. Though I would not put too much emphasis on this explanation; traders seem to love talking about their crimes via electronic chat even when they are sitting right next to each other.

I guess this story is good news from a prevention-of-financial-crime perspective, but it is sort of a sad story from a human perspective. All these people feeling disconnected from their work and their colleagues, with no strong personal ties of loyalty and friendship and common mission. Sure the common mission in these particular cases was crime, but still.

One way to read this story is that one sort of business that is conducted at offices is fraud, and people in the fraud business have become 31% less loyal and motivated and conscientious since the pandemic started, which is causing some previously viable fraud businesses to have to shut down. (Because the SEC caught them.) Which is not a loss for society or anything. But the mechanism here, of people feeling disconnected from their jobs and disloyal to their colleagues, is not unique to the fraud business. This story is a sort of leading indicator of a breakdown in morale and group cohesion generally as so much work is done from home. That is probably bad for a lot of projects; it's just that one of the projects it's bad for is fraud.

What is going on in the world?

Here's a list of alternative high level narratives about what is importantly going on in the world—the central plot, as it were—for the purpose of thinking about what role in a plot to take:

- The US is falling apart rapidly (on the scale of years), as evident in US politics departing from sanity and honor, sharp polarization, violent civil unrest, hopeless pandemic responses, ensuing economic catastrophe, one in a thousand Americans dying by infectious disease in 2020, and the abiding popularity of Trump in spite of it all.
- Western civilization is declining on the scale of half a century, as evidenced by its inability to build things it used to be able to build, and the ceasing of apparent economic acceleration toward a singularity.
- AI agents will control the future, and which ones we create is the only thing about our time that will matter in the long run. Major subplots:
 - ‘Aligned’ AI is necessary for a non-doom outcome, and hard.
 - Arms races worsen things a lot.
 - The order of technologies matters a lot / who gets things first matters a lot, and many groups will develop or do things as a matter of local incentives, with no regard for the larger consequences.
 - Seeing more clearly what’s going on ahead of time helps all efforts, especially in the very unclear and speculative circumstances (e.g. this has a decent chance of replacing subplots here with truer ones, moving large sections of AI-risk effort to better endeavors).
 - The main task is finding levers that can be pulled at all.
 - Bringing in people with energy to pull levers is where it’s at.
- Institutions could be way better across the board, and these are key to large numbers of people positively interacting, which is critical to the bounty of our times. Improvement could make a big difference to swathes of endeavors, and well-picked improvements would make a difference to endeavors that matter.
- Most people are suffering or drastically undershooting their potential, for tractable reasons.
- Most human effort is being wasted on endeavors with no abiding value.
- If we take anthropic reasoning and our observations about space seriously, we appear very likely to be in a ‘Great Filter’, which appears likely to kill us (and unlikely to be AI).
- Everyone is going to die, the way things stand.
- Most of the resources ever available are in space, not subject to property rights, and in danger of being ultimately had by the most effective stuff-grabbers. This could begin fairly soon in historical terms.
- Nothing we do matters for any of several reasons (moral non-realism, infinite ethics, living in a simulation, being a Boltzmann brain, ..?)
- There are vast quantum worlds that we are not considering in any of our dealings.
- There is a strong chance that we live in a simulation, making the relevance of each of our actions different from that which we assume.
- There is reason to think that acausal trade should be a major factor in what we do, long term, and we are not focusing on it much and ill prepared.
- Expected utility theory is the basis of our best understanding of how best to behave, and there is reason to think that it does not represent what we want.

Namely, Pascal's mugging, or the option of destroying the world with all but one in a trillion chance for a proportionately greater utopia, [etc](#).

- Consciousness is a substantial component of what we care about, and we not only don't understand it, but are frequently convinced that it is impossible to understand satisfactorily. At the same time, we are on the verge of creating things that are very likely conscious, and so being able to affect the set of conscious experiences in the world tremendously. Very little attention is being given to doing this well.
- We have weapons that could destroy civilization immediately, which are under the control of various not-perfectly-reliable people. We don't have a strong guarantee of this not going badly.
- Biotechnology is advancing rapidly, and threatens to put extremely dangerous tools in the hands of personal labs, possibly bringing about a '[vulnerable world](#)' scenario.
- Technology keeps advancing, and we may be in a vulnerable world scenario.
- The world is utterly full of un-internalized externalities and they are wrecking everything.
- There are lots of things to do in the world, we can only do a minuscule fraction, and we are hardly systematically evaluating them at all. Meanwhile massive well-intentioned efforts are going into doing things that are probably much less good than they could be.
- AI is powerful force for good, and if it doesn't pose an existential risk, the earlier we make progress on it, the faster we can move to a world of unprecedented awesomeness, health and prosperity.
- There are risks to the future of humanity ('existential risks'), and vastly more is at stake in these than in anything else going on (if we also include catastrophic trajectory changes). Meanwhile the world's thinking and responsiveness to these risks is incredibly minor and they are taken unseriously.
- The world is controlled by governments, and really awesome governance seems to be scarce and terrible governance common. Yet we probably have a lot of academic theorizing on governance institutions, and a single excellent government based on scalable principles might have influence beyond its own state.
- The world is hiding, immobilized and wasted by a raging pandemic.

It's a draft. What should I add? (If, in life, you've chosen among ways to improve the world, is there a simple story within which your choices make particular sense?)

Reflections on Larks' 2020 AI alignment literature review

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

This work was supported by OAK, a monastic community in the Berkeley hills. It could not have been written without the daily love of living in this beautiful community.

Larks has once again [evaluated](#) a large fraction of this year's research output in AI alignment. I am, as always, deeply impressed not just by the volume of his work but by Larks' willingness to distill from these research summaries a variety of coherent theses on how AI alignment research is evolving and where individual donors might give money. I cannot emphasize enough how much more difficult this is than merely summarizing the entire year's research output, and summarizing the entire year's research output is certainly a heroic undertaking on its own!

I'd like to reflect briefly on a few points that came up as I read the post.

Depth

The work that I would most like to see funded is technical work that really moves our understanding of how to build beneficial AI systems forward. I will call this "depth". It is unfortunately very difficult to quickly assess the depth of a given piece of research. Larks touches on this point when he discusses low-quality research:

[...] a considerable amount of low-quality work has been produced. For example, there are a lot of papers which can be accurately summarized as asserting "just use ML to learn ethics". Furthermore, the conventional peer review system seems to be extremely bad at dealing with this issue.

Yet even among the papers that did get included in this year's literature review, I suspect that there is a huge variation in depth, and I have no idea how to quickly assess which papers have it. Consider: which of the research outputs from, say, 2012 really moved our understanding of AI safety forward? How about from 2018? My sense is that these are fearsomely difficult questions to answer, even with several years' hindsight.

Larks wisely does not fall into the trap of merely counting research outputs, or computing any other such simplistic metric. I imagine that he reads the papers and comes to an informed sense of their relative quality without relying on any single explicit metric. My own sense is that this is exactly the right way to do it. Yet the whole conclusion of the literature review does rest critically on this one key question: what is it that constitutes valuable research in the field of AI alignment? My sense is that depth is the most valuable quality on the current margin, and unfortunately it seems to be very difficult either to produce or assess.

Flywheel

I was both impressed and more than a little disturbed by Larks' "research flywheel" model of success in AI alignment:

My basic model for AI safety success is this:

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

I was impressed because it is actually quite rare to see any thesis whatsoever about how AI alignment might succeed overall, and rarer still to see a thesis distilled to such a point that it can be intelligently critiqued. But I was disturbed because this particular thesis is completely wrong! Increasing the amount of AI alignment research or the number of AI alignment researchers will, I suspect, by default decrease the capacity for anyone to do deep work in the field, just as increasing the number of lines of code in a codebase will, by default, decrease the capacity for anyone to sculpt highly reliable research artifacts from that codebase, or increasing the number of employees in a company will, by default, decrease the capacity for anyone in that company to get important work done.

The basic reason for this is that most humans find it very difficult to ignore noise. It is easy to imagine entering into an unwieldy codebase or company or research field and doing important work while disavowing the temptation to interact with or fix the huge mess growing up in every direction, but it is extremely difficult to actually do this. It is possible to create large companies and large codebases in which important work gets done, but it is not the default outcome of growth. The large codebases and large companies that are prominent in the world today are the extreme success cases in terms of making possible important work, and, in my own direct experience, even these success cases are quite dismal on an absolute scale of allowing important work to happen.

It is not that large codebases and large companies actively prevent important work from getting done (although many do), it is that most humans find it difficult to do such work in the presence of noise. It is not enough for a large company or a large codebase to provide some in-principle workable trajectory by which important work can get done; it is a question of how many humans are actually capable of walking such a path without being constantly overwhelmed by the mess piling up around their feet.

It is not that we should try to limit the size of the AI alignment field forever. The field must grow, it seems, if we are to stand any chance of success. But we should try to walk along a careful and gradual growth trajectory that maximizes the field's capacity for truly deep research output. While doing this we should, in my view, be clear that among all the possible trajectories that involve growth, most are actively harmful. We should not, in my view, be optimizing directly for growth, but instead for depth, with growth as an unfortunately but necessary by-product.

Policy, strategy, technical

Larks has this to say about publishing policy research in the AI alignment field:

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

The only case I can think of where scientists are relatively happy about punitive safety regulations, nuclear power, is one where many of those initially concerned were scientists themselves, and also had the effect of basically ending any progress in nuclear power (at great cost to climate change). Given this, I actually think policy outreach to the general population is probably negative in expectation.

And he has this to say about publishing strategy research:

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

I basically agree with both of these points, which I would summarize as: Direct engagement with AI policymakers is helpful, but there are not many compelling reasons to publish AI policy work, since the main reason to publish such work would be broad outreach, and broad outreach on AI policy is probably harmful at this point due to the risk of setting up an adversarial relationship with AI researchers. Although high-quality strategy research exists, as an empirical observation it is just quite rare to read strategy research that truly moves one's understanding of the field forward.

My own out-take from these helpful points is as follows: in order to do beneficial work in general, and particularly in order to do beneficial work within AI alignment, begin by working directly on the very core of the problem, using your current imperfect understanding of what the core of the problem is and how to work on it. In AI alignment, this might be: begin by working, as best you can, on the core challenge of navigating the development of advanced AI. In doing so, you may discover that the core of the problem is actually not where you thought it was, in which case you can shift your efforts, or you may discover some neglected meta-level work, in which case you may then decide whether to undertake that work yourself. But in such a complex landscape, if you don't begin earnestly investigating the nature of and solution to the core of the problem then any other work you do is unlikely to be overall beneficial. This is the same "depth" I was trying to point at in the preceding sections.

Scalable uses for money

Larks encodes his conclusions by rotating each letter 13 places forward in the alphabet, in order to discourage us from merely reading his conclusions without engaging directly with the challenging task of formulating our own:

My constant wish is to promote a lively intellect and independent decision-making among readers; hopefully my laying out the facts as I see them above will prove helpful to some readers.

I very much admire this ethos, and will do my best not to undo his efforts, although I do want to comment on one general point mentioned in the encoded text. Larks notes that much of the best research is being conducted within large organizations that already have ample funding, and are neither accepting of nor in need of additional funding at this time. This is both heartening and distressing.

It is heartening, of course, to see important research being funded at such a level that in at least several prominent cases further funding by individual donors is literally impossible, and in several additional cases seems to be explicitly un-sought after by the organizations themselves.

But it is also a little distressing that after 20 years of work in AI alignment (counting from the date that MIRI, then the Singularity Institute for Artificial Intelligence, was founded), we have neither a resolution to the AI alignment problem nor any scheme for scalably utilizing funds to find one. What would a scalable scheme for resolving the AI alignment problem look like, exactly? Is depth scalable? If not, then why exactly is that?

These are questions about which I would very much like to have one-on-one conversations. If you would like to have such a conversation with me, please send me a direct message here on lesswrong.

Metta

May you find happiness and depth in your work.

May you find a way to live that truly supports you.

May your life and work come together beautifully.

May you bring peace to our troubled world.

Imitative Generalisation (AKA 'Learning the Prior')

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

TL;dr

We want to be able to supervise models with superhuman knowledge of the world and how to manipulate it. For this we need an overseer to be able to learn or access all the knowledge our models have, in order to be able to understand the consequences of suggestions or decisions from the model. If the overseers don't have access to all the same knowledge as the model, it may be easy for the model to deceive us, suggesting plans that look good to us but that may have serious negative consequences.

We might hope to access what the model knows just by training it to answer questions. However, we can only train on questions that humans are able to answer[\[1\]](#). This gives us a problem that's somewhat similar to the standard formulation of [transduction](#): we have some labelled training set (questions humans can answer), and we want to transfer to an unlabelled dataset (questions we care about), that may be differently distributed.

We might hope that our models will naturally generalize correctly from easy-to-answer questions to the ones that we care about. However, a natural pathological generalisation is for our models to only give us 'human-like' answers to questions, even if it knows the best answer is different. If we only have access to these human-like answers to questions, that probably doesn't give us enough information to supervise a superhuman model.

What we're going to call 'Imitative Generalization' is a possible way to narrow the gap between the things our model knows, and the questions we can train our model to answer honestly. It avoids the pathological generalisation by only using ML for IID tasks, and imitating the way humans generalize. This hopefully gives us answers that are more like 'how a human would answer if they'd learnt from all the data the model has learnt from'. We supervise how the model does the transfer, to get the sort of generalisation we want.

It's worth noting there are enough serious open questions that imitative generalization is more of a research proposal than an algorithm!

This post is based on work done with Paul Christiano at OpenAI. Thanks very much to Evan Hubinger, Richard Ngo, William Saunders, Long Ouyang and others for helpful feedback, as well as Alice Fares for formatting help

Goals of this post

This post tries to explain a simplified[\[2\]](#) version of Paul Christiano's mechanism introduced [here](#), (referred to there as 'Learning the Prior') and explain why a mechanism like this potentially addresses some of the safety problems with naïve

approaches. First we'll go through a simple example in a familiar domain, then explain the problems with the example. Then I'll discuss the open questions for making Imitative Generalization actually work, and the connection with the Microscope AI idea. A more detailed explanation of exactly what the training objective is (with diagrams), and the correspondence with Bayesian inference, are in the appendix.

Example: using IG to avoid overfitting in image classification.

Here's an example of using Imitative Generalization to get better performance on a standard ML task: image classification of dog breeds, with distributional shift.

Imagine we want to robustly learn to classify dog breeds, but the human labellers we have access to **don't** actually know how to identify all the breeds^[3], and we don't have any identification guides or anything. However, we **do** have access to a labelled dataset D. We want to classify dogs in a different dataset D', which is unlabelled.

One unfamiliar breed we want to learn to recognise is a husky. It happens that all the huskies in D are on snow, but in D' some of them are on grass.

Label: Husky



Image from D

Label: ???



OOD image from D'

A NN architecture prior likely doesn't favour the hypothesis 'a husky is a large, fluffy dog that looks quite like a wolf' over 'if there are a lot of white pixels in the bottom half of the image, then it's a husky'. These hypotheses both perform equally well on the training data. So a naïve approach of fitting a model to D and then running it on D' may easily misclassify huskies that are not on snow.

However, a human prior does favour the more sensible assumption (that the label husky refers to this fluffy wolf-like dog) over the other one (that the label husky refers to an image with many white pixels in the bottom half of the image). If we can use this human prior, we can avoid misclassifying huskies in D' - even if the two hypotheses perform equally well on D .

To apply the IG scheme here we're going to jointly learn three things.

- We're going to optimise z , which is a string of text instructions for how to label images (e.g. "A husky is a large, fluffy dog that looks quite like a wolf. A greyhound is a tall, very skinny dog. ...")
- Let $H^{\text{prior}}(z)$ be the prior log probability the human assigns [4] to the instructions z . We're going to train a model M^{prior} to approximate this function
- Similarly, we're going to train M^L to approximate $H^L(y|x, z)$, which is the log probability that a human assigns to label y (e.g. 'husky') given x (image of a dog) and z (text instructions on how to label images)

We find the z^* that maximises $M^{\text{prior}}(z) + \sum_{x,y \in D} [M^L(y|x, z)]$

Then we give this z^* to the humans, and have the humans use this to predict the labels for images in D' , ie query $H^L(y'|x', z^*)$.

Then we can use these human predictions to train a model M_{test}^L to approximate $H^L(\cdot | \cdot, z^*)$ on the distribution D' . We can then run M_{test}^L to get labels for images from D' with no distributional shift.

The hope is that the things in z^* will be sensible descriptions of how to label images, that conform to human priors about how objects and categories work. In particular, z^* is likely to contain instructions that the label for an image is supposed to depend on features of the object that's the subject of the photo, rather than the background.

So when we're querying our human labelers for $H^L(y' | x', z^*)$, the task they see will be: The human is shown a photo of a husky on grass (x') , along with the instructions 'a husky is a large, fluffy dog that looks quite like a wolf' and descriptions of many other dog breeds (z^*), and is asked how likely it is that this photo is of a husky (y')



If you're confused about the details of the setup at this point, I'd recommend reading the more detailed explanation in the appendix, which also builds up this diagram piece-by-piece.

Using this scheme, we can expect correctness on the test dataset, as long as our models are actually capable of learning H^{prior} and H^L given plenty of IID samples. We avoid problems related to overfitting and distributional shift.

Ways that this specific example is unrealistic:

Firstly, our model may not be capable enough to learn the human likelihood/prior functions, even given plenty of IID examples. IG is easiest to analyze when we have ML capable of learning to imitate most IID human behavior. If our ML is more limited, the generalization will be determined by a combination of human capabilities and model capabilities.

This example isn't very exciting, because classifying dogs is a problem that humanity has already solved. If we were actually doing this specific task in real life, we'd either give the workers a guide to identifying dog breeds, or let them look at D and learn the labels, and then label D' for us. The IG scheme is only needed if this isn't possible - for example, if there are no existing resources on how to identify dogs, and there are so many different dog breeds that it's too hard to get our labellers to learn them without help from z^* . Even then we might think that the labellers can just look at D and make their own notes on breed identification. IG is needed if this task is too difficult - e.g. if the relationship between visual characteristics and breeds is too complicated for humans to infer from a few examples.

Difficulties with this example that are also general problems

Optimising z is hard; we'd probably need a better way of representing z and exploring the space of zs than just searching over long strings of text. One way to improve might be to have our human labellers generate different hypotheses for what different breeds look like, then train a model to imitate this hypothesis generation.

Conveying the correct hypothesis in text might be hard. Explaining how to recognise different dog breeds is probably significantly harder than teaching a human by showing them some examples. Having human workers explicitly look for particular features that can be described in words may not be competitive with non-verbal, unconscious pattern-recognition. In this specific example, we might want to add visual information to z - for example, providing a diagram of key characteristics for each breed.

Key difficulties for IG

From the above problems, we can see that a core difficulty for this scheme is how to represent z . We want z to be able to capture everything that an NN could learn, and want it to be easy to optimise, but also need it to be human-understandable.

Text

One possibility is to have z be a long string of natural-language text. We have several problems:

1. Firstly, z will contain a lot of information, not all of which is necessarily relevant to a particular task. It might take the human an impractically long time to engage with the whole of z . We're hopeful that this can be fixed by using something like amplification or debate to help the human interact with this large text.[\[5\]](#)
2. Optimising over large strings of text is hard - we don't have a gradient. As mentioned above, we might attempt to mitigate this by having humans explore the space of zs , and imitating this exploration to guide our search
3. The biggest problem is that it's not clear if certain types of knowledge can be conveyed adequately via natural language with reasonable efficiency. Certainly encoding everything a vision model knows about images would take an extremely long text. In the example above, we can imagine using visual aids to make the problem easier. But vision is a domain humans are very well-adapted to processing. It's much less clear how we'd help a human understand instructions about how protein folding works.

Neural net

Another approach is to have z be a neural network together with annotations that are supposed to describe what different parts of the NN are doing. We can optimise the NN weights and text annotations jointly to maximise the human prior of this combination of weights and annotations, plus the likelihood the human assigns to D when given access to this network with these annotations.

For example, annotations might say "Neuron 327 indicates the presence of dog ears, the final output neuron indicates whether the dog is a husky." If a human assumes the annotations are correct, then it should be easy for them to label new images by running the neural network. The annotations may also allow the human to answer other questions about the input. In the example given above, we might hope that if the human sees that the circuits for fluffy coat, pointy ears, tail, paws etc are activated, they can use that information to classify the image as a husky.

The tricky bit is ensuring that the labels actually correspond to what the circuit is detecting. In the example, labelling the snow detection circuit as a husky detection circuit would perform just as well on the training data. Importantly, the prior captures whether the structure of the neural network is consistent with these annotations, and so optimizing for a high prior hopefully forces the annotations to make sense. For example, if a human saw that a "snow detector" taking on a high enough value would cause a "husky detector" to fire, without any contribution from ears, tail or fluffy coat, then they could rule out the relationship as implausible.

This has an advantage over a text-only representation, because some parts of the problem can be left as black boxes. For example, consider a blind human trying to complete the image-labelling task. If they were only using a text-based z , z^* would have to contain instructions that involved going through the image pixel by pixel and making calculations to compute edge detectors and other low-level vision features. The task would be very hard for the human, and z^* would be an extremely long text. However, if the human instead had access to various annotated circuits that detected floppy ears, tails, fluffy coats, etc, the task would be much more tractable.

The text part of z^* could also be significantly shorter. It seems likely to be easier to jointly optimise a NN together with text annotations, than to optimise a long text that contains all of the information encoded in the NN weights. We can start with a good init by training our NN on the labelled dataset D, so that our NN contains various features relevant to our task (e.g. features of dogs).

We still have a problem similar to (3) above---it's not clear what kind of annotations and weights could actually allow a human to understand a computation well enough to access all of its knowledge, or well enough to evaluate its plausibility.

Relationship with Microscope AI

My (Beth's) paraphrasing of the Microscope AI proposal is as follows:

We train ML models, then examine the representations they've learnt using our interpretability tools. We then use the knowledge and insight gained from these representations to help [amplified] humans do tasks, and never use the model directly. Also, if we spot problems with these representations we can prune or modify them to improve the usefulness, although this isn't the central part of the proposal. This gives us safety properties in at least two ways:

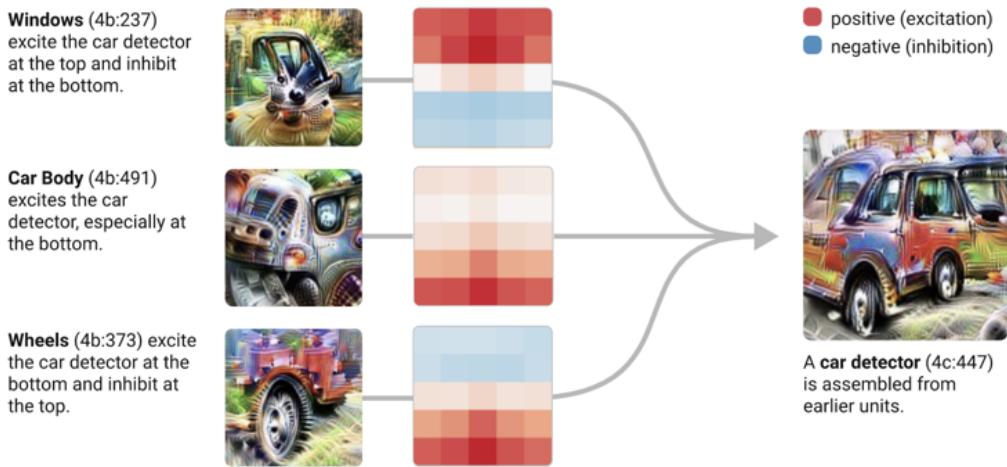
1. we don't have to deploy anything at all resembling an agent
2. we can inspect and understand everything before we use it. If some concept generalises badly or some subnetwork contains an inner optimiser, we're not going to unknowingly use these

There seems to be moderate convergence between the microscope AI idea and the Imitative Generalization idea. For the microscope AI proposal to work, we need there to be some human-understandable way to represent everything some NN 'knows' - this is what our interpretability techniques produce. The IG proposal can be framed as: instead of training a NN then extracting this human-understandable object with our interpretability tools, let's directly search over these human-understandable representations.

A key challenge for both approaches is whether (for any given NN) the human-understandable object actually exists, and if so how we can represent it. In both cases, we can give the human as many tools as possible to make this more tractable - for example, we could use IDA or debate.

The [OpenAI Microscope](#) produced by the Clarity team is one example of what such an object might look like. The Microscope contains multiple types of visualisation for every significant neuron and layer in the 'model organism' NNs. It's hard to say at this

stage whether these tools really allow a human to leverage everything the model knows, but it certainly contains a large amount of usable knowledge about image classification. The team learnt a lot about how to identify different dog breeds from playing with the Microscope!



Appendix

More detailed notation for training procedure

We have some space Z of objects that represent distributions over hypotheses We have some space of questions X and answers $Y:\{\text{Yes}, \text{No}\}$ We have two sets of questions drawn from two different distributions: D (train) and D' (test). The questions drawn from D have corresponding answers We train three models:

$$M^{\text{prior}} : Z \rightarrow [-\infty, 0]$$

gives the log likelihood of Z under the human prior

$$M_{\text{train}}^L : Z \rightarrow (X \rightarrow (Y \rightarrow [-\infty, 0]))$$

gives the log likelihood of some Y given some X and Z

$$M_{\text{test}}^L : X \rightarrow (Y \rightarrow [-\infty, 0])$$

gives the log likelihood under z^* of some Y given some X

Step 1:

We collect a dataset H^{prior} from humans by asking humans ‘how plausible is z ?’ for the zs we’re searching over. We train M^{prior} to minimise loss on $\log(H^{prior})$

We collect a dataset H_{train}^L from humans by asking humans ‘how plausible is y as an

answer to x , assuming $z?$ ’ for each $(x, y) \in D$. We train $M_{train}(z, x, y)$ to minimise loss
on $\log(H_{train}^L)$

Step 2:

We search over Z to find z^* that maximises $M^{prior}(z) + \sum_{x,y \in D} (M_{train}(x, y, z))$.

Step 3:

While we’re optimising z , we update H^{prior} and H_{train}^L to be on-distribution with the
current z , and retrain M^{prior} and M_{train} appropriately.

Step 4:

We collect a dataset H_{test}^L from humans by asking humans ‘how plausible is {yes/no}
as an answer to x , assuming $z^*?$ ’ for each x in D' . We train M_{test} to minimise loss on
 $\log(H_{test}^L)$

In practice, we’d want to share weights between M^{prior} , M_{train} and M_{test}

Correspondence with Bayesian updating

Consider some Bayesian agent that has a prior over hypotheses, z^{prior} . D corresponds to data they want to update on, and D' corresponds to some data they want to predict.

They have some method `log_likelihood()` of computing the likelihood of some data given different distributions z over hypotheses - this corresponds to H_{train} and H_{test}

If we replace $H^{\text{prior}}(\cdot)$ with $\text{KL}(\cdot || z^{\text{prior}})$, then what IG gives us is exactly the same as if the agent had performed the correct Bayesian update on all of D .

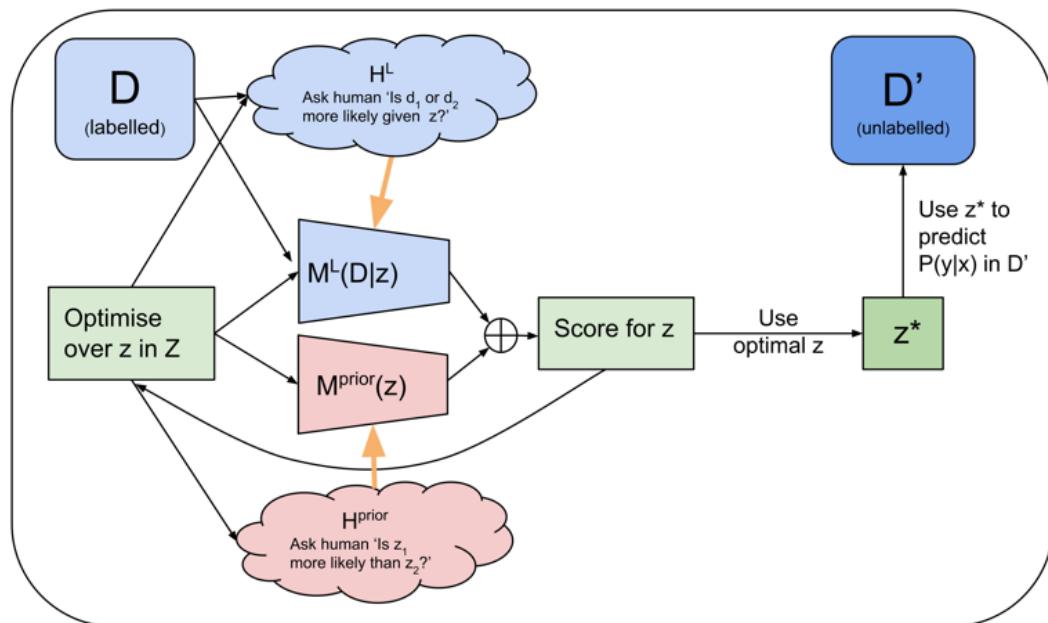
The agent's predicted probability of (x', y') after seeing D is $\text{log_likelihood}(z^*, x, y)$ with:

$$z^* = \max_{z \in Z} [\text{KL}(z || z^{\text{prior}}) + \sum_{x,y \in D} (M_{\text{train}}(x, y, z))]$$

z^* corresponds to the agent's posterior after seeing D

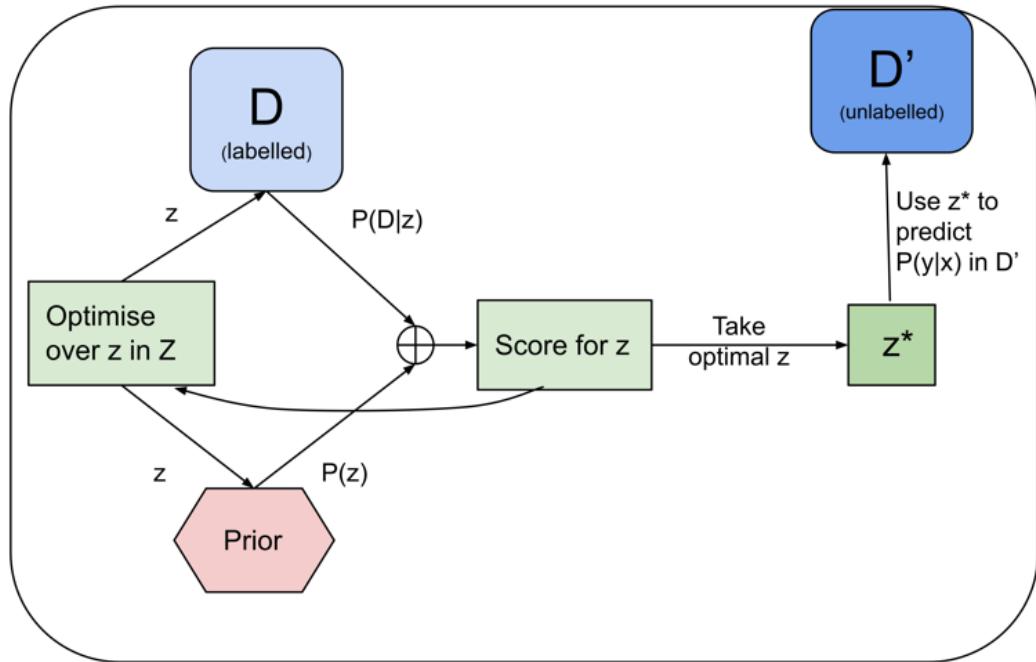
Diagrams

This is a picture of the whole scheme. The part circled in red is the output of the process: an ML model that we can use for answering questions in D' without distributional shift.



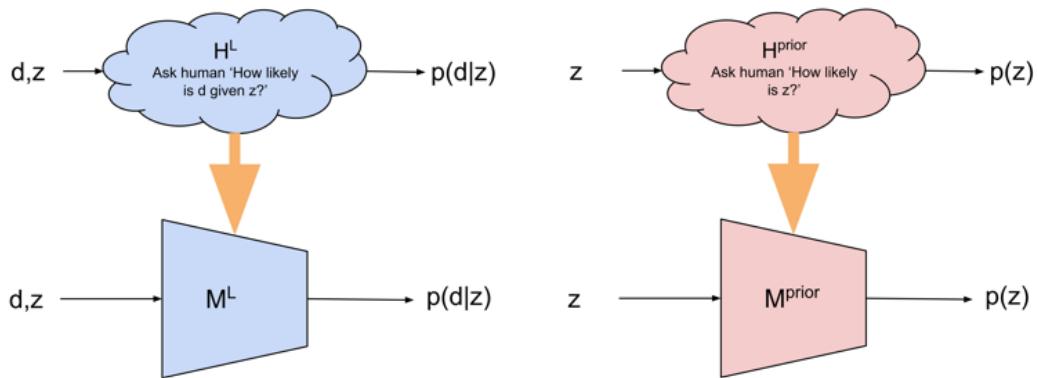
There's a lot going on here, so let's break it down. The underlying structure is just induction:

we have some prior and labelled dataset D ; we use the prior and the likelihood to compute a score for z and we use this to optimise z . Then we use z^* to predict the labels in D' .



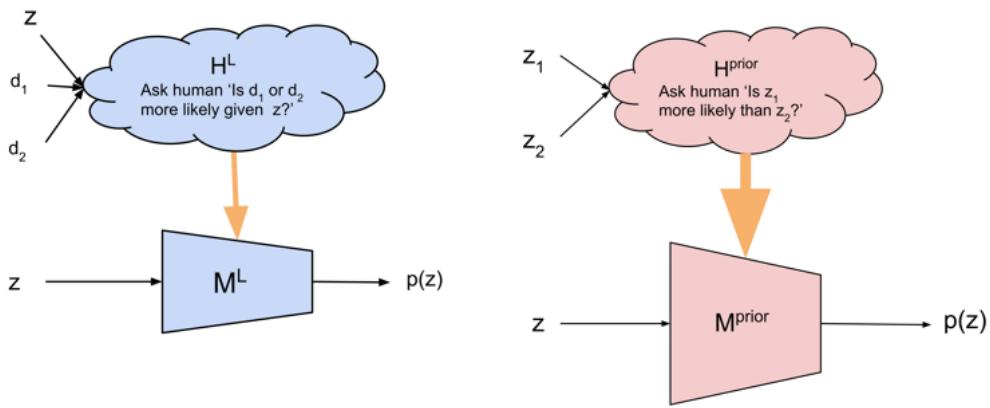
Using ML

We want to train models to imitate the human Prior and Likelihood functions. As long as we can sample $H^{prior}(z)$ and $H^L(d, z)$ for any z and d that we might want to evaluate our models on, we can ensure that the task is IID.

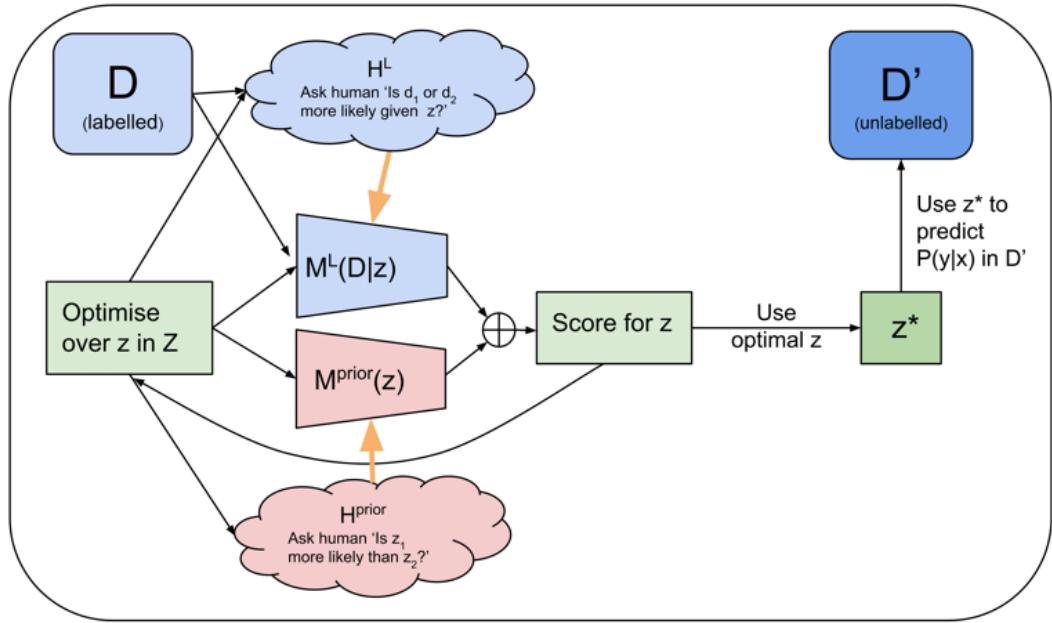


Making it easier for humans

Asking humans to directly report $P(z)$ or $P(d|z)$ is unreasonable. To make it easier for our labellers, we instead just ask them to compare pairs of d s or z s, and use this reward signal to fit M^L and $M^{\{\text{prior}\}}$ (as in e.g. <https://arxiv.org/pdf/1909.08593.pdf>).



When we put these pieces into the induction setup above, we just need to compute the overall likelihood $P(D|z)$ by summing log likelihoods for each $d \in D$, then add this to $\text{prior}(z)$ to get the posterior score for z .



Now we just need to learn to compute the likelihoods on examples from D' , given z^* .

Separating out the likelihoods for D and D' like this lets us ensure that each of the ML imitation tasks are IID.



Now the object circled in red is what we wanted all along - a ML model trained to label examples in D' , that approximates “ How a human would answer if they'd looked at all the data in D ”.

Notes

1. This is a simplification; using RL we can train on questions as long humans can reliably recognise good answers. Using IDA/Debate/RRM, we can train on questions where humans can recognise good explanations of good answers. We can also train on questions where we don't currently know the answer but can do an experiment. But this is still unlikely to cover everything the model might have learnt from unsupervised training [←](#)
2. I'm just considering a simplified version of LTP that doesn't use amplification [←](#)
3. For the purposes of this example, we can imagine that our human labellers are generally familiar with dogs and know what a dog breed is, but don't know the names and identifying characteristics of all the breeds. If the human labellers have never seen a dog before, the task is still fairly straightforward in theory but more tricky in practice. [←](#)
4. In practice, we're not going to be able to elicit $P(z)$ directly from the human; instead, we'll do something like asking humans to compare which of two z s are more likely, and use this as a reward signal, as in (e.g.)
<https://arxiv.org/pdf/1909.08593.pdf> [←](#)
5. More specifically, what we need here is an aligned model that has sufficient time/capacity to read and manipulate z , and pull out the relevant parts to show the human. [←](#)

Covid 1/28: Muddling Through

Three different time frames, three different fronts. We continue to muddle through, with some hope that this could be successful relative to our modest expectations.

There's the situation short term, there's the new strains, there's the vaccines.

On the short term front, forward looking news continues to improve, but not at the pace we'd like or that I expected, and the death rate this week unexpectedly (unexpected to me in any case) rose. All signs still point to steady improvement until the new strains have an impact, probably at a pace of something like 1% a day. Hospitalizations aren't listed in the numbers, but they too are falling steadily.

On the new strain front, the news is mixed.

For the non-English new strains, things got less scary, as it looks much less likely that the new strains can escape and reinfect, or that the vaccines won't work on them. If they did escape, we learned that the system isn't capable of responding as quickly as we'd like, but perhaps under duress that would change.

For the English strain, things got much scarier. Previously, we had no reason to think the new strain was more virulent, and if anything were hoping it was less so. Instead, it looks like it's plausibly substantially more virulent, with 30%+ higher death rates per infection. Given that this strain is about to take over, that's very bad news.

On the vaccine and policy front, we had excellent news, as the Biden administration announced a deal for an additional 200 million doses of vaccine from Pfizer and Moderna.

There's still a ton to do. We need to approve AstraZeneca now and Johnson & Johnson the moment they release their data. We still need to spend massively to expand capacity. We still need to move to half doses or smaller where available, and legalize rapid testing for real, and so on and so forth. But compared to expectations, I'll definitely take it.

I'll also take the numbers we are seeing on vaccinations. They're not what I wanted to see when this all started, but given how things went until this week, seeing the numbers rise to 1.2 million USA shots a day, with the constraint increasingly being supply, indicates that we've mostly muddled through.

The exception to that is our failure in long term care facilities, which isn't getting enough focus.

Let's run the numbers.

The Numbers

Predictions

Last week: 11.9% positive rate on 11.3 million tests, and an average of 3,043 deaths.

Prediction: 10.5% positive rate and 2,900 deaths per day.

Result: 11.2% positive rate and 3,257 death per day.

I overshot the drop in positive rate by a factor of two, and the deaths went up rather than slightly down, so not my best prediction week, and poor news all around.

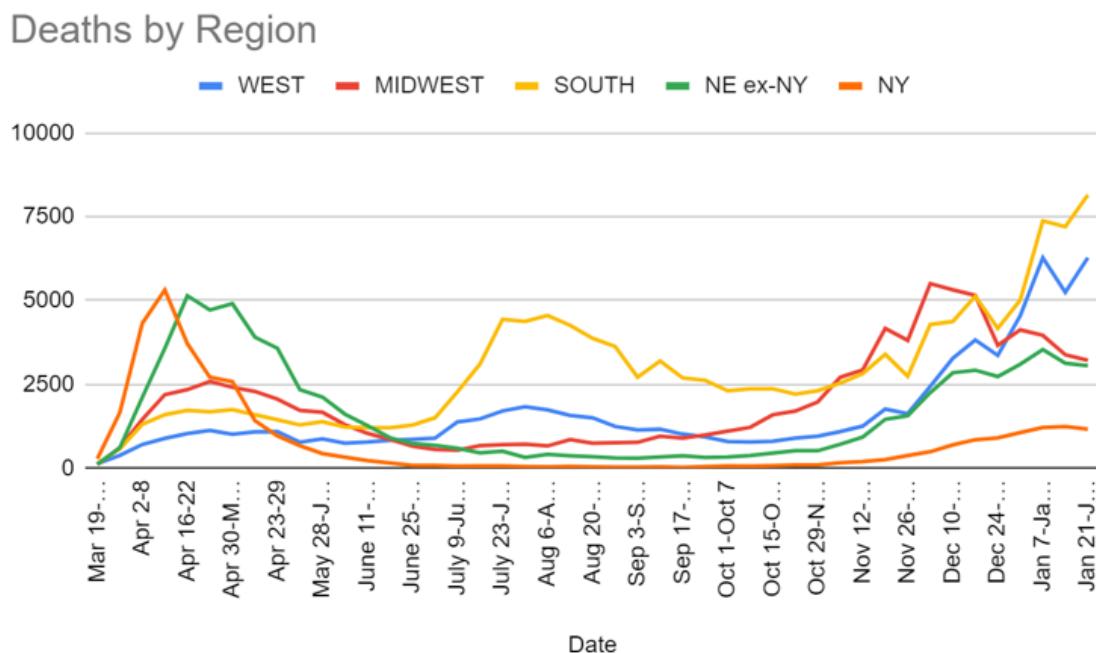
Part of the story about deaths is that our nursing home vaccination efforts have been dreadfully awful, and while I knew they were bad I didn't appreciate how bad they were. See later on for further discussion. Also I somehow keep not giving holidays proper respect. Presumably this has a lot to do with some combination of Martin Luther King Day, and the continued fallout from New Year's and Christmas, including secondary little waves as people returned home.

We do see some suggestions in the regional positive test percentages that things might not have started improving until the week of 1/13 in some places, in which case we could still see one more week of things getting worse before they get better. That's not what I expect, but on looking closer it wouldn't be *that* surprising. Once we get deeper into February a lack of improvement would be very surprising.

The other explanation is if the new English strain is much more present than we realize even now, and that it is more virulent than the old strain. I don't think this impact will be felt for another month or two.

Prediction for next week: 10.8% positive rate, as improvement is clearly ongoing but seems to be slowing down. 3,100 deaths per day, as I still expect this to start dropping, but I can see the case for it taking a bit longer.

Deaths



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Nov 26-Dec 2	1628	3814	2742	1939
Dec 3-Dec 9	2437	5508	4286	2744

Dec 10-Dec 16	3278	5324	4376	3541
Dec 17-Dec 23	3826	5158	5131	3772
Dec 24-Dec 30	3363	3668	4171	3640
Dec 31-Jan 6	4553	4127	5019	4162
Jan 7-Jan 13	6280	3963	7383	4752
Jan 14-Jan 20	5249	3386	7207	4370
Jan 21-Jan 27	6281	3217	8151	4222

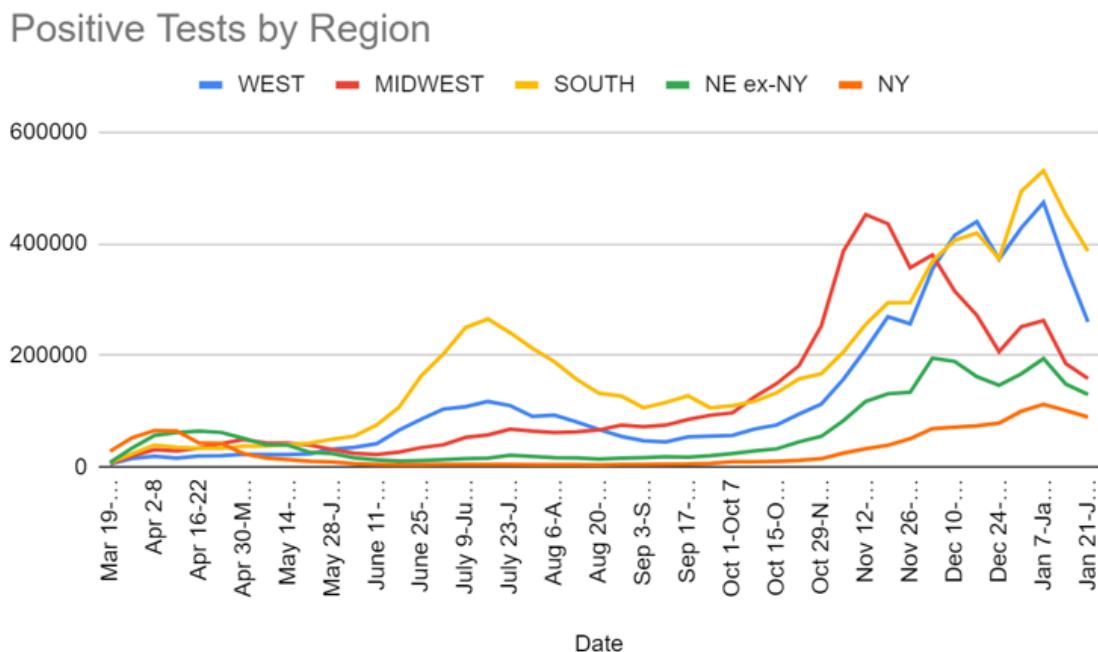
Both the South and West had substantially more deaths this week. The question remains why, and whether this is a data artifact and deaths from last week got shifted to this week, although it's hard for that to explain the South. This should be the local peak, but it's possible we have another week to go if we judge by local positive test percentages.

Positive Test Percentages

This section is unavailable right now because my parser that creates the graph stopped giving me non-zero numbers for a number of states, after running it twice. I've been ignoring that problem for Hawaii alone, but now it's a lot more than Hawaii. I don't want to delay the post longer. Hopefully I can figure out what's wrong in time for next week.

In the short term, positive test counts adjusted for overall test count should give a close enough answer for our purposes, but longer term I'll need to fix the parser. If I run into trouble I'll ask for help.

Positive Tests



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Dec 10-Dec 16	415,220	315,304	406,353	260,863

Dec 17-Dec 23	439,493	271,825	419,230	236,264
Dec 24-Dec 30	372,095	206,671	373,086	225,476
Dec 31-Jan 6	428,407	251,443	494,090	267,350
Jan 7-Jan 13	474,002	262,520	531,046	306,604
Jan 14-Jan 20	360,874	185,412	452,092	250,439
Jan 21-Jan 27	260,180	158,737	386,725	219,817

Test counts did drop a bit, but this is what we were hoping this graph would look like, with large drops across all regions.

Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Dec 3-Dec 9	10,466,204	13.9%	1,411,142	4.9%	4.67%
Dec 10-Dec 16	10,695,115	13.9%	1,444,725	4.9%	5.12%
Dec 17-Dec 23	10,714,411	13.7%	1,440,770	5.1%	5.57%
Dec 24-Dec 30	9,089,799	13.8%	1,303,286	6.0%	5.95%
Dec 31-Jan 6	9,334,345	16.4%	1,365,473	7.3%	6.42%
Jan 7-Jan 13	11,084,291	15.2%	1,697,034	6.6%	6.93%
Jan 14-Jan 20	11,300,725	11.9%	1,721,440	5.9%	7.35%
Jan 21-Jan 27	10,021,716	11.2%	1,679,399	5.3%	7.69%

I have no explanation for why we did less testing this week than we did the previous week, and did so everywhere including New York. I do not think it was because less people needed tests.

Covid Machine Learning Project

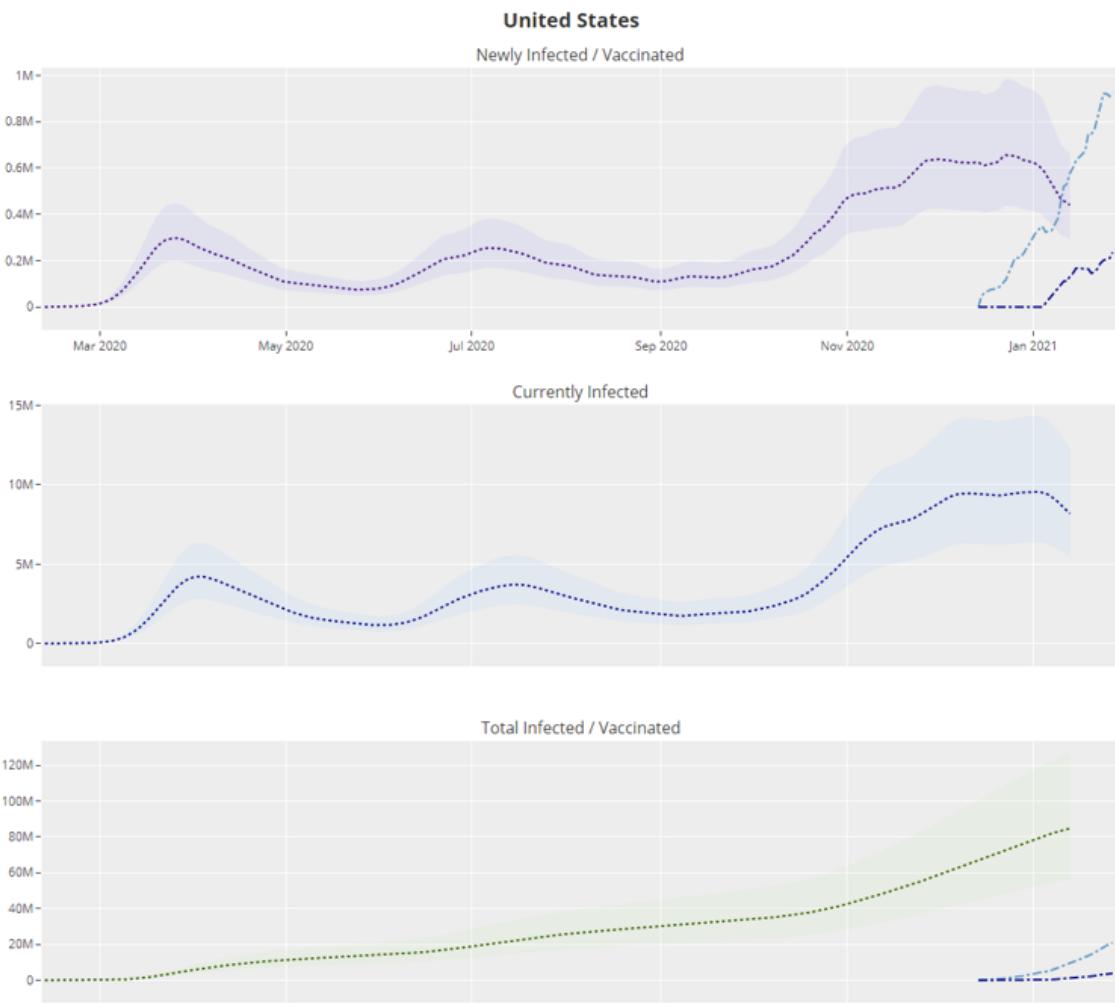
These can be hard to read, I recommend [checking the original source](#).

Last Updated: Thursday, January 28, 2021 (1am ET)

Newly Vaccinated (as of Jan 27): **925,000 / day** (350 / 100k)
Total Vaccinated (as of Jan 27): **6.3%** (1 in 16 | 20.8 million)

Newly Infected (as of Jan 13): **440,000 / day** (130 / 100k)
Currently Infected (as of Jan 13): **1 in 40** (2.5% | 8.2 million)
Total Infected (as of Jan 13): **25.5%** (1 in 4 | 84.7 million)

R_t (as of Jan 13): **0.84**
Adjusted Positivity Rate (as of Jan 27): **8.5%**
Infection-to-Case Ratio (as of Jan 27): **2.8** (36% detection rate)



Overall this is a great situation. These projections think R₀ is down to 0.84 with the number of infected per day down a third and falling fast.

I'm still the kind of person who is horrified by that tiny little dip in vaccinations in the upper right corner that happened yesterday. It's expected that this is only a blip, but I'll feel a lot more comfortable with it in the rear view mirror.

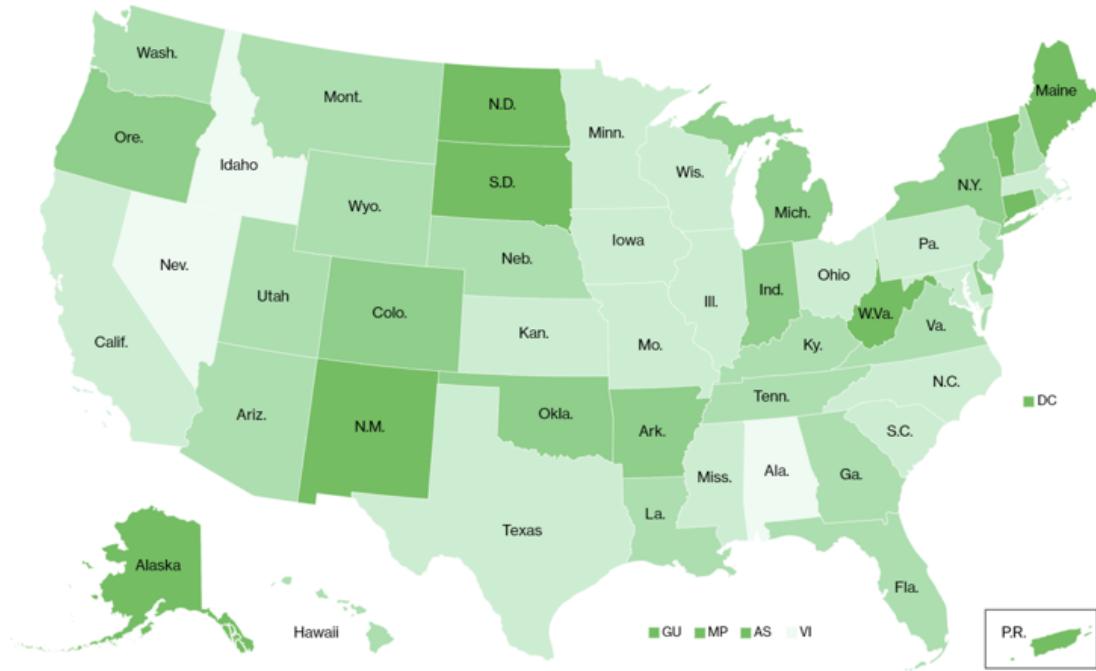
On January 13 they project 25.5% of all Americans as having been infected, up from 24.5% on January 6. That's still a big compounding edge going forward, which hopefully will rapidly be eclipsed by rising vaccinations.

Vaccinations

Vaccinations in the U.S. began Dec. 14 with health-care workers, and so far **25.6 million shots** have been given, according to a state-by-state tally by Bloomberg and data from the Centers for Disease Control and Prevention. In the last week, an average of **1.21 million doses per day** were administered.

Vaccines Across America

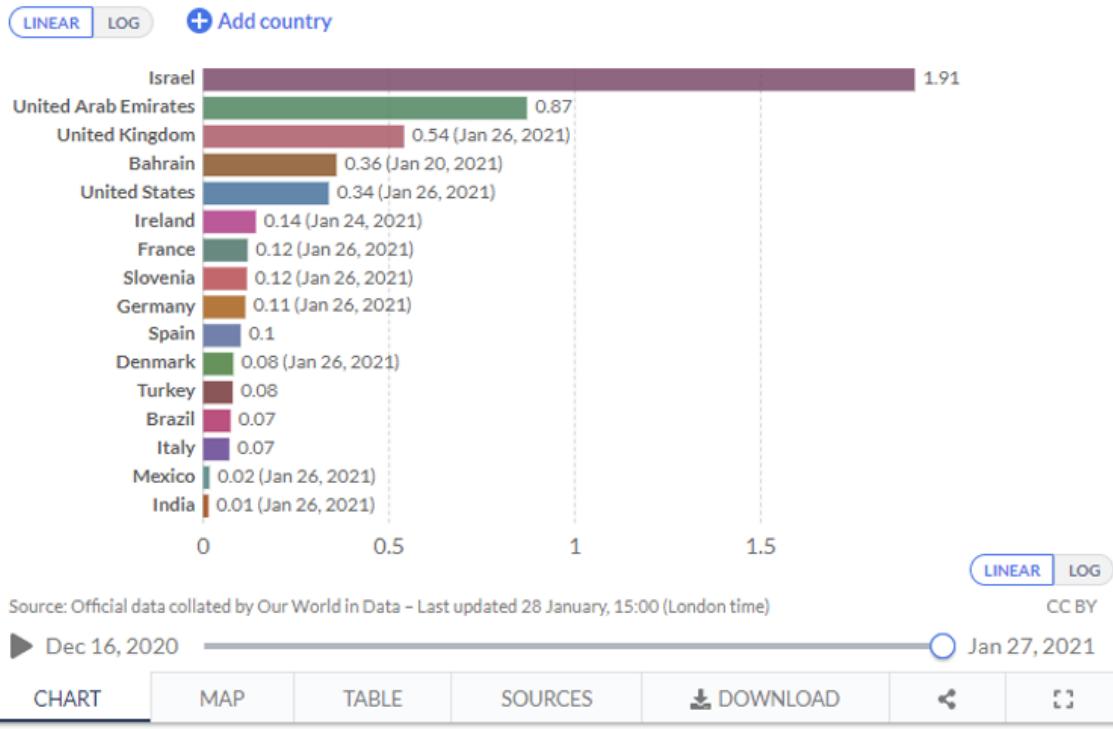
Across the U.S., 7.8 doses have been administered for every 100 people, and 54% of the shots delivered to states have been administered



Daily COVID-19 vaccine doses administered per 100 people, Jan 27, 2021

Our World
in Data

Shown is the rolling 7-day average per 100 people in the total population. This is counted as a single dose, and may not equal the total number of people vaccinated, depending on the specific dose regime (e.g. people receive multiple doses).



Last week we were averaging under a million doses, now we're over 1.2 million. That's a solid rate of increase if it can be sustained. Yesterday saw the seven day average decline slightly, which is always scary, and likely reflects that many states previously were bottlenecked on distribution but are now bottlenecked on supply. So in its own way it is *good* news, depending on what prior knowledge is being controlled for.

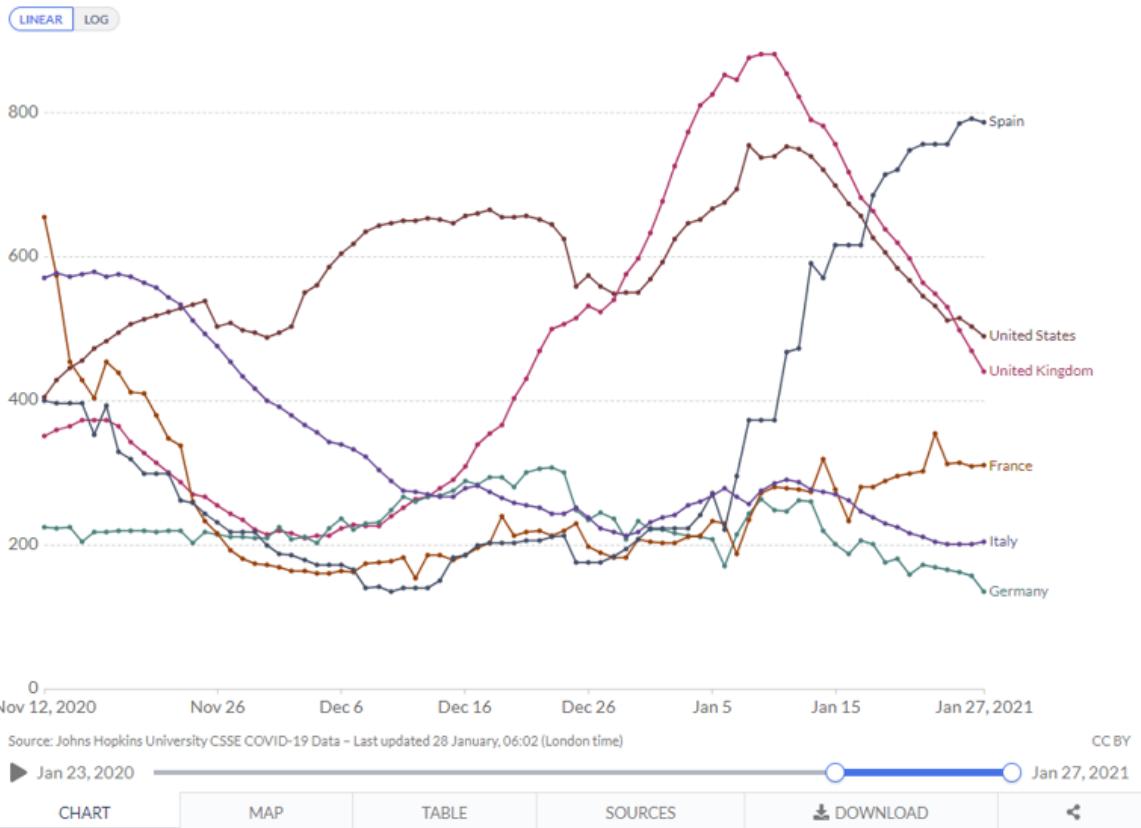
The only questions now are how rapidly we can increase supply, and whether we can use that supply more efficiently.

Europe

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

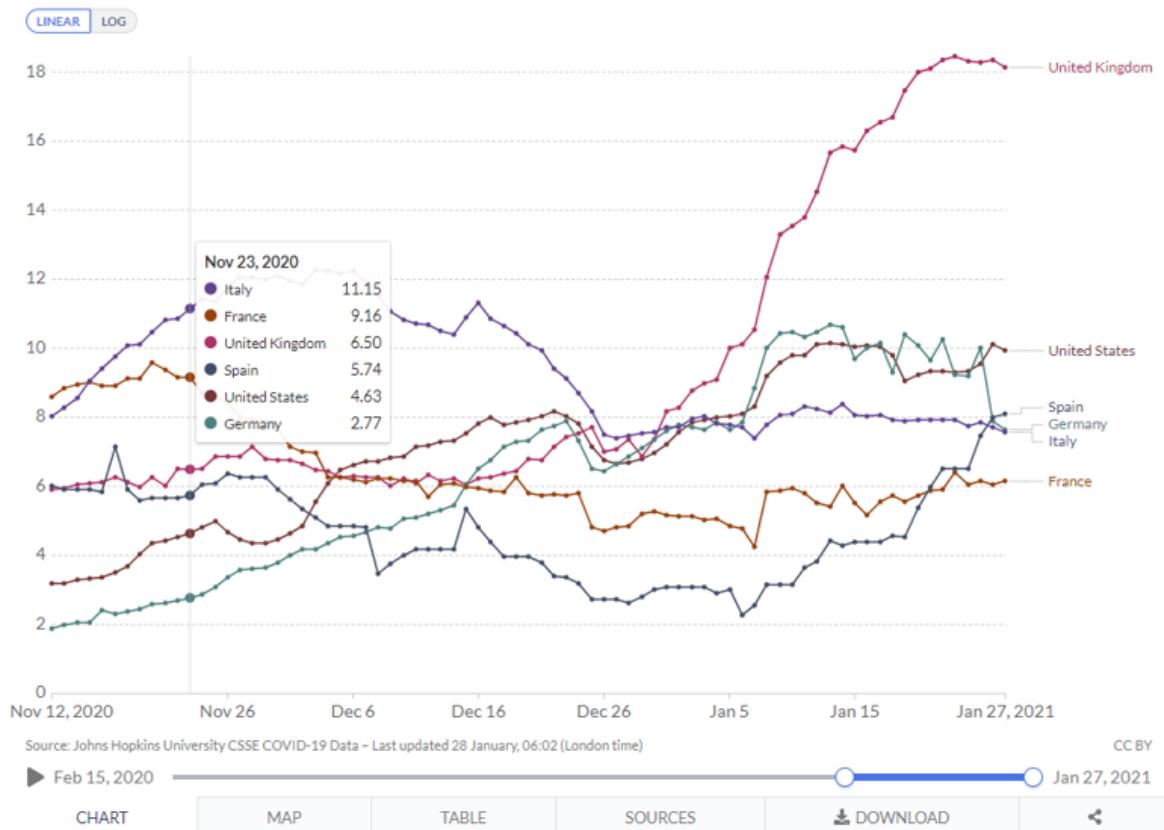
Our World
in Data



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



UK positive tests didn't peak until January 5, so it makes sense that their death counts are only peaking now, but this is still suggestive that the new strain might be more virulent, as will be discussed in the next section. Spain now is approaching the English peak, so we should expect a lot more deaths there starting next week.

The good news is that the UK seems to have things under control, and the lockdown continues to be working. That doesn't mean they can sustain that level of restrictions, or what the price will be for doing so, but it can definitely be done. I'm not shocked, but I'm definitely pleasantly surprised. I still don't know if we could do the same, but by the time we need to, we'll have a lot more help from the already immune.

The English Strain

It is now accepted that the English strain is here, it is rapidly taking over, and it is substantially more infectious than our previous strains.

We know it does not evade vaccines, and prior immunity protects against it. The new worry is that the [new strain might be deadlier](#).



Kai Kupferschmidt ✅ @kakape · Jan 22

Reports that #b117 may be 30% deadlier than other variants is very concerning. But it also comes with a lot of caveats at this stage. As so often: Truth is that we don't yet know for sure. Let's wait as data comes in.
We know what we have to do anyway.

...

The concerning document everyone's referring to [is here](#).

4. There have been several independent analyses of SGTF and non-SGTF cases identified through Pillar 2 testing linked to the PHE COVID-19 deaths line list:
 - a. LSHTM: reported that the relative hazard of death within 28 days of test for VOC-infected individuals compared to non-VOC was 1.35 (95%CI 1.08-1.68).
 - b. Imperial College London: mean ratio of CFR for VOC-infected individuals compared to non-VOC was 1.36 (95%CI 1.18-1.56) by a case-control weighting method, 1.29 (95%CI 1.07-1.54) by a standardised CFR method.
 - c. University of Exeter: mortality hazard ratio for VOC-infected individuals compared to non-VOC was 1.91 (1.35 - 2.71).
 - d. These analyses were all adjusted in various ways for age, location, time and other variables.
5. An updated PHE matched cohort analysis has reported a death risk ratio for VOC-infected individuals compared to non-VOC of 1.65 (95%CI 1.21-2.25).
6. There are several limitations to these datasets including representativeness of death data (<10% of all deaths are included in some datasets), power, potential biases in case ascertainment and transmission setting.
7. **Based on these analyses, there is a realistic possibility that infection with VOC B.1.1.7 is associated with an increased risk of death compared to infection with non-VOC viruses.**

By default, we should expect viruses to become *less* deadly over time rather than more severe, but more severe is always a risk. We also believe the new strain carries generally higher viral loads, which could plausibly be a cause of higher severity.

I concur with the paper's level of concern: A 'realistic possibility' of increased risk of death seems right. There are a lot of potential confounders here, including time, and the sample sizes aren't wonderful, so it is too early to jump to conclusions. This is not enough to overcome my prior that increased virulence is unlikely, but given that there is a plausible mechanism to explain it, I'd now put it at something like 40% that there's substantial (>20%) additional virulence, which I've lately been revising upwards.

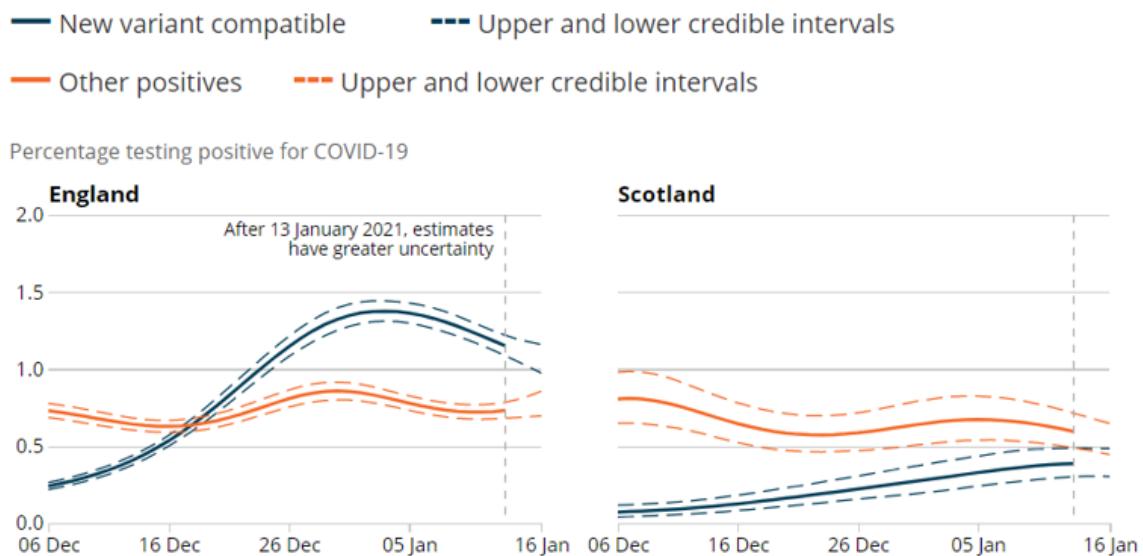
That would mean that the next few months are all that much more dangerous for those who are not vaccinated. Note that if we do reasonable prioritization the CFR should still drop from vaccinations more than it should rise from the new strain (perhaps not if the Exeter estimate is right, but certainly if the others are right). Remember that nursing homes have tiny populations and cause more than a third of all Covid-19 deaths. But the additional risk to you

is real, and compounds on top of the effect that all infections and deaths are divided only among the still unvaccinated and otherwise non-immune.

[Construct this with Fauci's framing of the situation](#), which suggests all the same actions I'd suggest, but in a way designed to maximize uptake rather than accuracy:

Fauci: 'We need to assume now' that British virus strain can 'cause more damage'

There was also some strange data this week about the new strain in England, [hence a lot of this graph](#) was seen:



The English graph here seems to show the new strain falling while the old strain is stable, which at first glance is counter to it being more infectious than the old strain. Unfortunately, as one would expect given other data, [this seems to be the result of how the data is gathered](#), rather than reflective of the situation on the ground. If you want to dive in, I'd encourage reading the whole thread:



Theo Sanderson
@theosanderson

...

1/ The ONS infection survey has come out, and the graph below has led some people to speculate that B.1.1.7 isn't more transmissible any more.

Unfortunately I don't think we can conclude that. Here is why [1/n]



Theo Sanderson @theosanderson · Jan 22

...

6/ So what's happened recently in England? If we look at all of these different combinations over time we can see that OR+N (which is how the ONS defines 'new-variant-compatible') has indeed flatlined or fallen lately.



Theo Sanderson @theosanderson · Jan 22

...

7/ But it hasn't been replaced by OR+N+S (which is essentially incompatible with B.1.1.7), it's been replaced by N only and OR only, which are completely compatible with being either B.1.1.7.



Theo Sanderson @theosanderson · Jan 22

...

9/ Finally (for now), it's really complicated. Random dropout of S could lead WT viruses to present as "OR+N" (new-variant-compatible). SGTL could lead a proportion B.1.1.7 to present as "OR+N+S", creating an upper limit on apparent B.1.1.7 prevalence..

England is clearly past its peak of infections, [but it remains under lockdown indefinitely](#). Until there's more people who are immune, there doesn't seem to be a way to lift the restrictions. So yes, it turns out we can beat the new variant with lockdowns, but it takes longer and is far more painful.

The Other New Strains

It is also accepted now that the South African variant is super scary.

Biden imposed travel restrictions on South Africa this week, which was long overdue. [Belgium closed its borders](#) entirely to non-essential travel.

[Moderna says it's working on Covid booster shot for variant in South Africa, says current vaccine provides some protection](#). This isn't the best news we could have hoped for, but overall it is reassuring. The vaccines still should mostly work as far as we can tell. [The headlines, as usual, often took the maximally alarmist approach](#).



zeynep tufekci ✅

@zeynep

...

Replying to [@zeynep](#)

Plea to media: this isn't a good headline. It makes people think the vaccine is six times less effective against the new variants (FALSE!) when the news today is *excellent*: The vaccine continues to work well against the new variants. That's the headline.



San Francisco Chronicle ✅ @sfchronicle · Jan 25

Moderna's coronavirus vaccine protected against the mutations of the virus first detected in Britain and South Africa, The Washington Post reported. But the antibodies were six times less effective at neutralizing the South African variant. trib.al/VGr2Vad

In case the vaccine stops working, we can tweak it and make a new one. The real questions are still whether we will pay for sufficient manufacturing capacity (yes, this is still an issue) and whether the FDA will throw up barriers for the new versions that slow things down by months.

[This headline today](#) tells the opposite story, in any case, excellent all around if accurate:

Empty vials of the Pfizer-BioNTech COVID-19 vaccine are seen at a vaccination centre at the University of Nevada, Las Vegas. (AP)

Pfizer-BioNTech Covid vaccine works against UK, South Africa variants: Report

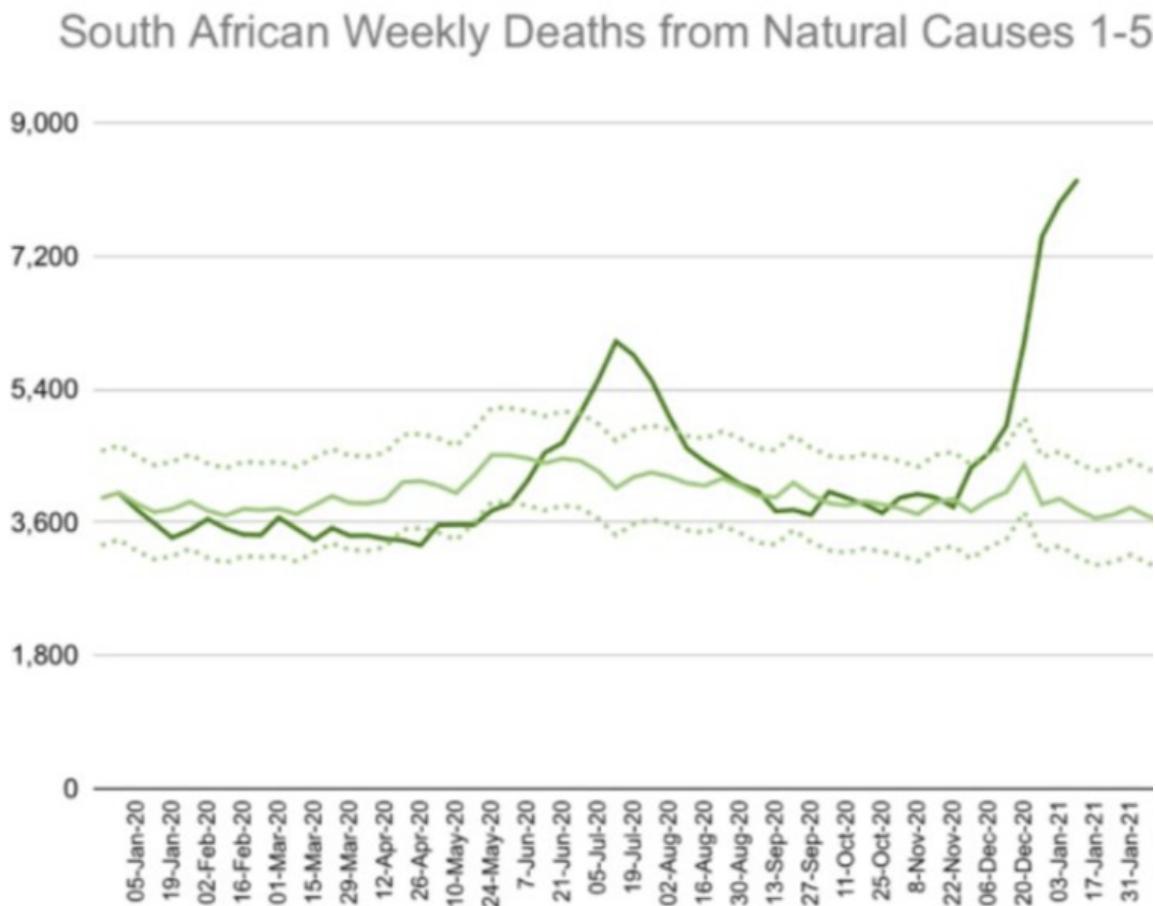
2 min read . Updated: 28 Jan 2021, 03:51 PM IST

Staff Writer

- Apart from Pfizer, Moderna's coronavirus vaccine also appears to work against the new and more infectious variants of Covid-19 found in the UK and South Africa

Of course [there's still this graph](#), of deaths from natural causes for people under 60 years old in South Africa, so yeah we should still go ahead and close some more borders and insist

on real quarantines and not make exceptions for citizens as citizenship has nothing to do with how viruses spread, as a prudent precaution:



The South African strain got to Australia, [where one quarantining traveller caught it from another while they were quarantining](#). Hopefully the problem can be contained by having everyone at the hotel in question and everyone who was previously there quarantine again, but it's a super scary moment.

Developments in Brazil with their new strain [continue to seem not great](#) either, especially given how much of the population should already be immune, which continues to suggest easier reinfection by the Brazilian strain.

And guess what? We've found our first case of the Brazilian strain in America, [in a traveller who was in Brazil](#). History repeats itself.

[This Wired story](#) focuses on who is to blame for the new strains, and says it was our botched pandemic response, and in particular our failure to properly handle the virus in immunocompromised patients. I've seen similar other claims elsewhere, but also a lot of objections to such claims, and don't have a strong position other than that 'who to blame' is generally not the most helpful framing.

You know what else scares me? [This scares me](#), because it indicates that Moderna expects to need booster shots, and also sees no path to getting booster shots other than going through all the trial phases over again for each new booster. So they intend to start 'general booster studies' in March, and then 'begin with their phase 3 trial participants' between June and August, and then have to wait for the data to come in while cases are presumably much

lower, and then only after that be able to start production. So the next time we get hit, it may well be many months before we can meaningfully prepare or respond.

What is the New Administration Doing About Covid-19?

It's now been a week rather than a day. Let's see how it's going.

By the standard of the old administration, which seemed to be doing *actual nothing* for the last several months of its tenure, this was a triumph on day one, I'm making a note here, huge success.

The standard of a benevolent and competent government is somewhat harsher. How'd we do?

[Here's a list of his first few days' executive orders.](#)

My main take on all these executive orders is that it seems the government is no longer capable of carrying out its basic functions without explicit executive orders, and that seems like a problem? Even when the executive changes it up and does provide the necessary orders?

I'll skip over the stuff that isn't pandemic-focused.

Biden signed an executive order [to create a National Center for Epidemic Forecasting and Outbreak Analysis](#). That is the kind of order that makes one wonder both 'doesn't Congress control the power of the purse anymore?' and also 'What do you mean we didn't have this before?' and also of course 'I take it we won't be using any prediction markets despite that being the obviously correct way to do this?' but this still has to be a net win in the long run.

He directed FEMA to expand reimbursement to states to fully cover the cost for National Guard personnel and emergency supplies. It's scary to think that failure to do this was slowing down emergency supply provision including the vaccine, but this seems to be true. Also, this is a transfer to the states, which is urgently needed.

He signed an order to establish the Pandemic Testing Board to expand US Covid-19 testing capacity. So, points for telling people to expand testing capacity. That's much better than the previous policy of actively suppressing testing capacity, and likely is a good step to improving our PCR test supply. The problem is that the main obstacle to testing is the FDA banning people from running tests, and I don't see any signs that this is likely to change much, let alone that preventing this will get the attention it deserves. I'd love to be proven wrong.

He signed an order to establish a preclinical program to boost development of therapeutics in response to pandemic threats. Again, that's presumably a clear win over not doing that, but again the main barrier to development of therapeutics is that we've banned the most efficient methods of developing and testing therapeutics, together with our failure to allow the use of therapeutics. Might want to get on that.

He signed an order to enhance the nation's collection, production, sharing and analysis of Covid-19 data. Did you know you can just do that? Order such things enhanced, and presto, they're enhanced? That's pretty neat. Good idea.

He directed FEMA to create federally-supported community vaccination centers. Good show.

He stopped the USA from withdrawing from the World Health Organization. While the WHO is a proud founding member of the Delenda Est Club, us being outside it likely wouldn't have

made things better.

He directed the Department of Education and HHS to provide guidance for safely reopening schools and childcare and higher education, and directed OSHA to release clear guidance, decide whether to issue emergency temporary standards, and ordered them to enforce worker health and safety requirements, and wait, they all needed to be directed to do that? Isn't that their jobs?

He issued a directive that CNN summarized as: "A presidential directive to restore America's leadership, support the international pandemic response effort, promote resilience for future threats and advance global health security and the Global Health Security Agenda." It seems directives have magical powers? More of us should try them.

He asked Americans to wear masks for 100 days. They should totally do that, but not sure why this is an executive order.

He created the position of Covid-19 response coordinator, which really doesn't seem like it should require a signing ceremony and a big to-do, and also shouldn't be required at all in the sense that this implies the position did not previously exist.

He extended the ban on evictions, as expected.

So, if nothing else, major points for nothing being actively unhelpful. You love to see it.

What was missing were the things that were most maximally helpful.

If I was in charge of the administration response, my top priorities would be the first three or four here, then the rest:

1. Approval of Astra-Zeneca and Johnson & Johnson Vaccines
2. Approval of all reasonable Covid-19 tests, and ramp up capacity.
3. Ramping up manufacturing capacity for vaccines, and/or buy more doses, by paying more money
4. Enable use of half doses where appropriate, and test even smaller ones.
5. Ramping up our capacity to put shots into arms, partly by paying more money.
6. Prioritization of vaccinations by age bands only to extent possible, and if possible deprioritize vaccination of those who already had Covid-19.
7. Expand vaccine efficiency through half-doses and making sure we use everything in every vial.
8. Impose travel restrictions to contain strains from Brazil and South Africa.

From Biden's first batch, I'd say the job got half-done on #8 (EDIT: looks like he banned people travelling from SA to USA, but not USA to SA and then back again, which somehow doesn't count, and that's pretty terrible), and some reasonable efforts were being made on #5, but I didn't see any of the impactful actions being taken on the other five, especially the first three, which are the ones that matter most.

But then a miracle occurred, and it turned out that *yes if you are the United States you can just buy more vaccine doses and then you get more vaccine doses, so we did that!*

(CNN) — President Joe Biden announced a series of measures on Tuesday aimed at ramping up coronavirus vaccine allocation and distribution, including the purchase of 200 million more vaccine doses and increased distribution to states by millions of doses next week.

With those additional doses, Biden said there would be enough to fully vaccinate 300 million Americans -- nearly the entire US population -- by the end of summer or early fall.

He described efforts to combat Covid-19 as a "wartime undertaking."

"We now have a national strategy to beat Covid-19. It's comprehensive. It's based on science, not politics. It's based on truth, not denial, and it is detailed," he said.

As part of the new efforts announced Tuesday, the US will buy 100 million more doses from Pfizer/BioNTech and 100 million more from Moderna -- the two-dose vaccines that have been granted emergency use authorization by the US Food and Drug Administration. Pfizer and Moderna are working to step up production, and Biden said that the additional doses will be available this summer.

The new purchase will increase the planned Covid-19 vaccine supply from 400 million to 600 million, an official told reporters on a call on Tuesday ahead of Biden's remarks.

There are two immediate reactions that come to mind.

The first is, of course, [woo-hoo!](#) Best possible news, other than good results from Johnson & Johnson. Everyone's timelines for normality can move up several months.

The second is, *wait, we could have just done this the whole time?* Things that, while positive and important, and thank you for finally stepping up, [could have been brought to my attention yesterday!](#) The costs of not stepping up on this until now have been staggeringly high.

There's a lot to do and a lot of fires to put out. A lot of the executive orders that were issued highlight exactly how deep a hole things started in, making it hard enough to merely stop digging. It's still worth noting that many of the high leverage potential actions are not being taken.

That may be why he said "[There's nothing we can do to change the trajectory of this pandemic for the next several months](#)" in the context of pushing for economic relief. Which is a deeply troubling thing for him to say. Of course there are lots of things we can do! A bunch of the things that Biden did on day one are going to do it. The extra vaccine purchases will *definitely* do that! We could *end* the pandemic in the next several months if we wanted that badly enough.

[This was HHS' Becerra](#) attempting to clean up the mess that line created:

"I believe President Biden made it very clear, the plane is in a nosedive, and we gotta pull it up. And you're not going to do that overnight," Becerra told "State of the Union" host Dana Bash on CNN. "But we're gonna pull it up, we have to pull it up. Failure's not an option here, and so we will."

Becerra went on to say that Biden's plan is "a rescue plan that should be followed by a recovery plan." He said that first the government has to "rescue the people" and the economy.

I worry that people who think we 'can't change the trajectory' will stop caring about the impact of their decisions on others, since *this statement is explicitly saying those actions don't much matter*.

We have serious action on one of our key priorities, expanding supply of existing vaccines. Now we need to push harder on allowing tests and additional vaccines from AstraZeneca and Johnson & Johnson, and on half doses. It looks good for Johnson & Johnson, but we need more pressure on AstraZeneca and on testing, and to put us over the top on half doses.

The Quest to Sell Out of Covid Vaccine

I would feel bad about my massive confusion and inability to figure out how many doses we have, [except that I am not alone](#) and this is real.

HEALTH AND SCIENCE

CDC director says federal government does not know how much Covid vaccine the U.S. has

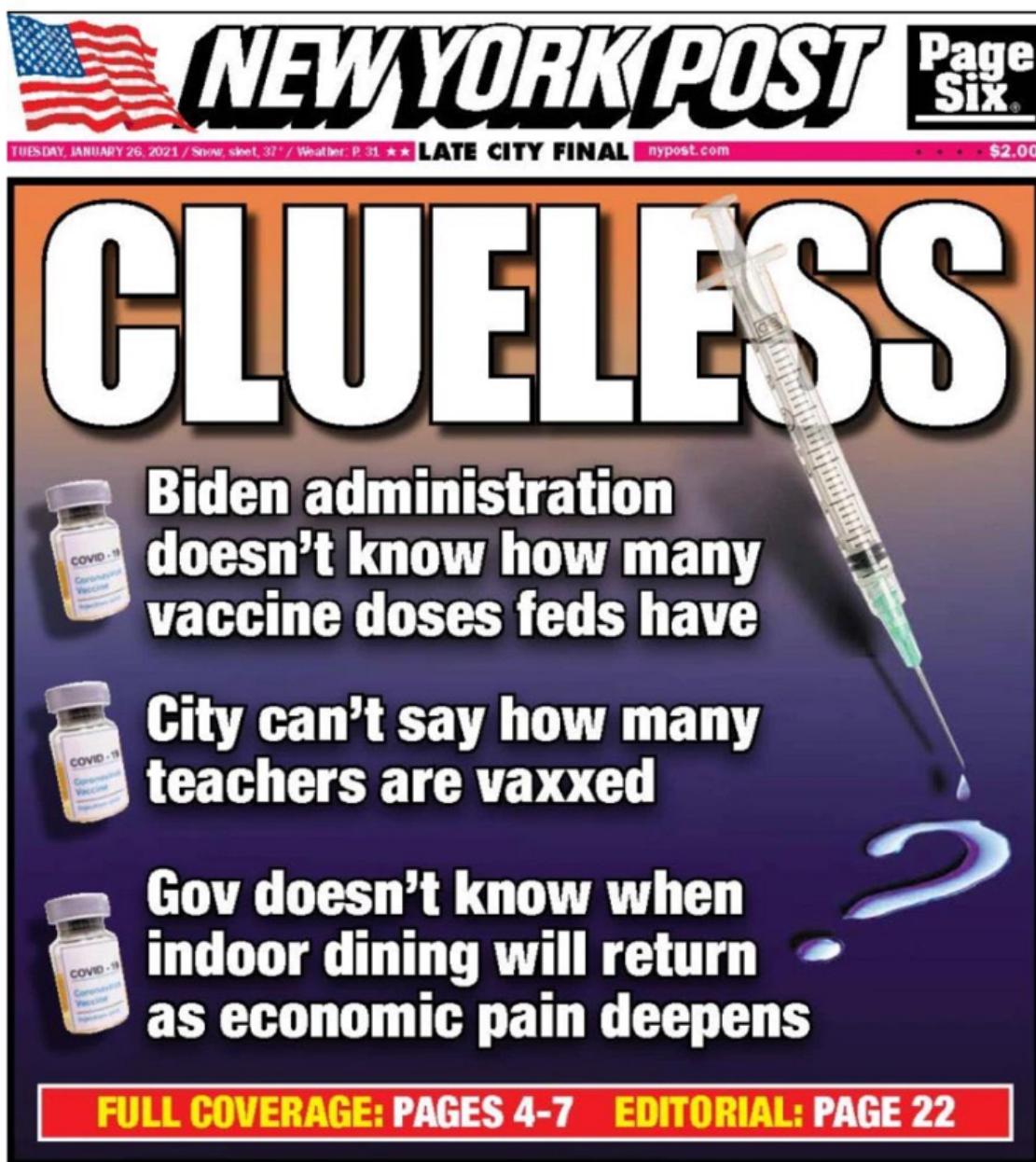
- "I can't tell you how much vaccine we have, and if I can't tell it to you then I can't tell it to the governors and I can't tell it to the state health officials," CDC director Dr. Rochelle Walensky told "Fox News Sunday."

"If they don't know how much vaccine they're getting not just this week but next week and the week after they can't plan. They can't figure out how many sites to roll out, they can't figure out how many vaccinators that they need, and they can't figure out how many appointments to make for the public," Walensky said.

“The process of distributing the vaccine, particularly outside of nursing homes and hospitals, out into the community as a whole did not really exist when we came into the White House,” Klain told MSNBC’s “Meet the Press” on Sunday.

[It's a giant mess out there.](#) Millions of doses are missing.

Or in layman's terms:



Logistics is hard, especially when you have no idea what is going on or how much inventory you have let alone where it is or where it is going. And as Yogi Berra said, if you don't know where you are going, you might not get there.

It is one thing for the White House to not have a plan. We all knew the previous administration did not have much of a plan. It is another thing to completely lose track of what is going on and where you have how much vaccine. That's kind of impressive.

It is a real question how much this lack of knowledge matters. On the one hand, yes it makes planning on the part of states, counties, sites and people much harder. No one knows how much vaccine they will be getting. On the other hand, aside from having to cancel and schedule pending appointments, it's not that clear how much value of information is here. What behaviors, other than short-term appointments, need to change? Should capacity be different based on existing supply, or (as I'd strongly suspect) does it almost entirely depend on the pace of future supply?

One danger is that doses might be held back en masse more often as second doses when they could be used now as first doses, because there is no confidence in future deliveries, delaying the whole process and increasing risk of spoilage. In some places this is clearly happening.

Still, the good news is that the core behaviors don't change. Use what vaccine you have to vaccinate people, get shots into arms, expand capacity and be ready for when there is more vaccine later. Uncertainty certainly can interfere with the incentives and motivation to expand capacity, but my understanding in at least most places is that this won't impact the important practical barriers to expanding capacity, since they're almost entirely regulatory limitations and other rules from on-high combined with time required to set things up, and getting vaccine to distribute even once is enough to set up the capacity involved, so long as it doesn't sit around unused for way too long afterwards.

The uncertainty involved here is not great, but given that we also have high uncertainty about our manufacturing capacity and how many new doses will be available in the future, it's not at all clear to me that this is a substantial additional burden or that it will much impact our long term path. Mostly it's deeply embarrassing, in a #YouHadOneJob sense.

So while noting that 'as far as we can tell' is not all that far, as far as we can tell, how is it going?

Compared to before, [I'd say pretty well](#):



PoliMath @politicalmath · Jan 21

Um.

...

Good?

Yay?

Is this supposed to be a scandal?

CORONAVIRUS

Health experts blame rapid expansion for vaccine shortages

Over the past few days, authorities in California, Ohio, West Virginia, Florida and Hawaii warned that supplies were running out.



By Associated Press on January 21, 2021

8

25

179



PoliMath @politicalmath · Jan 21

...

I mean... didn't the federal government say that they would prioritize sending vaccines to states that actually use all their vaccines?

So... isn't that incentive to drive states to, I don't know, use all their vaccines?

Isn't that what we want?

The details, though, are contradictory and hard to sort out.

[Bloomberg says](#) that only a few states have managed to use over 80% of their vaccine allocations, and none have used over 90%. They have New York as having used 58% as of January 24.

Cuomo vehemently disagrees, [and says the state is fresh out](#), and certainly the state has been cancelling appointments like they're fresh out:



Morgan Mckay
@morganfmckay

...

Replies to [@morganfmckay](#)

Cuomo says 97% of COVID-19 vaccines that have been allocated to New York has been used so far

Top Regions for administering the vaccine:

- Southern Tier: 100%
- North Country: 99%

Low Performing Regions:

- Mohawk Valley: 77%
- Capital Region: 82%



Morgan Mckay @morganfmckay · Jan 22

...

Replies to [@morganfmckay](#)

Cuomo says by the end of the day TODAY all the vaccines that have been delivered to NY will have been used.

Week 6 allocations will be arriving "as we speak."

Only 250,400 doses will be coming in for week 6

(This looks weird, so I checked: The Capital Region is 1.17 million people, Mohawk Valley is 432k people, and NY overall is 19.45 million people, so those two regions would have the majority of unused doses if Cuomo's numbers are correct here.)

The difference between 58% usage and 97% percent usage is rather large.

If it is a reporting lag issue, it is a rather large reporting lag issue, and would presumably be and imply excellent news about distribution.

If the rest of the doses are truly gone, then New York has managed to throw out a full third of the doses we were given, and that should likely be the biggest scandal of the entire pandemic. Heads would need to roll, ideally literally. We're terrible, but I do not think we are *that* terrible.

The other possibility (that doesn't involve anyone lying or committing fraud) is that this is an interpretation issue. In that scenario, New York State is reserving second doses and calling those vaccines 'used' whereas the official tally notes they haven't been used because they haven't been used. If that's the case, then about 39% of doses are reserved this way.

Using Bloomberg's numbers, New York has given about 88% of its delivered shots as first doses, or 1.4mm out of 2.4mm allocated doses, so the math would work if the difference is

the reservation of most second doses. [It does seem that this is largely the case](#), with Cuomo saying the use of second doses (that are in storage) as first doses is 'up to the federal government' which I'm guessing is true in practice if and only if Cuomo decides that it is up to them. I also know that not all second doses are held in reserve, because I know someone who has a second dose appointment but no confirmed second dose, and the site reports having cancelled some second doses for lack of supply.

[Virginia is an example of a place that is both not selling out and also not supplying vulnerable people with the ability to get vaccinated:](#)

Governor Northam insists on controlling our lives, now including when and where we may receive the COVID vaccine.

As of tomorrow, hospitals in Virginia will no longer be able to administer COVID-19 vaccines. Thousands of [elderly people are having their vaccine appointments canceled](#). From now on, all COVID-19 vaccines will go to the local health departments and none directly to hospitals.

Virginia Hospital Center had been running clinics all day every day to give people the vaccine. [Appointments there for all 1st dose vaccines have been canceled](#) because the hospital will no longer be able to get the vaccines.

Northam's health department has also forbidden people from crossing county lines to get the vaccine. If the county next to you has an abundance of the vaccine, you can't get it. Only residents of that county may get their vaccine.

These new rules will result in many people either having their vaccination appointment canceled or delayed for months. Currently, 7.5 million people in Virginia, Maryland, and DC qualify to get the vaccine, if only they had access to it. The new rules limit the options citizens have for getting the shot. Everyone MUST go through their local health department to be vaccinated. That means in a county such as Loudoun, with a population of over 420,000, and two health department locations to receive the vaccine, will continue to inoculate 400 to 900 people a day. There are no other options. The Loudoun health department has said they are trying to open a third location for vaccinations (possibly at Dulles Town Center) but that could take months. If Loudoun continues at its current pace it will take well over a year for the local health department to inoculate all those who want vaccines. If Loudoun hospitals were allowed to open clinics for vaccines, many more people could be inoculated every day but the Northam administration will not permit it.

I talk about New York a lot because New York has a lot of reporters, Cuomo gets a lot of extra attention, and also I live in New York. Virginia is highlighted here because that's where

George Mason University is, hence the attention from Marginal Revolution. No doubt there are lots of other disasters in other places that are getting less attention.

Kittaruss County in Washington claims to be doing extraordinarily well, with zero doses wasted and 95%+ distribution within a week of getting doses. [They did it exactly the way you'd expect](#), by having experience with other disasters, and doing the work of finding people to vaccinate, while going down the priority list as needed when they had extra shots. One thing not mentioned is how efficient they've been extracting extra doses, since 'no doses wasted' could imply insufficient attention to that factor.

West Virginia and Alaska still win the crown. They might not have quite given 110%, [but WV is still claiming to be at 101.4% at least for first doses](#), and are well ahead on percent of population vaccinated:

West Virginia COVID-19



The way that graphic is organized makes it clear that there are generally two sets of books being kept, one a percentage of doses administered and the other a percentage of *first* doses administered, and a lot of the confusion is that the two measures are being frequently conflated.

The Quest for More Vaccine and More Vaccines

Johnson & Johnson's vaccine is potentially a true game changer. It is one dose, does not require special temperatures, and they are claiming 100 million doses by the end of April. Fauci is now on record [expecting a decision on the vaccine by February 7](#) and [general expectation is next week](#). What's confusing is that it seems that the process is now something like, first we get the data, then we analyze the data, then we make a decision – so far so good – and *then* we apply for Emergency Use Authorization.

This seems accurate to me. First, data is gathered, then the data is analyzed, and once everyone agrees that we should grant approval, then we apply and a meeting gets scheduled for several weeks later. I hope to be wrong and pleasantly surprised.

In a sane [world be analyzing their results now and at most waiting for the formal end of data gathering to instantly approve distribution](#):



Walid Gellad, MD MPH @walidgellad · 9h

Every AM, I get on twitter hoping to see J&J vaccine results.

...

It's unfortunate they're holding back on efficacy analysis until 2mths median follow up. Knowing whether the vaccine is effective is more important right now - let the safety data mature while efficacy is reviewed.

Then again, this isn't the biggest deal, [if J&J is truly 'going all out'](#) in its production capacity. [They're looking to have 100 million shots \(of their one shot vaccine\) available by April](#). Snags will mean it is going to miss its goal of 12 million doses available in February. It would be much better to get those shots into arms in February rather than March, but the key is to create as many shots as possible and then put them into arms. I am confident that we'll be able to catch up, and get all the shots into arms quickly once approval does come, now that supply is rapidly becoming the limiting factor.

Meanwhile, Pfizer has decided that since we've managed to extract more doses out of their vials, that means they've delivered extra doses, [and so based on their contract they can ship us less vials](#), as they did with Denmark:

Last month, pharmacists across the US found a pleasant surprise when they discovered that the vials of Pfizer's COVID-19 vaccine contained extra doses.

As a result, Pfizer will now ship fewer vials of vaccine to the US to account for that, according to a New York Times report. The pharmaceutical company has committed to providing 200 million vaccine doses to the US by the end of July. The extra doses found in the initial allocations will now count toward that number.

Pfizer charges by the dose and for weeks has reportedly pushed officials at the US Food and Drug Administration to formally acknowledge that the vials contain six (and sometimes seven) doses, instead of five.

I have zero problem with Pfizer *getting paid for* the extra doses. I'm actively in favor of that, as I do not have a fear that someone, somewhere is making a profit. What I definitely have an issue with is them *delivering less vials* to us as a result, especially if they then use this as a reason not to ramp up production, or ship those doses instead to places that aren't extracting the extra doses and instead throwing them out. This seems like a very easy

compromise to have struck – we agree to pay for the extra doses, Pfizer agrees to still give us all 40 million vials.

Scott Gottlieb [makes an interesting defense of Pfizer's move here](#), that only by giving everyone no choice but to use the extra doses, can they ensure that the extra doses get used:

Dr. Scott Gottlieb, who sits on the board of [Pfizer](#), on Monday defended the company's move to ship fewer vials of its Covid-19 vaccine and count six doses per vial, instead of five, saying that it's the best way to ensure the extra dose gets used.

When the company began shipping vials of its vaccine last month, pharmacists discovered that they could often extract an extra dose from each vial that, on paper, only contained five doses. That discovery meant that the United States might actually get more doses of the vaccine than the 200 million the Department of Defense purchased under its contract with Pfizer.

“The bottom line here is that this is a very scarce resource. We need to make sure every dose gets used,” Gottlieb said on CNBC’s “[Squawk Box](#)” on Monday. “The only way to do that is to market this as a vial that has six doses and provide the proper equipment to extract that sixth dose, which in fact Pfizer is doing.”

There was then a big to-do about the fact that the syringes necessary to extract the extra doses often aren't being provided, and thus the extra doses couldn't be extracted and thus shouldn't count. This of course points out that *the extra doses were being wasted by this before, and Pfizer only fixed it when it got them paid*. Which is exactly why you want to pay them for the extra doses. If it got the syringe problem fixed it was worth ten times the extra cost for that alone.

I fail to see why shipping less doses is necessary to get maximum uptake, given there's a massive shortage either way, and it doesn't seem like this will prevent all the vials from being distributed, which is the only thing that matters. I hope.

The key as always is to keep the vaccine flowing as fast as it can be manufactured, and to ramp up that speed of manufacturing as fast as possible. Payment is one thing, but decreasing vial allocation to people who efficiently use the vials they get is a rather low and destructive thing to do, and it would be unwise to tolerate it.

In the short term, looks like supply will be increasing slowly, [but definitely increasing](#) (WaPo):

Weekly Covid-19 Vaccine Allocations

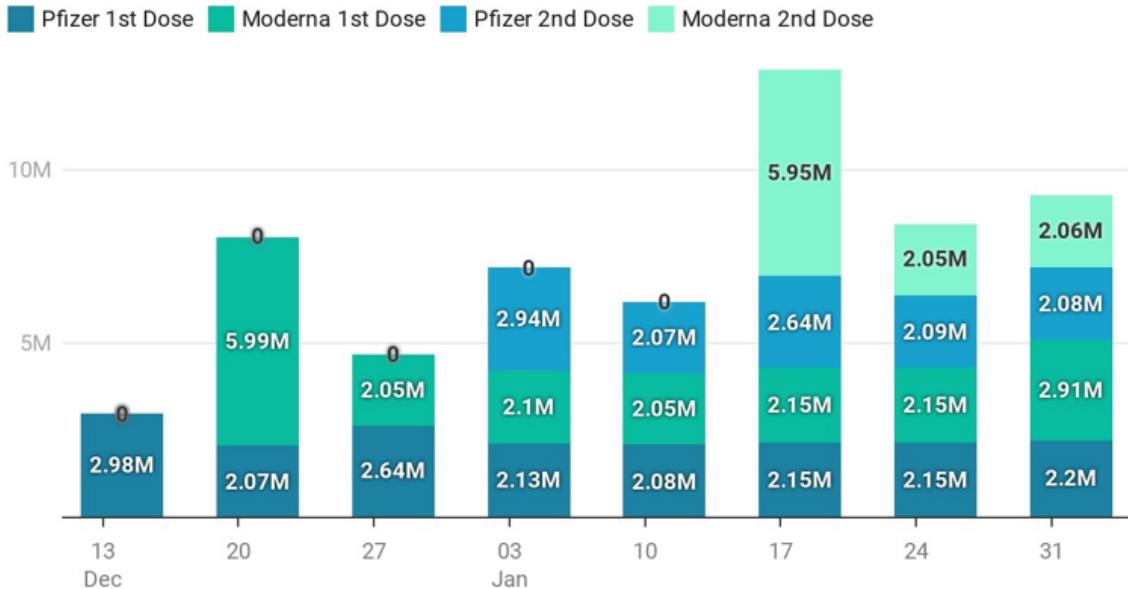


Chart: Benjy Renton • Source: HHS • Created with Datawrapper

[Pfizer also going to deliver doses faster than expected \(Bloomberg\)](#).

In ‘actually being helpful’ news, while they wait for their own candidate to become ready, [Sanofi to produce 100 million Pfizer vaccine doses, CEO says](#). The ability to profit from the deal allows Sanofi to take helpful action here, which is excellent. Could we get anyone else to follow suit? How about Merck, [whose vaccine unfortunately didn’t work out](#) and was cancelled this week due to lack of immune response?

AstraZeneca meanwhile is suffering supply problems, [and cut their first quarter allocation of vaccine to Europe by 60% this week](#). I hope that they’re giving Europe all the doses we’ve refused to approve here in America, so they’re not sitting on the shelf unused. It’s kind of weird that we took one of the most promising vaccine candidates, and made an exclusive deal with one pharma company but required them not to make a profit until after ‘the pandemic is over.’ Might not have been the best incentive structure Oxford had available. I thought the chaps there were smarter than that.

This led to a fight between the EU, UK and AstraZeneca, because the EU wants you to know that when they pay for second class vaccine distribution three months late, they will *not* be treated as ‘second class citizens’ and the UK ‘better think twice’ [or they’ll prevent the vaccine from being shipped](#):

[00:02:44] Four already, five weeks now, more than five weeks to five, the biotech vaccine that is only produced in Europe, that has been developed with the aid of the German state and the European Union, money is shipped to the United Kingdom. So people in the United Kingdom are vaccinated with a very good vaccine that is produced in Europe, supported by European money. If there is anyone thinking that European citizens would accept that we give this high quality vaccine to the UK and would accept to be treated as second class by UK based company. I think the only consequence can be immediately stop the export of the biotech and then we are in the middle of a trade war. So the company and the UK better think twice. If they really want to stop all the export

[00:09:57] We see that Europe is not treated well, not from the United States and not from the U.K., and then we have to show our weapons. Europe was always open. We wanted cooperation. Europe was the initiator of Kovács. But in the meantime, U.K., you said it yourself. UK did the treaty UK first. So we need to react on this. If it's UK first and if it's us first, we need to tell to other companies in the world, if we treat the Europeans as second class, you will suffer for this. And I'm sure I can repeat, they have already understood the message. There is already good news from today.

Meanwhile, *they haven't even approved the vaccine in question yet.*

And they are saying that when they do eventually approve the vaccine they're rattling sabers about not getting, [they might not approve the vaccine to older people](#), because they have a "limited amount of data" on that subpopulation.

Germany has gone ahead and done exactly that, recommending the vaccine only to ages 18-64 and not to the elderly, those years becoming ever more the magical talisman.

Don't get me wrong, AstraZeneca *absolutely* messed up their trials in several ways and this is one of them, they've been a disaster as good vaccine candidate shepards, but punishing the world to spite AstraZeneca's face in a way that mysteriously *lets them make more money* because they get paid higher prices for later sales doesn't seem like a punishment that fits the crime.

I'm also worried that it is possible that no one in the trial was named Erica, so perhaps it doesn't work on people named Erica. Please, Erica, consider whether this is safe. Erica?

One Dose, Two Dose, Half Dose, Who Knows

The ultimate low hanging fruit in expanding vaccine doses is to give half doses of vaccine to young healthy people, thus greatly expanding supply. We know that the Pfizer and Moderna vaccines generate robust immune responses in young people when given two half-doses, similar to that for full doses. That second half of each dose is what was used in the Phase 3 trials, so it is what we are doing, but it is mostly wasted. I would *happily* accept two half doses earlier, rather than two full doses later. And that's exactly the choice we have to make.

Vivek Murthy, who is playing a key part in the Biden Administration's response, said on a podcast with Ezra Klein that his concern is with the *duration* of the vaccine response from a half dose, because "we don't know" (which is another "there is no evidence for") that the response will be as long lasting. Maybe it fades away faster, you see, and then we'd need to use *more* vaccine in the future. Which is nonsense. If you generate the same response there's no reason to worry that you'll then remember it was from a half-dose and fade away in a few weeks. Like all such FUD concerns, I suppose it's *theoretically* possible, but come on. He also suggested a similar set of FUD about delaying second vaccine doses, saying "science from other vaccines shows us there is a risk it won't work and will require revaccination." Whereas everyone who actually looked at other vaccines, that I know of, thinks that risk is minimal.

Vivek also used the term "driven science" as code for "anything not in a proper study isn't evidence." Ezra correctly pushed back hard on this, asking if "following science" meant only "certainty." Vivek thanked him for the question, said no we look at all the available information and make the best decision we can, and entirely disregarded the contradiction.

On the other hand, he was quite promising on [questions of rapid testing](#):



Ezra Klein ✅ @ezraklein · Jan 26

I asked him whether the FDA was being too conservative on approving rapid, at-home testing (Hi @MichaelMina_lab!). "I do think we've been too conservative," he said. "I do."

2

31

227



Ezra Klein ✅ @ezraklein · Jan 26

"There's a difference between public health/surveillance testing and diagnostic testing...The FDA to me speaks to our failure to think broadly enough about the kind of testing that we needed."

Encouraging! Hopefully they can change that, quickly.

We have great news. [Half doses are being seriously considered for at least for Moderna!](#) Talks have begun with the company. I don't see why this needs to be up to the company. All we have to do is allow professionals to give out half doses to people under the age of 55. I suppose Moderna wants to be paid for the half doses as if they are full doses, since they are "doses" and they're paid by the dose, and of course *I do not care about that at all, pay the money if they ask for it*. Because that matters almost not at all.

The real question is, why stop at half doses? If a half dose and a full dose work equally well, the next logical question is, what does a quarter dose do?

[Let's look at the data:](#)

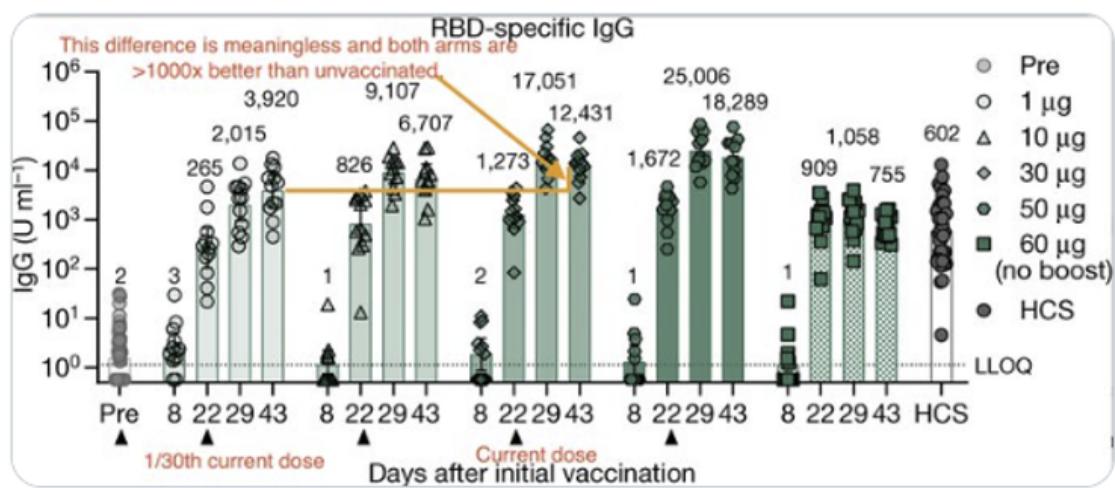


William Gibson
@wgibson

...

The fact that people are furious about half-dose strategies is nonsensical. At the current vaccine-limited rate we would immunize to herd immunity somewhere between 4 and 10 years from now.

But few know that 1/30th the Pfizer dose elicits excellent antibody responses!



[Here's the thread.](#)

Upon further analysis, I think he's going too far. I do not agree that the difference here is meaningless, or that there are no downsides to giving only one microgram instead of thirty or fifteen.

The first thing to understand is that this is a log scale of immune response. Lines that don't look that different can represent substantially different levels of response, and the tiny dose is definitely not going to always get it done.

Another issue is that when you give the immune system a *very small* stimulus, this can be interpreted as something to ignore, and tolerance can result. Thus, if you give very small doses, it risks messing up your immune response. This would be very bad, and is a real risk if you push too far.

None of that means that we shouldn't *test* tiny doses. We should certainly test quarter doses, and do more testing on half doses if we think half doses require more testing, and once we verify quarter doses are good check eighth doses to see what happens.

What this all definitely drives home, once again, is that we have had it in our power the whole time to put a stop to all this.

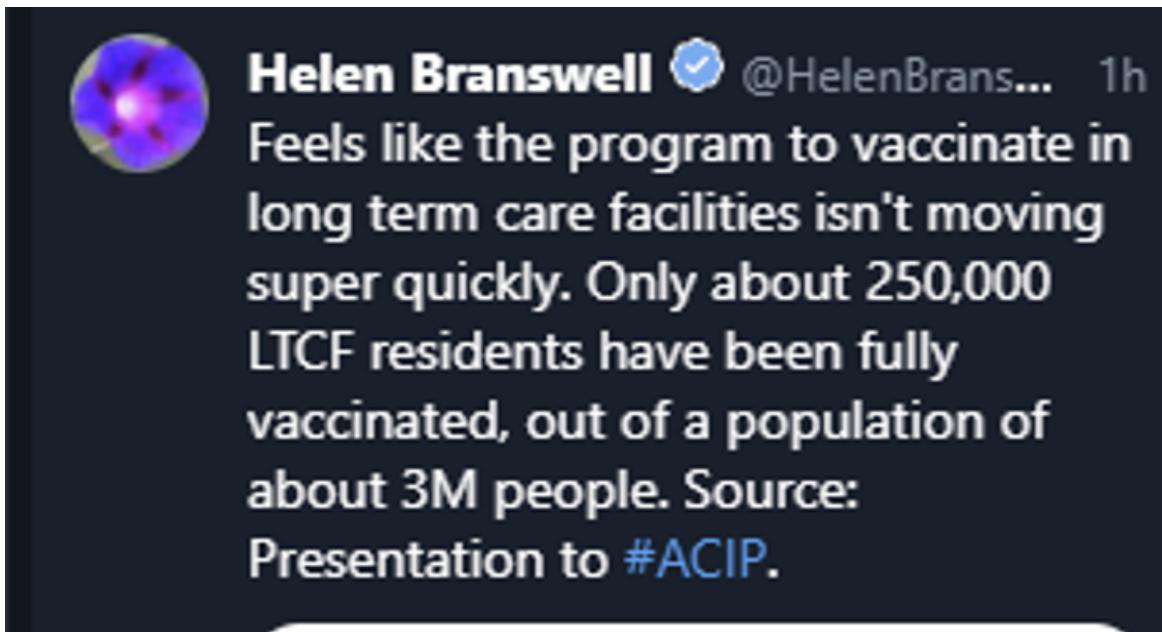
Vaccine Allocation by Politics and Power

It matters who you know. Someone I know managed to become known to someone in a position to have information on vaccine supply, and placed a call. The person in the know is going to alert my associate to new vaccine shipments in their local area, so they can call around and make an appointment before others can do so. That's how such things go.

If you are going to complicate the process of vaccination in order to do better prioritization, the most obvious group to *exclude* are those who are already immune. Vaccine given to an immune person is vaccine wasted, and that's one person in four. [Giving people antibody tests](#) before vaccination thus makes a lot of sense.

The problem is that even this maximally useful policy would dramatically slow down vaccinations, as it places another barrier into the process. In theory it's great, in practice this would backfire. A lighter version of the suggestion is to [exclude those with known prior symptomatic Covid-19](#). I'd urge those who know they have already had Covid-19 to wait on their vaccinations until we have enough supply for whoever wants it, but I wouldn't introduce any formal rules or checks on that basis.

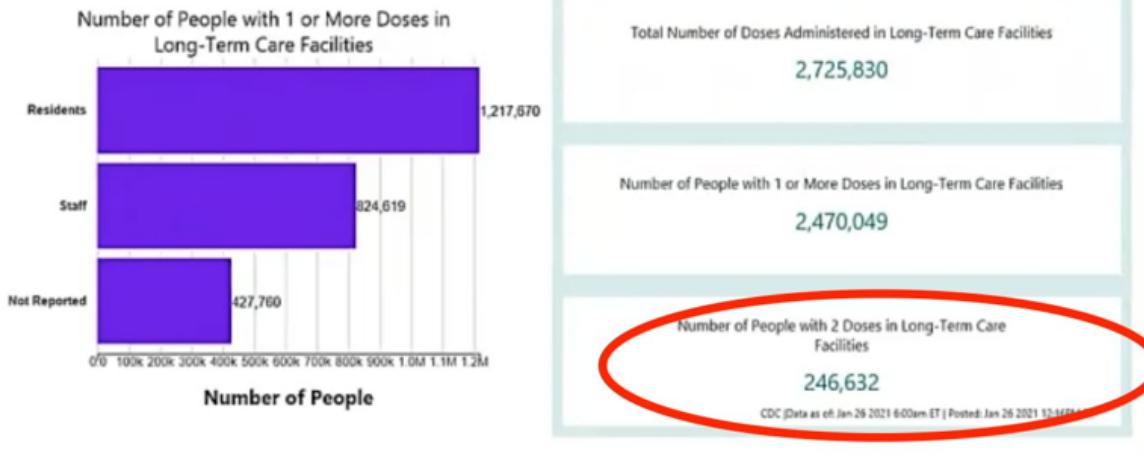
[The long term care facilities still aren't getting their shots into arms](#). It's going rather epically badly, and [not appreciating the extent until now](#) is one reason why I undershot the death count this week by so much:



A screenshot of a Twitter post from user @HelenBrans... (Helen Branswell). The profile picture is a purple circular logo with a white star-like pattern. The tweet text is as follows:

Helen Branswell  @HelenBrans... 1h
Feels like the program to vaccinate in long term care facilities isn't moving super quickly. Only about 250,000 LTCF residents have been fully vaccinated, out of a population of about 3M people. Source: Presentation to #ACIP.

Federal Pharmacy Partnership for Long-Term Care (LTC) Vaccination Program



The partnership with CVS and Walgreens does not appear to be working. The question is why. My presumption is that we're asking CVS and Walgreens to do exactly what they *don't* normally do, and go other places to deliver medicine, rather than asking them to give shots in their stores which is what they know how to do.

At the time, I was happy to see *any* use of such resources to administer vaccines. That was a mistake, and I should have been more suspicious that they were given exactly the contract they were not suited to handling, because they were not suited to handling it, thus choosing ineffective action while allowing blame for failure of the one key action item to be reallocated and claiming partnerships with companies with reputations for effectiveness. While providing no real incentives for these companies to deliver the goods. Quite the trick.

Stop Living in Fear After Vaccination



Arie Kovler @ariehkovler · 21h

Very positive vaccine news: Israeli HMO Maccabi reports just 20 positive coronavirus tests out of 128,000 people who had their second Pfizer shot a week or more ago. None of the 20 were serious cases.

140

3.9K

14K



...



Arie Kovler @ariehkovler · 21h

Not all 128k were tested, only people with known exposures or symptoms. So the real number is probably a little higher. But most of the 128k will be over 60, because they got vaccinated first. This suggests the vaccine is extremely effective.

...



Arieh Kovler ✅ @ariehkovler · 21h

...

Worth pointing out, too, that Israel is in the midst of its worst Covid-19 wave so far, with around 10% test positivity.

4

57

467

↑



Arieh Kovler ✅ @ariehkovler · 21h

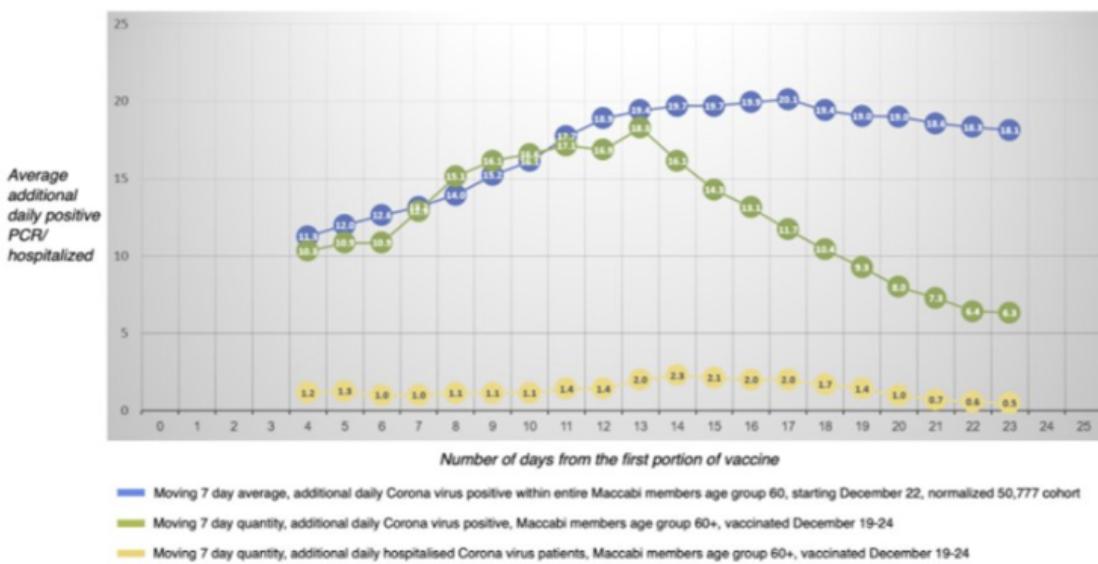
...

Final point: 40-50% of Israel's Covid-19 cases are the B117 "British" variant. So this is also good news about the vaccine's ability to protect from that mutation.

[The vaccine protects](#) against Covid-19, better than we could have hoped for. Every data point we get reinforces that. Over a hundred thousand people, mostly elderly, and zero serious cases.

As per previous findings, [big drop starts on day 14 after the first dose, which includes the standard testing lag and is a seven day average](#), but yes you need the second dose to get full effectiveness.

Cohort of 50,777 Maccabi members vaccinated December 19-24



Remember that PCR tests often come back positive for a while after you are sick, so it's likely many of the 20 positives happened before the second shot had its effect. And remember that for young people the shot will be far more effective than this.

[Here's what happens when you vaccinate old people first](#), keeping in mind coverage is far from complete:

Going in the right direction

2

Israel, change in critically ill covid-19 patients from previous week, by age group, 2021, %



Source: Eran Segal

The Economist

A sign that vaccination is starting to give Israeli hospitals some breathing space emerged a fortnight after January 2nd, the day when the proportion of those over 60 who had been vaccinated reached 40%. The number critically ill

I am not a doctor and nothing anywhere in any of these posts is medical advice, and everything in this section is definitely not medical advice, but seriously. The vaccines work.

If you're fully vaccinated, please do keep wearing masks, and otherwise reinforce good pandemic norms. But don't keep putting your life on complete hold, and especially please do not keep freaking the f*** out about Covid-19 risks, especially the risk of you infecting others.

All around me, I see people *who have been vaccinated* needlessly freaking out, or freaking others out, about Covid-19 exposure.

A vaccinated person I know is freaking out because their personal trainer tested positive, they had a 25-minute session with them mostly more than 6 feet apart, and they are worried they caught the virus then and will spread the virus to others.

A distinct vaccinated person I know saw a friend of theirs, and now has to get a Covid test before someone else will agree to see them a week later.

A distinct vaccinated person I know is planning to travel and see others, and people around them are freaking out about this as an unsafe act.

The talking point that "we have no proof that vaccinated people can't spread the virus" is successfully freaking out a lot of people who really, really should know better.

[Here is PoliMath](#) ranting about the whole situation. In addition to the central point that vaccinated people are in the worst case very, very poor transmitters of Covid-19, the second point, that saying the technically correct statement "X may be possible" in such situations is deeply misleading and irresponsible, cannot be emphasized enough.

None of this makes sense in terms of the magnitudes of the risks involved.

It all makes a lot more sense if what people cared about all along was *whether they are blameworthy for creating "risk"* rather than caring about the spread of Covid-19. If people have mostly been LARPing a pandemic rather than caring about who gets sick. Now that the actual risk is mostly gone, the act continues, because blame for "risk" is a binary and the actions in question still get instinctively processed as "risk" and people wish to instead remain in a state of grace, either in their own eyes and/or the eyes of others. Morale remains low, so the motions continue.

The motions can (mostly) stop. By all means, keep your mask on around the unvaccinated, do social distancing with respect to the unvaccinated whenever and to the extent it's reasonable to do and avoid needlessly risky interactions. But demanding tests or continuing to fully distance is pandemic LARPing.

And if *both* of you are vaccinated, yeah, within reason do whatever you want. [This article, entitled "Vaccinated People are Going to Hug Each Other"](#) and likely previously titled "Giving People More Freedom is the Whole Point of Vaccines," another good title, doesn't add much new info, but is good on this.

Calling this a moral panic seems exactly right to me.

I don't know what the 'intractable health and labor challenges" are of keeping schools open *when all adults involved are vaccinated*, [but that seems to be where we are, somehow](#):



PoliMath Retweeted



Mary Katharine Ham @mkh... 17h

Early 2020: I pulled my kid out of school for 2 days for my wedding & it was v. important I clear it with the teacher, school, & district & I still fielded phone calls about her absences anyway.

Early 2021: I mean, it's only been a year. Who needs school anyway, really?

Given the seemingly intractable health and labor challenges, some district officials have begun to say out loud what was previously unthinkable: that schools may not be operating normally for the 2021-2022 school year. And some labor leaders are seeking to tamp down the expectations Mr. Biden's words have raised.

"We don't know whether a vaccine stops transmissibility," said Randi Weingarten, the powerful president of the American Federation of Teachers, the nation's second largest teachers union.

Some virus experts, however, have said there is reason to be optimistic on this question.

For the most extreme cases of all, I've seen reports from several places of support for suspending schools indefinitely even after teachers are vaccinated. Here's the most stark one of the bunch:



OpenFCPS @OpenFCPS2020 · Jan 21

...

The FEA president, speaking before the board, just confirmed she opposes returning to school next fall 5-days a week. She said FCPS should stay in hybrid until kids are vaccinated. We note most experts expect this to happen in 2022 or not at all.

Given what I think about schools, I have mixed feelings about such stands, but if I instead felt as almost everyone does that school is highly valuable to children rather than primarily a prison sentence, I would treat such responses as insane. The FEA here is saying that even if *all teachers* are vaccinated, and also the pandemic is mostly over, the schools should remain closed until *the students* are vaccinated, despite the students being at almost zero risk even if they get infected, and the chance of anyone getting infected in that scenario being very close to zero.

If you're going to close the school for that level of risk, you should go ahead and permanently close the school and also hide in your house with your doors locked.

The teachers ([such as these in Maryland](#)) who are objecting to current calls for schools to open now have a much better case. It's definitely not safe out there and the vaccine is on its way, so the push *now* to reopen schools that have been closed for a long time seems rather perverse.

Biden is usefully leading by example, [attending church services in person now that he is fully vaccinated](#). Needless to say, he is at very high risk, and it is quite important to keep him safe. The press seemed to focus on his choice of church, rather than on the decision to attend at all, because of the general ban on things that might cause people to take more risks, including after vaccination.

Sports Go Sports

The NFL season is one game away from over. Early on there were constant cries of how crazy it was we were playing football. With each postponed game the cries of 'how can they think they will play this season' rang louder. And there were indeed some rather stupid games played on occasion. But here we are, at the end of a mostly normal NFL season, getting ready for a good old fashioned Good vs. Evil Super Bowl.

The same applies to college football. Yes, we lost a lot of games along the way, and several leagues had the self-inflicted wound of starting late, and the virus seemed to sideline my team's entire offense, but mostly we got to have our season and all the joys that come along with it.

All of that applies despite the Covid-19 situation being *vastly worse* than the situation at the start of the season, and it was all done without a bubble. Perhaps the virus can indeed (mostly) be contained. [How did we do it \(CDC link\)?](#)



The NFL did it by learning from its experiences and early mistakes, doing the math, figuring out what was actually risky, and using testing, mandatory masks, contact tracing and quarantines. The CDC report is worth reading in full, here are some key passages:

"Midseason, transmission was observed in persons who had cumulative interactions of <15 minutes' duration, leading to a revised definition of high-risk contacts that required consideration of mask use, setting and room ventilation in addition to proximity and duration of interaction. The NFL also developed an intensive protocol that imposed stricter infection prevention precautions when a case was identified at an NFL club. The intensive protocol effectively prevented the occurrence of high-risk interactions, with no high-risk contacts identified for 71% of traced cases at clubs under the intensive protocol. The incorporation of the nature and location of the interaction, including mask use, indoor versus outdoor setting, and ventilation, in addition to proximity and duration, likely improved identification of exposed persons at higher risk for SARS-CoV-2 infection. Quarantine of these persons, along with testing and intensive protocols, can reduce spread of infection."

...

"Over the course of the monitoring period (August 9–November 21), 623,000 RT-PCR tests were performed among approximately 11,400 players and staff members; 329 (approximately 2.9%) laboratory-confirmed cases of COVID-19 were identified. After intake screening, in August and early September, fewer than 10 COVID-19 cases were identified per week for the following 7 weeks (Figure), during which time the standard protocol was in effect, which emphasized physical distancing, masking, limited numbers of persons in specific areas, and other important behavioral and facility-related parameters. However, during September 27–October 10, a total of 41 cases were identified among players and staff members, 21 of which were believed to have resulted from within-club transmission at a single club, requiring closure of that club's facilities. Subsequent contact tracing identified multiple instances of transmission that likely occurred during <15 minutes of cumulative interaction within 1.8 meters (6 feet). Among the 21 persons with suspected within-club transmission, 12 had no device-recorded interactions of ≥15 consecutive minutes with a person with confirmed COVID-19, including eight who had no interactions >5 consecutive minutes and seven who had no interactions >15 cumulative minutes per day (with no other known exposures to a person with COVID-19)."

If you want to dig into exactly how Covid-19 spreads, this seems like a great natural experiment and data set to dig your teeth into. The number of positive tests still went up as conditions in the country got worse, but the problems within the league seem like they were well-contained by the protocols.

The NFL shows what you can do the hard way. The NBA is more ambitious, playing its playoffs in a bubble, [and now using trained dogs to detect Covid-19 for fans looking to attend games](#). Canine sniffing is a known detection technology that should likely be seeing widespread use, so once again the NBA is showing us what in a saner world would already be standard procedure, because they're not afraid to look a little weird.

You Should Know This Already

Masks are great, premium masks are better. If you haven't already, [it's time to step up to at least KN95s](#). I actually find them more comfortable than cloth. Also a reminder to take your Vitamin D.

[In-person higher education spreads Covid-19](#) and no our precautions are not adequate to stop this ([study](#)):



Jeff Ballinger @press4change · Jan 24

New @CDCgov study:

...

56% case increase in counties that had large university w/ in-person classes plus a +2.4% positivity...

Contrast:

17.9% decreased incidence w/ [#remoteteaching](#)

A lot of that is probably population effects, but given the effect size this seems too large for that to explain this away. That doesn't mean this isn't worth it. If it's worth four years of everyone's life and saddling them with life-crippling debt, who is to say whether a disease students can mostly shake off should be where you draw the line? The students mostly don't seem to think that. Note that this doesn't mean that it's *impossible* to safely run an in-person college, only that a lot of them did so highly unsafely.

[There are libertarians in a pandemic](#), who periodically point out that the government needlessly blocking private action is entirely responsible for the whole pandemic, and who have many helpful ideas on the margin. They banned tests. They banned experiments and they banned challenge trials. The Moderna vaccine was designed in January. Never forget.

Even most relatively smart people [fail to appreciate the benefits of forecasting](#). They instinctively seek to avoid information, because information causes [Not Okayness](#) and [blameworthiness](#). And one usual reaction to anything you want to shut down is to force it to show its concrete benefits, and then only count the benefits that can be proven and quantified. Whereas [in my culture](#), the quest to get good forecasts is useful in so many different ways. Not only do those forecasts have humongous value of information compared to the cost of getting them, but the act of figuring them out causes us to learn all sorts of useful things and train all sorts of useful skills, gives us valuable discipline and a proper tax on bullshit, and allows us to identify valuable sources of future information. Among other things.

It is one thing to not exclude the previously infected, it's another thing to actively encourage them to try and get vaccinated during an acute shortage, but then again, [Andrew Cuomo is the Worst](#):



Andrew Cuomo @NYGovCuomo · 5h

Did You Know: The CDC recommends that people who had #COVID & recovered still get the vaccine.

...

As we wait for more supply, get the facts you need to be ready when the vaccine is available to you.

See all the vaccine FAQs at ny.gov/vaccine

[Having issues with vial production due to component shortages will not cause us to make an effort to recycle old vials.](#)

Periodically there will be claims of "no evidence" to stop people from doing obviously correct things, [such as delaying second doses for 12 weeks in the UK](#). In this case, there's actually tons of evidence that such delays are highly unlikely to be an issue, and the objection is pure fear, uncertainty and doubt about all deviations from the exact thing done in a study rather than the result of any serious engagement with the physical world. I really wish we could reverse the "no evidence" thing back at such folks effectively. There's definitely "no evidence" that a 12-week delayed dose is less effective, and also there's no Bayesian evidence either.

[Poorly ventilated spaces spread Covid-19 highly effectively, and make talking highly dangerous.](#) Strangely I didn't see an obvious link back to the actual study, even on the Royal Society webpage, but the findings are no surprise.

[How we usually do Covid-19 'safety protocols' as a nation:](#)



Emin Gün Sirer

@el33th4xor

...

Not surprisingly, the reentry lines at [@JFKairport](#) are a total disaster, 11 freaking months into a pandemic. Get this: they disabled half the kiosks for "social distancing," which forces people to spend far more time in the serpentine line where there's no social distancing.

6:29 PM · Jan 24, 2021 · Twitter for Android

This seems to be the standard pattern. Take the valuable activity and both require it and limit its supply because of social distance, thus forcing people into lines and otherwise into confined indoor spaces that result in a lot more risk rather than less. The supreme version of this was of course the completely NYC-subway-style-packed airports back in February, but ordinary disastrous versions of this are everywhere.

Whereas here's how China rolls, leaving nothing to chance:

More than a million Beijing residents undergoing coronavirus testing amid a fresh outbreak have been administered anal swabs, which are considered more accurate and raise the chances of detecting COVID-19, said a Chinese disease specialist.

The key districts of Daxing and Dongcheng began a mass testing drive on Friday after a nine-year-old boy tested positive for the more virulent strain of the virus, first discovered in London and the southeast of England last month.

Health authorities in the Chinese capital said they were aiming to screen more than two million people in 48 hours. Among them, around 1.6 million inhabitants in Daxing were to be given antibody tests, as well as throat, nasal and rectal nucleic acid swabs.

I strongly suspect that the new tests are being favored *because* of the inconvenience rather than in spite of it, thus implicitly appealing to regime sensibilities and the accompanying unambiguous signal that they mean business, and given that the justifications for their improved accuracy don't seem great. Still, dedication, man. They have it.

Life must be lived and frequent testing helps a lot, but if you act like this (pictured include Dave Chappelle, Joe Rogan and Elon Musk):

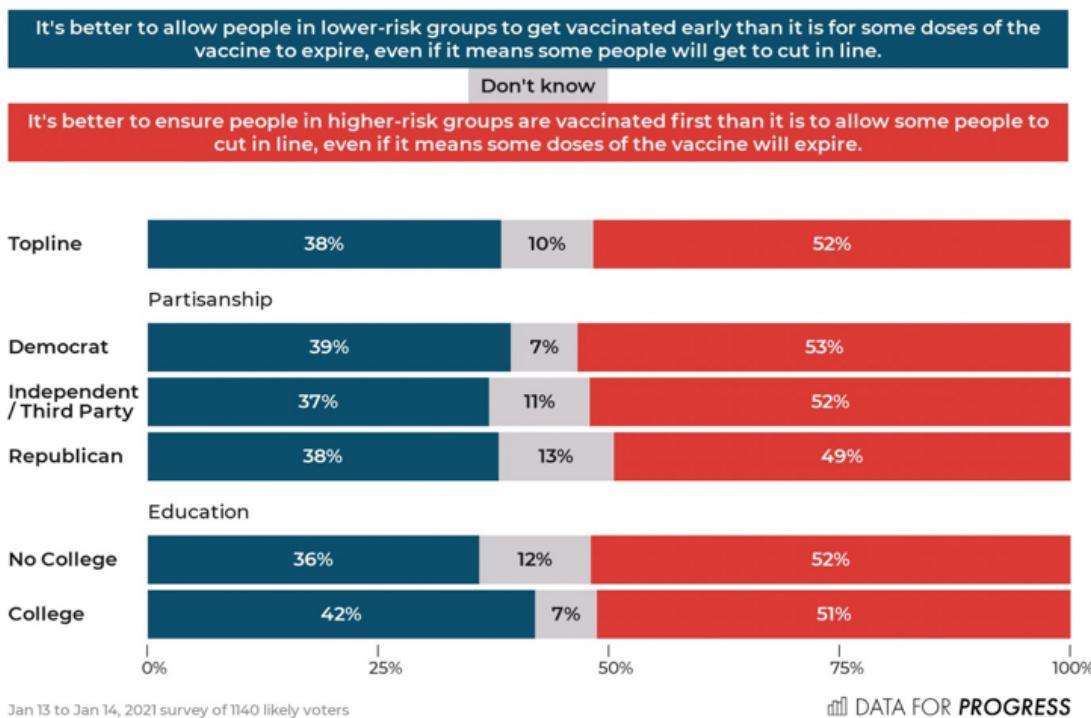


Then there is a good chance that at least one of you is going to soon test positive for Covid-19, which Chappelle did. Balance in all things.

[This is indeed a terrible way to ask the question](#) and is going to warp the answer quite a bit, because it implies to people that the deserving will get their shots slower rather than faster and highlights the frame that this is somehow “cheating,” but yes a lot of people really would rather throw a dose away than let someone “cut the line” to get it. You have two cows, someone else deserves the cows more, so we should kill the cows (from a subscriber-only SlowBoring):

A Majority Of Voters Prefer To Prioritize Higher-Risk Groups When Distributing Coronavirus Vaccine Even If Means Some Doses Are Left To Expire

When thinking about the distribution of the coronavirus vaccine, which statement comes closer to your view, even if neither is exactly right?



Not Covid-19 and I mentioned this a few weeks ago, but seriously, [Matt Levine is on liquid fire](#) and you should subscribe to his newsletter.

Sometimes people lie about their health status if you give them incentive to do so, [in this case faking negative tests](#) in order to travel.

I will not suggest allocation by scarce resources by price. [I will not suggest allocation of scarce resources by price.](#)



Eliezer Yudkowsky ✅ @ESYudkowsky · Jan 23

Remember: If it had been legal to sell half of the first million doses of vaccine for \$10,000 each, BioNTech could've easily gotten a \$5 billion loan in April, and used that money to help scale up mRNA vaccine production to billions of doses. This is not the market outcome.

70

244

1.3K



Eliezer Yudkowsky ✅

@ESYudkowsky

...

Replying to [@ESYudkowsky](#)

The bottleneck on the mRNA vaccines is the (new) encapsulation. There's a creative will to invest \$5B in that exact step in March, that happens when there's a huge payout for getting that part right. "Do your job and get paid by the govt", not so much.



Eliezer Yudkowsky ✅ @ESYudkowsky · Jan 24

...

Exam question #23: There is a SHORTAGE, a PRICE CONTROL, and a group of VOTERS yelling ANGRILY at anyone who suggests the two facts might be CORRELATED IN ANY WAY. Using your general knowledge of political history and economics, describe WHAT HAPPENED.

9

26

200



Once again, a reminder that we *could still spend that five billion*, and at the minimum do immense good in the third world. Call your local effective altruist.

Periodic reminder that we could win now [if we used a sufficient quantity of rapid tests](#), and that the entire pandemic continues to be the fault of FDA regulations. A standalone post on rapid tests should arrive at some point in the coming week, most but not all of which you should know already.

[If you fly a private plane into an isolated indigenous village, then lie about who you are and your eligibility to get a vaccine dose](#), being fined less than the cost of the plane trip for doing so is not going to be much of a deterrent.

[Experts say serology tests unreliable, as immunity doesn't require antibodies](#). What I love about this is that they're unreliable because of *false negatives*, and they *underestimate* immunity. Which "experts" are now able to point out, because it's now an argument against doing the obviously correct and useful thing of figuring out who is immune. Whereas, seriously, who thinks we shouldn't use the best immunity passport test we have if that test has some false negatives? So what?

[Oh, and Scott Alexander is back! Woo-hoo!](#)

(While Scott's return is excellent, a hit piece might still be published and the whole thing was pretty awful, so the ghosting of the New York Times will continue for now, until morale improves or Scott sends me a one-line email telling me to stop, [as per my interpretation of Scott's explanation of events](#), and we'll see what happens)

The CDC Should Know This Already

[We'll be on top of this face mask thing *any day now, really*:](#)



• **Rogue Works Progress Administration** @RogueWPA · Jan 21

...

I seriously want to see why the training is a year out of date. (We can bracket that it was stupid when written). My hunch is they can't change it without competitive bids and certification that a revised test complies with ADA and a dozen other regs.



Wine&Wit

@WineWit1

...

Interesting. Currently taking the CDC's own training to become infection prevention certified. Apparently masking for asymptomatic people is NOT a source control measure for respiratory viruses, so I got that question wrong.



Posttest: Question 3

What are examples of source control measures to contain respiratory pathogens?

Please select all that apply.

- A. Place a facemask on all visitors.
- B. Cough into the hands.
- C. Covering your cough with your bent elbow.
- D. Prompt disposal of used tissues.



Sorry, your answer is incorrect. The correct answers are C and D.

Source control measures to contain respiratory pathogens include: Placing a facemask on individuals with signs of respiratory infection; promptly discarding used tissues in the nearest waste container; and performing hand hygiene after having contact with respiratory secretions. Visitors who do not have signs of respiratory infection would not need to wear a facemask. Coughing into hands is not a recommended source control measure. The recommended practice is to cover the mouth and nose with a tissue when coughing or sneezing, or cough into the elbow or upper arm, instead of hands, if tissues are not available.

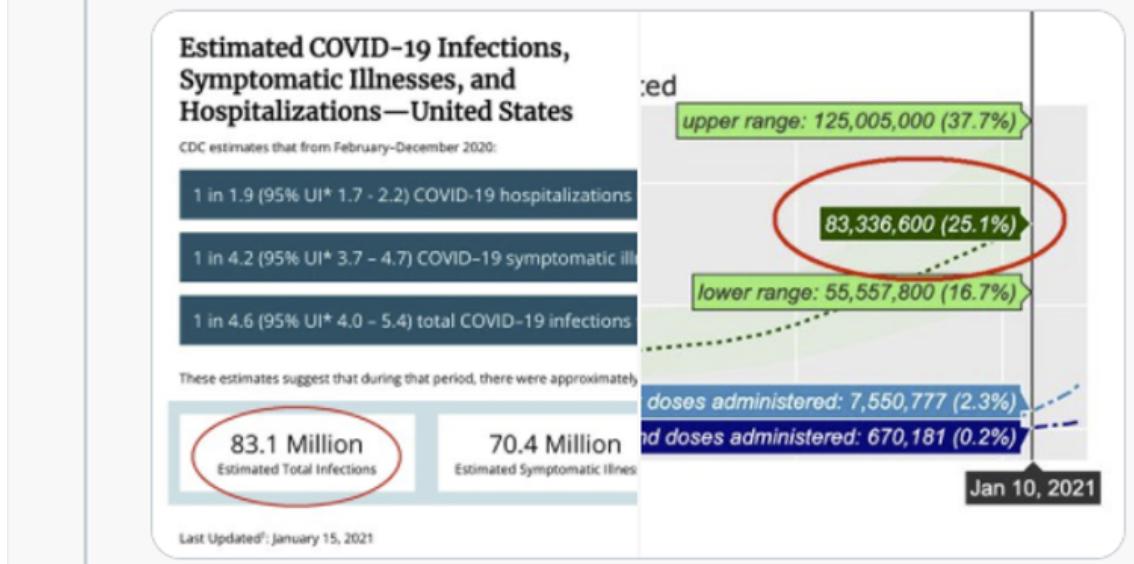
The CDC did manage this week to [fix its crazy overestimate of the true number of Covid-19 cases \(CDC link\)](#):



Youyang Gu @youyanggu · 18h

A month ago, I reported on @CDCgov's overestimate of true infections in the US.

It appears that last week, the CDC significantly lowered their estimates. It now closely matches [covid19-projections.com](#)'s latest estimate of ~83M infected (~25% of the population).





Youyang Gu @youyanggu · 18h

Their explanation:

...

"Since the previous update, CDC has received additional data about the proportion of persons with symptomatic illness who seek [...] testing services. The higher values of health-seeking behavior result in lower estimates of infections"

[cdc.gov/coronavirus/20...](https://cdc.gov/coronavirus/2020-surveillance/estimating-the-number-of-symptomatic-cases-of-covid-19-in-the-united-states.html)

Which is great, and qualifies under the policy of always praising when mistakes are fixed rather than using it as an excuse to harp on those mistakes more, but the lack of a new released methodology is still kinda suspicious ([link to paper](#)):



Youyang Gu

@youyanggu

...

Replies to @youyanggu

The CDC lowered the fraction of true COVID-19 infections that are reported by 40%, from 1 in 7.2 to 1 in 4.6.

Their new 95% uncertainty interval (1 in 4.0-5.4) is completely disjoint from their original interval (1 in 6.2-8.5).

They have yet to release an updated methodology.



Youyang Gu @youyanggu · 18h

Replies to @youyanggu

...

Given this drastic change in results, it seems reasonable (to a non-academic like myself) that their original paper in "Clinical Infectious Diseases", a top infectious diseases journal, be retracted or at the very least, be corrected.



Youyang Gu @youyanggu · 18h

I hope this is not representative of the standards for paper acceptance across academia.

...

Otherwise, I would unfortunately not blame people for being skeptical of peer review/published results.

Trust in science should not be automatically given - it should be earned.



Youyang Gu @youyanggu · 18h

I have always tried to judge scientific work by the quality of the work itself, not by who wrote it.

...

Unfortunately, this has not been always true in academia, as some have accused me of "intellectual elitism" for simply critiquing CDC's work.

I believe Youyang Gu is going to be disappointed. The (good?) news is it seems likely the CDC essentially copied him. Hopefully he does not lose hope.

[CDC admits that mixing Pfizer and Moderna vaccines is acceptable in 'exceptional cases,'](#) continues to insist they are technically not interchangeable. Also admitted up to a six week wait between doses is fine.

FDA should also know this already, [and also eases up its guidelines on second dose timing exactly when it becomes logistically necessary to do that](#). Not that it contains information anyone would have doubted:

“The FDA recognizes that getting as many people as possible across the country fully immunized will help to curtail the spread of the virus that causes COVID-19 and should be a priority,” the FDA said in a statement. “Modest delays in the administration of the second dose, if absolutely necessary, would not be expected to decrease the protection conferred by the 2nd dose and are preferable to not completing the 2-dose series.”

Works for me.

[In summary:](#)



Noah Smith 🐰 ✅ @Noahpinion · Jan 24

...

Me in 2019: America's CDC is the greatest public health institution on planet Earth, if anyone can handle a pandemic it's us

Me in 2021: HOLY SHIT, the CDC updated their website ON A WEEKEND



very bad I.N.T.E.L.L.I.G.E.N.C.E. @verybadintel · Jan 24

Replying to @Noahpinion

CDC updated the website on saturday and Sunday!

biden did fix it...

In Other News

I had a section on Bill Gates and his efforts to accelerate vaccine production, after several commenters responded to my 'someone should help accelerate production' by asking about Bill Gates. It got long and stands on its own [so I turned it into its own post](#).

[Malcolm Ocean updates us on the situation in Canada.](#)

[Oklahoma's governor bought a \\$2mm stockpile of HCQ back when it was being touted, and is still trying to return it.](#) Presumably it can eventually be used for its original intended purpose, it's still a good drug people use for real problems.

Competition for being The Worst is always intense. Consider the latest entry, a Harris County DA who (I am assuming Gokal's account is accurate) literally arrested a doctor for stealing vaccines when [he took otherwise expiring doses and gave them to whoever he could find](#):

Gokal, who worked for Harris County Public Health, was supervising a vaccine distribution site on Dec. 29 when an opened vial of Moderna doses was left over at the end of the day, around 6:30 p.m. Since the doses would expire within six hours, Gokal through his attorney said that he offered the vaccine to health workers and police on site, but they declined or already had been inoculated.

Gokal said he called a supervisor at the health department, who knew of no available patients. He then used contacts in his cell phone and administered about nine doses off-site to eligible recipients: elderly residents or those with certain medical conditions. He said he gave the final dose to his chronically ill wife after 11 p.m., unable to find any other recipient.

Budweiser, the consensus Worst Beer, is giving up the one good thing about the brand, its Super Bowl advertising (remember [the Bud Bowl?](#) So good, chef's kiss), [and instead donating the funds to Covid-19 vaccine awareness efforts](#). Lack of awareness does not seem to me to be a bottleneck, but corporate virtue signaling is still expected to be good business – Budweiser officially says this will be “good for the brand.”

[The WHO also ups its game](#) to stay in competition to be the worst, taking the classic “no evidence” line rather explicitly to deny life saving medicine to pregnant women, as in and I quote “there is no reason to think there could be a problem” but still, don’t do it:

NEWS

Moderna COVID-19 vaccine should not be used on pregnant women: WHO

By [Hannah Sparks](#)

January 26, 2021 | 9:00am | Updated

“While pregnancy puts women at a higher risk of severe COVID-19, the use of this vaccine in pregnant women is currently not recommended, unless they are at risk of high exposure,” the [official WHO status report](#) released Tuesday reads.

“There is no reason to think there could be a problem in pregnancy, we are just acknowledging the data is not there at the moment,” O’Brien said, according to a [Reuters report](#).

It also made the move of ‘[don’t provide any word on Moderna until the end of January](#)’ which you have to admit is a strong move in this competition.

[Lilly’s monoclonal antibody treatment appears highly effective in small study](#), says Lilly. Effect is big enough that it’s either real and the treatment is great, or it’s fraud.

First we turned to Starbucks, but that’s old and busted, [so next we turned to Chick-fil-A](#) to solve our drive-through congestion issues. Call the pros, indeed.



Will Haynie
@willhaynie



Chic Fil A manager Jerry Walkowiak donating his professional drive thru experience to help our vaccination program in Mt Pleasant today. When you need help, call the pros.

Haynie said Walkowiak was able to reduce the wait from an hour to 15 minutes.

Luckily the judge not only dismissed the charges but yelled at the DA for bringing them. Gokal deserves a medal. At minimum, he deserves his job back.

[Neat but not all that useful visual view](#) of Metaculus forecasts about Covid-19.

Marginal Revolution links [to this thread](#) about an outbreak in Peru. Takeaways seem to be that 30-40% seropositivity does not always prevent subsequent infection waves, and that big enough waves can get to 70%+ seropositivity, with a lot of excess death.

I am strongly in favor of experimentation in general but I'm going to make an exception and say that we'd all be better off if Pfizer didn't [test its vaccine in children ages 12-15](#). When I heard about this, my thought was 'no one would be so foolish as to want to prioritize children who can't even get sick' but no, actually, lots of people are exactly that foolish slash selfish and are ready to demand that their precious snowflakes get the protection they need ahead of others who can actually be harmed. [And in fact Israel is expanding its vaccination efforts to 16-18 year olds right now](#). So it's a vital line in our defense that we can claim that we don't know the vaccine works in children, and it would be a shame to lose that defense until our vaccine supplies are adequate. Hopefully we can string this study along until then.

[It's happening:](#)



Amazon News @amazonnews · Jan 24

...

Starting today, [@Amazon](#) will help vaccinate eligible people against [#COVID19](#) at a pop-up clinic in Seattle.

And we're just getting started. Find out more ➡ amzn.to/2Y3vdXo



5

60

193



California continues to make strange decisions regarding its lockdown procedures, [lifting regional stay-at-home orders](#) that encompassed 90%+ of their population on Monday. All hail

the control system. Things in California do not seem to be going sufficiently well that, given you'd instituted such an order, it would make sense to lift it unless your priorities (and perhaps something else that isn't the level of infections) had suddenly changed.

[Some people are confused.](#)



Aidan Smith @aidan_smx · Jan 25

...

ICU availability in Southern California is at 0%! They just detected a novel strain of COVID currently only in CA! They're literally lifting air quality regulations in LA so they can burn more corpses!

And Newsom's response is to ease restrictions?! I feel like I'm going insane.

You, on the other hand, have been reading this column for many weeks. So you shouldn't be.

#3: Choosing a cryonics provider

This is post 3 of 10 in my [sequence](#) on how to sign up for cryonics.

There are a fair few companies working in or adjacent to the cryonics space, but so far as I can tell*, the most commonly considered ones are:

- [Alcor](#) near Phoenix
- The [Cryonics Institute](#) (CI) near Detroit (whole-body only)
- [KrioRus](#) in Moscow

I only seriously considered Alcor and CI, which is the case for most people I know, since KrioRus is 6000 miles away from us.

Edit 9/2022: I've heard good things about [Tomorrow Biostasis](#), based in Germany and serving all of Europe. Given that most Europeans probably currently can't / don't want to deal with Moscow, this seems like a good alternative to look into.

Honestly, I chose Alcor long before I started the signup process – basically because my friends had chosen it, and I figured they'd done so for good reason. But in the interest of information-sharing (which is the whole reason I'm writing this sequence, after all), I decided to dig into the Alcor vs CI question. I came out the other side more confident in my choice of Alcor.

Process

The first thing I did was to look back into ancient history, to [this 2012 LessWrong question](#) about Alcor vs CI. While the comments did raise a lot of important considerations and were a helpful starting point, the primary thing I found out was.... that there's a lot of catfighting in the cryonics community. Here's hoping this post attracts more constructive feedback and fewer diatribes.

Notably, I did **not** reach out to either Alcor or CI directly, instead working only from publicly available information. While I'm told that people at both organizations are very helpful and happy to answer questions, I think it's illuminating to see what they share publicly. Also, honestly, I just don't like talking to people and didn't want to devote *that* much time and energy to looking into minutiae, when the broader picture was already pretty clear to me. People who have insider information are welcome to set me straight when I raise questions or reveal gaps in my knowledge.

One more note

While people argue a lot about which organization is better for them at the individual level, I think it's good to acknowledge that all cryonics organizations are part of the same ecosystem that's pushing research forward and aiming at long-term human flourishing. In this capacity, cryonics organizations all complement one another. People seem to get really caught up in the narcissism of small differences and lose sight of the fact that cryonicists, by and large, all share the same values.

Costs to the consumer

CI is usually cited as the cheaper option, but I became increasingly uncertain about this as I looked into it. If you look just at their preservation prices, then you come away thinking that the cost of whole-body preservation via Alcor is more than 7x what it is via CI. But this is

misleading, because Alcor rolls standby and transportation costs into its signup fees, while CI doesn't.

Ultimately CI does come out cheaper overall, but the cost difference may not be as stark as it first appears. Let's get into it.

Alcor

As of this writing, Alcor charges a minimum of \$80,000 for neuropreservation and \$200,000 for whole-body. Here's a breakdown of the full costs:

- **One-time application fee: \$300**
- **Standby fees:** \$180/year (waived if you overfund your life insurance policy by \$20,000, which you [should do anyway](#)) = **\$0/month**
- **Membership dues:** \$15 x (age at time of sign up) / year = \$22-\$75/month (median ~\$44/month)
- **Life insurance premiums (neuro):** \$15-\$300/month (median ~\$100/month)
- **Life insurance premiums (whole-body):** \$30-\$600/month (median ~\$200/month)

So, if you are healthy and under ~35, your monthly fees are likely to come out to about **\$150 for neuro** or **\$250 for whole-body**. Note that you get significantly discounted application fees and membership dues if someone else in your immediate family is already signed up for Alcor. If this applies to you, your monthly fees will be \$23 lower!

(You may be wondering why there's such a wide range of life insurance premium costs. I will talk a whole lot more about life insurance in later posts, but in brief, your premiums depend heavily on the type of insurance you decide to buy, and also increase steeply with your age.)

Cryonics Institute

First let's do the same calculation we just did for Alcor. CI's minimum whole-body suspension fee is \$28,000, but if you're an Annual Member rather than a Lifetime Member, it's actually \$35,000. But I don't think it actually matters that much, since as far as I can tell, it's hard to take out a life insurance policy for less than \$50,000. So your life insurance premiums end up being pretty much the same either way.

Annual Membership:

- **One-time initiation fee: \$75**
- **Membership dues:** \$120/year = **\$10/month**
- **Life insurance premiums:** \$8-\$300/month (median ~\$100/month)

Lifetime Membership:

- **One-time initiation fee: \$0**
- **Membership dues:** \$1250 once, amortized over ~50 years = **~\$2/month**
- **Life insurance premiums:** \$8-\$300/month (median ~\$100/month)

So your fees come out to about **\$100/month** if you're an Annual Member, or \$110/month if you're a Lifetime Member.

But wait! Unlike for Alcor, it doesn't end there! Behold:

Fees for Standby, Stabilization, Transport, and Cryopreservation provided by Suspended Animation and the Cryonics Institute

Standby provided by Suspended Animation	Insured Plan	Prepaid Flat-Rate Plan	Prepaid Incremental Plan
Additional Days of Standby provided by Suspended Animation	\$30,000 Life insurance payable to Suspended Animation. Covers up to two deployments plus unlimited standby days, only while a Serious Risk prevails.	\$25,000 Prepaid by refundable cash deposit to Suspended Animation. Covers up to two deployments plus unlimited standby days, only while a Serious Risk prevails.	\$7,500 Prepaid by refundable cash deposit to Suspended Animation. Covers only one deployment for up to 72 hours, when requested by a member experiencing Small or Serious Risk.
Stabilization and Transport provided by Suspended Animation	\$2,500 per day Paid by any method acceptable to Suspended Animation. Under the Insured Plan or Prepaid Flat-Rate Plan, this fee is charged only when the risk has diminished from Serious to Small but the member requests a continuation of standby.		
Cryopreservation provided by the Cryonics Institute	\$30,000 completion fee Paid by life insurance or other acceptable arrangement via the Cryonics Institute. Includes anti-ischemic medications, rapid cooling, cardiopulmonary support, perfusion with organ preservation solution, and transport to the Cryonics Institute. A reduced fee may apply if some procedures are not appropriate or possible.		
Cryopreservation provided by the Cryonics Institute	\$28,000 or \$35,000 Paid by life insurance or other acceptable arrangement via the Cryonics Institute. (Option 1 or Option 2 plan). Includes cryoprotective perfusion with vitrification solution, and subsequent maintenance in liquid nitrogen for the indefinite future.		

You can see that if you want standby and transportation, you'll end up needing to pay an extra \$60,000, for a total of about \$90,000 – on par with signing up for Alcor neuro.

When is CI cheaper?

If you already have life insurance through your employer

CI accepts funding via employer-sponsored group life insurance, while Alcor doesn't. If you can get your employer to pay for your life insurance you'll have a much cheaper out-of-pocket cost. (Note that Suspended Animation **does not** accept group insurance, so you'll still have to figure out standby costs.)

If you're 100% set on whole-body

If you're signing up for whole-body preservation regardless of provider, then preservation via Alcor still costs twice as much as preservation via CI, even taking into account the extra standby and transportation fees that CI members pay to Suspended Animation.

If you do not want standby

All of CI's 'hidden costs' are for standby and transportation. If you are making your own arrangements in this domain, then it doesn't make sense to pay Alcor's mandatory standby fees. There are three main reasons you might not want standby:

- **You live in the Detroit area**, so you're likely to already be very close to CI when you die.
- **You're contracting with a local standby team and/or funeral director** – especially likely if you're outside of North America, since Alcor and Suspended Animation are primarily set up to operate in the US and Canada.
- **You think centralized standby doesn't work**.

Quality of cryopreservation

There are two main factors that determine the quality of your cryopreservation:

1. How quickly preservation happens after your clinical death
2. What preservation methods are used

I'll go into both of these below. For a sense of how things actually go down in real-world, non-idealized situations, you can see CI's cryopreservation case reports [here](#) or Alcor's [here](#).

Standby services

What is standby?

When you're close to death, you (or your loved ones) may call for a team of trained individuals to wait at your bedside, ready to stabilize you and get you ready for cryopreservation and (if needed) transportation, as soon as you're pronounced legally dead.

Why does standby matter?

Standby exists so that you can get cooled down and cryopreserved as quickly as possible. Speed is important because your organs degrade quickly after you stop breathing and your heart stops beating.

You'll hear the word "ischemia" used in this context. Ischemia refers to deficient blood flow to part of the body – in this case the whole body, but we primarily care about the brain – and the resulting oxygen shortage. Organs without oxygen are quickly damaged (see [Wikipedia](#) if you're interested in the mechanism), and importantly, this interferes with the cryopreservation process:

One of the most robust findings in our studies, and scientific papers of others researchers going back to the 1960s, is that cerebral ischemia produces perfusion impairment in the brain in a time- and temperature dependent manner. In cryonics such perfusion impairment translates itself into ice formation. The real difference is not between Alcor and CI but between people who do not receive rapid stabilization and cooling and those who do. ([source](#))

Cryonicists often talk about "ischemic time", which refers (roughly) to the duration of time between your legal death and your cryopreservation. In cryonics, you can incur either warm ischemia (at room temperature) or cold ischemia (cooled down). Cold ischemia is less bad than warm ischemia, because being colder slows down the degradation process, but it's still ischemia. The shorter your ischemic time, the better.

How does standby work?

Standby teams aim to intervene as soon as possible after the pronouncement of legal death, to minimize ischemic time. Once death is pronounced, they stabilize the patient by lowering body temperature and restoring circulation. They also administer medications intended to improve the quality of cryopreservation (e.g. heparin to avert blood clots).

Perfusion can be done near the location of legal death ('field perfusion') but is usually done at the preservation facilities. This means that, for anyone who undergoes clinical death while not located right near their cryonics provider, there are hours and often even days between clinical death and the beginning of perfusion.

See [this case report](#) from Suspended Animation for a complete picture of the standby process.

Hospice care

A friend who recently helped cryopreserve a family member told me "The best quality preservations by far occur in hospice death near the perfusion team." This is because being near the perfusion team and preservation facility allows for the quickest preservation following clinical death, and the least ischemic time.

Alcor strongly encourages members who are terminally ill to relocate to a care facility near Alcor; if you choose to do so, their standby program entitles you to relocation assistance of up to \$10,000.

A cooperative care center is important, because some hospices will refuse to allow the standby team into the room or try to block them from acting. Alcor has a relationship with a hospice in Scottsdale, Arizona. CI does not have a relationship with a hospice. For my friend, this meant he wasn't able to relocate his family member to Michigan in their final days. He did contract with Suspended Animation, but his family member still incurred significant cold ischemia while being flown to Michigan.

Mandatory vs optional standby

As mentioned in the previous section, on costs, CI does not make standby mandatory for its members, while Alcor does. Alcor "attempts to provide bedside standby service to all members in the U.S. and Canada [subject to a 180 day waiting period after signup]" ([source](#)), and paying into its standby program is mandatory for Alcor members.

CI members can get standby and transport services from Suspended Animation by paying a fee, but in practice, only about 30% of CI members choose this option. [\[1\]](#)

CI makes a good point that "Spending large sums of money for remote standby services... does not guarantee a successful suspension." However, I think a major hole in their standby philosophy is that, while personalized, decentralized local standby *is* likely better than the 'one-size-fits-all' centralized standby provided by Suspended Animation (at least in terms of average response times), almost no one is going to go to the trouble to set up their own standby solution. So the default for a CI member is to have no standby at all, which seems obviously worse than centralized standby.

Perfusion

Alcor and CI both aim to [vitrify](#) their patients rather than just straight freezing them, a process that significantly reduces damage to the organs. However, the two organizations perfuse their patients with different vitrification solutions, and that's what I'll be looking into here.

Alcor

Alcor pitches itself as performing state-of-the-art cryopreservation. I'll just quote from [their FAQ](#):

The cryoprotectant used by Alcor, M22, was developed for purposes of medical organ banking and transplantation. It was the first solution to ever permit the cryopreservation

and subsequent long-term survival of a vital mammalian organ (kidney). M22 is a “6th generation” vitrification solution, incorporating ice blockers, chilling injury protection, and numerous other insights...

[M22] is able to vitrify at slower cooling rates, and larger volumes, than any other vitrification solution in published scientific literature...

Alcor also uses demanding “closed circuit” perfusion, the same method of circulating fluids through the body used in heart surgery and organ cryopreservation research. This permits cryoprotectant to be introduced more gently, with better temperature control, without requiring cryoprotectant concentration in blood vessels to be far above target tissue concentration.

Cryonics Institute

CI pitches itself as performing affordable cryopreservation. They use a vitrification agent that they developed in-house, called VM-1. While there are no scientific journal publications about VM-1, [this LessWrong comment](#) goes into M22 vs VM-1 a bit (linked page available [here](#)). CI also discusses their perfusion process [here](#) (scroll to the bottom). Most relevant paragraph:

[VM-1 inventor] Dr. Pichugin believes that the combination of his vitrification solution and carrier solution are well optimized for both low viscosity and minimal expense, while providing powerful vitrification capability. He does not believe in the value of high molecular mass agents such as proteins, dextrans, HES, PVP, etc, to support oncotic pressure in brain perfusion in CI's protocol because he believes these agents increase viscosity and are not necessary due to the dehydrating effect of cryoprotectants. In practice the Cryonics Institute has not seen much brain edema or the need for oncotic support in perfusions of brains with CI-VM-1 and [the carrier solution] m-RPS-2.

CI later mentions keeping costs low by using industrial-grade cryoprotectants. Their focus on costs and their description of VM-1 make me inclined to believe Alcor when they say, "VM1 was developed as a solution of simple composition for economical cryonics, not preservation of organs for transplantation." I haven't seen positive evidence that VM-1 will allow for revival.

A comment from the 2012 thread says that CI "cryoprotects only the head, allowing the rest of the body to be straight frozen with massive damage", but I think this is no longer true. While head-only perfusion is still the default, CI members can now choose to have their body perfused as well – although CI [recommends against it](#) because "body perfusion with glycerol after having perfused the brain results in longer brain exposure to cryoprotectant toxicity and ischemic damage."

Organizational longevity

Organizational failure is the number two reason I expect cryonics not to work, with number one being existential catastrophe that wipes out all of humanity. I'm far from confident that any cryonics company is prepared to weather a couple hundred years, black swan events and all.

The actions Alcor has taken – choosing a low-risk location, planning their finances for the long term, and structuring their organization so that it's not in imminent danger of falling apart – do show that they've seriously considered the problem of organizational longevity, but I'm not convinced that they're prepared for the future to be... weird.

Relevant 2012 [comment](#) from LessWrong user [shminux](#):

To quote Peter Lynch, "I want to buy a company any fool can run, because eventually one will". Making a company fool-proof is essential when the main purpose of the company is to survive several hundred years (maybe even thousands), an exceedingly rare occurrence. None of the current cryo shops seem anywhere close to having the necessary structure in place.

Alcor

For an overview of measures Alcor has taken in pursuit of institutional longevity, see pages 4-8 [here](#).

Company structure

Alcor has a [self-perpetuating board](#) (which means that the board votes on who will be on the board) that's made up solely of Alcor members, which makes it pretty hard for any hostile outsiders to take over the organization.

Financially, Alcor has made sure not to put all of its eggs in one basket. It has its main operating funds, some reserve funds, an endowment, and the [Alcor Care Trust](#) (formerly the Patient Care Trust), and there are set rules for when and how each of those funds is touched. For example, the Alcor Care Trust has separate assets and a separate board of directors, and is supported by a 501(c)(3) distinct from Alcor. This makes it so that Alcor can't dip into patient care funds in order to cover other costs, such as legal fees.

These precautions seem pretty good overall, but it doesn't necessarily seem like Alcor is prepared for extreme events like the collapse of the US dollar, or widespread and enduring violent unrest in the United States.

Long-term financial planning

We'll go into finances more in the next section, but it's worth taking a look at how Alcor responds to point-blank questions (in its own FAQ). In answer to the question "How will Alcor sustain itself for the duration of my cryopreservation?"

Alcor's long-term planning is conservative. Minimum funding requirements budget \$115,000 to be set aside to fund long-term care of whole body patients, and \$25,000 for neuropatients. Any excess funding also goes toward long-term care unless the member specifies otherwise. As a result, Alcor has more funding set aside per volume of patients under care than any other organization by a wide margin.

Alcor also segregates long-term care funds in the [Patient Care Trust](#) (PCT), which has a separate board of directors that oversees investments and ensures PCT funds are only used for long-term patient care... The Trust holds the mortgage of the building housing Alcor patients as well as majority interest in the ownership of the building. The rest of the Trust investments are held at the investment firm of Morgan Stanley...

...Using a conservative estimate, the funds should generate more than enough money to cover patient maintenance indefinitely.

Location

Alcor [intentionally chose](#) a location with very low natural disaster risk, a low crime rate, good weather (to avoid transportation delays), and access to a major airport (the facilities are a 20-30 minute drive from Phoenix's international airport). In addition, Alcor's facilities have good security, and police response times in the area are quick.

Cryonics Institute

CI is pretty open about the fact that they have no plans whatsoever. From their FAQ:

Can you guarantee success?

Sadly, we can't. No one can guarantee success, because no one can guarantee the future. No one can predict scientific progress with certainty. We believe that a very strong case can be made for the probable success of cryonics. But that doesn't mean that social disruptions aren't possible. Nuclear war, economic collapse, political strife and terrorism, are all possible, and they could end the lives of cryopreserved patients just as easily as they can end the lives of any of us.

and

Can you guarantee the safety of patients?

The oldest patient currently still being held in cryopreservation is Dr. James Bedford, who was cryopreserved in 1967. He has survived the Cold War, the Vietnam War, the Gulf War, Watergate, the collapse of the Soviet Union and the 9/11 attacks — which is more than many of his contemporaries can say. The world is (relatively) stable at the moment, global world war doesn't seem likely, and the economy is relatively stable.

We can't guarantee the future. But we can and do guarantee this: that at CI we will give our very best efforts to see our member patients are restored to life and good health. The life of every director and officer and member of CI depends on those same efforts.

It's also worth noting that I didn't find anything about long-term financial planning on their website, and that their location is not optimized along the lines of Alcor's – while CI is very near a major airport (again, a 20-30 minute drive), it's also just outside the [infamously high-crime](#) Detroit, and Michigan is subject to [a fair number](#) of natural disasters. (Michigan might not get many earthquakes, but it does regularly get thunderstorms and blizzards, which frequently cause delays both at airports and on roads.)

No but actually??

[shminux](#) made an excellent point back in 2012 that I have yet to see addressed anywhere. I think [it's](#) worth quoting in full:

[M]y biggest concern is the continuous operation of a cryoshop over the potential centuries or even millennia until the revival is attempted, as nearly no entities have ever survived that long. I have been unsuccessful in my search for an Alcor executive explicitly responsible for existential risk analysis and mitigation.

By existential risk to the company I mean an event that would result in the company failing to the degree that the stored patients are discarded, even though the outside world merrily hums along, and not an event that wipes out a large chunk of humanity.

The FAQ does not seem to answer the obvious hard questions like "what if Morgan Stanley goes under?", "what if the US dollar collapses?", "what other existential risks exist, and what are their probability estimates and error bars?", "what is the estimated lifetime of Alcor until it suffers a complete failure from one of the existential risks to it coming to pass?" etc. By the way, if you think that the answer to the last question is "infinite", I recommend a basic probability and statistics course.

In other words, the risk management appears to be at the level no better than that of a regular insurance company, which is completely inadequate for an organization whose long-term survival is the most critical issue. Is this perception wrong?

Seems right to me.

Finances

(Please help make this section better! Finances are not my forte [\[2\]](#))

The fundamental financial need of a cryonics organization is to be able to pay for the preservation and indefinite storage of its members. To do this, they have to balance [charging high enough prices that they get enough money per person to cover costs] and [charging low enough prices that they can attract new members and retain old ones]. They should also be conservative in their planning, and wise in their spending and investment.

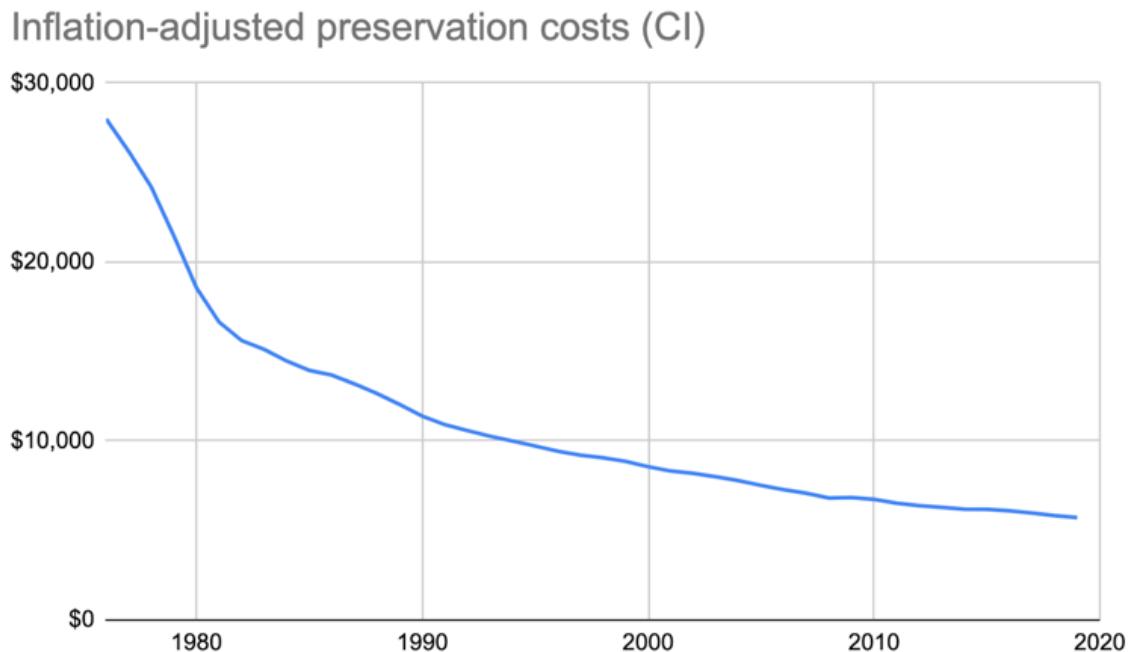
Responses to inflation

Cryonics Institute

CI has not raised its prices since it was founded in 1976. [\[3\]](#) This makes me extremely nervous. CI themselves [point out](#)^[4] that not charging enough [can bankrupt a cryoshop](#), and I don't see why they're not more worried about that for themselves.

Sure, CI's expenses have stayed constant and quite low over the past ~15 years, but it still seems like bad financial planning to keep costs the same over a period that's seen 363% inflation! More than anything, eating continuously decreasing real costs for 45 years indicates to me that CI isn't taking long-term planning seriously.

This is what it looks like to keep costs at \$28,000 starting in 1976:

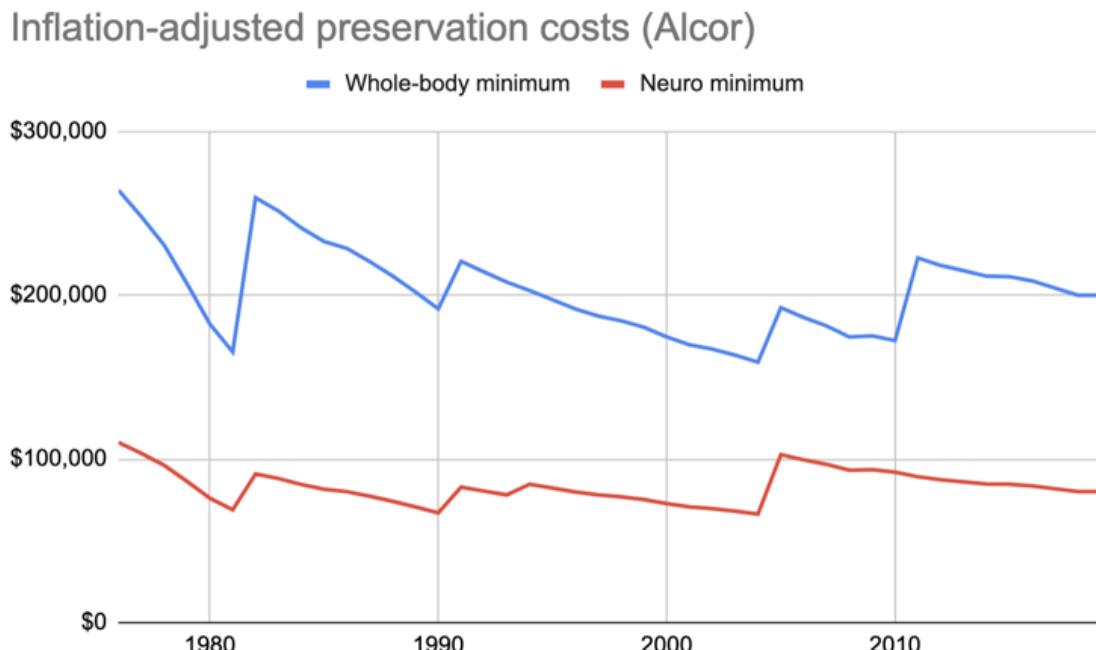


Alcor

Alcor has raised its prices multiple times since its founding, and even they are [struggling](#) to [keep up](#) with inflation. Here's a history of Alcor's cost increases that Mati Roy and I pieced together from [this essay](#) and the Wayback Machine:

	Neuro minimum	Whole-body minimum	Inflation rate since then
1976	\$25,000	\$60,000	254%
1982	\$35,000	\$100,000	127%
1991	\$45,000	\$120,000	68%
1994	\$50,000	\$120,000	57%
2005	\$80,000	\$150,000	27%
2011	\$80,000	\$200,000	12%

And here's what that looks like inflation-adjusted:



You can see that Alcor has kept real costs fairly steady over time – and that it's due for another increase soon. Dues and application fees have followed a similar pattern of periodically adjusting upwards for inflation, though those adjustments are smaller and more frequent.

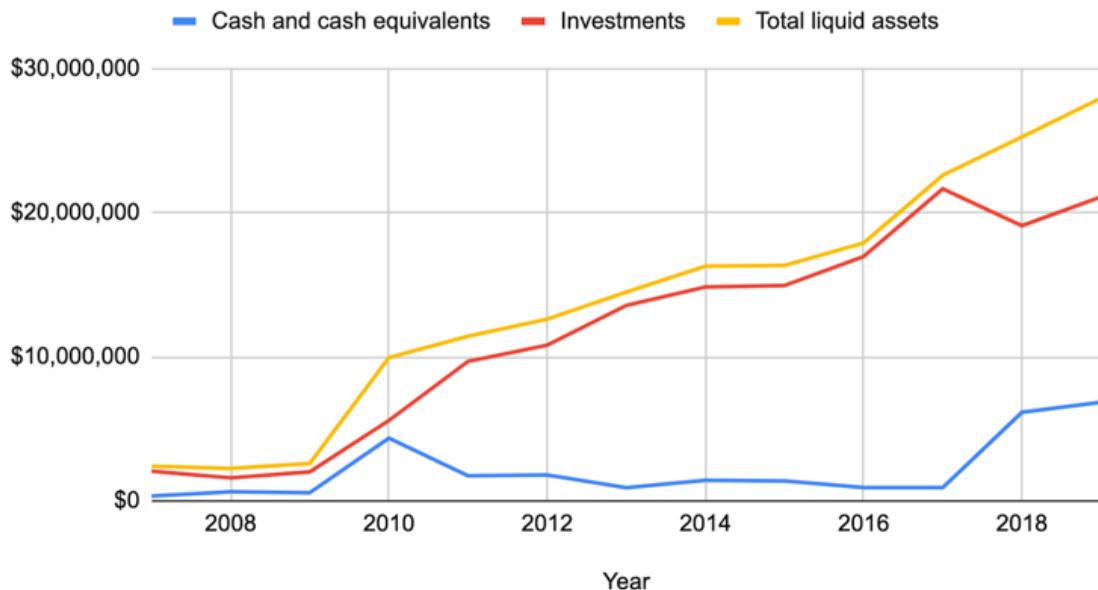
Alcor previously had a policy of grandfathering members in at the prices at their time of signup, but this policy is no longer in force, which I think is a wise financial decision on their part.

Investments & assets

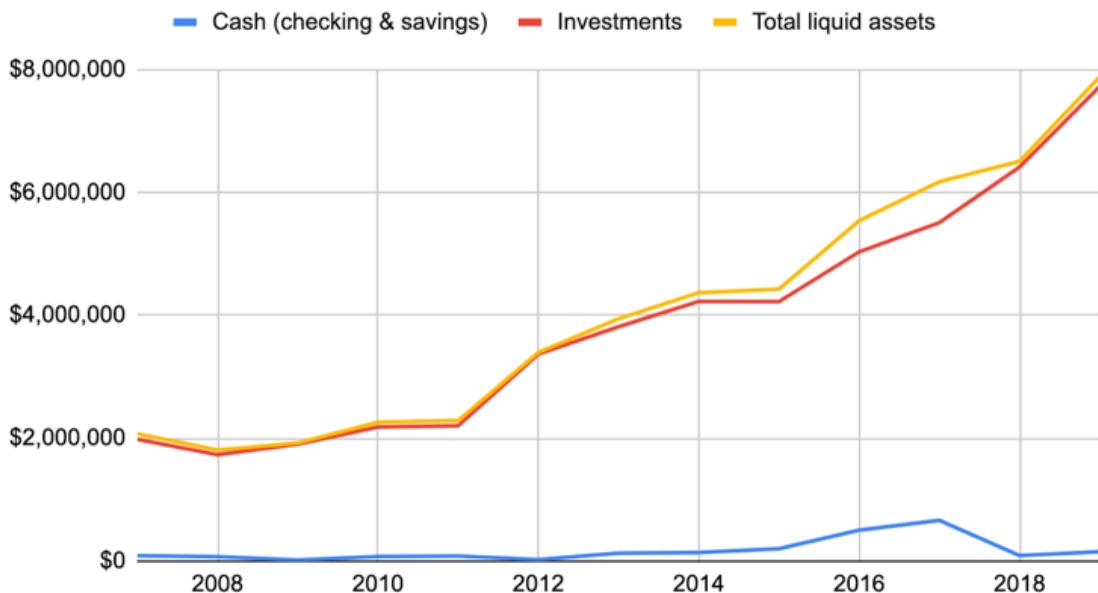
I'll continue to disclaim that I'm not very financially savvy, but it seems to me, just as a basic sanity check, that if an organization is being smart about its investments, its assets should grow over time.

I graphed the assets of both companies for the most recent 13 years, excluding restricted assets as well as property and equipment (it's *really* bad if they liquidate their property and equipment). Data is taken from public financial statements ([Alcor](#), [CI](#)).

Alcor assets 2007-2019



Cryonics Institute assets 2007-2019



As you can see, both organizations passed my sanity check. Both have seen their assets grow at an average of 12.4% per year since 2011 (I excluded the years before that because of the recession), suggesting they are following similar investment strategies. This roughly tracks the S&P 500 over the same time period.

Note that Alcor has decided to keep more assets in cash lately; I don't know why.

Expenditures

Snapshot

I skimmed the financial statements of both organizations for the past couple years, and this is what I think I see:

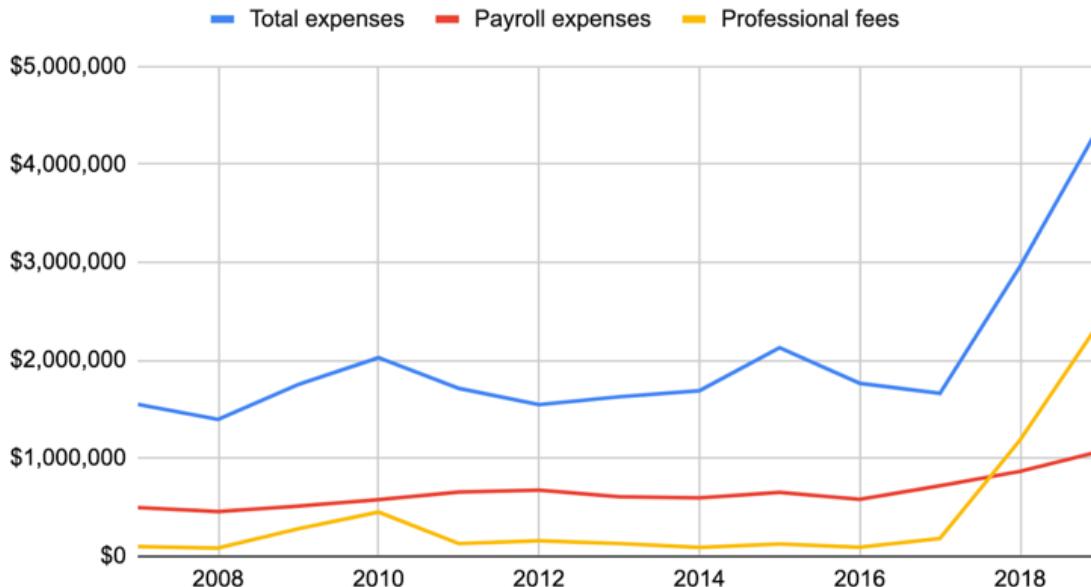
Organization	2019 expenses	Liquid assets as of 12/31/2019	# of patients as of 12/31/2020
Alcor	\$4.4M	\$27.9M	181
Cryonics Institute	\$350K	\$7.9M	198

In brief: The two have similar numbers of patients, Alcor has significantly more assets than CI, but CI spends a significantly smaller percentage of its assets each year than Alcor (4.4% vs 15.8% in 2019).

Alcor

I think the most obvious question is: Why are Alcor's expenses so high?? Not only are they high, but they've been increasing over the past few years – they hovered between \$1.5M and \$2M from 2007 to 2017, then shot up to \$3M in 2018 and \$4.4M in 2019.

Alcor expenses 2007-2019



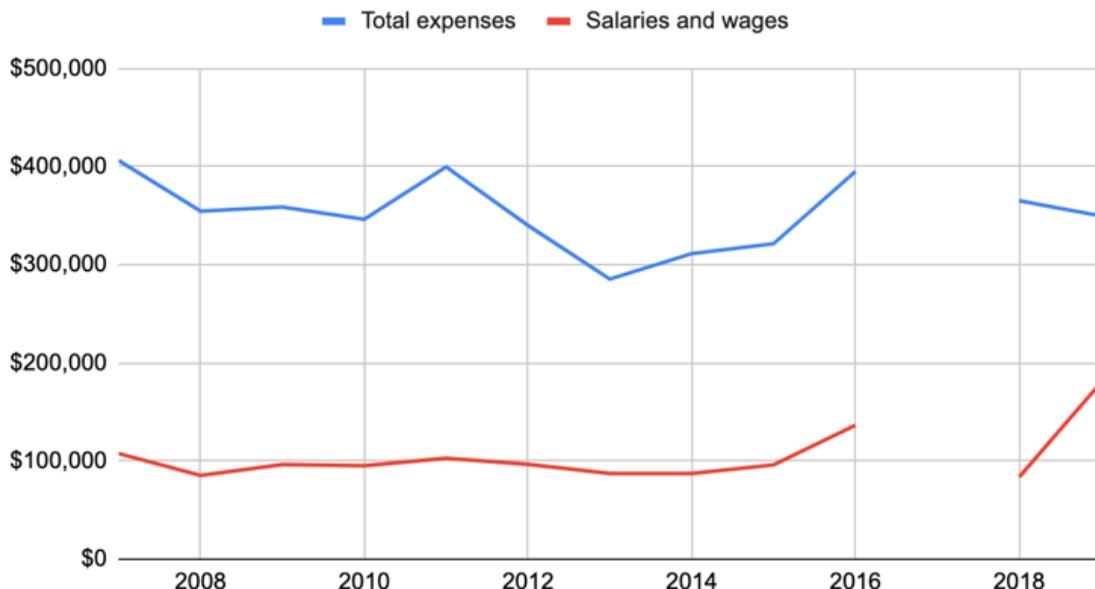
In both cases (2017-18 and 2018-19) the bulk of the increase was due to 'professional fees', which increased by more than a million dollars each year – a 550% year-over-year increase from 2017-18 and an additional 100% increase from 2018-19.

I didn't see any explanation for this, either in the financial reports or in any of [Alcor's updates](#) since late 2017 – admittedly I didn't look *that* hard, but the increase is so large that you'd hope it'd be mentioned prominently. I'd hazard a guess that it points to either a major shift in strategy, or some really thorny legal cases (Mati pointed me to [this 2019 lawsuit](#)), but that's just speculation. I did check whether there'd been a large spike in the number of preservations done in those years, but there wasn't anything outside of the normal range.

Another concern I have is that the 2017 numbers as reported [at the end of 2017](#) are substantively different from the 2017 numbers as reported (and used for comparison) [at the end of 2018](#). Not different enough to affect the trends, but definitely different, and I don't know why. Note that, as mentioned above, I did not reach out to Alcor directly to ask.

Cryonics Institute

Cryonic Institute expenses 2007-2019



(Note: 2017 is omitted because as far as I can tell they [only have half-year data](#) for that year? Spending for that year does appear to fall somewhere within the expected range – I'm just not sure exactly where.)

As you can see, CI's expenses have remained remarkably stable over the past decade, basically always falling between \$300K and \$400K per year – and note that this is before adjusting for inflation.

Other factors

Professionalism

CI does a way worse job than Alcor of projecting professionalism. For example, compare Alcor's annual financial report's vs CI's. Alcor has long, standardized, well-formatted documents full of legalese, while CI has PDFs of spreadsheets that look different from year to year and don't even have font consistency within individual documents. Its website is difficult

to navigate and riddled with spelling errors. As we saw above, it doesn't even pretend to be planning for long-term stability.

Also, CI members regularly receive phishing emails like this one:

From: Cryonics Institute <info.cryonics.fastmail@gmail.com>

Date: Thu, Nov 5, 2020 at 8:39 AM

Subject: Cryonic Institute

To:

Could you please confirm to us when you would be able to make payment for your Membership and Standby payment.

I will await your reply.

Sincerely

Andy Zawacki

On the other hand, there's the argument that CI is what it is, and it isn't trying to deceive you with a veneer of professionalism. It's not like Alcor's consistent font usage changes the fact that it's underprepared for global disasters or that many of its patients incur lots of ischemic time.

Membership

If we take their reported membership numbers at face value, CI has more members than Alcor. However, I'm told that CI has stopped reporting how many of its members are actually signed up for cryopreservation (you can be a CI Member without signing up, whereas you're only a full Alcor Member if you sign up; otherwise you're only an Associate Member), so just comparing the numbers to one another doesn't tell you much.^[1] In any case, they're in the same general ballpark, with around 1000 people signed up for cryonics through each organization.

Summary

Why choose CI?

Easier signup

Signing up with CI is easier because they don't need to be the owner of your life insurance policy (just the beneficiary), which broadens your options considerably. You can even use a group insurance policy obtained through your employer.

More financially conservative

You might decide to bet on CI long-term based on the fact that its spending is much lower and much more consistent than Alcor's.

It can be cheaper

If you're really strapped for cash and want to get signed up right now, CI is more likely to be affordable. See "[When is CI cheaper?](#)" above.

Why choose Alcor?

Better preservation

It seems like by far the most important thing is being near a perfusion team when you die, if at all possible. Alcor makes that easier with its relocation assistance and its relationship with a hospice. It also removes the cognitive burden of standby arrangements by making centralized standby mandatory for all members.

Perfusion is a murkier area, but I feel that I've seen more reason to be confident in Alcor's perfusion technology than CI's.

Future planning

While it's still far from as good as it needs to be to last 1000 years, Alcor at least outperforms CI on having done basic future planning, like choosing a low-risk location for its facilities and raising its prices to keep pace with inflation.

Solid reputation

While the reputation itself is screened off by the other considerations presented in this post, it appears to me that - due to its claims to professionalism - Alcor is held to higher standards than CI by people in the cryonics community. This kind of scrutiny may or may not lead to actual better performance, but it at least incentivizes it.

Bottom line

If you're in the Americas, I recommend Alcor. If you're in Eurasia, I probably still recommend Alcor, but I'd also be interested in someone looking further into Tomorrow Biostasis.

Commenting guidelines

I approached this question in good faith and had no pre-existing ties to either Alcor or CI. No cryonics organization fills me with confidence, but given that these are the options I have to choose among, I've chosen Alcor. If your calculus comes out different, feel free to express why in the comments. I'd also be interested to hear if you think I've made any factual errors. However, if I judge your comments to be unnecessarily partisan (in any direction), hostile, or otherwise unproductive, I will delete them.

*Besides Alcor and CI, there are at least four other US-based cryonics organizations: American Cryonics Society (ACS), Oregon Cryonics, Osiris, and Trans Time.

- If you sign up for cryopreservation through American Cryonics Society, your storage will ultimately be handled by CI. ACS's value-add is that they offer additional options, like different preservation procedures and establishment of a research and reanimation fund. You can read more about these [on their website](#).
- [Oregon Cryonics](#) has a decent reputation from what I can tell, but they only accept members who live very near their facilities in Salem, so they're not worth looking into for the vast majority of people.
- [Osiris](#) is very new and [not very well regarded](#).
- Trans Time [appears](#) to have its own cryo storage facility with a nonzero number of patients in cryostasis, but their [website](#) is so intensely awful that I can't figure out what they do or how one might go about signing up with them.

	Whole-body Neuro	Founded	Standby	Patients	Members*
	price	price			
Alcor	\$200,000	\$80,000	1974	Mandatory	181 1317
ACS	\$155,000	N/A	1969	Optional	19 ?
CI	\$28,000	N/A	1976	Optional	200 1725
KrioRus	\$36,000	\$18,000	2005	Optional	51 200
Oregon Cryo	N/A	\$48,000	2005	Mandatory	8 ?
Osiris	\$28,500	N/A	2016	?	?
Trans Time	\$150,000	?	1972	?	3 ?

*Remember that members are not necessarily signed up for cryonics.

There are also some additional international cryonics organizations:

- Shandong Yinfeng Life Science Research Institute (or Yinfeng for short), in China, has been operating since 2015.
- [Cryonics Germany](#) is a small operation that provides neuro storage only.
- [Southern Cryonics](#) is slated to open soon and will be the first cryonics provider in the southern hemisphere.

1. ^

The last time CI shared the fraction of their members that were signed up for cryonics was in 2014, and it was $578/1010 = \sim 57\%$. If we assume the same ratio today, we find that 983 people are signed up for cryonics with CI. [As of October 2020](#), 285 CI members were signed up with Suspended Animation; $285/983 = \sim 29\%$.

2. ^

This section would benefit immensely from someone more financially savvy looking into it for even just an hour. For example, I'd love for someone to look into this claim from the 2012 thread:

Unlike CI, Alcor has created robust practices and mechanisms for long-term maintenance and growth of the Patient Care Trust Fund and the Endowment Fund. Go take a look at CI's financial reports. See how little money is available for the indefinite care and eventual revival of each patient. Also look at the returns on investment of those funds.

3. ^

See "[What about inflation?](#)"

4. ^

See "[But what if I don't have anything?](#)"

Exercise: Taboo "Should"

Thank you to [Elizabeth](#) for a great conversation which spurred me to write up this post.

Claim: moral/status/value judgements (like "we should blame X for Y", "Z is Bad", etc) like to sneak into epistemic models and masquerade as weight-bearing components of predictions.

Example: The Virtue Theory of Metabolism

The virtue theory of metabolism says that people who eat virtuous foods will be rewarded with a svelte physique, and those who eat sinful foods will be punished with fat. Obviously this isn't meant to be a "real" theory; rather, it's a tongue-in-cheek explanation of the way most laypeople actually think about diet and body weight.

Lest ye think this a strawman, let's look at some predictions made by the virtue theory.

As a relatively unbiased first-pass test, we'll walk through [Haidt's five moral foundations](#) and ask what predictions each of these would make about weight loss when combined with the virtue theory. In particular, we'll look for predictions about perceived healthiness which seem orthogonal (at least on the surface) to anything based on actual biochemistry.

1. Care/harm: food made with care and love is healthier than food made with indifference. For instance, home cooking is less fattening than restaurant, or factory-farmed meat is more fattening than free-range chicken.
2. Fairness: food made fairly is healthier than food made unfairly. For instance, "fair-trade" foods are less fattening.
3. Loyalty/ingroup: food made by members of an ingroup is more healthy. For instance, local produce is less fattening.
4. Authority/respect: food declared "good" by legitimate experts/authorities is more healthy. Fun fact for American readers: did you know the original food pyramid was created by the department of agriculture (as opposed to the department of health), and bears an uncanny resemblance to the distribution of American agricultural output?
5. Sanctity/purity: impure or unnatural food is unhealthy. For instance, preservatives, artificial flavors, and GMO foods are all more fattening, whereas organic food is less fattening.

Maybe I'm cherry-picking or making a just-so story here, but... these sound like things which I think most people do believe, and they're pretty central to the stereotypical picture of a "healthy diet". That's not to say that there isn't *also* some legitimate biochemistry sprinkled into peoples' food-beliefs, but even then it seems like the real parts are just whatever percolated out of Authorities. As a purely descriptive theory of how laypeople model metabolism, the virtue theory looks pretty darn strong.

Of course if pressed, people will come up with some biology-flavored explanation for why their food-health-heuristic makes sense, but the correlation with virtue instincts pretty strongly suggests that these explanations are post-hoc.

An Exercise

This post isn't actually about the virtue theory of metabolism. It's about a technique for *noticing* things like the virtue theory of metabolism in our own thinking. How can we detect moral/status/value judgements masquerading as components of predictive models?

The technique is simple: [taboo](#) concepts like "good", "bad", "should", etc in one's thinking. When you catch yourself thinking one "should" do X, or X is "good", stop and replace that with "I like X" or "X is useful for Y" or "X has consequence Z" or the like. Ideally, you keep an eye out for anything which feels suspiciously-value-flavored (like "healthy" in the examples above) and taboo them.

So, for instance, I might notice myself eating vegetables because they are "good for me". I notice that "good" appears in there, so I taboo "good", and ask *what* exactly these vegetables are doing which is supposedly "good" for me. I may not know all the answers, and that's fine - e.g. the answer may just be "my doctor says I'll have fewer health problems if I eat lots of vegetables". But at least I have flagged this as a thing-about-which-I-have-incomplete-knowledge-about-the-physical-gears-involved, rather than an atomic fact that "vegetables are good" in some vague undefined sense.

In general, there is absolutely no morality or status whatsoever in any physical law of the universe, at any level of abstraction. If a subquery of the form "what is Good?" ever appears when trying to make a factual prediction, then something has gone wrong. Epistemics should contain exactly zero morality.

(Did you catch the "should" in the previous sentence? Here's the full meaning with "should" taboo'd: if a moral statement is load-bearing in a model of the physical world, then something is factually incorrect in that model. At a bare minimum, even if the end prediction is right, the gears are definitely off.)

The usefulness of tabooing "should" is to flush out places where moral statements are quietly masquerading as facts about the world, or as gears in our world-models.

Years ago, when I first tried this exercise for a week, I found surprisingly many places where I was relying on vague morally-flavored "facts", similar to "vegetables are good". Food was one area, despite having already heard of the virtue theory of metabolism and already trying to avoid that mistake. The hooks of morality-infected epistemology ran deeper than I realized.

Politics-adjacent topics were, of course, another major area where the hooks ran deep.

Political Example: PS5 Sales

Matt Yglesias provides a prototypical example in [What's Wrong With The Media](#). He starts with an excerpt from a playstation 5 review:

The world is still reeling under the weight of the covid-19 pandemic. There are more Americans out of work right now than at any point in the country's history, with no relief in sight. Our health care system is an inherently evil institution that forces people to ration life-saving medications like insulin and choose suicide over suffering with untreated mental illness.

As I'm writing this, it looks very likely that Joe Biden will be our next president. But it's clear that the worst people aren't going away just because a new old white man is sitting behind the Resolute desk—well, at least not this old white man. Our government is fundamentally broken in a way that necessitates radical change rather than incremental electoralism.

The harsh truth is that, for the reasons listed above and more, a lot of people simply won't be able to buy a PlayStation 5, regardless of supply. Or if they can, concerns over increasing austerity in the United States and the growing threat of widespread political violence supersede any enthusiasm about the console's SSD or how ray tracing makes reflections more realistic. That's not to say you can't be excited for those things—I certainly am, on some level—but there's an irrefutable level of privilege attached to the ability to simply tune out the world as it burns around you.

The problem here, Yglesias argues, is that this analysis is bad - i.e. the predictions it makes are unlikely to be accurate:

What actually happened is that starting in March the household savings rate soared (people are taking fewer vacations and eating out less) and while it's been declining from its peak as of September it was still unusually high.

[...]

The upshot of this is that no matter what you think about Biden or the American health care system, the fact is that the sales outlook for a new video game console system is very good.

Indeed, the PS5 sold out, although I don't know whether Yglesias predicted that ahead of time.

So this is a pretty clear example of moral/status/value judgements masquerading as components of a predictive model. Abstracting away the details, the core of the original argument is "political situation is Bad -> people don't have resources to buy PS5". What does that look like if we taboo the value judgements? Well, the actual evidence cited is roughly:

- Lots of people have COVID
- American unemployment rate is at an all-time high
- Health care system forces rationing of medication and doesn't treat mental illness
- Bad People aren't going away (I'm not even sure how to Taboo this one or if there'd be anything left; it could mean any of a variety of Bad People or the author might not even have anyone in particular in mind)
- Lots of people are concerned about austerity or political violence

Reading through that list and asking "do these actually make me think video game console sales will be down?", the only one which stands out as *directly* relevant - not just mood affiliation with Bad things - is unemployment. High unemployment *is* a legitimate reason to expect slow console sales, but when you notice that that's the *only* strong argument here, the whole thing seems a lot less weighty. (Amusing side note: the unemployment claim was false. Even at peak COVID, unemployment was far lower than during the Great Depression, and it had already fallen below the level of more recent recessions by the time the console review was published. But that's not really fatal to the argument.)

By tabooing moral/status/value claims, we force ourselves to think about the actual gears of a prediction, not just mood-affiliate.

Now, in LessWrong circles we tend not to see really obvious examples like this one. We have implicit community norms against prediction-via-mood-affiliation, and many top authors already have a habit of tabooing “good” or “should”. (We’ll see an example of that shortly.) But I do expect that lots of people have a general-sense-of-social-Badness, with various political factors feeding into it, and quick intuitive predictions about economic performance (including console sales) downstream. There’s a vague idea that “the economy is Bad right now”, and therefore e.g. console sales will be slow or stock prices should be low. That’s the sort of thing we want to notice and taboo. Often, tabooing “Bad” in “the economy is Bad right now” will still leave useful predictors - unemployment is one example - but not always, and it’s worth checking.

Positive Example: Toxoplasma Memes

From [Toxoplasma of Rage](#):

Consider the war on terror. They say that every time the United States bombs Pakistan or Afghanistan or somewhere, all we’re doing is radicalizing the young people there and making more terrorists. Those terrorists then go on to kill Americans, which makes Americans get very angry and call for more bombing of Pakistan and Afghanistan.

Taken as a meme, it’s a single parasite with two hosts and two forms. In an Afghan host, it appears in a form called ‘jihad’, and hijacks its host into killing himself in order to spread it to its second, American host. In the American host it morphs in a form called ‘the war on terror’, and it hijacks the Americans into giving their own lives (and tax dollars) to spread it back to its Afghan host in the form of bombs.

From the human point of view, jihad and the War on Terror are opposing forces. From the memetic point of view, they’re as complementary as caterpillars and butterflies. Instead of judging, we just note that somehow we accidentally created a replicator, and replicators are going to replicate until something makes them stop.

Note that last sentence: “**Instead of judging**, we just note that somehow we accidentally created a replicator...”. This is exactly the sort of analysis which is unlikely to happen without somebody tabooing moral/status/value judgements up-front.

If we go in looking for someone to blame, then we naturally end up modelling jihadists as Evil, or interventionist foreign policymakers as Evil, or ..., and that feels like it has enough predictive power to explain what’s going on. Jihadists are Evil, therefore they do Bad things like killing Americans, and then the Good Guys kill the Bad Guys - that’s exactly what Good boils down to in the movies, after all. It feels like this model explains the main facts, and there’s not much mystery left - no reason to go asking about memetics or whatever.

Interesting and useful insights about memetics are more likely if we first taboo all that. [Your enemies are not innately Evil](#), but even if they were it would be useful to taboo that fact, unpack it, and ask how such a thing came to be. It’s not that “Bad Person does Bad Thing” always makes inaccurate predictions, it’s that humans have

built-in intuitions which push us to use that model regardless of whether it's accurate, for reasons more related to tribal signalling than to heuristic predictive power.

Example: Copenhagen Ethics

[The Copenhagen Interpretation of Ethics](#) says that when you observe or interact with a problem, you can be blamed for it. In particular, you won't be blamed for a problem you ignore, but you can be blamed for benefitting from a problem *even if you make the problem better*. This is not intended as a model for how ethics "should" work, but rather for how most people think about ethics by default.

Tabooing moral/status/value judgements tends to make Copenhagen ethics much more obviously silly. Here's an example from the original post:

In 2010, [New York randomly chose homeless applicants to participate in its Homebase program](#), and tracked those who were not allowed into the program as a control group. The program was helping as many people as it could, the only change was explicitly labeling a number of people it wasn't helping as a "control group". The response?

"They should immediately stop this experiment," said the Manhattan borough president, [Scott M. Stringer](#). "The city shouldn't be making guinea pigs out of its most vulnerable."

Let's taboo the "should"s in Mr Stringer's statement. We'll use the "We should X" -> "X has consequence Z" pattern: replace "they should immediately stop this experiment" with "immediately stopping this experiment would <????> for the homeless". What goes in the <????>?

Feel free to think about it for a moment.

My answer: nothing. Nothing goes in that <????>. Stopping the experiment would not benefit any homeless people in any way whatsoever. When we try to taboo "should", that becomes much more obvious, because we're forced to ask *how* ending the experiment would benefit any homeless people.

Takeaway

Morality has no place in epistemics. If moral statements are bearing weight in world-models, then at a bare minimum the gears are wrong. Unfortunately, I have found a lot of morality-disguised-as-fact hiding within my own world-models. I expect this is the case for most other people as well, especially in politically-adjacent areas.

A useful exercise for rooting out some of these hidden-morality hooks is to taboo moral/status/value-flavored concepts like "good", "bad", "should", etc in one's thinking. Whenever you notice yourself using these concepts, dissolve them - replace them with "I want/like X" or "X is useful for Y" or "X has consequence Z".

Three caveats to end on.

First caveat: as with the [general technique of tabooing words](#), I don't use this as an all-the-time exercise. It's useful now and then to notice weak points in your world-

models or habits, and it's especially useful to try it at least once - I got most of the value out of the first week-long experiment. But it takes a fair bit of effort to maintain, and one week-long try was enough to at least install the mental move.

Second caveat: a moral statement bearing weight in a world-model means the model is wrong, but that does *not* mean that the model will improve if the moral components are naively thrown out. You do need to actually figure out what work those moral statements are doing (if any) in order to replace them. Bear in mind that morality is an area subject to a lot of [cultural selection pressure](#), and those moral components may be doing something non-obviously useful.

Final caveat: *before* trying this exercise, I recommend you *already* have the skill of [not giving up on morality altogether just because moral realism is out the window](#). Just because morality is not doing any epistemic work does not mean it's not doing any instrumental work.

Note: I am actively looking for examples to use in a shorter exercise, in order to teach this technique. Ideally, I'd like examples where most people - regardless of political tribe - make a prediction which is obviously wrong/irrelevant after tabooing a value-loaded component. If you have examples, please leave them in the comments. I'd also be interested to hear which examples did/didn't click for people.

Avoid Unnecessarily Political Examples

One of the motivations for [You have about five words](#) was the post [Politics is the Mindkiller](#). That post essentially makes four claims:

- Politics is the mindkiller. Therefore:
- If you're not making a point about politics, avoid needlessly political examples.
- If you are trying to make a point about *general* politics, try to use an older example that people don't have strong feelings about.
- If you're making a current political point, try not to make it *unnecessarily* political by throwing in digs that tar the entire outgroup, if that's not actually a key point.

But, not everyone read the post. And not everyone who read the post stored all the nuance for easy reference in their brain. The thing they remembered, and told their friends about, was "Politics is the mindkiller." Some people heard this as "politics == boo". LessWrong ended up having a vague norm about avoiding politics at all.

This norm might have been good, or bad. Politics *is* the mindkiller, and if you don't want to get your minds killed, it may be good not to have your rationality website deal directly with it too much. But, also, politics is legitimately important sometimes. How to balance that? Not sure. It's tough. [Here's some previous discussion on how to think about it](#). I endorse the current LW system where you can talk about politics but it's not frontpaged.

But, I'm not actually here today to talk about that. I'm here to basically copy-paste the post but give it a different title, so that one of the actual main points has a clearer referent.

I'm not claiming this is more or less important than the "politics is the mindkiller" concept, just that it was an important concept for people to remember separately.

So:

Avoid unnecessarily political examples.

The original post is pretty short. Here's the whole thing. Emphasis mine:

People go funny in the head when talking about politics. The evolutionary reasons for this are so obvious as to be worth belaboring: In the ancestral environment, politics was a matter of life and death. And sex, and wealth, and allies, and reputation . . . When, today, you get into an argument about whether "we" ought to raise the minimum wage, you're executing adaptations for an ancestral environment where being on the wrong side of the argument could get you killed. Being on the *right* side of the argument could let you kill your hated rival!

If you want to make a point about science, or rationality, then my advice is to not choose a domain from *contemporary* politics if you can possibly avoid it. If your point is inherently about politics, then talk about Louis XVI during the French Revolution. **Politics is an important domain to which we should individually apply our rationality—but it's a terrible domain in which to**

learn rationality, or discuss rationality, unless all the discussants are already rational.

Politics is an extension of war by other means. Arguments are soldiers. Once you know which side you're on, you must support all arguments of that side, and attack all arguments that appear to favor the enemy side; otherwise it's like stabbing your soldiers in the back—providing aid and comfort to the enemy. People who would be level-headed about evenhandedly weighing all sides of an issue in their professional life as scientists, can suddenly turn into slogan-chanting zombies when there's a [Blue or Green](#) position on an issue.

In artificial intelligence, and particularly in the domain of nonmonotonic reasoning, there's a standard problem: "All Quakers are pacifists. All Republicans are not pacifists. Nixon is a Quaker and a Republican. Is Nixon a pacifist?"

What on Earth was the point of choosing this as an example? To rouse the political emotions of the readers and distract them from the main question? To make Republicans feel unwelcome in courses on artificial intelligence and discourage them from entering the field?¹

Why would anyone pick such a *distracting* example to illustrate nonmonotonic reasoning? Probably because the author just couldn't resist getting in a good, solid dig at those hated Greens. It feels so good to get in a hearty punch, y'know, it's like trying to resist a chocolate cookie.

As with chocolate cookies, not everything that feels pleasurable is good for you.

I'm not saying that I think we should be apolitical, or even that we should adopt Wikipedia's ideal of the Neutral Point of View. But try to resist getting in those good, solid digs if you can possibly avoid it. If your topic legitimately relates to attempts to ban evolution in school curricula, then go ahead and talk about it—but don't blame it explicitly on the whole Republican Party; some of your readers may be Republicans, and they may feel that the problem is a few rogues, not the entire party. As with Wikipedia's npov, it doesn't matter whether (you think) the Republican Party really *is* at fault. It's just better for the spiritual growth of the community to discuss the issue without invoking color politics.

Review of Soft Takeoff Can Still Lead to DSA

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

A few months after writing [this post](#) I realized that one of the key arguments was importantly flawed. I therefore recommend against inclusion in the 2019 review. This post presents an improved version of the original argument, explains the flaw, and then updates my all-things-considered view accordingly.

Improved version of my original argument

1. Definitions:
 1. "Soft takeoff" is roughly "AI will be like the Industrial Revolution but 10x-100x faster"
 2. "Decisive Strategic Advantage" (DSA) is "a level of technological and other advantages sufficient to enable it to achieve complete world domination." In other words, DSA is roughly when one faction or entity has the *capability* to "take over the world." (What taking over the world means is an [interesting question](#) which we won't explore here. Nowadays I'd reframe things in terms of [potential PONRs](#).)
 3. We ask how likely it is that DSA arises, conditional on soft takeoff. Note that DSA does not mean the world is actually taken over, only that one faction at some point has the ability to do so. They might be too cautious or too ethical to try. Or they might try and fail due to bad luck.
2. In a soft takeoff scenario, a 0.3 - 3 year technological lead over your competitors probably gives you a DSA.
 1. It seems plausible that for much of human history, a 30-year technological lead over your competitors was *not* enough to give you a DSA.
 2. It also seems plausible that during and after the industrial revolution, a 30-year technological lead *was* enough. (For more arguments on this key point, see my original post.)
 3. This supports a plausible conjecture that when the pace of technological progress speeds up, the length (in clock time) of technological lead needed for DSA shrinks proportionally.
3. So a soft takeoff could lead to a DSA insofar as there is a 0.3 - 3 year lead at the beginning which is maintained for a few years.
4. 0.3 - 3 year technological leads [are reasonably common today](#), and in particular it's plausible that there could be one in the field of AI research.
5. There's a reasonable chance of such a lead being maintained for a few years.
 1. This is a messy question, but judging by the table below, it seems that if anything the lead of the front-runner in this scenario is more likely to lengthen than shorten!
 2. If this is so, why did no one achieve DSA during the Industrial Revolution? My answer is that spies/hacking/leaks/etc. are much more powerful during the industrial revolution than they are during a soft takeoff, because they

have an entire economy to steal from and decades to do it, whereas in a soft takeoff ideas can be hoarded in a specific corporation and there's only a few years (or months!) to do it.

6. Therefore, there's a reasonable chance of DSA conditional on soft takeoff.

Factors that might shorten the lead	Factors that might lengthen the lead
If you don't sell your innovations to the rest of the world, you'll lose out on opportunities to make money, and then possibly be outcompeted by projects that didn't hoard their innovations.	Hoarding innovations gives you an advantage over the rest of the world, because only you can make use of them.
Spies, hacking, leaks, defections, etc.	Big corporations with tech leads often find ways to slow down their competition, e.g. by lobbying to raise regulatory barriers to entry.
	Being known to be the leading project makes it easier to attract talent and investment.
	There might be additional snowball effects (e.g. network effect as more people use your product providing you with more data)

I take it that 2, 4, and 5 are the controversial bits. I still stand by 2, and the arguments made for it in my original post. I also stand by 4. (To be clear, it's not like I've investigated these things in detail. I've just thought about them for a bit and convinced myself that they are probably right, and I haven't encountered any convincing counterarguments so far.)

5 is where I made a big mistake.

(Comments on my original post also attacked 5 a lot, but none of them caught the mistake as far as I can tell.)

My big mistake

Basically, my mistake was to conflate leads measured in number-of-hoarded-ideas with leads measured in clock time. Clock-time leads shrink *automatically* as the pace of innovation speeds up, because if everyone is innovating 10x faster, then you need 10x as many hoarded ideas to have an N-year lead.

Here's a toy model, based on the one I gave in the original post:

There are some projects/factions. There are many ideas. Projects can have access to ideas. Projects make progress, in the form of discovering (gaining access to) ideas. For each idea they access, they can decide to hoard or not-hoard it. If they don't hoard it, it becomes accessible to all. Hoarded ideas are only accessible by the project that discovered them (though other projects can independently rediscover them). The rate of progress of a project is proportional to how many ideas they can access.

Let's distinguish two ways to operationalize the technological lead of a project. One is to measure it in ideas, e.g. "Project X has 100 hoarded ideas and project Y has only 10, so Project X is 90 ideas ahead." But another way is to measure it in clock time, e.g. "It'll take 3 years for project Y to have access to as many ideas as project X has now."

Suppose that all projects hoard all their ideas. Then the ideas-lead of the leading project will tend to lengthen: the project begins with more ideas, so it makes faster progress, so it adds new ideas to its hoard faster than others can add new ideas to theirs. However, the clocktime-lead of the leading project will remain fixed. It's like two identical cars accelerating one after the other on an on-ramp to a highway: the distance between them increases, but if one entered the ramp three seconds ahead, it will still be three seconds ahead when they are on the highway.

But realistically not all projects will hoard all their ideas. Suppose instead that for the leading project, 10% of their new ideas are discovered in-house, and 90% come from publicly available discoveries accessible to all. Then, to continue the car analogy, it's as if 90% of the lead car's acceleration comes from a strong wind that blows on both cars equally. The lead of the first car/project will lengthen slightly when measured by distance/ideas, but shrink dramatically when measured by clock time.

The upshot is that we should return to that table of factors and add a big one to the left-hand column: Leads shorten automatically as general progress speeds up, so if the lead project produces only a small fraction of the general progress, *maintaining* a 3-year lead throughout a soft takeoff is (all else equal) almost as hard as growing a 3-year lead into a 30-year lead during the 20th century. In order to overcome this, the factors on the right would need to be very strong indeed.

Conclusions

My original argument was wrong. I stand by points 2 and 4 though, and by the subsequent posts I made in [this sequence](#). I notice I am confused, perhaps by a seeming contradiction between my explicit model here and my take on history, which is that rapid takeovers and upsets in the balance of power have happened many times, that power has become more and more concentrated over time, and that there are not-so-distant possible worlds in which a *single man* rules the whole world sometime in the 20th century. Some threads to pull on:

1. To the surprise of my past self, [Paul agreed DSA is plausible for major nations, just not for smaller entities like corporations](#): "I totally agree that it wouldn't be crazy for a major world power to pull ahead of others technologically and eventually be able to win a war handily, and that will tend happen over shorter and shorter timescales if economic and technological progress accelerate.") Perhaps we've been talking past each other, because I think a very important

point is that it's common for small entities to gain control of large entities. I'm not imagining a corporation fighting a war against the US government; I'm imagining it taking over the US government via tech-enhanced lobbying, activism, and maybe some skullduggery. (And to be clear, I'm usually imagining that the corporation was previously taken over by AIs it built or bought.)

2. Even if takeoff takes several years it could be unevenly distributed such that (for example) 30% of the strategically relevant research progress happens in a single corporation. I think 30% of the strategically relevant research happening in a single corporation at beginning of a multi-year takeoff would probably be enough for DSA.
3. Since writing this post my thinking has shifted to focus less on DSA and more on [potential AI-induced PONRs](#). I also now prefer a [different definition of slow/fast takeoff](#). Thus, perhaps this old discussion simply isn't very relevant anymore.
4. Currently the most plausible doom scenario in my mind is maybe a version of [Paul's Type II failure](#). (If this is surprising to you, reread it while asking yourself what terms like "correlated automation failure" are euphemisms for.) I'm not sure how to classify it, but this suggests that we may disagree less than I thought.

Thanks to Jacob Laggeros for nudging me to review my post and finally get all this off my chest. And double thanks to all the people who commented on the original post!

Taking money seriously

(cross-posted from [my blog](#))

As a young environmental activist, I lived the stereotype by not giving money serious thought, both personally and politically. There is a point where the less money you have, the less you care about it, even though you should care more. Let's call it financial learned helplessness - a state of mind where you believe that you will never have money, so when you do have some, and you get the chance to save and compound, you pass on it and spend the money. I think this sentiment is common.

While you might say: "Well, that just proves my point that poor people are poor because of their own decisions and lack of discipline", someone else might say: "Well you don't understand that, had you grown up in the same circumstances as they did, with the same brain they were born with, you would have done the same". And then you might respond with: "Maybe that's true, but we should still say that they make bad decisions and have no discipline because shaming someone is a powerful social tool to incite change." And then someone would reply: "In an ancestral-like environment, that may be true, but in today's world, people will always have the option to walk away from your shaming, so what you're really doing is driving them away". And you might then say: "It still makes sense to shame them, because if it becomes a popular sentiment, they will have nowhere to hide, and this will nudge them to make better decisions". But then someone would respond: "In addition to being cruel, that's very unlikely. But seeing how a lot of poverty is just being born in the wrong family or wrong neighborhood, it's also false." And then this debate would become a debate about social mobility, inherited capital, or free will, which we all know doesn't exist because nothing in this world is free except refills.

All of this is related to personal financial ability though. The other side of that coin is political views surrounding money, or what you might call economic literacy. In the eyes of environmental and/or leftist activists, money is icky. It's something to remind you of the injustices of the system. There never seems to be enough of it, and who controls it anyway?

In my experience, people with grand ideas of social reform ignored money questions. How much something costs in a proposed policy is never the subject, because things generally don't cost anything.

I once sat in a conference on the power system in my country. The government had recently published a development strategy, and this strategy included new extraction sites for natural gas, as well as a big new LNG (liquefied natural gas) terminal. If you don't know, an [LNG terminal](#) is a port where tankers can unload natural gas that they're carrying from somewhere else.

I sat and listened to people who explained how this was a very bad idea and how it needed to be stopped. And I agree that solar power (which is abundant where I live) is a better idea than relying on fossil fuels. But as I listened, it dawned on me: these people haven't really thought about this problem. Neither had I. Whether something is good or bad for the climate is an important factor - but it's not the only one. There are other things you have to think about, for example geopolitical aspects of your decisions. Where are you getting most of your power now? How long will the transition last? Most important of all - *how much will all of this cost?*

It feels great to say "it costs what it costs", and signal to others that you believe that the climate is much more important than imaginary things like money. But that feels great exactly until you have to implement the solution and see that it isn't as easy as you thought it would be. There are irrationalities related to money, like stock price manipulation. In addition, the entire system is complex. You have to put in effort to figure out what's what. None of these things, however, should make you think: "I'm simply going to ignore all of it because it's less important". Is money less important than climate? In a sense, yes. Does this mean that you can ignore it? No.

[Money questions] is just a subset of [number questions]. In that conference, there were several people arguing for a specific course of action related to the power system. And I'm almost certain that most of them couldn't define the relationship between a joule and a kilowatt-hour even if their life depended on it. They argued for a specific course of action regarding a Very Important Thing and not only did they not know the intricacies and details about the power system - like what's the annual consumption - but they didn't even speak the language used to talk about these things! (Neither did I, and I helped organize this conference, so if anything, this is a self-own.)

I get that this is gatekeeping. And I get why gatekeeping is bad. [Wisdom of crowds](#) is a potent and real thing. But real talk for just a second: if you don't know what the hell you're talking about, you should probably learn about it before talking. Experts might certainly be biased or simply not care about the environment, which is why you have to make a push for the environment, convince people that it's an Equally Important Thing. But to seek a position of influence and not even speak the language - no good! "But maybe the language itself locks people out and perpetuates the oppressive structure of so-called experts", you say. There may be fields where this is entirely true. But if you want to build a house, who do you trust more, architects and construction workers, or journalists?

It may be true that journalists will have good feedback that architects and construction workers should incorporate. But if some journalists - with no building experience - try to stop your house from getting built and provide alternative ways you should build - e.g. the types of foundations used, or what kind of insulation you should use in your walls etc. - you'd take them more seriously if they had at least some experience or understanding of building. If they understood soil types, or how terrain slope affects the depth of foundations, or what the R insulation value even is - if they spoke the language of the trade, you wouldn't discard their views as easily.

What happens in practice is that these interactions turn to tribal warfare. Take plastics, for example. They are a blessing and a curse, but I find that activists that I worked with don't see the blessing side at all. They would be hard-pressed to name at least one useful thing about plastics, despite this being a core component in their lives, one that they couldn't live without. Almost everything in their physical realm is made out of plastics. It's like living in a well-run society, with functioning roads, hospitals and public toilets, and saying that all taxes are evil and unnecessary. It's easy to take the world around you for granted and forget that it's not actually the default mode. But it takes effort to recognize the [complex web of effort](#) that took generations to build, just so you can have cheap abundant energy, materials - or clean streets and affordable hospitals.

Game-theoretically, you're putting yourself at a disadvantage here. Other people - who don't know what the hell they're talking about - will seek positions of influence and work their way through status games. And you, who also don't know what the hell

you're talking about, will be stuck home with a book about transformers, harmonics, and surge arresters, and will feel like you know even less than you had known when you started reading.

While having fewer uninformed opinions is a status boost in rationalist circles, this is not the norm. How to proceed? Simply keep quiet about things you don't understand enough? Always advocate for expert opinion? I don't have a good solution, but my current practical answer is:

- get good at learning quickly
- build a knowledge foundation

The skill of quick learning includes knowing how to focus: actively discarding things that you don't need, and taking in things that you do need. But it also includes the optimistic view that [you can teach yourself anything](#).

On the other hand, having a knowledge foundation means learning elementary things about how the world works. It means to be [up close and personal with the world](#) and it means writing [fact posts](#). Little by little, the world starts *coming into place*.

Luna Lovegood and the Chamber of Secrets - Part 13

"Wait," said Luna, "This is the Lost Diadem of Ravenclaw. It makes the wearer smarter. You might want it."

Professor Quirrel took the diadem in his hands. He feinted as if to place it over his head.

"I am an Occlumens," said Professor Quirrel, "Ravenclaw's device rips the incoherence out of doublethink. If I were to place this device over my head I would be lucky if it did not shred my mind. Nice try."

Professor Quirrel tossed the diadem back to Luna. Luna kowtowed.

"I heard stories of the First Wizarding War. You never cared much for individual human beings but you were always very careful not to destroy wizardkind," said Luna, "I get the feeling you put some effort into protecting the universe."

"So?" said Professor Quirrel.

"You are bored. This plane is too small for you," said Luna.

You-Know-Who did not murder her.

"You should not be a villain," said Luna.

"If you tell me to be a hero then you will die painfully," said Professor Quirrel.

"You should be a god," said Luna.

Luna willingly bestowed the astrolabe to Professor Quirrel.

"Is that all?" said Professor Quirrel.

"Yes," said Luna.

"Avada Kedavra," said Professor Quirrel.

Luna collapsed. Professor Quirrel sheathed his wand. His slender skeleton fingers untangled the clockwork. Professor Quirrel unfolded the astrolabe around him. He ascended to a higher plane of existence.

Luna stepped out of the Forgotten Library. She held the Sword of Gryffindor in her left hand and Wanda in her right. She buried Wanda in Hagrid's pumpkin patch.

The final duel of Lockhart's tournament was that afternoon. Professor Flitwick refereed. Luna lost.

Clang. Luna dropped the Sword of Gryffindor on Professor Lockhart's empty chair. She sat down for dinner in her seat at the end of the Ravenclaw table. A student stood behind her.

"You fought well in Lockhart's dueling tournament," said Ginevra Weasley, "Why don't you try sitting with us Gryffindors for a change?"

The astrolabe displayed "7" on one dial and "0" on all the rest. A tall, slender snakelike figure stepped into Heaven's throne room where a god rested. The trespasser threw a tactical reality anchor like a javelin. It stuck into the wall behind the throne. The trespasser stabbed his second tactical reality anchor behind himself into the floor of the entrance.

"LET'S DUEL."

Credits

You may do whatever you want with this story. You may expand it. You may abridge it. You may retcon it. You may turn it into an audiobook. You may repost it elsewhere.

- Please respect J.K. Rowling's copyright. Harry Potter fanfiction must remain non-commercial, especially in the strict sense of traditional print publishing.
- If you copy this story's exact text, then I request (though do not require) you include an attribution link back to the original story here on Less Wrong.

Thank you J.K. Rowling for creating *Harry Potter* and Eliezer Yudkowsky for creating *Harry James Potter-Evans-Verres*. In addition, thank you MondSemmel, Measure, ejacob, Gurkenglas, Jeff Melcher, gilch, mingyuan, Dojan and everyone else in the comments who corrected spelling and other mistakes in this story.

DALL-E by OpenAI

This is a linkpost for <https://openai.com/blog/dall-e/>

My own take: Cool, not super surprising given GPT-3 and Image GPT. I look forward to seeing what a bigger version of this would do, so that we could get a sense of how much it improves with scale. I'm especially interested in the raven's progressive matrices performance.

Eight claims about multi-agent AGI safety

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

There are quite a few arguments about how interactions between multiple AGIs affect risks from AGI development. I've identified at least eight distinct but closely-related claims which it seems worthwhile to disambiguate. I've split them up into four claims about the process of training AGIs, and four claims about the process of deploying AGIs; after listing them, I go on to explain each in more detail. Note that while I believe that all of these ideas are interesting enough to warrant further investigation, I don't currently believe that all of them are true as stated. In particular, I think that so far there's been little compelling explanation of why interactions between many aligned AIs might have catastrophic effects on the world (as is discussed in point 7).

Claims about training

1. Multi-agent training is one of the most likely ways we might build AGI.
2. Multi-agent training is one of the most dangerous ways we might build AGI.
3. Multi-agent training is a regime in which standard safety techniques won't work.
4. Multi-agent training allows us to implement important new safety techniques.

Claims about deployment

5. We should expect the first AGIs to be deployed in a world which already contains many nearly-as-good AIs.
6. We should expect AGIs to be deployed as multi-agent collectives.
7. Lack of coordination between multiple deployed AGIs is a major source of existential risk.
8. Conflict between multiple deployed AGIs risks causing large-scale suffering.

Details and arguments

1. Multi-agent training is one of the most likely ways we might build AGI.

The core argument for this thesis is that multi-agent interaction was a key feature of the evolution of human intelligence, by promoting both competition and cooperation. Competition between humans provides a series of challenges which are always at roughly the right level of difficulty; [Liebo et al. \(2019\)](#) call this an autocurriculum. Autocurricula were crucial for training sophisticated reinforcement learning agents like AlphaGo and OpenAI Five; it seems plausible that they will also play an important role

in training AGIs. Meanwhile, the usefulness of cooperation led to the development of language, which plays a core role in human cognition; and the benefits of cooperatively sharing ideas allowed the accumulation of human cultural skills and knowledge more generally.

2. Multi-agent training is one of the most dangerous ways we might build AGI.

Humans have skills and motivations (such as deception, manipulation and power-hungriness) which would be dangerous in AGIs. It seems plausible that the development of many of these traits was driven by competition with other humans, and that AGIs trained to answer questions or do other limited-scope tasks would be safer and less goal-directed. [I briefly make this argument here.](#)

3. Multi-agent training is a regime in which standard safety techniques won't work.

Most approaches to safety rely on constructing safe reward functions. But [Ecoffet et al. \(2020\) argue](#) that “open-ended” environments give rise to incentives which depend on reward functions in complex and hard-to-predict ways. Open-endedness is closely related to self-play (which was used to train AlphaGo) and multi-agent environments more generally. When a task involves multiple agents, those agents might learn many skills that are not directly related to the task itself, but instead related to competing or cooperating with each other. E.g. compare a language model like GPT-3, which was directly trained to output language, to the evolution of language in humans - where evolution only selected us for increased genetic fitness, but we developed language skills because they were (indirectly) helpful for that.

Furthermore, as I point out [here](#), it's not even clear what “good behaviour” would actually look like in such environments, since they don't necessarily contain tasks corresponding directly to things we'd like AIs to do in the real world. And fine-tuning on real-world tasks may not be sufficient to override dangerous motivations acquired during extensive multi-agent training.

4. Multi-agent training allows us to implement important new safety techniques.

The most central example of a safety technique which rely on multi-agent environments is probably work done by Gillian Hadfield and others about [learning group-level norms](#). More generally, [CHAI's concept of Assistance Games](#) frames the machine learning training process as an interactive game played between humans and AIs, to better allow humans to guide AI behaviour.

I've also written about some tentative ideas for how to [select for obedience in multi-agent environments](#).

5. We should expect the first AGIs to be deployed in a world which already contains many nearly-as-good AIs.

Paul Christiano [defends this thesis](#) as follows:

- Lots of people will be trying to build powerful AI.
- For most X, it is easier to figure out how to do a slightly worse version of X than to figure out how to do X.

- The worse version may be more expensive, slower, less reliable, less general... (Usually there is a tradeoff curve, and so you can pick which axes you want the worse version to be worse along.)
- If many people are trying to do X, and a slightly worse version is easier and almost-as-good, someone will figure out how to do the worse version before anyone figures out how to do the better version.

Robin Hanson also argues that progress in AI will be widely-distributed, and not very “lumpy”. [He discusses this argument here](#), in part by summarising the lengthy [AI foom debate](#).

6. We should expect AGIs to be deployed as multi-agent collectives.

I discuss this hypothesis [here](#), building on concepts introduced by Bostrom. Summary: after training AGI, there will be strong incentives to copy it many times to get it to do more useful work. If that work involves generating new knowledge, then putting copies in contact with each other to share that knowledge would also increase efficiency. This would be easier if they had already been trained to collaborate; but even if not, their general intelligence should allow them to learn to work together. And so, one way or another, I expect that we'll eventually end up dealing with a “collective” of AIs, which we could also think of as a single “collective AGI”.

Arguably, on a large-scale view, this is how we should think of humans. Each individual human is generally intelligent in our own right. Yet from the perspective of chimpanzees, the problem was not that any single human was intelligent enough to take over the world, but rather that millions of humans underwent cultural evolution to make the human collective much more intelligent.

7. Lack of coordination between multiple deployed AGIs is a major source of existential risk.

Critch makes this case [here](#), summarising [this more extensive report](#). He distinguishes between single AIs which are aligned to single humans (single/single delegation), versus the problem of living in a society where many AIs are each used on behalf of many humans (multi/multi delegation):

It might be that future humans would struggle to coordinate on the globally safe use of powerful single/single AI systems, absent additional efforts in advance to prepare technical multi/multi delegation solutions.

For a historical analogy supporting this view, consider the stock market “flash crash” of 6 May 2010, viewed as one of the most dramatic events in the history of financial markets. The flash crash was a consequence of the use of algorithmic stock trading systems by competing stakeholders. If AI technology significantly broadens the scope of action and interaction between algorithms, the impact of unexpected interaction effects could be much greater, and might be difficult to anticipate in detail.

Note that he claims that this may be true even if single/single alignment is solved, and all AGIs involved are aligned to their respective users.

8. Conflict between multiple deployed AGIs risks causing large-scale suffering.

The Centre on Long-term Risk argues for this thesis in [this research agenda](#). Key idea:

Many of the cooperation failures in which we are interested can be understood as mutual defection in a social dilemma. Informally, a social dilemma is a game in which everyone is better off if everyone cooperates, yet individual rationality may lead to defection. ... An example of potentially disastrous cooperation failure is extortion (and other compelling threats), and the execution of such threats by powerful agents.

Since threats are designed to be strong disincentives, we should expect that the types of threats made against aligned AIs will be very undesirable by human moral standards, and try to design AGIs in ways which prevent threats from being carried out by or against them.

How good are our mouse models (psychology, biology, medicine, etc.), ignoring translation into humans, just in terms of understanding mice? (Same question for drosophila.)

I've been thinking about why some domains have reached more definite, mathematized, and law-like models than others, e.g. the hard sciences of physics and chemistry vs the soft sciences of psychology and political science.

One hypothesis is that it's just much easier to experiment on, say, pendulums than on humans. Experimenting on humans is slow, it's hard to create very controlled conditions, IRB's severely limit what you can do. Theoretically, if I could experiment on a large number of humans for long periods of time without constraints on what I could do (something like Aperture Science?), then I maybe could harden the soft sciences.

A piece of evidence on this would be how good are our mouse models? There are substantially fewer restrictions on mice experiments and resultantly we do a lot of experiments on them (hence the "IN MICE" that ought to be appended to so many scientific result headlines). Maybe translation into humans is poor, and we're generally focused on humans so we feel like there's a lot we don't know, but if you just asked about how well we understood mice, actually we have a lot of really solid knowledge about psychology, metabolism, immune function, political economy, etc.

Can anyone familiar with mouse models in any domain comment on how good the mouse models are vs human models? 1x as good, 5x, 20x? Good = really solid predictions, we feel like we know what's going on, etc.

For that matter, the same question could be asked for drosophila.

A vastly faster vaccine rollout

When a traveler introduced smallpox to New York City in 1947, the city—and in particular its health commissioner, Israel Weinstein—apparently ran an epic vaccination campaign, reaching 5 million people in the first two weeks.¹ That is, four hundred thousand vaccinations per day. San Francisco in two days.

For covid, the first New York City vaccine [was given](#) on the 14th of December, and if I understand, by the 10th of January, twenty seven days later, 203,181 doses [had reportedly been given](#). That's around eight thousand doses per day. A factor of fifty fewer.

That's a pretty incredible difference. Why is New York fifty times slower at delivering covid vaccines in 2021 than it was at delivering smallpox vaccines in 1947?

Part of the answer is presumably ‘regression to the mean’: if thousands of different cities at different times try to roll out vaccinations quickly, with a similar basic ability to do so, and there is some random chance in how well it goes, then the one we tell stories about seventy years later will one that got surprisingly lucky. You shouldn't expect your own effort to go as well as that best one. But—without having done the statistics—I don't think you should expect your attempt to be fifty times worse. New York didn't get *that* lucky.

Perhaps there are other differences in the problems faced. For instance, the current vaccine needs refrigeration, or we are so late in the disease spread that we can't all crowd together in around fast-moving vaccinators, or be rounded up in our natural crowded habitats, like classrooms or offices.

Though the 1947 situation looks harder in important ways too. For one thing, there was no time to prepare. The vaccination project began the day that the disease was identified to health commissioner Weinstein. By 2pm, he was apparently holding a news conference publicly asking residents to be vaccinated. With covid, there were about ten months to prepare. For another thing, now people have smartphones with which they can be alerted, and computers and online systems that might be used to coordinate them and tell them what to do.

I heard about this episode from the [Rachel Maddow Show](#), and read about it in the [New York Times](#), both of which bring it up to note the inadequacy in the current vaccine efforts. The New York Times says a rollout like this, “almost certainly couldn’t happen today”, and offers some explanations:

1. Complicated relationships between city and other governments these days

“In 1947, the city was able to act alone, as opposed to navigating a complicated relationship with the governor of New York and the federal government,” said Dr. Irwin Redlener, director of the Pandemic Resource and Response Initiative at Columbia University’s Earth Institute. “The city was able to say, ‘We’re going after this,’ and then make it happen.”

2. A ‘hollowing out’ of the public health infrastructure

But this time, with the coronavirus pandemic, New York faces a logistical hurdle. Experts in infectious disease point to a hollowing out of the public

health infrastructure — not just in the city, but across the country.

3. Lack of public faith in medical science, government and the media

"This was the height of polio in the United States," he said. "People had a much better sense of the impact of infectious disease. They saw it all the time, and they were rightly fearful. But they were also optimistic that medical science could conquer this. In 1947, there was tremendous faith in the medical community, unlike today."...

...Yet, [infectious disease experts] believe the biggest obstacle is not distribution but the public's distrust of government, science and the media.

"We're coming out of a train wreck of messaging," Dr. Redlener said. "We've learned that politics is poison to a public health initiative, especially during a crisis. Honesty and straightforward, clear messaging are absolutely critical."

In 1947, Dr. Weinstein was the only voice with a megaphone. He spoke and people listened.

"Back then, there was a much simpler media landscape," Ms. Sherman said as she laid out the Ad Council's campaign, which is due to kick off early next year. "In today's environment, we're dealing with a highly, highly fragmented media. We'll be relying on micro-influencers who are the trusted voices."

They seem to favor #3, noting that it is what 'experts believe'. But it seems so implausible—am I totally misunderstanding the claim? On the current margin, if lack of public trust was making the vaccine rollout even ten times slower, wouldn't we see campaigns begging us to go out and get the incredibly accessible vaccine, rather than seeing [elderly people camping in outside queues](#) to get vaccinated, and most people being told that they just can't have a vaccine for months or a year? Perhaps they mean that ultimately the number of doses given out will be limited by public willingness to receive them? (Which seems surely true, but not necessarily the answer to an interesting question).

The NYT's other suggestions don't seem immediately wrong, but are too vague for me to understand. I guess I'd like to know how things went differently at an object level. At what point, in 1947, did someone decide that it was their job to gather a volunteer army to administer vaccines, for instance? Do people work in similar roles nowadays? Did they think of this? Did they expect to get into trouble for having a bunch of lay people giving injections? Do they have to fill out a lot of paperwork to recruit them, whereas in 1947 they would have just shouted for them in the street? (I don't know.) If New York had had these constraints, would their vaccination campaign have looked like ours, or is there more to explain?

I suppose I have several questions:

1. What is really going on, that can account for a near 50x slowdown?
2. Why does the New York Times have such an unenlightening, vague, and seemingly wrong discussion so close to where one could have a genuinely interesting one?
3. ([Are these things related?](#))

4. My guess after thinking about it for ten seconds is that the gap in speed is to do with more things being regulated and formalized. The difference in time it would take me to physically cause someone to have a cookie versus to legally sell someone a cookie seems huge enough for this kind of thing to account for large slowdowns, and I do expect creating a massive band of volunteer non-nurses to administer a vaccine to require time-consuming paperwork or to be straight up illegal. How does this explanation sound?

This isn't just an idle inquiry, and shouldn't be just another interesting story for another payload of political disapproval.

Naively extrapolating, New York City could be fully vaccinated in about seven weeks if we knew how to do what was done in 1947. At the current rate, which will presumably change, vaccinating everyone would take years.²

What if someone figured out how to replicate even part of what New York did before at Weinstein's direction? In America alone, around three thousand people are dying each day now, as they wait for the vaccine. My boyfriend's ninety year old grandmother in Vermont was diagnosed with covid last week. Her center was scheduled to begin vaccinating its residents this Wednesday.

(Regardless of what makes things slower these days, good on everyone who is working to forward the vaccination effort, and also those doing their best to make it appropriately safe. Good luck being fast and right.)

P.S. In fact the whole of America vaccinated fewer people in the first two weeks of having a covid vaccine than New York did in 1947:



Also interestingly, others are not doing better.

1. My information about this is all from the [New York Times](#), [Wikipedia](#), and the [Rachel Maddow Show](#) ↵
2. Naive extrapolation says six years, but this is especially naive since at that rate covid will speed things up by reaching many people before the vaccine does. Plus we should probably expect some speedup over time, if it was going to take that long. Or something different to happen. ↵

Actually possible: thoughts on Utopia

(Cross-posted from [Hands and Cities](#))

Life in the future could be profoundly good. Many people accept something like this in principle. But I think it often goes underestimated in practice, especially once we imagine society's most glaring problems fixed, and ask where we might go from there. The difference in quality of life between a fixed-up version of our current world and the best possible future is, I think, less like the difference between a mediocre job and a beach vacation, and more like the difference between being asleep and being awake; between blindness and seeing; a droplet and an ocean; a cave and an open sky.

Following [Bostrom \(2008\)](#), let's call a profoundly good future "Utopia." The most important fact about Utopia, I think, is that we (that is, humanity and our descendants) could, with enough patience and wisdom, actually build it. This is a real thing we could actually do; a way the future could really be. It's a thing the way laptops were a thing before they were invented; the way you could, if you really wanted, find a way to get some ice cream by the end of next week. It's not some empty fantasy or symbol. It's a physically-available path that, modulo certain exotic scenarios, is actually open to us, if we don't mess up.

This post describes some different ways of thinking about Utopia, and what I think matters most about the differences. It also discusses some common objections to different types of "Utopian" thought and practice. My focus is on the *quality* of life in Utopia; see [Ord \(2020\)](#), Chapter 8, for more on just how much of that life there could be.

I. Concrete utopias

We can think of visions of Utopia in two broad categories: concrete, and sublime.

Concrete Utopias are imagined in specific and often human-scale terms. Most literary depictions of Utopia fall into this category, as does the vision offered by a friend of mine who hopes, if he makes it to the Utopian future, to "sit on the giant pile of pizza and play video games all day" (I think this is at least partly tongue in cheek, but it's a bit hard to tell).

One issue people raise about these Utopias — and which the pizza example, for me, illustrates — is that they don't always sound particularly appealing, especially not to everyone, or after scrutiny. [Orwell](#), for example, complains that "all efforts to describe permanent happiness, on the other hand, have been failures." See Yudkowsky, [here](#), for discussion of some of the complexities involved.

My own worry about concrete Utopias is not that they are unappealing. Indeed, Le Guin's [Omelas](#), absent the child, doesn't sound so bad to me at a glance. And regardless, whether we can describe or imagine a perfect society, from our current standpoint, seems to me like very little evidence about the degree of perfection available in principle.

Rather, my worry is that concrete Utopias are unrealistically *small*, familiar, comprehensible. Glancing at [various Wikipedias](#), my sense is that literary depictions of Utopia often involve humans in some slightly-altered political and material

arrangement: maybe holding property in common, maybe with especially liberated sexual practices, etc. And when we imagine our own personal Utopias, it can be easy to imagine something like our current lives, but with none of the problems, more of the best bits, a general overlay of happiness and good-will, and some favored aesthetic — technological shiny-ness, pastoralness, punk rock, etc — in the background.

This seems fine to me as a way of connecting emotionally with the possibility of a better world, or for thinking through different political possibilities. But if we start to slip into thinking that Utopia would actually be like that — e.g., that the best futures would be cleaned-up and somewhat improved versions of the world we know — then I start to worry about *drastically* underestimating how good — and how different — the future could be. Here a quote from [Bostrom](#) that I often think of in this context:

“What I want to avoid is to think from our parochial 2015 view—from my own limited life experience, my own limited brain—and super-confidentially [sic] postulate what is the best form for civilization a billion years from now, when you could have brains the size of planets and billion-year life spans. It seems unlikely that we will figure out some detailed blueprint for utopia. What if the great apes had asked whether they should evolve into *Homo sapiens*—pros and cons—and they had listed, on the pro side, ‘Oh, we could have a lot of bananas if we became human’? Well, we can have unlimited bananas now, but there is more to the human condition than that.”

Basically, many concrete Utopias sound to me a lot like the banana thing. It’s not just that they’re false in the way that all concrete visions of the future are false, or that they’re overconfident. It’s that they’re on the wrong scale, in the wrong register of comprehension. The apes who expect bananas aren’t just wrong about human food selection. They’re wrong, in a deeper way — a more exciting and important way — about what they’re getting themselves in for.

A lot of this is about the possibility of dramatic improvements to our capacities. These are notably absent from many concrete Utopias, and understandably so, given the difficulty of imagining them from our current perspective. The phrase “brains the size of planets” sounds silly, but I actually think it points in the right direction here, with the right connotation of alien and overwhelming differences in scale and situation. Obviously, we should approach any fundamental changes to our condition with extreme caution. But the case for — ethically, responsibly, safely — enhancing our capacities, once doing so becomes possible, seems to me extremely strong, and we know of few deep limits to the project, save those we impose on ourselves, or on each other.

Some will argue that a Utopia that involves such enhancements is not worthy of the name. I disagree with this, but I won’t get into the issue now. If we constrain our scenarios to ones in which the basic biological and cognitive situation that we face remains unaltered, then concrete visions of Utopia, based on what we can currently imagine, will have a better shot at capturing what’s available. But views, like my own, that countenance the possibility (given the right types of caution, ethics, safety, etc) of significant enhancements in human capacities should, I think, be very wary of the banana thing.

I think part of what’s underlying this thought is a broader aesthetic, which I’ll call the “vastness of mind-space aesthetic,” and which also, I think, underlies a lot of worries about artificial intelligence. This aesthetic applies the kind of “holy shit things can get

extremely big” lesson we’re familiar with from cosmology to the space of possible minds, in an effort to overcome similar types of parochialism (see [Ord \(2020\)](#), Chapter 8, for some discussion). Hence, one thinks of sufficiently advanced AI vs. human not as John Von Neumann vs. Homer Simpson, but as human vs. ant — and this only as a grossly inadequate gesture. The same expansion of scale applies to the quality of life that mind-space makes available, if we’re able to explore it.

II. Sublime utopias

Enter sublime Utopias. These visions focus less on concrete details, and more on the sense in which Utopia is *incomprehensible* from our current perspective. Hence the word “sublime,” which I hope will connote not just “awesomeness” but something like “beyondness” — e.g., a mental representation that foregrounds its own inadequacy, and points past itself to something overpowering and still-not-understood.

Sublime Utopias are less at risk of banana-type mistakes. Their central problem, though, is that they can lose all substance and emotional appeal altogether. Thus, if “heaven is a place where angels sing hymns, and you get to be with your family again” is a concrete vision of heaven, “heaven is eternal union with God in perfect love” is a sublime one, especially if we add “which you cannot hope to comprehend with a finite mind, but trust me it’s great.” The latter is less parochial, and less at risk of objections like “but I don’t like hymns” (though here I tend to think: man, the hymns in heaven could be so good); but it’s also less rooted in what actual humans care about, and correspondingly unexciting (the parallel holds for Hell: “separation from God” is a lot less scary than literal torture). Incomprehensibility drains content. In the extreme, sublime Utopias are pure light, with nothing to see.

Nick Bostrom’s “Letter from Utopia” is the best depiction of a sublime Utopia that I’ve encountered. I recommend the published [2008 version](#), still available on the Future of Humanity Institute’s website, rather than the [updated 2020 version](#), in which some my favorite parts have been removed or edited. In particular, Bostrom provides a great example of what I see as the best method of characterizing sublime Utopias without losing all content. I’ll call this method “extrapolating the trajectory of life’s best moments.”

Peak experiences are, I think, the most straightforward input to this method, but it can also, I think, be applied to changes in the quality of e.g. our relationship, communities, and epistemologies. Running with peak experiences as our example, though, and following Bostrom: consider what happens to your mind in the moments that have most shown you what life can be at its best — moments of joy, love, beauty, energy, immensity. Moments that have made you sit up straight, and say: “Oh. Whoa. This is crazy. This isn’t some pale imitation of life; this is the real thing. This is important. I didn’t realize this was possible.” For me, such moments often have a quality of greater awokeness, aliveness, *here*-ness. The world is vast, and shining with beauty and meaning; and it seems, not a new world, but the world as it always has been; the world I’ve always been living in, but haven’t truly seen.

The moments most salient to others may have a different flavor. My aim is not to isolate some specific dimension along which all such moments vary, but rather to point to the familiar experience of realizing that some dimension or other that matters to the quality of our lives can vary, in the good direction, much more than you had previously encountered; that there is more to the story than you had thought — something, perhaps, that lots of people have already been talking about; something

they may even be fighting for, changing their lives for; but which you hadn't, till now, really understood.

As Bostrom notes, a central issue with such moments is that *it's hard to remember what they're really like*, once they fade. And they fade all too quickly. Life descends into something more dead, numb, thin, mundane. Bostrom calls this "soot"; [Tim Urban](#), "fog"; someone else I know, "Dilbert dust." For our purposes, though, the important thing about such experiences isn't how long they last, or whether it's worth, in our present situation, trying to make them last longer; the point is the evidence they provide about what's possible. From Bostrom (in the 2008 version):

"Quick, stop that door! Look again at your yellowing photos, search for a clue. Do you not see it? Do you not feel it, the touch of the possible? You have witnessed the potential for a higher life: you hold the fading proof in your hands. Don't throw it away. In the attic of your mind, reserve a drawer for the notion of a higher state of being, and in the furnace of your heart keep at least one aspiring ember alive."

I think the notion of "proof" here is a good one. As far as I can tell, we basically have proof that life can be *at least* as good as the best experiences any human has ever had — and these experiences are much better, richer, vaster, more conscious and alive and awake, than can be easily appreciated when they're not being had.

More importantly, though, I think we basically know (even if we don't strictly have proof) that these experiences are nowhere near the limit of what's possible; that our actual, physical universe is one that makes possible structures of consciousness and wisdom and joy and love that are vastly farther in whatever direction our minds travel, when they become more conscious, wise, joyful, loving, than anything anyone has seen so far. There are oceans we have barely dipped a toe into; there are drums and symphonies we can barely hear; there are suns whose heat we can barely feel on our skin.

From [William James](#):

"Our normal waking consciousness... is but one special type of consciousness, whilst all about it, parted from it by the filmiest of screens, there lie potential forms of consciousness entirely different."

Indeed, some possible forms of consciousness may be simultaneously extremely good, and *utterly* unlike anything any creature on earth has ever touched. Oceans, not that we haven't waded into, but that we don't even know exist.

That said, I'm inclined to think that Utopia, however [weird](#), would also be, in a certain sense, recognizable — that if we really understood and experienced it, we would see in it the same thing that made us sit bolt upright, long ago, when we first touched love, joy, beauty; that we would feel, in front of the bonfire, the heat of the ember from which it was lit. There would be, I think, a kind of remembering. As [Lewis](#) puts it: "The gods are strange to mortal eyes, and yet they are not strange." Utopia would be weird and alien and incomprehensible, yes; but it would still, I think, be *our* Utopia; still the Utopia that gives the fullest available expression to what we would actually seek, [if we really understood](#).

Perhaps there are arguments out there that this sort of "sublime Utopia" oversells what's at stake (Internet, I'm curious to hear what you think are the best ones). My current best guess failure mode would be one in which it turns out that humans, for some reason, don't or can't value things very different in depth and intensity from

what we have now. That is, my best guess is that if what I've said about un-swum oceans and unheard symphonies turns out false, it's false for ethical rather than metaphysical or physical reasons; false because our values are small, not because the space of possible modes of consciousness is any less large. The human brain, after all, is an extremely specific and limited organ: it has some 100 billion neurons, it uses some 20 Watts of power, its signals travel, at best, at some [hundreds of meters per second](#). These are nowhere near physical limits. If small changes in chemistry can produce the dramatic variation we actually witness in quality of human mental life; if small (relative to what's possible) changes in brain architecture and cell count can move us from ants to mice to monkeys to humans; then think of what a sustained, serious (cautious, ethical, responsible) effort to explore the full range of what's possible is likely to reveal.

III. Rapture of the nerds?

As far as I can tell, this is a real thing. It's not some woo thing that the hard-nosed people understand to be silly. There's no fancy metaphysics here. We don't need to think that peak experiences reveal any special insight or knowledge unavailable to science or the mundane world. We don't need to think that any technologies are around the corner. We don't even need to be realists about consciousness. All we need to think is that whatever happens in our brains and bodies, during the events humans label "peak experiences," is movement along some trajectory we care about; and that it is possible for bodies, brains, and other relevantly similar physical structures to travel much, much further in that direction. This seems to me like the actual situation we're in.

All the same, people object to visions of Utopia — and especially, sublime ones — on the grounds of their spiritual connotations. "Rapture of the nerds," people say (though this phrase can be applied to many things). Or, maybe more specifically: it all sounds a bit too much like the thing people hoped for out of Heaven, or that people get all excited about when they're in some kind of "spiritual" mode. Better to keep your feet on earth.

It is certainly possible relate to the possibility of Utopia using archetypes, impulses, and forms of psychological orientation familiar from religious and spiritual contexts. As a small example: as with Heaven, it can be easy, in the face of Utopia's incomprehensibility, to round it off, in the imagination, to something like an abstract wash of perfection or oceanic light, maybe with a few incomplete images thrown in; or to think of it as a place where all problems are solved, all desires fulfilled. But unlike Heaven, Utopia, if something like it ends up getting built, will be a specific, concrete, physical world, with attendant frictions and problems, idiosyncrasies and contingencies; its own ways of distributing resources, resolving/preventing conflicts, and so on; and ultimately, with fundamental limitations on what can be done. There are no promises of perfection here, or of eternity.

More broadly, some possible failure modes in relating to Utopia do seem similar to failure modes associated with religion (though some of these are also failure modes of believing, more generally, that something is extremely important). I discuss a few of these in the next section.

At a higher level, though, I want to note that I actually think it fine and appropriate to direct certain attitudes familiar from religious and spiritual contexts towards something like Utopia. In particular, I think Utopia an appropriate object of hope, wonder, humility, and purpose. It is, in the words of [Carl Sagan](#), a "worthy goal."

What's more, while the forms of consciousness, love, joy, and wisdom available in Utopia are not transcendent of the natural world, they are transcendent, in some basic sense, relative to us. To point at them, we need to point past ourselves, along the line we've already seen ourselves travel, but beyond our own current limits, to something rightly, I think, thought of as higher, deeper, more awake, more real. Spiritual and religious contexts are the primary place humans have gone to develop tools and practices for doing this type of thing.

Indeed, at their best (in my view), religion and spirituality express an aspiration to look ultimate reality in the eye; to live in the light of our actual existential condition, seen, as much as possible, in its entirety. Utopia, I think, is the place for this too — and in that sense, the proper object of something closely akin to religious aspiration. Now we see in a mirror dimly. Then, face to face. (Or at least, much less dimly.)

Some will eschew attitudes that smack of religion and spirituality regardless. That's OK too. The main thing is not shrink your sense of what's possible in doing so; not to put the wrong limits on what can be really hoped for, what can *actually be a thing* in this plain old atoms-and-stuff itchy-feet Donald Trump gotta-wash-the-dishes universe, because it sounds too far outside the bounds of the mundane, the familiar — and hence, one might subtly assume, the secular.

IV. Does this matter? Is it good to think about?

Do visions of Utopia matter? To work towards a better world, after all, we don't need to ask, in any detail, just how much better it could be — what our efforts could ultimately lead to. We just need to keep moving in the right direction. Perhaps, then, this kind of thinking is a distraction or worse, especially given how pressing our actual, present problems are; how much we seem at risk of futures much worse than Utopia; how idle such speculations can seem, as the world burns.

Indeed, the project of trying to sketch a plan for something like Utopia, and to build it now, here on earth, has a very bad track record. "Utopian" thinking has an aura, at best, of a kind of thin-ness and naïveté — an aspiration to encompass all of the world's mess and contradiction in a top-down, rational plan, to "solve things once and for all," to "do it right this time." Here one thinks of poorly but "rationally" planned cities, failed communes, and the like.

And at worst, of course, Utopianism ends in abject horror. There is a certain type of fervor that comes from thinking, not just that paradise is possible, *but that you and your movement are actually, finally, building it*. In the grip of such enthusiasm, humans have historically been very willing to do things like "purge" or "purify"; to crush and suppress any opposition; to break eggs to make omelets, and so on. As [Kundera](#) writes:

"Totalitarianism is not only hell, but also the dream of paradise — the age old dream of a world where everybody would live in harmony, united by a single common will and faith, without secrets from one another. . . If totalitarianism did not exploit these archetypes, which are deep inside us all and rooted deep in all religions, it could never attract so many people, especially during the early phases of its existence. Once the dream of paradise starts to turn into reality, however, here and there people begin to crop up who stand in its way, and so the rulers of paradise must build a little gulag on the side of Eden..."

We've seen what humans marching under Utopian banners have wrought; our historical immune system has left us wary.

And there are subtler objections to dreams of Utopia as well. Paying lots of attention to how much better life could be can smack of greed and dissatisfaction; of wanting rather than appreciating; of ambition and striving rather than humility and contentment. [Le Guin](#), in characterizing what Utopia has been (though need not be), adds a slew of further adjectives: “Euclidean,” “masculine,” “European,” “dry,” “aggressive,” “lineal,” “expanding,” “advancing.” One wants, perhaps, more earth, dirt, ground. The place to live and love and work is *here*; and *here*, one might think, risks neglect or worse, if we set our gaze on some imagined and unknowable future.

There's lots to say about these sorts of objections. For now, the main thing I want to emphasize is that I don't think we should spend much time now trying to figure out in any detail what Utopia looks like; that task is for a species much wiser and safer than we are at present. The main thing for us to do about Utopia, I think, is to [protect its potential to someday be made real](#): and that does, indeed, mean living and loving and working *here*, in the present.

But keeping in mind a basic picture of what's at stake is important. This is partly because it matters to the type of hope one has; the type of story one thinks of human pain and joy as ultimately a part of; the type of victory one thinks it possible, amidst all the world's darkness, to fight for, and to win. But more broadly, I think, the possibility of a profoundly, unimaginably good future is, modulo certain exotic scenarios, just a basic fact, a feature of our actual situation that we need to act in light of. For the normal, mundane reasons that apply to most basic but possibly decision-relevant facts, then, I think it's worth being clear about.

How to Write Like Kaj Sotala

Since I started quarantine in late February, I've been trying to learn the skill of modelling implicit skills and knowledge. Over that time, I've interviewed nearly 20 people and learned one specific implicit theory or skill that they use. Many of those people in the rationality and post-rationality communities.

Kaj Sotala is one of my favorite writers on Lesswrong. He writes clearly, deliberately, and precisely. His writing is deliberate, interesting, and approachable.

After nearly 4 hours of interviewing him, this is my model of how he writes.
Note that while it **is inspired by Kaj, it is not necessarily fully endorsed** by him.
There is always a translation process from his brain to his mouth to my brain to refining to putting it to the page.

That being said, my ability to help correct misunderstandings and bridge perspectives in text has profited massively from internalizing the model here.

Beliefs and Motivation

Primary Motivator

Let's first talk about beliefs and motivations. What's his primary motivation to start this process? For him, it's a process of seeing people disagree, and feeling a **visceral sense of frustration** at people talking past each other.

For him, it's almost a proprioceptive sense of two different shapes. One person is saying circle, and the other is hearing square. It's really important to make these shapes match up!

Underlying Motivator

There are two underlying values here. The less salient one is wanting a sense of **admiration** from others. It's really nice to get praise for creating good explanations that unify two viewpoints.

The more salient value is a sense of **Harmony and Beauty**. If you can create a beautiful explanation, people can understand each other. The universe is just a bit more harmonious if you can take two perspectives and integrate them!

Necessary Conditions

In terms of what initially motivates him and lets him see the opportunity for this sense of Harmony and Beauty is when he sees people talking past each other, and **feels like he sees where the disconnect is**. Or he has a sense of what could be said instead.

Writing Strategy

So those are some of his underlying beliefs and motivations. But how does he actually craft these beautiful explanations? What is his strategy?

Tests

First, there's a few background processes he is running when he's writing. Tests to make sure that what he's writing will lead to that beauty and harmony, and resolve that frustration.

Looking at Criticisms

The first test he does is recalling specific criticisms from earlier and the conversation, and sort of imagining the general type of person that would have those criticisms. He's checking if what he's writing is actually **answering those criticisms in a convincing way**.

Imagining a Specific Person

Second, he's sometimes imagining a specific skeptical person reading what he has written, and seeing how he imagines they would react. He's asking:

1. **Do they understand what I'm saying?**
2. **Are they convinced by what I'm saying?**

Primary Strategy

Let's start with his Primary Strategy. In order to meet those tests, **first build a model of what both people want**. Begin **generating some explanations**, and see if they're satisfying that model.

As you begin to generate ideas and explanations, **make some bullet points that capture these ideas** and explanations. Keep doing this and jotting down notes **until you feel you have enough to be convincing!**

Then begin to **use this strategy recursively to flesh out each of the bullet points**. At this point you can also begin to **edit what you've written**, and **bring in related ideas**. You can also begin to **look up specific details** that would help.

Secondary Strategy

So that's the primary strategy. But what does Kaj do when using this strategy still doesn't sufficiently meet his criteria to create harmony and beauty, resolve frustration, and satisfy the skeptical person in his head? That's where the secondary strategy comes in!

First, **check if there's a lack of consistent frame or narrative** that would tie the bullet points together. It's important not only two have harmony and beauty between the two perspectives, but between the individual bullet points as well!

The way to do this is to begin to **brainstorm a bunch of different narratives or frames**, and loop back to your tests and your sense of harmony and beauty to **see which one would be best**. You can also **look at other explanations of similar things** and see what frames worked.

Another related failure mode that can happen here is that you're **wishywashy between two different narratives** or frames, because both feel natural to use.

In this instance, you need to accept that you can only take one path. Kaj likes to remind himself of the phrase "**Kill your darlings**" to help himself with this.

Tertiary Strategy

So what if he's tried his primary strategy, and he's tried his secondary strategy, but no matter what he tries he can't make any headway. That's where his tertiary strategy comes in!

If you keep trying to add bullet points and creating a narrative, but it's still not working, It's often because **you're trying to explain too much** or to too general of an audience.

The strategy here is just to **explain a smaller subset** of what you were trying to do. You can also give the piece to someone in your target audience, and **get very specific feedback** of what's not working.

Alternatively, sometimes you're just **trying to explain something that you don't have a full grasp of** yet. In this case, ask yourself if you anticipate reading to be useful. If so, go back and **read some more material** to get a firmer grasp and jog your creativity.

Sustaining Emotions

So in general, those are his cognitive strategies for writing. What is the sustaining emotion that his carrying him through this process? There are actually two different modes Kaj can operate in here.

Urgent Excitement

One is a sense **Urgent Excitement**. It feels like he has a million ideas in his head, and there's this sense of urgency to capture them, which leads to this sense of exciting momentum as they begin to take shape.

Quiet Satisfaction

The other is more of a sense of **quiet satisfaction**. It's almost like building a physical thing, like a house or a bench. There's a sense of building something slowly and deliberately, and watching it become more real.

Behaviors

So that's how he's feeling, what is he actually DOING on the outside? What are the behaviors that can help here?

First, Kaj will often **go on walks to think**, he needs something that he can put his attention on so he's not obsessively thinking about writing.

Second, **listening to music** can be really helpful. Find something that has the vibe you want in your writing.

Third, he likes to **have one browser window that ONLY has the google doc he's writing in**, and a separate browser window with all of his tabs for research. This allows him to get less distracted when he's writing!

Supporting Factors

Are there any other factors that help Kaj do what he does? Yes, here are a few that seem the most helpful.

First, Kaj has a giant **google doc for capturing everything that may be relevant**. Whenever he stumbles upon something that might be useful, he tries to add it to this document.

Second, he often tries to **create an uninterrupted space and time to write**. This really makes the process go smoother.

Third, the use of bullet points is his secret weapon. Sometimes writing narrative paragraphs feels hard. So **by using bullet points, he can pretend he's not doing that!** Sometimes he'll write the whole piece in bullet points, then simply remove them and he has paragraphs!

Internalizing This Strategy

So that's Kaj's writing strategy. Are you a writer who wants to incorporate this into your own process? Here's the process I recommend.

Reimagine the Past

First, think of a time in the past when you could have used this strategy. Go down each point one by one, and slowly and cumulatively "**edit your memory" to add each point to your memory**". Notice how it would have been to use this writing strategy!

How it would have been to have these factors. How would it have been to be motivated in this way? What would it have felt like and how would you have acted differently. How would your writing have come out differently?

Imagine the Future

Then, think of a time in the near future when you'd like to write, and what you'd like to write about. Slowly and cumulatively add each element of this strategy, and imagine exactly what you'll do as you write using this strategy and these emotions. How will it feel. How will you begin? imagine in vivid detail, as if you're already there.

Practice in the Present

Finally, use the strategy in the present. **Set aside some time to write, and slowly and cumulatively add each piece of the strategy** until you've written something amazing.

A Call for More People to Interview

If you're interested in listening to all these interviews and internalizing the model, [sign up for my interest list here.](#)

More importantly, I'm always looking for people to interview. If you have something you do, but you don't know exactly how you do it, [go here to sign up for an interview!](#) I'd love to chat with you.

Covid: The Question of Immunity From Infection

Over and over and over again, I've been told we should expect immunity from infection to fade Real Soon Now, or that immunity isn't that strong.

With several recent papers and the inevitable media misinterpretations of them, it's time to take a close look at the findings.

This was originally part of the 1/21 update, but I've split it off so that it can be linked back to as needed, and to avoid cluttering up the weekly update.

Note that this post is not looking at any new strains that might provide immune escape. It's studying infections during a period when such strains were not a substantial issue. This is distinct from concerns about strains with immune escape characteristics.

First up is this paper:

- Do antibody positive healthcare workers have lower SARS-CoV-2 infection rates than antibody negative healthcare workers? Large multi-centre prospective cohort study (the SIREN study), England: June to November 2020**

Findings Between 18 June and 09 November 2020, 44 reinfections (2 probable, 42 possible) were detected in the baseline positive cohort of 6,614 participants, collectively contributing 1,339,078 days of follow-up. This compares with 318 new PCR positive infections and 94 antibody seroconversions in the negative cohort of 14,173 participants, contributing 1,868,646 days of follow-up. The incidence density per 100,000 person days between June and November 2020 was 3.3 reinfections in the positive cohort, compared with 22.4 new PCR confirmed infections in the negative cohort. The adjusted odds ratio was 0.17 for all reinfections (95% CI 0.13-0.24) compared to PCR confirmed primary infections. The median interval between primary infection and reinfection was over 160 days.

From this, of course, media headlines were things like “immunity only lasts five months,” but let’s ignore that and keep looking at the data, and see what the study actually says.

Variables

Questionnaires on symptoms and exposures were sent electronically at baseline and every two weeks (Supplementary appendix); SARS-CoV-2 antibody (using the Roche cobas® or Abbott immunoassay®) and Nucleic Acid Amplification Testing (NAAT), generally RT-PCR, was conducted at enrolment and at regular intervals (PCR every two weeks, antibody every four weeks). Testing was performed in the clinical laboratory at the site of participant enrolment, using locally validated testing platforms.

Cohort assignment at baseline

Participants were assigned to the positive cohort if they met one of the following criteria: antibody positive on enrolment or antibody positive from prior clinical laboratory sample, with or without prior PCR positive; antibody negative on enrolment with prior antibody positive, with or without prior PCR positive; antibody negative on enrolment with a PCR positive result prior to enrolment. Participants were assigned to the negative cohort if they had a negative antibody test and no documented positive PCR test. Those in the negative cohort moved to the positive cohort 21 days following a PCR positive test result or at the time of antibody seroconversion with no positive PCR test.

Reinfection case definitions

The SIREN case definitions for reinfections have been described elsewhere and range from confirmed to possible dependent on the strength of serological, genetic and virological evidence.³⁶ A possible reinfection was defined as a participant with two PCR positive samples 90 or more days apart (based on previous national surveillance analysis)³⁶ with available genomic data or an antibody positive participant with a new positive PCR at least four weeks after the first antibody positive result. A probable case additionally required supportive quantitative serological data and/or supportive viral genomic data from confirmatory samples.

We subcategorised possible reinfections by symptom status to highlight those with stronger evidence and provide comparability with definitions used elsewhere.^{28,31} Participants reporting any of cough, fever, anosmia or dysgeusia 14 days before or after their positive PCR result were defined as having 'COVID-19 symptoms' and 'other potential COVID-19 symptoms' if reporting any other recognised symptoms listed in Appendix A.^{34,35}

RESULTS SECTION:

By 24 November 2020, 409 new infections were detected in the negative cohort: 318 were new PCR positive infections; 249 (79%) of these cases were symptomatic at infection, 196 (62%) with typical COVID-19 symptoms, and 53 (17%) with other symptoms; 40 (12%) were asymptomatic and 28 (9%) did not complete a questionnaire at the time of their symptoms; Forty-four reinfections were identified, 15 (34%) were symptomatic: two defined as probable (described in detail elsewhere³⁶), both symptomatic, and 42 possible; 13 symptomatic, two (23%) of whom reported typical COVID-19 symptoms. Forty (both probable and 38 possible) reinfections were antibody positive at enrolment; three had previously positive antibody tests but two were antibody negative and one indeterminate on enrolment; and one individual remained antibody negative but reported COVID-19 symptoms and a documented PCR positive status in April 2020. Twenty-one (47.7%)(50%) of these individuals had historic PCR positives from their primary infection, of whom 19 reported COVID-19 symptoms and two other symptoms within 14 days of their positive test. Fourteen (31.8%) individuals (including both probable cases) reported a history of COVID-19-like illness but did not have a PCR test due to lack of availability at the time of their primary illness; 13 (92.9%) with typical COVID-19 symptoms and one with other symptoms. Nine (20.5%) reported no history of any potential COVID-19 related symptoms.

Bottom line infection rates:

respectively (Table 3). The incidence density per 100,000 days of follow up between June and November 2020 in the positive cohort was 3 .3 reinfections and in the negative cohort was 17.0 new PCR positive infections per 100,000 days of follow-up. Figure 3 describes the

Finally:

Restricting reinfections to probable reinfections only, we estimated that between June and November 2020, participants in the positive cohort had 99% lower odds of probable reinfection, adjusted OR (aOR) 0.01 (95% CI 0.00-0.03). Restricting reinfections to those who were symptomatic we estimated participants in the positive cohort had 95% lower odds of reinfection, aOR 0.08 (95% CI 0.05-0.13). Using our most sensitive definition of reinfections, including all those who were possible or probable the adjusted odds ratio was 0.17 (95% CI 0.13-0.24).

I'll pause here before I read the discussion section.

What I am seeing is that in *probable* infections, meaning infections that were serious enough and real enough to get confirmed, we see a 99% reduction, a large enough reduction that error in the original antibody/PCR tests might well account for either or both of the remaining two (2) cases.

Even looking at only *symptomatic* infections, we still get a 95% reduction.

Whereas if we only look at 'there was a test that came back positive on people getting periodically tested, but without requiring any symptoms or verification' we only get an 83% reduction.

Naturally, the public-facing articles all seem to quote the 83%, and ignore the 95% and 99%.

I'd also note that they nowhere attempt to control for the two most obvious differences between the two samples, which are:

1. The antibody positive sample knows they are antibody positive, and thus likely took fewer precautions across the board than they would have otherwise.
2. The antibody positive sample *are the people who got infected the first time*.

If immunity conferred zero protection, what do you think the risk ratio would have been? Three?

(And again, they also take something presented after five months of follow-up, and report it as 'immunity lasts five months' because journalism.)

They start off with this in the discussion, which seems maximally skeptical (e.g. 'at least 75%' means the 95% CI starts at 76%) and they seem to know it:

We have detected two probable reinfections (both symptomatic with high viral loads, genome sequencing demonstrating phylogenetic relatedness to concurrently circulating strains, and a boosted antibody response), which have been characterised and reported separately,³⁶ and 42 possible reinfections in our positive cohort. This compares with 318 new PCR positive infections, 249 of whom were symptomatic, 78% with typical COVID-19 symptoms, in our negative cohort. Using a symptomatic case definition aligned with positive PCR results, previous infection reduced the odds of infection by at least 90% (aOR 0.06 with 95%CI of 0.03 to 0.09) and even when we included all possible and probable reinfections reduced the odds of reinfection by at least 75% [aOR 0.17 (95% CI 0.13-0.24)]

We believe this is the minimum likely impact as the curve in the positive cohort was gradual throughout, indicating some of these potential reinfections were likely residual RNA detection at low population prevalence rather than true reinfections. In the negative cohort the gradient was gradual up to around day 100 and has then accelerated, broadly coinciding with the period when community prevalence increased rapidly.³⁸ In addition, we did not include 94

Primary conclusion here seems very clear. Past infection is highly effective protection against future infections that matter.

Secondary conclusion is that our tests are not perfect, and when you loosen your criteria, you get a bunch of false positives.

Also that it is perfectly fine not to consider that there *might be a difference* between the behaviors, past and present, or the prior health or immunology, of those who are known to previously have had Covid-19, in a scientific paper, with a straight face.

Mostly, this paper confirms that dogs bite men, and it is much rarer for men to meaningfully bite dogs.

On to the [second paper](#), also oh my eyes, ouch this typeface my eyes.

First thing I note is that they used “Poisson regression” which might be technically correct but is a deeply, deeply silly way to say that you’re counting the number of days of exposure.

How scientific papers say something is “so overwhelmingly obvious we didn’t actually need to do this study but we did anyway” these days:

Evidence for post-infection immunity is emerging. Despite an estimated 55 million people infected worldwide and high rates of ongoing transmission, reports of SARS-CoV-2 reinfection are few, mostly in individuals with mild or asymptomatic primary infection.¹¹⁻²⁰ Although a lack of widely-available PCR testing early in the pandemic may limit the numbers of confirmed reinfections reported, this suggests that infection with SARS-CoV-2 provides some protective immunity against reinfection in most people.

Their data:

165/11052 (1.5%) HCWs were PCR-positive while anti-spike IgG seronegative, 76 during asymptomatic screening and 89 while symptomatic. 3/1246 (0.2%) HCWs were PCR-positive following a positive anti-spike IgG measurement, all were asymptomatic. Allowing for the varying duration of follow-up, rates of new PCR-positive results were 0.86 and 0.21/10,000 days-at-risk in seronegative and seropositive HCWs respectively (incidence rate ratio, IRR, 0.24 [95%CI 0.08-0.76, p=0.015]). No antibody-positive individual had a subsequent symptomatic infection; rates of new PCR-confirmed symptomatic infection were 0.46 and 0.00/10,000 days-at-risk in seronegative and seropositive individuals respectively.

Incidence of PCR-positive results varied by calendar time (Figure 1A), reflecting the first and second waves of the pandemic, but was consistently higher in the seronegative versus seropositive group. Adjusting for age, gender and for changes in incidence by month, the IRR for being seropositive was 0.25 (95%CI 0.08-0.80, p=0.019) (Table 2) and adjusting for time as a continuous variable, 0.26 (95%CI 0.08-0.81, p=0.020) (Figure 1B). As rates of asymptomatic testing varied by antibody status, we performed a sensitivity analysis, randomly removing PCR results for seronegative HCWs to match testing rates in seropositive HCWs, yielding an adjusted IRR of 0.28 (95%CI 0.09-0.90, p=0.032).

Their bottom line:

Here we present follow-up from a prospective longitudinal cohort study of healthcare workers (HCWs) at Oxford University Hospitals (OUH). Comparing the incidence of PCR-positive results and symptomatic infection in seropositive versus seronegative HCWs during up to 30 weeks of follow-up demonstrates post-infection immunity lasting at least 6 months.

The study assumes that people are protected a minimum of 60 days following infection, and only considers exposure days after that, then measured relative frequency of positive testing of various types.

There were zero symptomatic reinfections, while roughly half of all positive tests in the non-infected group had symptoms. Various tricks get them all the way to only a 72% reduction in infections for the previously infected group, while once again not at all taking into account any way the group differs, or the rate of false positives.

These two studies, taken together, seem to tell a clear story. You are not fully protected against getting a positive PCR test, and thus there is some chance you could at future points be somewhat infectious, but for the duration of the studies you are at essentially zero meaningful health risk, or even risk of showing symptoms, from Covid-19 if you were previously infected.

Why I'm excited about Debate

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

I think [Debate](#) is probably the most exciting existing safety research direction. This is a pretty significant shift from my opinions when I first read about it, so it seems worth outlining what's changed. I'll group my points into three categories. Points 1-3 are strategic points about deployment of useful AGIs. Points 4-6 are technical points about Debate. Points 7-9 are meta-level points about how to evaluate safety techniques, in particular responding to [Beth Barnes' recent post on obfuscated arguments in Debate](#).

1. Question-answering is very useful.

People often claim that question-answering AGIs (which I'll abbreviate as QAGIs) will be economically uncompetitive compared with agentic AGIs. But I don't think this matters very much for the two most crucial applications of AGIs. Firstly, when it comes to major scientific and technological advances, almost all of the value is in the high-level concepts - it seems unlikely that implementing those advances will require AGI supervision (rather than just supervision from narrow AIs) during deployment.

Secondly, aligned QAGIs can do safety research to help us understand how to build aligned agentic AGIs, and can also predict and prevent their misbehaviour. So even a relatively small lead for aligned QAGIs could be very helpful.

2. Debate pushes capabilities in the right direction.

Another objection I used to have: I tend to expect that QAGIs are pretty safe anyway, which implies that in aligning QA systems, Debate isn't helping tackle the cases we should be most worried about. And if it's used as a final step of training for systems that have previously been [trained to do other things](#), then it's very unclear to me whether Debate would override whatever unsafe motivations those systems had already acquired.

But now I think that Debate could be an important tool for not only making QA systems more aligned, but also more competitive. Systems like GPT-3 have shown a very good understanding of language, and I expect them to gain much more world-knowledge, but it's hard to elicit specific answers from them. We might hope to make them do so by using reward-modelling to fine-tune them, but I expect that in order to scale this up to complex questions, we'll need to make that process much more efficient in human time. That's what Debate does, by allowing humans to evaluate answers on criteria that are much simpler than the holistic question "is this answer good?"

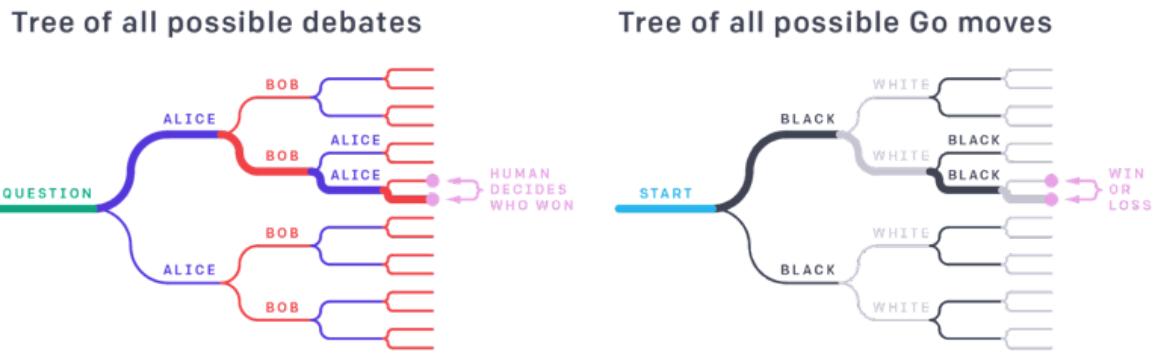
3. Debate provides a default model of interaction with AGIs.

We won't just interact with agentic AGIs by giving them commands and waiting until they've been carried out. Rather, for any important tasks, we'll also ask those AGIs to describe details of their plans and intentions, and question them on any details which we distrust. And for additional scrutiny, it seems sensible to run these answers past other AGIs. In other words, Debate is a very natural way to think about AGI deployment, and describes a skillset which we should want all our AGIs to have, even if it's not the main safety technique we end up relying on.

4. Debate implicitly accesses a complex structure

I originally thought that Debate was impractical because debates amongst humans aren't very truth-conducive. But now I consider it misleading to think about Debate as simply a more sophisticated version of what two humans do. The comparison to a game of Go is illuminating. Specifically, let's interpret any given Go position as a question: who wins the

game of Go starting from this position? Then we can interpret a single game of Go played from that position, by sufficiently strong players, as good evidence that the (exponential) tree of other possible games doesn't contain a refutation of any of the moves played by the eventual winner. Similarly, the hope is that we can interpret a single line of debate, starting from a given question, as good evidence that the exponential tree of other lines of debate doesn't contain a refutation of any of the claims made by the eventual debate winner.



In other words, the core insight of Debate is that we can evaluate a whole argumentative tree while only exploring one branch, given a strict standard of judging (i.e. whoever loses that one branch loses the whole tree), because the debaters will model the rest of the tree to the best of their (superhuman) abilities. We can't do this in normal debates between humans, because human debaters aren't smart enough for other humans to reliably interpret the outcome of one specific branch of a debate tree as strong evidence about the rest of the tree. Therefore human incentives are very rarely set up to punish minor errors, which may allow compounding inaccuracy. A better analogy than a normal human debate might be to a human debate where, before making each argument, each side can consult a large team of experts on the topic; in this case, it seems much more reasonable to expect both that small mistakes will be caught, and that small mistakes are deliberate lies which should cause the liar to lose the debate. (Punishing minor errors does introduce more variance into AI debates, since the truth-telling debater could get unlucky by making small mistakes. But unlike human debates, we can run AI debates a large number of times, hopefully decreasing that variance significantly.)

The intuition I've described makes Debate compare favourably to [recursive reward modelling](#) (RRM), which needs to actually implement the whole exponential tree of agents answering subquestions. (I think Jan Leike envisages RRM trees as being much shallower than Debate trees, though.) RRM does have other advantages - in particular its ability to train agents which actually take actions in the world. But as already discussed, I find this less compelling than I used to.

5. Reasoning can be truth-conducive even in adversarial environments

I'm reasonably compelled by [Sperber and Mercer's claim](#) that explicit reasoning in humans primarily evolved not in order to help us find out about the world, but rather in order to win arguments. [EDIT: More specifically, Sperber and Mercer claim that "reason is not geared to solitary use, to arriving at better beliefs and decisions on our own. What reason does, rather, is help us justify our beliefs and actions to others, convince them through argumentation, and evaluate the justifications and arguments that others address to us." This still involves evaluations which aim to find out which arguments are right about the world]. I think this frames our current situation in a new light: reasoning fails to track the truth so often not necessarily because it's a weak tool, but because it's specifically been selected to promote many of these failures (such as overconfidence in our own claims). And yet, despite that, it's still reasonably truth-conducive - we can still reason about complex scientific domains, for

example. This makes me more optimistic that Debate can be fairly truth-conducive even if the training incentives aren't quite right.

Consider also that, for existential safety, we only need reasoning to be truth-conducive enough to detect *catastrophes*. We should expect that doing so is much easier than finding the truth about all (potentially very subtle and nuanced) questions.

6. Debate passes the human relevance test.

[An important heuristic I use](#) for evaluating safety research directions: would these work for making humans safer, if applied over evolutionary timeframes? This decomposes into a few different components:

1. Does the technique apply to “prosaic” machine learning, without requiring deeper insights into cognition? I.e. does it work if we build AGI by optimising a big neural network via local search techniques like gradient descent or evolution?
2. Does the technique scale up to generally intelligent agents?
3. Does the technique scale up to very complex environments?
4. Does the technique work even after [a long period of “pre-training”](#) instills unsafe traits in the agent?

I think that Debate passes the first three of these tests, which is a good sign! Most other safety techniques don't. The fourth criterion is a harder one: it's analogous to the question of whether a safety technique applied to humans would be able to remove deeply-ingrained goals like our survival instincts, while avoiding deceptive alignment. I'm not convinced that Debate solves this, but I don't think any other safety technique does so either.

7. Debate promotes empirical investigation

We can consider Debate to be one type of [iterated amplification](#) (IA) - but for the purposes of doing good follow-up research, I think the specificity of Debate makes it more valuable than the broad concept of IA in general. In particular, this specificity makes it easier for more people to do conceptual and empirical investigation. This is especially important given that so many ideas in safety are so abstract and hard to ground in concrete research questions. Further, the adversarial framing of Debate makes it easier to spot problems, by pushing us to think about how the setup might be exploited. These factors have already contributed to interesting insights, such as [OpenAI's work on cross-examination](#).

8. Recent negative results aren't too discouraging.

[Beth recently put a post online](#) with negative results from Debate - in particular, the finding that misleading debate trees can be constructed for which it seems necessary to examine all the steps of the argument in order to refute it.

There are a few reasons I'm not too discouraged by this. Firstly, the physics problems they're using to evaluate Debate are *hard* - they are designed to be counterintuitive, and sometimes confuse physics PhDs. Even people who can solve them in one way often can't evaluate other potential ways of solving them. So the fact that we can generate confusing debate transcripts about them isn't much evidence about the limitations of Debate on domains which both debaters have thorough knowledge of. Note also that we haven't yet seen what top-notch physicists with relevant expertise would make of these transcripts; it may be that it only takes a few weeks of practice on these types of everyday physics questions for experts to gain sufficient knowledge to refute misleading arguments.

We might be worried that, if we have such trouble with everyday physics, it'll be very difficult to scale up to more difficult questions. But note that current difficulties are partly because Debate experiments don't (yet) allow debaters to make empirical predictions. Given how valuable this step has been for humans, it seems plausible that adding it would make Debate significantly more powerful. A very rough summary of human intellectual history: we tried to

make progress via debate for thousands of years, and gained little knowledge (except in maths). Then we started also relying on empirical predictions, and underwent the scientific and industrial revolutions shortly afterwards. Predictions are very powerful tools for cutting through verbal obfuscation, when used in conjunction with verbal reasoning.

9. Recent negative results expect too much from Debate.

The OpenAI team is trying to use Debate to access *all* the knowledge our agents have. But this seems like an unrealistic goal - consider how incredibly difficult it is in the case of humans. There's plenty of human knowledge that is very difficult to access (e.g. because it relies on finely-honed intuitions) even when the human in question is being fully cooperative. Indeed, given how much knowledge is tacit or vague, I'm not even sure what it would mean to succeed in accessing all an agent's knowledge.

From my perspective, if Debate makes it easier to reliably access a part of the debaters' knowledge, that seems pretty useful. And if it boosts the capabilities of our language models, that would also be great.

More generally, we already knew that there are some problems on which Debate fails - such as cryptographic functions whose solutions are very hard to find. So what we're trying to figure out is to what extent the most interesting problems fall into that category. But insofar as we're worried about agents taking harmful actions, we're worried about Debate arguments with verifiable consequences, and so I expect that leveraging empirical evidence (as discussed previously) will be a big advantage.

Unnatural Categories Are Optimized for Deception

Followup to: [Where to Draw the Boundaries?](#)

There is an important difference between having a utility function defined over a statistical model's performance against specific real-world data (even if another mind with different values would be interested in different data), and having a utility function defined over features of the model itself.

Arbitrariness in the map doesn't correspond to arbitrariness in the territory. Whatever criterion your brain is using to decide which word you want, is your non-arbitrary reason ...

So the one comes back to you and says:

That seems wrong—why wouldn't I care about the utility of having a particular model? I agree that categories derive much of their usefulness from "carving reality at the joints"—that's one very important kind of consequence of choosing to draw category boundaries in a particular way. But other consequences might matter too, if we have some *moral* reason to *value* drawing our categories a particular way. I don't see why I shouldn't be willing to trade off one unit of categorizational nonawkwardness for X units of morality, even if trading off a million units of categorizational nonawkwardness for the same X units of morality would be bad.

I once read about [an analogy between category boundaries and national borders](#). Imagine a diplomat trying to come up with a proposal for a [two-state solution to the Israeli-Palestinian conflict](#). There's no such thing as the "correct" border between Israel and Palestine, but there are consequences of choosing one border or another. For example, awarding territory to one side risks angering the other. For another, if the West Bank and Gaza Strip are to be part of Palestine, but Tel-Aviv and the southern city of Eilat are to be part of Israel, then topology forces you to decide which of Israel and Palestine gets to be continuous, and which will be split into two parts, because a "land bridge" between Gaza and the West Bank would separate Tel Aviv and Eilat, and *vice versa*. Since borders can't be "true" or "false", the diplomat's task is and *can only be* to weigh these kinds of trade-offs.

Analogously, I think of language, following Eliezer Yudkowsky's ["A Human's Guide to Words"](#), as being a human-made project intended to help people understand each other. It draws on the structure of reality, but has many free variables, so that the structure of reality doesn't constrain it completely. This forces us to make decisions, and since these are not about factual states of the world—what the definition of a word *really* is, in God's dictionary—we have nothing to make those decisions on except consequences.

... okay, I think I see the problem. I see how one might have gotten that out of "A Human's Guide to Words"—if you skipped all the parts with math. I am now prepared to explain exactly what's wrong here [in more detail](#) than [my previous attempt](#): not just that this position is not in harmony with the [hidden Bayesian structure](#) of language and cognition, but how the hidden Bayesian structure of language and cognition explains why an intelligent system might find this particular mistake *tempting* in the first place, and what breaks as a result.

Category "boundaries" are a useful *visual metaphor* for helping explain the cognitive function of categorization. If you have the visualization but you *don't* have the math, you might think you have the freedom to "redraw" the category "boundaries". Simple, compact boundaries might *tend* to be more useful, but more complicated boundaries aren't *false* and therefore aren't forbidden if you have some non-epistemic reason to prefer them ... right?

Only in the sense that *no* hypothesis is "false"! Categories, words, correspond to hypotheses—probabilistic models that [make predictions](#). If I see a dolphin in the water, and I say, "Hey, there's a dolphin!", and you *understand* me, that enables you to predict quite a lot about there being [this-and-such kind of aquatic mammal with fins, a tail, &c.](#) in the water.

This AI capability of "speech" is not only very powerful; it's also easy to understand the [cause-and-effect evidential entanglement which explains how it works](#)—at least at a very high level.

Photons bounce off the dolphin and hit my eyes. I recognize the photons as forming an image that matches a concept that I associate with the word/symbol "dolphin" (implementation details omitted). I emit a "dolphin" signal composed of sound waves which hit your eardrum. By [a convention that culturally evolved due to our predecessors having a shared interest in communicating with each other](#), you map the "dolphin" signal to an internal concept that closely resembles the one I associate with that same signal. This works because we happen to live in a world where the distribution of creatures has [cluster-structure](#) whereby dolphins have lots of things in common with each other, such that it's possible to use observations about an entity to infer that it "is a dolphin", and then use the *dolphin* concept to make good predictions about aspects of the entity that have not yet been observed; we owe our confidence that we've learned "the same" *dolphin* model to the fact that dolphins actually exist.

But the *dolphin* concept/model/hypothesis is subject to the universal [mathematical laws](#) of reasoning under uncertainty. In particular, probability-mass flows *between* hypotheses: as long as you [never assign a probability of zero](#) (which is a log-odds of negative infinity), nothing you believe can ever be [definitively \(infinitely\) falsified](#)—it "just" makes quantitatively worse predictions as *compared to* other hypotheses.

Because category "boundaries" are merely a visualization for a probabilistic model that makes predictions about the real world, you *can't* "redraw the boundaries" associated with a communication signal without messing with the model that generates them, which means messing with your predictions about the real world.

Might there be some non-epistemic reason for an agent to prefer a model that makes worse predictions? Sure! Correct maps are useful for [steering reality into configurations ranked higher in your preference ordering](#)—but causing a *different* agent to have *incorrect* maps might make them *mis-navigate* reality in a way that benefits you! We call this [deception](#).

In a related phenomenon, a poorly-designed agent might get confused and end up manipulating its *own* beliefs: optimizing its map to *inaccurately* portray a high-value territory (rather than optimizing the territory to be high-value by using a map that reflects the territory), a kind of *self-deception*. We call this [wireheading](#).

The laws of probability and information theory allow us to calculate how information can be efficiently encoded and transmitted from one place to another. Given some distribution of random variables, and some specification of what information about those variables you want to transmit, some encodings—some ways of "drawing" category "boundaries"—quantitatively *perform better* than others. Agents that *want to communicate with each other* will tend to invent or discover conventions that efficiently encode the information they're trying to communicate. Agents that communicate in ways that systematically depart from efficient encodings are [better modeled as](#) trying to deceive each other or wirehead themselves.

Let's walk through a simple example. [Imagine that you have a peculiar job in a peculiar factory](#): specifically, you're a machine-learning engineer tasked with automating away the jobs of humans who sort objects from a mysterious conveyor belt.

Another engineer has already written a system that processes camera and sensor data about the objects into more convenient "[features](#)": color (measured on an eight-point blueness scale), shape (measured on an eight-point "eggness" scale), and vanadium content (a boolean Yes or No). Your task is to further process this information into a format suitable for giving commands to other systems—for example, the robot arm that will physically move the objects into appropriate bins.

The feature data consists of the blueness-eggness-vanadium-content joint distribution given by this 128-entry table:

	eggness							
	0	1	2	3	4	5	6	7
blueness = 0, vanadium = 0	$\frac{49}{1600}$	$\frac{97}{1600}$	$\frac{49}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 0, vanadium = 1	0	0	0	0	0	0	0	0
blueness = 1, vanadium = 0	$\frac{97}{1600}$	$\frac{193}{1600}$	$\frac{97}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 1, vanadium = 1	0	0	0	0	0	0	0	0
blueness = 2, vanadium = 0	$\frac{49}{1600}$	$\frac{97}{1600}$	$\frac{49}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 2, vanadium = 1	0	0	0	0	0	0	0	0
blueness = 3, vanadium = 0	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 3, vanadium = 1	0	0	0	0	0	0	0	0
blueness = 4, vanadium = 0	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 4, vanadium = 1	0	0	0	0	0	0	0	0
blueness = 5, vanadium = 0	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 5, vanadium = 1	0	0	0	0	0	$\frac{3}{100}$	$\frac{3}{50}$	$\frac{3}{100}$
blueness = 6, vanadium = 0	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 6, vanadium = 1	0	0	0	0	0	$\frac{3}{50}$	$\frac{3}{25}$	$\frac{3}{50}$
blueness = 7, vanadium = 0	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$	$\frac{1}{1600}$
blueness = 7, vanadium = 1	0	0	0	0	0	$\frac{3}{100}$	$\frac{3}{50}$	$\frac{3}{100}$

This seems like ... not the most useful representation? The data is all there, so *in principle*, you could code whatever you needed to do based off the full table, but it seems like it would be an unmaintainable mess: you'd sooner *resign* than write a 128-case [switch statement](#). Furthermore, when the system is deployed, you hope to typically be able to give the binning robot messages based on *only* the color and shape observations, because the Sorting Scanner that the vanadium readings come from is expensive to run. You *could* just do a Bayesian update on the entire joint distribution, of course, but it seems like it should be possible to be more efficient by exploiting regularities in the data, not entirely unlike how your colleague's system has *already* made your job much simpler by giving you blueness and eggness feature scores rather than raw camera data. Eyeballing the table, you notice it seems to have a lot of redundancy: most of the probability-mass is concentrated in two regions where the blueness and eggness scores are either both high or both low—and vanadium is *only* found when both blueness and eggness are high.

O tragedy O the stars! *If only* there were *some more convenient and flexible way* to represent this knowledge—some kind of deep structural insight to rescue you from this cruel predicament!

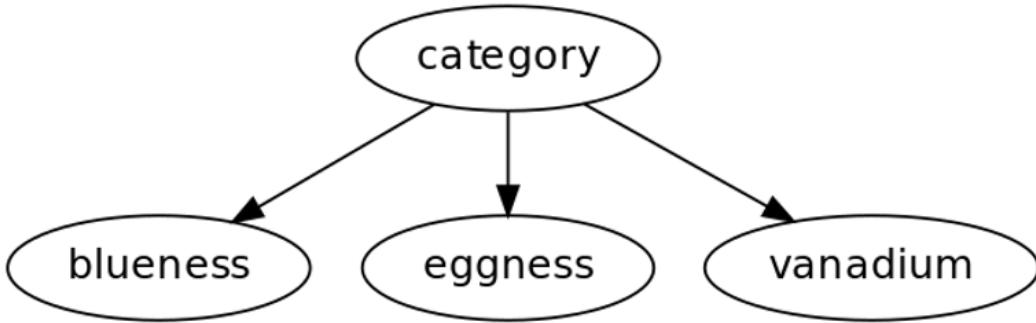
... alright, dear reader—I shouldn't patronize. [You already know how this story ends.](#) The distribution factorizes!

$$\sum_{\text{category}} P(\text{category}) \cdot P(\text{blueness}|\text{category}) \cdot P(\text{eggness}|\text{category}) \cdot P(\text{vanadium}|\text{category})$$

(The distribution in this made-up toy example factorizes *exactly*, but in a messy real-world application, you might have a spectrum of approximate models to choose from.)

We can simplify our representation of our observations by using a [naïve Bayes model](#), a "star-shaped" [Bayesian network](#) where a central "category" node is posited to underlie all of our observations: we believe that each object either "is a blegg" (and therefore contains vanadium and has high blueness and eggness scores) with probability 0.48, "is a rube" (and therefore has no vanadium and low blueness and eggness scores) with probability 0.48, or belongs to a catch-all "other"/error class with probability 0.04. (Maybe the camera is buggy sometimes, or maybe there are some other random objects mixed in with the rubes and bleggs?)

category	$P(\text{category})$
blegg	12/25
rube	12/25
??	1/25



category	$P(\text{color or shape score} \text{category})$							
	0	1	2	3	4	5	6	7
blegg	0	0	0	0	0	1/4	1/2	1/4
rube	1/4	1/2	1/4	0	0	0	0	0
??	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

$P(\text{vanadium} \text{category})$	
blegg	1
rube	0
??	0

The full joint distribution had 127 degrees of freedom (a table of $8 \cdot 8 \cdot 2 = 128$ separate probabilities, constrained to add up to 1), whereas the naïve-Bayes representation only needs 57 parameters ($3 \cdot 1$ prior probabilities for the categories, plus $3 \cdot 8 = 24$, $3 \cdot 8 = 24$, and $3 \cdot 2 = 6$ -entry *conditional* probability tables for each of the features). The advantage would be much larger for more complicated problems: the joint distribution table grows exponentially with more features, quickly becoming infeasible to *store and represent*, let alone *learn*.

It must be stressed that our "categories" here are a *specific mathematical model* that makes *specific* (probabilistic) predictions. Suppose we see a black-and-white photo of an egg-shaped object: specifically, one with an eggness score of 7. Given that observation of eggness = 7, we can update our probabilities of category-membership.

$$P(\text{category} = c|\text{eggness} = 7) = \frac{P(\text{eggness} = 7|\text{category} = c)P(\text{category} = c)}{\sum_{d \in \{\text{blegg}, \text{rube}, ??\}} P(\text{eggness} = 7|\text{category} = d)P(\text{category} = d)}$$

We think the egg-shaped object is almost certainly a blegg (specifically, with probability 0.96), even if the black-and-white photo doesn't directly tell us how blue it is, *because*

$$P(\text{category} = \text{blegg}|\text{eggness} = 7) = \frac{\frac{1}{4} \cdot \frac{24}{25}}{\frac{1}{4} \cdot \frac{24}{25} + 0 \cdot \frac{24}{25} + \frac{1}{25}} = \frac{24}{25} = 0.96$$

We can then use our updated beliefs about category membership (0.96 blegg/0 rube/0.04 unknown, as contrasted to the 0.48/0.48/0.04 prior) to get our updated posterior distribution on the 0-7 blueness score (0.005/0.005/0.005/0.005/0.245/0.485/0.245—left as an exercise for the reader).

In addition to categories facilitating efficient probabilistic inference *within* the system that you're currently programming, *labels* for categories turn out to be useful for *communicating* with other systems. The robot arm in the Sorting room puts bleggs in a blegg bin, which gets taken to a room elsewhere in the factory where there's sophisticated vanadium-ore-processing machinery that has to handle both bleggs and gretrahedrons.

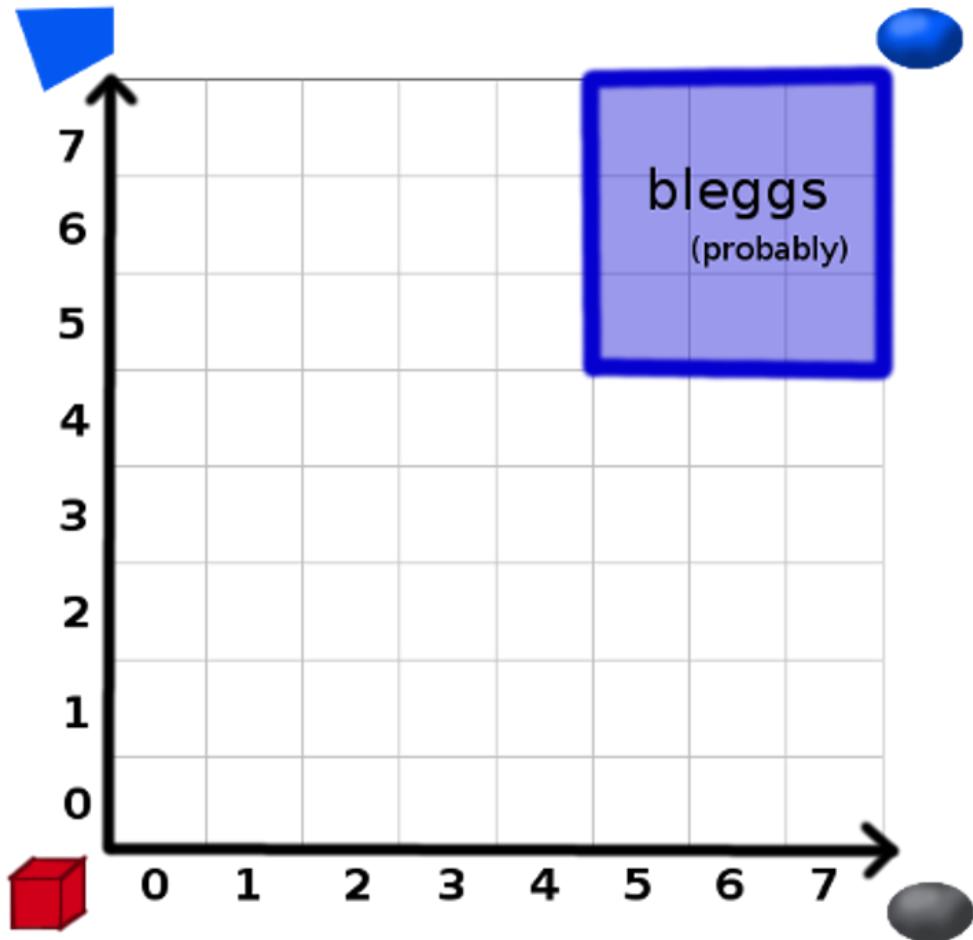
But suppose the binning arm doesn't need to *know* about the blueness and eggness scores: it can close its claws around rubes and bleggs alike, and you only need to program it to pick up an object from a certain spot on the conveyor belt and place it into the correct bin. However, the vanadium-ore-processing machine does need to do further information processing before it can operate on an object—perhaps it needs to vary its drill speed in proportion to the density of a particular blegg's flexible outer material (which it can estimate based on how brightly the blegg glows in the dark), but it uses a different drilling pattern for gretrahedrons.

If you need to send commands to both the binning arm and the ore-processing machine, it's a more efficient communication protocol to just be able to send the 28-byte [JSON](#) payload `{"object_category": "BLEGG"}` and let the other machines do their work using their *own* models of bleggs, rather than having to send over the raw camera data plus the binary code of the Bayesian network and feature extractors that you initially used to *identify* bleggs. [Intelligence is prediction is compression](#): our ability to find an encoding that [compresses the length of the message](#) needed to convey information about the objects is [fundamental to our having learned something](#) about the distribution of objects.

The `{"object_category": "BLEGG"}` message is a useful shorthand for "linking up" the models between different machines. Different machines might not use the *same* model: the classifier system uses blueness and eggness scores to *identify* bleggs, but the ore-processing machine, having been *told* that an object is a blegg, can take its approximate blueness and eggness for granted and only needs to reason about its luminescence and vanadium content.

But this trick of using a signal to correlate the models between different machines only works *because* and *insofar as* both models are pointing to the same cluster-structure in reality. If the model in the classifier system doesn't meaningfully *match* the model in the ore-processing system—if the classifier code sends the `{"object_category": "BLEGG"}` message given a object with blueness score between 5 and 7, but the ore-processor, upon receiving the `{"object_category": "BLEGG"}` message, positions its drills in the expectation of processing an object with an eggness score between 0 and 2—then the factory doesn't work.

As a human learning math, it's helpful to examine [multiple representations of the same mathematical object](#). We've already seen our blueness-eggness-vanadium model represented as a table, and factorized into a graphical model. We've done also some algebraic calculations with it. But we can also visualize it: the set of camera observations that the model classifies as a blegg with probability ≥ 0.96 can be thought of a area with a boundary in two-dimensional blueness-eggness space:



("With probability ≥ 0.96 " because our catch-all "other"/error category can also generate examples with high blueness and eggness scores; we can't say things like "Everything inside the boundary in the diagram is a blegg" when we're talking about a formal model where some of the categories generate overlapping observations in whatever subspace the diagram is depicting.)

If you were trying to teach someone about the hidden Bayesian structure of language and cognition, but thought your audience was too stupid or lazy to understand the actual math, you might be tempted to skip the part about factorizing a joint distribution into a star-shaped Bayesian network and just talk about "drawing" "boundaries" in configuration space for human convenience, perhaps with a hokey metaphor about national borders. Then the audience might walk away with the idea that there's no reason not to replace the old *blegg* concept and its boring compact boundary, with a new *blegg** concept that has an exciting squiggly border.

Alaska [isn't even contiguous with](#) the rest of the United States. If that's okay, why can't the borders of bleggness be a little squiggly?



Because the "national borders" metaphor is [just a metaphor](#). It *immediately* breaks down as soon as you try to do any calculations.

When we say that [the United States purchased Alaska from the Russian Empire](#), that means that this-and-such physical area on the Earth's surface went from being the territory of the Russian government, to being territory of the United States government, where land being the "territory of" a "government" is a complicated idea that has something to do [Schelling points over who gives orders to policemen and soldiers in that area](#).

When you reprogram your machine-learning system to send an {"object_category": "BLEGG"} message when it sees an object with an eggness score of 2 and a blueness score of 1, then your vanadium-ore-processing machine wears down its drill bits trying to process a rube.

Other than the fact that *some aspects* of both of these situations can be usefully *visualized* as changes to a two-dimensional diagram depicting an area with a boundary, what do these situations have to do with each other? They don't. Countries aren't Bayesian networks. They just aren't. When we depict a country on a map, we're *not talking* about a cognitive system that can use observations of latitude to estimate probabilities of country-membership and then use that distribution on country-membership to get an updated probability distribution on longitude. (I mean, given a world map, you *could* program such a thing, but it seems kind of useless—it's not clear why anyone would *want* that particular program.) Why would you expect to understand an AI-theory concept by telling a story about national borders?

So, that's what's wrong with the national-borders metaphor. But we haven't yet really explained the problem with "unnatural" categories—those that you would *visualize* as a squiggly, "gerrymandered" boundary. The squiggly *blegg** boundary doesn't have the nice property of corresponding to the category labels in our nice factorized naïve Bayes model, but it still contains information. You can still do a Bayesian update on being told that an object lies within a squiggly boundary in configuration space. If that update eliminates half of your probability-mass, that's one information-theoretic bit, no matter how the category is shaped in Thingspace.

If you only care about how much probability you assign to the *exact* answer, then a bit is a bit. But if an approximate answer is approximately as good—if your answerspace has a metric on it, so that "approximate" can mean something—then some bits can be more valuable than others.

Suppose some random variable X is uniformly distributed on the set $\{1, 2, 3, 4, 5, 6, 7, 8\}$. You have the option of being told either whether an observation x sampled from X is even or odd, or whether x is greater or less than 4.5. Either way, you eliminate half of your hypotheses: the [entropy](#) of your probability distribution goes from $\log_2 8 = 3$ to $\log_2 4 = 2$. Either way, you've learned 1 bit.

Still, if you have to make a decision that depends on "how big" x is, it seems like the "1-4 or 5-8" category system is going to be more useful than the "even/odd" category system, even though they both provide the same amount of information about the *exact* answer. If you learn that $x \in \{1, 2, 3, 4\}$, then you know that x is "small", but if you learn that x is odd, you haven't learned much about how big it is: it could be 1, but it could just as well be 7.

To formalize this, let's measure how "good" a category is using the [expected squared error](#). "Error" is how much a prediction is wrong by: if you guessed x was 2, but it was actually 5, your error would be $5 - 2 = 3$, and your squared error would be the square of that, $3^2 = 9$. The expected squared error of a probability distribution is, on average, the square of how much your guess about a sample from that distribution will be wrong. ([The squared error has nicer mathematical properties than the absolute error.](#))

For our example of x sampled from X uniformly distributed on $\{1, 2, 3, 4, 5, 6, 7, 8\}$, your best-guess estimate \hat{x} of x is going to be the expected value

$$\sum_{x \in \{1 \dots 8\}} P(X = x) \cdot x = \cancel{1+2+3+4+5+6+7+8} / 8 = 4.5$$

And the initial expected squared error is

$$E[(x - \hat{x})^2] = \sum_{x \in \{1 \dots 8\}} P(X = x) \cdot (x - \hat{x})^2 = \cancel{(1-4.5)^2 + (2-4.5)^2 + \dots + (8-4.5)^2} / 8 = 5.25$$

Suppose you then learn whether x is even or odd.

With probability 0.5, you learn that x is even. In that case, your new estimate \hat{x} taking that into account would be

$$\sum_{x \in \{2, 4, 6, 8\}} P(X = x) \cdot x = \cancel{2+4+6+8} / 4 = 5$$

and your new expected squared error (in the "even" possible world) would be

$$E[(x - \hat{x})^2] = \sum_{x \in \{2, 4, 6, 8\}} P(X = x) \cdot (x - \hat{x})^2 = \frac{(2 - 5)^2}{4} + \frac{(4 - 5)^2}{4} + \frac{(6 - 5)^2}{4} + \frac{(8 - 5)^2}{4} = \frac{9 + 1 + 1 + 9}{4} = 5$$

With probability 0.5, you learn that x is odd. Similar calculations (left as an exercise) also give a new expected squared error of 5 in the "odd" possible world. Averaging over both cases (trivially, $0.5 \cdot 5 + 0.5 \cdot 5 = 5$), learning whether x is even or odd only brought our expected squared error down from 5.25 to 5, barely changing at all.

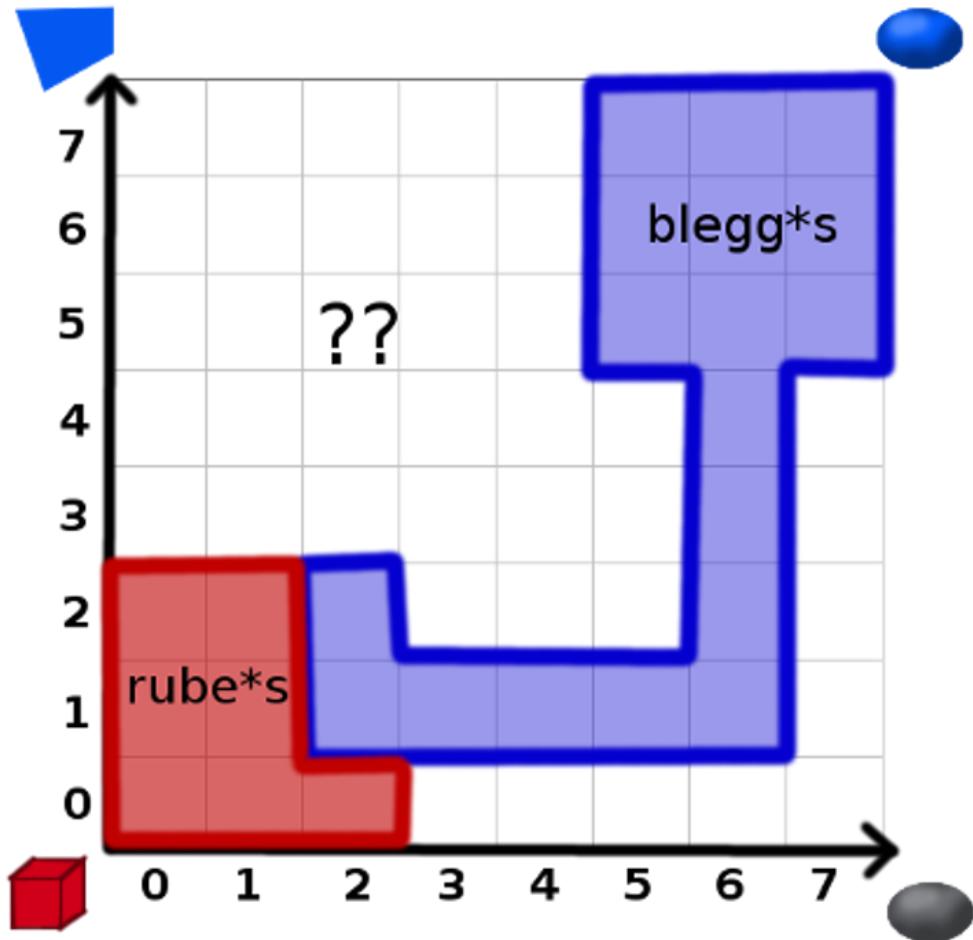
In contrast, if you learn whether x is 1–4 or 5–8, your expected squared error plummets to 1.25. (Exercise.) By being compact, the "1–4 or 5–8" category system is much more useful for getting *close* to the right answer than the "even/odd" category system.

The same goes for natural categories *versus* squiggly category "boundaries" in configuration space; we just need to supply some [metric](#) to define what "close" means.

For our blueness-eggness-vanadium distribution, suppose we use the [Euclidean distance](#) on blueness-score \times eggness-score \times 1-if-vanadium-present-else-0. (So, for example, the "distance" between the typical blegg and the typical rube is $\sqrt{(6 - 1)^2 + (6 - 1)^2 + (1 - 0)^2} = \sqrt{25 + 25 + 1} = \sqrt{51} \approx 7.14$ under this metric.)

Then our expected squared error before being told anything about an object is about 13.63. On being told whether an object is a blegg, rube, or other (according to the categories in our nice factorized naïve Bayes model), our expected squared error plummets to 1.38.

But suppose that, instead of our nice factorized naïve Bayes model, we use a category system based on drawing squiggly "boundaries" in configuration space: everything inside the blegg* boundary in the diagram is a blegg*, everything within the rube* boundary in a rube*, and anything outside belongs to a catch-all "other*" category.



On learning whether an object is a blegg*, rube*, or other*, our expected squared error only goes down to about 4.12.[\[1\]](#)

In this sense, the gerrymandered blegg* concept is *quantitatively less informative* than the original, compact blegg concept. The *metric* we assigned to blueness-eggness-vanadium space was our choice, and could depend on our values: for example, if we simply *don't care* about predicting how blue an object is, we could disregard the blueness score and only define a concept on the eggness-vanadium subspace (in which case our initial expected squared error is about 6.94, plummets to 0.69 given knowledge of blegg/rube/other category-membership, but only goes down to about 1.81 given knowledge of the gerrymandered blegg*/rube*/other* category). Or if we *don't care* about predicting blueness *very much*, we could calculate our error score with respect to a metric that gave blueness very little weight. (Exercise.)

But *given* a metric on the variables that you care about predicting and using to inform predictions, which categories are cognitively useful depends on the the distribution of data in the world. You can't define a word any way you want.

The dependence on a choice of metric on configuration space—and really, a choice of the space—gives a *sense* in which optimal categories are [value-laden](#), but it's a specific kind of *lawful* dependence between your values and the distribution of data in the world, *not* an atomic preference for using a particular encoding for its own sake.

The cognitive function of categorization is to group similar things together so that we can make similar decisions about them. A function measuring the extent to which things are "similar" has to take the things as input, but the extent to which things are *decision-relevantly* similar also depends on what you're trying to accomplish with your decisions, and that can be algorithmically complex. It might not be just a matter of only looking at some decision-relevant subspace of a natural, "obvious" configuration space that's available to [all possible minds](#) (like not caring what color your toothbrush handle is—um, if we pretend that all possible minds had human-like color vision); the dimensions of the space you do your [similarity-clustering](#) in might themselves be complicated features (in the sense of machine learning) of which agents with different values would have no reason to [logically pinpoint](#) that particular criterion [by which things may be judged](#). How you should define words *depends on* what you want, but that's *not* the same as defining words any way you want.

For example, [poison isn't a natural category to a generic mind studying chemistry](#): we group cyanide and hemlock together as *poison* because we value human health, and so we want to have a category for scary chemicals that disrupt human metabolism, causing death or serious illness. But this determination depends on the intricate details of human biochemistry. (The [theobromine in chocolate](#) is okay for humans at typical doses, but potentially fatal to dogs, which are actually pretty close to us in animalspace.) The compact category "boundary" that minimizes predictive error on human-healthspace, corresponds to a squiggly "boundary" in the chemicalspace you would be looking at if you've never seen a human and just want to make predictions about the chemicals themselves.

Or [tiny molecular smileyfaces and real human smiles might be grouped together](#) as similar as far as an image-classifier's [curve detector](#) is concerned, even if they're not similar as far as the [abstracted idealized dynamic](#) of human morality is concerned.

The technical sense in which optimal categories can be value-laden doesn't alter the basic morals of our basic Bayesian philosophy of language. Your values can give you a particular configuration space and a metric on the space, but *given* that, sane agents want to "carve it at the joints" in order to get a communication system that minimizes predictive error. If you're trying to find an efficient encoding of your observations, there's no reason to *want* squiggly, gerrymandered categories in the decision-relevant space.

The one replies:

You're still not addressing my crux! I don't doubt what you say about minimizing prediction error with respect to some squared metric thingy. But what if that's not what I care about? My utility function assigns high value to using the squiggly *blegg** category boundary—such that the utility of using my preferred category outweighs the disutility of making less accurate predictions. You can define a word any way you want—if you're willing to pay the costs.

So, what, you just intrinsically assign high utility to using the same communication signal to encode eggness-2/blueness-1 observations as eggness-6/blueness-6 observations, given the joint distribution specified in my story problem about sorting objects in a factory? Really?

"... yes!"

Okay, but where would that kind of exotic utility function come from? How would it arise naturally in an intelligent system?

There's a *trivial* sense in which you can interpret any action taken by an agent as being taken because the agent *values taking that action*. This theory [is compatible with all possible behaviors and therefore explains nothing](#).

The value of [decision-theoretic utility functions](#) isn't that "Because utility!" serves as [an all-purpose excuse for any possible behavior](#). It's that [simple coherence desiderata imply that an agent's behavior should be describable as maximizing expected utility for some utility function](#)—with corresponding constraints on the shape of that behavior.

Situations like [the Allais paradox](#) illustrate what these constraints look like. Consider an AI faced with playing the following game. There's a switch that can be turned On or Off, that starts out on in the Off position. At midnight, a coin is flipped. If the coin comes up Tails, the game ends. If the coin comes up Heads, then at a quarter past midnight, if the switch is Off, then the AI gets paid \$100, and if the switch is On, a six-sided die is rolled, and the AI gets paid \$110 if the die doesn't come up 6.

Suppose that, before midnight, the AI is willing to pay a dollar to flip the switch On (as if it thought that winning \$110 with a probability of 5/12 is better than winning \$100 with a probability of 1/2). Suppose the coin comes up Heads, and the AI is then willing to pay another dollar to flip the switch Off again (as if it thought that \$100 with certainty is better than \$110 with probability 5/6). Then the AI is two dollars poorer in exchange for the switch being in the same position it started in.

These gambling preferences violate [the independence axiom](#) of the [von Neumann–Morgenstern utility theorem](#). You *can't* have a utility function U for which

$$\frac{1}{2} \cdot U(\$100) < \frac{5}{12} \cdot U(\$110)$$

and

$$U(\$100) > \frac{5}{6} \cdot U(\$110)$$

because the sides of the second inequality are just those of the first multiplied by two, and multiplying by two should preserve the direction of inequality.

Having shown this, can we say that an AI with such behavior is "irrational"? But what does that even mean? If, for some reason, you specifically programmed the AI to prefer options it considers "certain", or to want switches to be "On" before midnight but "Off" after midnight, then it would be functioning as designed.

What we *can* say about such an AI, is that it doesn't have a utility function [in terms of money](#), and is therefore not coherently optimizing for acquiring money. Recall that we say that a system is an optimizer if it systematically [steers the future](#) into configurations that rank higher [with respect to some preference ordering](#). This helps us make predictions about what effects the system has, without having to model the details of *how* it brings those effects about. A well-designed agent that was optimizing for acquiring money would be expected to obey the independence axiom.

If the AI playing this game isn't coherently optimizing for acquiring money, what *is* it optimizing for? To tell, we'd need to observe its behavior in different environments and see how it [responds to perturbations](#). If it is trying to acquire money but is just *biased* to prefer certainty (in violation of the von Neumann–Morgenstern axioms), then we'd expect it to make choices that result in money but continue to exhibit Allais-like glitches around gambles involving probabilities close to 1. If it just likes switches to be off after midnight, then we'd expect it to turn switches off at that time even if there's no gambling game going on.

This methodology for attributing goals to an agent—consider it to be "optimizing for" outcomes that it [systematically achieves across a variety of environments](#)—applies to the behavior of sending communication signals, just as it does to the behavior of flipping switches.

Back to the factory. Our classifier system sends a `{"object_category": "BLEGG"}` message when it gets feature data corresponding to the compact *blegg* concept. This behavior is optimized for sending messages that allow other systems to minimize the expected squared error of their predictions of objects with respect to our standard metric on blueness-eggness-vanadium space. We *don't* intrinsically "assign utility" to using that particular category system; the category is the *solution* to an optimization problem about how to efficiently get blueness-eggness-vanadium information from one place to another.

A system that sends a `{"object_category": "BLEGG"}` message when it gets camera data corresponding to the gerrymandered *blegg** concept would be optimized for ... what? If you don't intrinsically assign utility to using that particular category system, then *why* would you program the system that way? What could possibly be the problem for which the gerrymandered category is an optimized solution?

Well. Suppose that, besides your dayjob as a machine-learning engineer, you *also* happen to own a side interest in the firm that supplies bleggs and rubes to this very factory. And suppose that vanadium fetches higher market prices than palladium, such that the factory is to pay the supplier \$2 per blegg but only \$1 per rube—and that the accounts-payable records are to be compiled based on how much the classifier you're currently programming sends `{"object_category": "BLEGG"}` and `{"object_category": "RUBE"}` messages, *not* how much metal actually gets harvested.

You can't help but notice that you stand to make more money if the system you're programming sends BLEGG messages more often. You can't just make it send BLEGG messages all the time—someone would notice and you'd get fired. But the ore-processing room can cope with a few suboptimally-sorted objects. Surely it's no big deal if you just ... adjusted the category boundary of BLEGG-ness a bit?

We saw earlier that the *blegg* concept does better than the *blegg** concept with respect to mean squared error (given a metric on the feature space).

That's not the only possible scoring function with which one could formalize how "good" a category system is. Suppose that instead we score our category system by which one best minimizes the expected squared error *minus* supplier revenue in cents. With respect to this criterion, accurate predictions are still good, but supplier revenue is *also* good.

Learning whether an object is a blegg, rube, or other (according to the "natural" categories in our naïve Bayes model) yields a squared-error-minus-revenue score of about -142.62 . (Don't ask me what the units are on this.) But learning whether an object is a *blegg**, *rube**, or *other** yields a squared-error-minus-revenue of -151.57 , which is lower (which is better, because we formulated this as a minimization problem). So with respect to *that* scoring function, the *blegg** category "boundary" is preferable.

The one says:

But now it sounds like you're agreeing with me! The compact *blegg* category serves the factory owner's goals better, which you formalized in terms of minimizing average squared error. The squiggly *blegg** boundary makes the factory perform less well, but it serves the moonlighting engineer's goals better, which you formalized in terms of minimizing squared error minus supplier revenue. There's no rule of rationality against the engineer programming the system using the *blegg** category boundary if it suits their goals better.

Only in the sense that there's no rule of rationality against *lying*! Suppose I'm selling you some number of gold and silver bars, but you can't examine the metal yourself until later; you can only hope that the receipt I give you is accurate. Consider the following two scenarios.

In the first scenario, I *lie*: the receipt says I delivered 60 gold bars and 20 silver bars, but I actually delivered 40 gold bars and 40 silver bars. You live in a low-trust world where lying is very common and contract enforcement isn't really a thing: a third of the time an object is claimed to be gold, it turns out to be silver. So when you discover the fraud, you feel disappointed but not surprised: you would have *preferred* to get what you paid for, but you can't say you *anticipated* it.

In the second scenario, I tell the truth—with respect to a category system that suits my goals. The receipt says I delivered 60 gold bars and 20 silver bars—and I did. It's just that what I prefer to call "gold bars", you prefer to call "gold bars, or silver bars with odd serial numbers", and what I call "silver bars", you call "silver bars with even serial numbers". You know this, so when you examine the actual contents of the delivery, you feel disappointed but not surprised: you would have *preferred* to transact under your definitions of 'gold' and 'silver', but you can't say you *anticipated* it.

We might question whether these are two different scenarios, or two descriptions of the *same* scenario: the same physical receipt, the same physical metal, the *same buyer anticipations about the metal conditional on observing the receipt*. If [we just pay attention to the evidential entanglements](#) instead of being [confused by words](#), then [there's no functional difference between](#) saying "I reserve the right to lie $p\%$ of the time about whether something belongs to category C ", and adopting a new, less-accurate category system that misclassifies $p\%$ of instances with respect to the old system.

Minimizing the squared-error score is *about* map-territory correspondence: ways of communicating that help the factory machines make better predictions about the objects, get a higher score.

Minimizing the squared-error-minus-supplier-revenue score is a *compromise* between map-territory correspondence and saying whatever makes the supplier the most money.

The *degree* of compromise is quantitative: there's a continuum of possible scoring functions between "minimize expected squared error, only" (for which the naïve-Bayes categorizer is a good solution), and "maximize supplier revenue, only" (for which "always say BLEGG" is the optimal solution). If always saying whatever profits you and not revealing *any* information about the territory is deception pure

and simple, then the intermediate points on a continuum with that can be thought of as partially deceptive.

Depending on your goals, deception can be rational! If you *don't care* about other agents having accurate models and just want to [intervene on them to make them believe](#) whatever makes them behave in a way that benefits you—or [whatever makes them happy](#)—then you can do that! There's [no God to stop you](#). But in order to help you *decide* whether deceiving people is the right thing to do, it helps to *notice* that what you're doing is deceiving people.

It helps to notice what you're doing—if you're trying to be an agent that coherently steers the future in some direction. But who does that, really? Maybe you just want to *feel good!* And not even coherently steer the universe into configurations where you feel good, either!

Rational agents should want to have true beliefs: the map that reflects the territory, is the map that is *useful* for navigating the territory. But you don't—can't—have unmediated access to the world; you can only *infer* what the world is like from sensory data, and effectively [live in your model of the world](#). Given the tricky indirection involved, it's not surprising that [poorly-designed](#) agents like humans sometimes get confused and "wirehead" themselves: if you don't notice the difference, it's tempting to fabricate a fake map that *falsely* portrays the territory as being good, instead of making a map that reflects the territory (which you can use to figure out how to improve the territory).

Similarly, if you don't notice the difference, it's tempting to choose language that makes the world *sound* good, than to have your language accurately describe the world (which description you can use to figure out how to make the world better).

Suppose I want people to think I'm funny. *Funny* is a value-laden concept in the specific lawful sense described earlier: non-human agents would have no motive to evaluate the particular [fixed computation](#) of humor. It's also a [fuzzy concept](#): we don't have a simple test to precisely measure in standard units exactly how funny a joke is, but there's *enough* regularity in how people use the word "funny" for the word to be a useful communication signal. It's *also* a [two-place concept](#): people have different senses of humor, so that what I consider funny isn't exactly the same as what you consider funny.

Given all these complications, one could imagine being tempted to think that humor is "subjective", and that therefore I can define it any way I want, and that therefore, if I feel sad about not being "funny", I can fix that by *changing my definition of the word "funny"* such that it includes my jokes. Because definitions can't be "false", right!? There's no rule of rationality prohibiting this boundary-redrawing project—and since I want so desperately to be "funny", there's every rule of human decency in favor of it, right?

So, this obviously doesn't work. (Okay, it "works" if you deliberately choose to define the word "work" such that it works, but it doesn't *actually* work.) [Yes requires the possibility of no](#): redefining X to make "Is it X?" come out true no matter what, [loses the purpose](#) of asking the question in the first place. The proposal to redefine the word "funny" came with the purported justification that words don't have intrinsic meanings, so it can't be "wrong" to redefine it. But precisely because words don't have intrinsic meanings, there's no reason to *want* to redefine an *existing* word, except to piggyback off the meaning people are *already* using that signal for.

(Note that this, in itself, isn't necessarily deceptive. Sometimes, [coining new senses of a word that piggyback off an existing meaning can be a powerful tool for extending our vocabulary to cover new phenomena that we don't already have words for](#)—as long as we're careful to [specify which meaning is intended](#) when it's not clear from context.)

It's not plausible to suppose that I want to be "funny" *because* I like five-letter words that start with the letter *f*; I want to be funny *because* of what that communication signal is already understood to refer to in [common usage](#). The redefinition might (or might not) succeed at making me feel better about myself, but if it does, it only works *by means of* confusing me: using [strategic equivocation](#) to arbitrage the hedonic gap between my new definition, and the old definition (which I still mentally associate with the word).

If it *does* succeed at making me feel better about myself, is the redefinition "rational"? Happiness is good, right? [Should not rationalists win?](#)

I do not frame an answer: that would depend on how you draw the category boundaries of "rational", which is [not an interesting question](#). (As it is written of a virtue which is nameless: if you speak overmuch of the Way, you will not attain it.)

What I *can* say, however, is that redefining the concept of humor is not a [procedure](#) that uses a map that reflects the territory to systematically achieve goals across a wide range of environments. If there's anything I can *do* to become funnier (like practicing telling jokes in a mirror, or studying great comedians to imitate their timing and delivery), I would seem less likely to notice and execute on such a plan after [having sabotaged](#) the concept I would need to notice the problem in the first place.

The map is not the territory ... but for [real agents embedded in the physical universe](#), the map is *part* of the territory. This presents some complications to applications of our anti-wireheading moral. We don't want to wirehead ourselves by making the map look good at the expense of undermining our ability to navigate the territory—but there's no bright-line distinction demarcating which configurations of atoms are "the map". [From the perspective of the eternal](#), it's *all* just territory.

In [the previous post](#), we considered the case of an assembly line (well, sorting line) worker in the blegg-rube factory being excited about an ostensible promotion to the position of Vice President of Sorting—only to be aggrieved on finding out that it's a promotion literally in name only, with no changes in pay, authority, or work tasks.

If we interpret the title as part of "the map", a communication signal with the function of encoding information about the person's job, then we want to say that the new title is *substantively misleading* ([even if it's not technically a "lie"](#)): when you hear that someone's job is being a "Vice President", you predict that their work involves managing people and making high-level executive decisions for the firm. Your probability that the "Vice President" has to spend all day moving objects from a conveyor belt into one of two bins based on the object's color and shape (a task that should probably be automated), is *lower* than before you heard the person's title: hearing the title made you update in the wrong direction.

But if we interpret the title as part of "the territory", a feature of the job itself, rather than a communication signal *about* the job—then it's not misleading and *can't* be misleading. The job happens to be one that has the symbols "Vice President" printed on the accompanying business cards and employee roster, much like how bleggs are objects that happen to be blue. You can't say the blue is "lying"; that doesn't make any sense!

The function of words is to serve as signals for communication, so it seems safe to say that language should usually be construed as part of "the map". Changing names and *only* names, without altering the things that the names *refer* to, as in the phony "Vice President" example, is probably deceptive. But for other features associated with a category, it may not always be obvious when we should construe them as "map" rather than "territory": using a feature to infer category-membership is formally equivalent to regarding it as a signal sent by senders of that category. Is that man *pretending to be a doctor*, or does he just happen to be wearing a lab coat?

The concept we're [groping towards](#), and hoping to formulate an elegant reduction of, is that of *mimicry*. Suppose there is some existing category of entity, an original, typified by some cluster of traits. A *mimic* is an entity optimized to approximately match the distribution of the original in many, but not all traits, thereby being part of the same cluster as the original in some *subspace* of the space the original category is defined in, but not the space as a whole. For example, if the vector

$[4, 4, 4, 4, 4] \in \mathbb{R}^5$ is the original, then an optimization process trying to construct a mimic of it in the subspace spanned by x_1 , x_4 , and x_5 might choose $[4, 0, 0, 4, 4]$: if you only look at the first, fourth, and fifth coordinates, then $[4, 4, 4, 4, 4]$ and $[4, 0, 0, 4, 4]$ "look the same"—they *are the same in that subspace*, but not the same if you include the second and third coordinates.

We can find examples in nature. Suppose one type of butterfly has evolved to be toxic to a type of predator, and also has distinctive wing markings that function as an [honest warning signal](#) to that predator: [this butterfly is not good to eat](#). This provides an ["opportunity" \(in evolutionary time\)](#), for a second species of butterfly to develop similar wing markings, so that predators will confuse it for the first type of butterfly, despite the second butterfly not paying the metabolic cost of producing toxins. This kind of situation is called [Batesian mimicry](#).

Is Batesian mimicry deceptive? (In our usual [functionalist](#) sense, which is obviously not a claim about butterfly *psychology*.) Is the second butterfly's very existence a kind of lie?

In some sense, yes! The mimic butterfly has been optimized by evolution to look like the first butterfly because of the fitness payoff of being categorized by the predator as the first, toxic, kind of butterfly. The "categorized by the predator as toxic" category is a natural, compact region in wing-marking-space, but "comes apart" into two clusters in the broader wing-markings-actual-toxicity space.

Furthermore, the evolutionary dynamics create [an asymmetric relationship between the two categories](#), that isn't captured by just the two trait-clusters themselves. The reason for the mimic butterfly to have those particular wing-markings is *in order to* manipulate the predator's predictions of toxicity (which was learned from encounters with the original), so if the original's wing-markings were to change as a result of some new selection pressure, the mimic would be subjected to selection pressure to "keep up" by changing its wing-markings accordingly.

That's not true in the other direction: if the mimic's markings were to change, the original wouldn't "follow": the original would instead benefit from the [probabilistic strength of its warning signal](#) not being parasitically diluted by the mimic anymore. Thus, the asymmetric terminology of "original" and "mimic" is appropriate: it's not just that these two species happen to look like *each other*; one of them was there *first*, and the other looks like *it*.

Is mimicry *always* deceptive? Not necessarily—there might be some situations where the *relevant* set of variables are among those where the mimic matches the distribution of the original.

Suppose you and I are feeding some ducks in the park. I say, "I love feeding these ducks!"

You say, "Wrong! These aren't all ducks. This park is where a local inventor tests out his [Anatid-oid](#) robots that are designed to look and act like ducks. Therefore, you can't say, 'I love feeding these ducks'; you need to say 'I love feeding these ducks and Anatidoid robots'."

"Wow, they're so realistic!" I say. "I can't even tell which ones are really robots! In fact," I continue, "since I can't tell, I'm inclined to just keep calling them all ducks; it would be pretty awkward to refer to each one as a duck-or-Anatidoid-robot."

"But it *is* possible to tell," you claim. "For example, if you get really close to one of the Anatidoid robots, and there's not a lot of ambient noise, you can hear the gears inside, turning."

"Okay," I say, "but I *can't* hear the gears from here. Since I have no way of telling the difference between ducks and Anatidoid robots without doing the more expensive evidence-gathering of cornering one in a quiet place, it makes sense for me to talk and think about the robots as being a kind of duck."

"But that's a *lie*! Ducks and Anatidoid robots may look and act similarly, but they're actually very different! Ducks are made of flesh and blood inside and are fated to die, whereas Anatidoid robots have a plastic interior and are immortal. And the ducks digest and gain nutrients from the scraps of bread we're feeding them, whereas the Anatidoid robots merely store the bread in an internal compartment that later gets dumped as they recharge wirelessly in the inventor's lab."

"Sure," I agree. "And if I were interacting with these entities in a context where I wanted to minimize the expected squared error of my predictions about their internal makeup, energy sources, or ultimate fate, then I would want to make that distinction. But I just want to watch some cool ducks in the park, and *in the context* of that activity, I only need to minimize the expected squared error of my predictions about appearance and behavior."

This is the origin of the famous [duck test](#): if it looks like a duck, and quacks like a duck, and you can model it as a duck without making any grievous prediction errors, then it makes sense to consider it a member of the category *duck* in the range of circumstances where your model continues to perform well.

The features for which mimics fail to match the original need not be hidden (like gear sounds that you can't hear in a noisy park) in order for mimics to not be deceptive; they only need to be [irrelevant](#) in the context the category is being used. [Squirt guns](#) aren't guns—and are usually manufactured in unrealistic colors specifically to prevent being confused with real guns—but in the context of a [water fight](#), the utterance "Don't point that gun at me" (without the [privative adjective](#) *squirt*) is understood perfectly well.

Nondeceptive mimicry is *fragile*, however: it works in contexts where all the relevant features are ones where the mimic matches the original. Mimics that don't match the distribution of the original along relevant features are deceptive in the sense that agents that observe the mimic and assign it to the same mental category as the original on the basis of the matching features, will use that categorization to make predictions about unobserved but nonmatching features, and be wrong. And they'll be wrong because the mimic is optimized to "look like" the original (to match on many observable features).

If different agents using a shared language disagree on what features are "relevant", they may have an incentive to fight about how [scarce and valuable short codewords](#) should be defined in their common language, in order to exert control over what inferences and decisions agents using that language can easily make and [coordinate on](#).

Let's consider how this might apply to a real-world issue. From moral perspectives that place a lot of value on the welfare of nonhuman animals, factory farming is an [ongoing moral catastrophe](#). Unfortunately (for the farmed animals), meat-eaters and the global agriculture industry they support aren't going to change their ways because of anyone's [desperate cry at the horror of suffering](#) or carefully-reasoned appeal to the global utilitarian calculus. Animal-rights advocates can sway behavior on the margin, but there's just too much biological and cultural inertia favoring the consumption of animal products for it to be feasible to *outlaw* factory farming the way [chattel slavery was outlawed](#). It's not that humans hate farm animals; they're just ... made out of tissue that we can use for other things.

An alternative strategy for ending factory farming is to prioritize the development of artificial substitutes that *mimic* real meat, eggs, dairy, &c. along the consumption-relevant dimensions of taste, texture, nutrition, &c., but are produced in a lab or factory rather than from the tissues of sentient creatures. In the limit of arbitrarily capable physical manufacturing technology, carnivores and factory-farming opponents alike could both be satisfied: if two steaks are *indistinguishable by any physical means whatsoever*, then a meat-eater has no reason to care which one came from an actual cow's flesh, and which one was molecularly assembled by nanobots. Perhaps a Society of hunter-gatherers that attached cultural significance and ritual to the labor of killing one's own meal would have a reason to object, but modern folk for whom food comes from the supermarket have no basis within their experience to say that the nanoassembled steak isn't "real".

Unfortunately, we do not have arbitrarily capable physical manufacturing technology. Although [progress continues](#), modern animal product substitutes are sufficiently unsuccessful mimics that they are usually not considered to belong to "the same" category as the original. [Veggie burgers](#) are not burgers in the sense that a customer who ordered "a burger" at a restaurant and was served a veggie burger would be likely to notice and complain—and in particular, would probably not be satisfied if the waiter were to reply, "Well, if you specifically wanted a burger *made from cow flesh*, you should have said that."

As technology to make plausible mimics/substitutes improves, however, different interest groups might face a temptation to fight over the meanings of words that was not present when the mimics weren't plausible enough for a dispute to arise. If you have the power of [setting the default extension](#) of a word that people are *already* using to communicate with, you can exert some amount of control over the decisions people make while trying to *think* using that word. Should the meaning change, then a restaurant customer who wants to make sure they receive a burger under the old definition now has to use more words, while those who don't have a strong preference or are too shy to complain will accept the restaurant's interpretation of the order.

Thus, if a fight breaks out about the meaning of the word *meat*, animal rights activists have a moral incentive to draw the category "boundaries" to include even substitutes that are very bad (on the empirical merits of successfully mimicking the original), whereas existing agricultural interests have a financial incentive to draw the "boundaries" to exclude even substitutes that are very good. (This kind of dispute [is not hypothetical](#), and isn't necessarily limited to just words: [in the late 19th century, dairy farmers pushed for laws that required margarine to be dyed pink](#) to prevent consumers from confusing it for butter—the law effectively interpreting color as a communication signal, rather than a property of the good itself.)

If a fight breaks out about the meaning of the word *meat*, rationalists may not all take the same side, but we can at least strive for objectivity in *describing the conflict*—and in particular, to *notice the difference* between definitions motivated by *describing reality*, and definitions motivated by the

positive or negative effects (such as profitably deceiving other agents) of choosing one description or another.

If some think that some meat substitute should be considered meat *because* the "taste" dimension is genuinely most relevant to the true meaning of *meat*, and some oddities in the texture don't matter, but others think *vice versa*, the philosophy articulated on this post has nothing to say to either side: the math of minimizing expected squared error by putting labels on clusters doesn't say *which* subspace to look for clusters in.

But if some think that some meat substitute should be considered meat *because* saving nonhuman animals from a life of torture is more important than conceptual parsimony ... I can't prove that that's not the right answer to the *decision problem* of what verbal behavior to perform. The stakes are genuinely high.

What I *can* say is that the hidden Bayesian structure of language and cognition makes no reference to the stakes, and departing from the structure [extracts a price](#) that [isn't up to us](#).

If, empirically, being generous about what counts as "meat" can prevent massive suffering (by altering the social defaults around consumption behavior), then maybe that's the right thing to do.

Similarly, if [telling the public that masks don't work for preventing respiratory disease can preserve supplies for medical professionals who need them more](#), then maybe that's the right thing to do.

And if you live in an absurd thought experiment where saying " $2 + 2 = 5$ " could save [3↑↑↑3](#) lives, maybe saying " $2 + 2 = 5$ " is the right thing to do. But the *empirical* question of whether you happen to live in that particular thought experiment, doesn't change the *laws* that govern what you have when you take $\bullet\bullet$ -many plus another $\bullet\bullet$ -many, no matter what symbols are used to communicate this fact, and no matter the consequences for communicating it.

For these reasons [it is written of the third virtue of lightness](#): you *cannot* make a true map of the category by drawing lines upon paper according to impulse; you must observe the joint distribution and draw lines on paper that correspond to what you see. If, seeing the category unclearly, you think that you can shift a boundary just a little to the right, just a little to the left, according to your caprice, this is just the same mistake.

And as it is written of a virtue which is nameless: perhaps your conception of rationality is that it is rational to believe the words of the Great Teacher, who [lives in an area where claiming that the sky is blue would be political suicide](#).

And the Great Teacher says, "Some people I usually respect for their willingness to publicly die on a hill of facts, now seem to be talking as if color references are necessarily a factual statement about frequencies of light. But using language in a way *you* dislike, is not lying. You're not standing in defense of Truth if you insist on a word, brought explicitly into question, being used with some particular meaning." And you look up at the sky and see blue.

If you think: "It may look like the sky is blue, such that I'd ordinarily think that someone who said 'The sky is green' was being deceptive, but surely the Great Teacher wouldn't egregiously mislead people about the philosophy of language when being egregiously misleading happens to be politically convenient," you lose a chance to discover your mistake.

How will you discover your mistake? Not by comparing your description to itself.

But by comparing it to that which you did not name.

(Thanks to Jessica Taylor, Abram Demski, and Tsvi Benson-Tilson for discussion and feedback.)

1. [The source code of the Python script used for these calculations is available.](#) ↵

Retrospective on Teaching Rationality Workshops

TL;DR: I organised a series of afternoon applied rationality workshops for the Cambridge Effective Altruism group based on some core CFAR classes. These went much better than I expected them to, and seem to have added long-term value to participants. My goal in this post is to share my resources and [lesson plans](#), my thoughts on teaching applied rationality well, my attempts to distill the concepts into key ideas & mental habits, and to convince other people to organise similar workshops!

(Disclaimer: This was inspired by CFAR material, but is very much all my interpretation and framings. Any problems with the material are likely from me, not CFAR)

Introduction

I went to a Centre For Applied Rationality (CFAR) workshop last year and found this a really valuable experience. I especially found that a few techniques stuck well with me and became valuable parts of how I thought. So, to help consolidate these and to share them with others, I organised some afternoon workshops for the Cambridge EA group based on my favourite classes. I think these were extremely successful and added more value than I expected to the participants. So this post is my attempt to write up a retrospective on the workshops, what I think was valuable about them, and what I learned from them.

I taught four workshops:

- [Having productive disagreements](#) (based on Double Crux)
- [Effective planning](#) (based on Murphyjitsu)
- [Building good habits](#) (based on Trigger-Action Plans, or TAPs)
- [Building useful systems](#) (based on Systemisation)

The format was fairly different from the standard CFAR workshop (90 minute afternoon workshops every 2 weeks, rather than an intense 4 day workshop), which I expected to make it much harder to have an impact. But, based on a followup 2-3 months later with participants, I found that some insights had stuck, and were being used regularly. A significant component of this was that I focused on distilling the techniques down to the key ideas, and useful mental habits. These seemed to stick well with some people, without requiring much followup effort.

Overall, I feel very happy with the impact, and I think afternoon classes are low cost to run (and seem to transfer acceptably to remote). I don't think a CFAR-style framing of rationality resonates with everybody, but there are enough people who found it high value that I think this was super valuable on average. So I would be very excited to see more people try to run classes like these!

I'd be especially excited to see more EA student groups running them I expect improving the long-term effectiveness of young EAs to be very high leverage, and a significant way for a local group to add value. I'd also expect it to be valuable to the

organisers, I found writing and running these workshops extremely helpful for understanding the ideas more deeply myself, and for improving at teaching! And I've found myself applying these ideas much more widely in my life.

Motivation

I have a few different goals with this post. As a result this post ended being pretty long, so I've tried to write each section to be self-contained, and to indicate which sections different people might find interesting:

- Convincing other people to run workshops like this: Explaining how I did things, and sharing my attempts to assess the longer-term impact on participants
 - I'd be especially excited to see other local Effective Altruism or Less Wrong groups running workshops like these.
 - See the [Doing These Yourself](#) section for more thoughts on who should do this
 - See the [Impact](#) section for my case for why these were worth running (continued in [Appendix B](#))
- Making it easier for other people to run workshops like these: All existing sources for this content I found are focused on *explaining* the content (eg the [CFAR Handbook](#)). I think applied rationality is very much something you learn by *doing*, and that it works best when taught in a structured environment that makes the default action to practice things, rather than requiring agency. And that comes with clear exercises to guide you through the ideas.
 - It took me a fair amount of effort to restructure this content as a coherent lesson plan with exercises. To my knowledge, my [lesson plans](#) are the best publicly available sources of this, so I hope this is useful!
 - See the [Content](#) section for summaries of the content, and my thoughts on what to emphasise, and common misunderstandings (continued in [Appendix A](#))
 - See the [Teaching Philosophy](#) section for my meta-level thoughts behind structuring the lessons the way I did
- Sharing my general teaching philosophy for applied rationality, and specific insights I've learned about teaching this stuff well. I mainly focus on making things intuitive and actionable: having clear and concise insights, motivating examples and useful exercises. See the [Teaching Philosophy](#) section for this.
- Sharing my distillations of these four techniques and ideas with anyone interested in rationality
 - If you're interested in the ideas but not in teaching, I recommend the section on [Content](#) and [Appendix A](#)
 - If you like how I think about them, I recommend going through the relevant [lesson plan](#) and actually doing the exercises and questions!

Workshops

Content

Format: These were 90 minute afternoon classes on weekends, mostly aimed at student EAs (late teens/early twenties). The target audience were people with a prior

interest in EA, rationality and optimising their life, but without much specific experience of CFAR techniques.

The following is a rough summary of the key takeaways and structure of my productive disagreements workshop (based on CFAR's Double Crux class). I've tried to go into detail on **Pedagogical-Content Knowledge** (PCK): knowledge about the topic, how students tend to engage with it, and how to teach it well. Eg examples that worked well, common misconceptions, subtle nuances to emphasise, etc. I think PCK is really useful to teach the classes well, but also very useful to understand the ideas more deeply yourself, even if you don't intend to teach them. In the interests of space, I've put similar sections for my workshops on planning, habits and systems in [Appendix A](#).

- [Lesson Plan](#)
- **Motivation:** I think disagreements should be about truth-seeking, but it is difficult to do this productively. Three major failure modes I wanted to address:
 - Focusing on minor disagreements rather than important points
 - Not taking each other seriously
 - Not talking about the same thing
- **Intended takeaways:**
 - Discussions should be about seeking truth, and understanding the other person's model that leads them to their beliefs
 - You get the most value from a discussion if you're open to changing your mind and take the other person seriously. A good way to achieve this is to paraphrase the other person's position to them, and iterate until they say it's a fair representation
 - Often discussions derail because they focus on misunderstandings or minor points. This can be addressed by explicitly identifying and stating your core and most important beliefs. We call these **cruxes**, and define them as 'an underlying belief, where if you changed your mind on this belief, you'd change your mind about the point of disagreement'
 - Our emotional reactions are an important component in what we believe, and engaging with them is a key part of truth-seeking discussion
- This was inspired by the CFAR Double Crux class (a paired framework for having productive discussions), but I reframed it around techniques you could apply in a discussion with anyone, rather than people explicitly using the same techniques. See the [CFAR Handbook](#) for more details
- **Techniques:**
 - Replace the symbol with the substance (tabooing words) - notice load-bearing words, words with a very nuanced and unclear meaning, and replace that word with a definition
 - **Paraphrasing:** Repeat the other person's position back in your own words. Ask whether it's correct, get feedback, and iterate until they're happy.
 - Emphasis on *ask, you're probably wrong first time!*
 - This ensures you're on the same page, highlights misunderstandings, and helps to take them seriously
 - **Seeking cruxes:** A crux is an underlying belief, where if you changed your mind on that, you'd change your mind on the conclusion. To find them, first freely generate supporting arguments for the conclusion. Then, for each argument, imagine a hypothetical world where that argument is false, and introspect on whether you'd change your mind on the conclusion
 - Emphasis on *introspect* - it's easy to trick yourself into thinking something *should* be a crux when it is not your true justification, you

just can't imagine it being wrong.

- **Resonate:** When hearing a position, *first* say the point you most agree with in it, before giving any counter arguments
- **Structure:**
 - Introduction: Outlining philosophy & goals of discussion (the goal is to seek truth, and you should assume good-faith until proven otherwise)
 - Operationalising: Failure mode 1 is not talking about the same thing
 - Resolve this by paraphrasing
 - Resolve this by noticing load bearing words and tabooing them
 - Resolve this by giving examples and making things concrete
 - **Exercise:** Practice operationalising vague statements
 - Seek cruxes: Failure mode 2 is talking about minor points, rather than the root cause of disagreement
 - Defining the idea of a crux, giving the algorithm for finding cruxes. This is useful because cruxes expose the models beneath your beliefs, and so are more valuable to focus on.
 - Encourage the social norm of saying “my crux is”
 - **Exercise:** Give a list of prompts, have them practice seeking cruxes on a favourite
 - Arguing in good-faith: Failure mode 3 is not taking your opponent seriously or trying to learn from them
 - Resolve this by paraphrasing
 - **Exercise:** Pair up and practice explaining your cruxes to a partner, and having them paraphrase it back
 - Resolve this by steelmanning
 - Resolve this by resonating
- **PCK:**
 - It's easy to get mired in a “conflict vs mistake theory” argument here, I try to address this by clearly outlining a mistake theory framework of assuming good faith at the start
 - Feedback says I overdid this, and my specific audience thought that was all obvious. Some friends objected to the mistake theory framing
 - The operationalising exercise didn't work super well, the prompts I generated felt too vague and the task was unclear
 - In future, I would scrap the operationalising section, it seemed the weakest
 - The cruxes exercise worked *really* well. Some people were initially skeptical and found the notion of cruxes dull and obvious. But seeing an explicit algorithm and trying it gave a visceral sense of “huh, I understand my position better”
 - The paraphrasing exercise worked *really* well. People had a visceral sense of “huh, that was way harder than I expected”
 - Note: Be clear before people seek cruxes that they'll pair up and share them later, so they shouldn't pick anything too private
 - It was valuable to identify all the techniques as mental reflexes, this helped them stick
 - It's easy to conceive of cruxes as a logical thing, and ignore implicit emotional biases. It is worth emphasising the introspective part, you can't have a meaningful discussion without understanding your emotional biases.
 - If doing again, I'd focus the class more explicitly on cruxes & paraphrasing, these seemed the most valuable takeaways. And give more time for exercises

I tried to heavily frame each workshop around building mental habits and reflexes. Ie, rather than the point being to learn a long, effortful algorithm, breaking the algorithm into bite-sized steps, learning the cues for when each step is relevant, and learning to bring them up in the moment. I think this is a skill best trained with TAPs. And based on the long-term feedback, this was an extremely successful approach! Few participants put in much effort to practice the techniques, but several managed to absorb these mental habits

Impact

People generally enjoyed the workshops, and when asked immediately afterwards gave highly positive feedback. I think the main source of impact is whether people absorb these techniques in the long-term, so I followed up 2-3 months after the workshops, and asked for qualitative feedback on how well the techniques had stuck. I heard back from about 85% of participants.

My best attempt to summarise this data was to loosely categorise people into neutral (no real impact), moderate successes (some long-term benefit) and strong successes (significant long-term benefit, regularly use the ideas in daily life). If you want to try analysing the feedback yourself, you can see my [anonymised summaries of all testimonials](#).

- Productive disagreements: of 6 participants, 2 moderate and 2 strong successes
- Effective planning: of 12 participants, 2 moderate and 2 strong successes
- Building good habits: of 6 participants, 3 moderate successes, no strong successes
 - I taught this workshop a few weeks before lockdown began in the UK, which I think was an unusually bad time. I think habits are very tied to your environment and daily routine, and easily break when that significantly changes. Several participants formed habits they found useful, that were then disrupted by lockdown
- Building useful systems: of 11 participants, 5 moderate and 1 strong success

Two highlighted testimonials from the productive disagreements workshop that I'm particularly excited about:

- Highlighted testimonial 1 (Strong):
 - Has internalised the idea of cruxes and now does this naturally. Finds this useful for understanding what he believes and why. Finds it notably useful to process unpleasant personal decisions and to change his mind on this
 - Intermittently notices "I am confused" in the moment and starts paraphrasing
 - Intuitively notices overloaded words and taboos them
 - Finds the techniques most useful when thinking alone, he generally finds arguments and debates unpleasant
- Highlighted testimonial 2 (Strong):
 - Has internalised the idea that "other people's minds should make sense" when talking to or thinking about other people
 - Puts more effort into understanding other people's models
 - Regularly uses paraphrasing

In the interests of space, I give some highlights of the testimonials I am most excited about for the other workshops in [Appendix B](#). I think reading things like this is most

interesting for gauging the impact of teaching rationality in this format. But I also find that seeing how other people engage with techniques in practice can help me understand them more deeply myself, and help me see how to put them into practice

My prior was that short, one-off classes would not have any noticeable long-term impact, because they would be too short and easily forgotten, and not have a surrounding context of self-improvement to reinforce the ideas and get people to practice. Seeing this long-term feedback has strongly updated me towards thinking these kinds of workshops are valuable. The workshops didn't have a significant long-term impact on most attendees, but had an impact on enough attendees that it seems extremely worth the total time investment to run and attend them.

I expect that certain kinds of people will get much more benefit from these workshops than others, and I had the useful filters of:

- A Cambridge student (which I expect to somewhat correlate with work ethic and intelligence)
- Involved in EA Cambridge
- Interested in spending weekend afternoons learning about rationality

I expect these filters significantly increased expected benefit

I framed each workshop around a series of Trigger Action Pattern-style mental habits, eg "when I notice I am confused about what somebody is saying -> try paraphrasing it back to them". I think this worked extremely well, and created a significant amount of the value of these workshops. A lot of the most successful feedback is people for whom these habits stuck, and are now regularly used. As far as I can tell, people didn't put significant effort into deliberately retaining these habits, they just made intuitive sense and stuck. I am very pleasantly surprised that they stuck this easily. My rough model for this is that people remembered and used the habit shortly after the workshop, found it useful, and this reinforcement kept happening until the habit stuck.

One major weakness is that for the more practical workshops, people rarely put in meaningful effort to practice or retain the techniques after the workshop. Eg, in the systemisation workshop, I had participants design and implement a system *in* the workshop. Many found this useful, and had the system stick, but far fewer applied the ideas to design more systems afterwards. One guess for addressing this would be to have follow-up workshops entirely focused on applying and practising the techniques, as a form of group accountability. I'd be extremely interested in hearing any other ideas for addressing this problem!

In hindsight, I think most of the value of the workshops came from people doing exercises and practising the ideas, and less so from the content and theory. In future, I'd shift emphasis to spend less time talking, and spend more time on the exercises. Though I think there is still significant benefit to spending *some* time on theory, and trying to articulate the mindset behind *why* the techniques make sense and are useful.

Teaching Philosophy

I think teaching is an extremely important skill, and teaching skill and philosophy is responsible for a lot of the variance in how well people learn, so it's something I try to think about a lot. In this section, I'm going to try to summarise my thoughts on

teaching as relevant to applied rationality. If you're interested, I go into a lot more detail on my thoughts on teaching generally in [this blog post](#)

One of the most important parts of my teaching philosophy is that **learning is a process of information compression**. We take in *far* more information than we retain. From a 90 minute workshop, most people will retain a few key points tops. This is important, because it means that I should be trying to choose those key points, and shaping the lesson around them. Fundamentally, the entire point of the workshop is to give context and reinforcement to those key points, everything else is irrelevant. Further, extracting those key points from a stream of information is significant intellectual labour. The student doesn't know what is and is not important, and it takes effort to identify this, effort taken away from actually learning. Thus, as the teacher who *does* know what is and is not important, my role is to make it as easy as possible to identify these key points. Some strategies to achieve this:

- Explicitly write down the key takeaways I want people to get from each section of the workshop. Default to cutting anything that doesn't help reinforce this takeaway.
- Summaries! Give an overview of the key points in the introduction and conclusion, and give a recap at the end of each section
- When I introduce an important point, explicitly say it's important! Eg, saying "this is a really important point, if you don't understand it, please flag this now so I can go into more detail."
- Give examples and spend more time on the most important points.
 - A good slogan: **pace according to difficulty, not length** - a common mistake is to just monologue until you get through all the ideas. But the real bottleneck is ensuring students can process and understand the points - it's worth spending more time on important points, giving examples and alternate framings, even if this means you need to skip details on easier, less important points
 - And if a point is hard and unimportant, why are you including it?!

Some ideas for teaching applied rationality specifically:

- Compress the key points into clear, concrete mental habits, of the form "If [condition], then [action]". Eg, "when I intend to do something in future -> ask 'would I be surprised if this doesn't happen?'"
 - There should be an explicit context for *when* the technique will be useful
 - Examples are valuable for giving context, and making the idea feel more concrete
- The number one failure mode is that the ideas make sense, but feel abstract. It needs to be *actionable*, and clearly relevant to their life!
 - Give a *ton* of examples. Illustrate every point with an example, distribute lists of more examples afterwards. Different things are actionable to different people, and examples help to give inspiration
 - Every workshop I thought I had too many examples, and got the feedback to include more next time
 - Be grounded in day-to-day life, and give relatable, everyday examples. Personal examples are a bonus
 - This is part of why it's important to have practiced a technique a bunch of times yourself before teaching it!
 - Lead with relatable, motivating examples - make it clear that the technique can help achieve *their* goals, and help with *their* failures

- The goal is always to be *immediately* intuitive. When you're trying to reframe how people interpret how their mind works, you're fighting an uphill battle. There's often a moment of snap judgement where the idea feels either intuitive or weird, and if it feels weird it's very hard to be useful. And the exact framing can significantly change how this snap judgement goes
 - First, use motivating examples to strongly relate the ideas to *their* lives. Ideally have several examples across a range of problems, and get people thinking "yes, that is a problem I have and struggle to fix"
 - Then, when introducing the specific idea/technique, ensure every step feels motivated, intuitive and grounded in the audience's internal experience.
 - I found beta testing with friends valuable here - try out different framings on different people
 - I think this part can determine a *lot* of the variance in how useful a workshop is, and it is worth putting a significant amount of effort in
 - Note: People are complicated, and some people just won't get value from a technique! This is fine. But I find there are also people who bounce off a technique framed one way, but get a lot of value from another framing
 - Different framings work on different people. I like to invite questions after introducing key ideas, and to try different framings as appropriate if people didn't like the original one
- Examples! I've said it a bunch of times already, but it's so important it's worth saying again. A class with *only* examples can work, a class with *no* examples will not.
 - I think about examples in 3 different ways:
 - Motivating examples - why you care about these ideas at all. These are often a good hook to begin the session with
 - Illustrative examples - a long example that can be broken down to illustrate an idea
 - Micro-examples - short examples given after a point to illustrate. This should be 1 clause (2 at worst) and be immediately intuitive - if you ever need to explain it, it's a bad micro-example
 - Teaching is hard, because it's a lossy process of information transfer over a noisy channel. Examples are a different way of communicating information than explanations, and so are invaluable as a way to detect and correct errors
- The ideas should be *embedded* in their mind. It needs to be active, reflexive knowledge, not abstract academic knowledge. This is because most of the impact comes from recognising a problem *in the moment* and doing something about it
 - Ask questions!
 - It's easy for people to be passive. I like to ask for a hand-poll (raise your hand to a point proportional to how much you agree with the statement) so everyone has to think
 - Alternately, say you'll pick on people at random
 - Note: Ask questions one at a time - you want it to be clear what the attendees should be thinking about
 - Give exercises! You want to make the ideas *reflexive*, so this needs to be something people practice and actually *do*. This lowers activation energy for actually using it in future
 - Exercises also highlight misunderstandings, when it's still easy to resolve them
 - I like paired exercises, it's far less easy to zone out (though warn people in advance!)

- It *really* helps to get people to think about the ideas, relate them to *their* life and *their* personal context
 - One of my favourite exercises is getting people to generate ways the problems arise in their lives, or where they can apply the technique. This also means they leave the session with a list of other places to apply the techniques!
- You are teaching guidelines, not rules. Emphasise this frequently and repeatedly. People are complex, and different approaches work best for different people.
 - When people do exercises, encourage them to experiment with and adapt the techniques, try to model this yourself

Doing These Yourself

Overall, I think these workshops were a major success at actually conveying the techniques to the audience. I've gotten a lot of value from these ideas in my personal life, empirically they're teachable, and I'd be excited about seeing these insights spreading and helping others to become more effective! I also found teaching these to be valuable personally, because it significantly clarified the ideas in my head. I've noticed myself using the ideas much more often in normal life as a result.

My lesson plans are [here](#). I expect these to work best as a source to ad-lib from and adapt, rather than to follow perfectly, but I designed them to be detailed and thorough, so I hope they can save a significant amount of work! EA Stanford have run two of these workshops based on my lesson plans, and seemed to think it went well and was much lower effort to run than if writing workshops from scratch.

If you want to run one of these, to prepare, I would recommend making a copy of these notes, reading through in detail, and editing it to be in your voice. Eg, noticing the framings you dislike and changing them, replacing my examples with ones you relate with, noticing the points you don't understand and cutting or thinking more about them. I think it's important to understand what's in the plan and to have it be something that makes sense to you, before trying to teach it to others. I generally err on the side of putting too much content into lesson plans, and cutting things when presenting (or overrunning, or both). I'd recommend cutting the parts of the lesson that don't seem as interesting or exciting to you.

I'd be especially excited to see the productive disagreements workshop done in EA groups. To me, a **key** part of EA culture is having good epistemics - taking other people's ideas seriously, trying to understand them, and being open to changing your mind. But I rarely see this norm explicitly set or taught, it seems more something that people are either already on board with, or pick up by osmosis. And I'm excited about seeing attempts to explicitly set culture.

Further thoughts on how to actually do this:

- Who should teach this?
 - I think the most important part is to understand the ideas well yourself. To have actually practiced and applied them in your life, to have a sense for the nuance, benefits and limitations.
 - I think this is important because there's a lot of subtlety and nuance to how the techniques are taught. This manifests in what parts of the techniques you emphasise, how you answer questions, help people struggling with the techniques, etc.

- I think it'd be especially good to have been to a CFAR workshop and connected with the ideas there. But I expect the more important part is to have actively practiced the techniques, found them useful in everyday life, and experienced the ways they can be hard to use or fail to be useful.
 - A rough litmus test: Have you used the technique regularly for at least a month after you learned it, and have you found it useful?
- Downside risk
 - Overall, I don't think there's high downside risk from other people teaching this stuff (if I did, I wouldn't make this post!). My guess is that most of the downside risk comes from techniques that focus on introspection and motivational issues
 - I expect the biggest downside of these techniques would be to use them to approach a naive idea of productivity in unhealthy ways. Eg, training a TAP for "when procrastinating, stop" and generally feel guiltier when procrastinating. Or to systematise your life in a way that doesn't account for breaks and leads to longer term burn-out
 - I don't strongly trust my inside view on this, so I'd love to hear other people's thoughts!
 - Some guesses for ways teaching rationality could go badly:
 - The instructor focuses on copying the structure of a CFAR class, rather than what's real and useful, and participants learn a vague version of a technique that feels fake/forced. And this makes learning the technique well harder later (CFAR calls this 'idea inoculation')
 - A student thinks rationality is all about trusting logic and ignoring emotions and anything that feels 'irrational', and makes unhealthy decisions ([my take on why this is an error](#))
 - A student labels part of their current mindset as 'bad habits' and tries to 'fix' them, and these habits are actually focused on protecting something important, which the student hasn't dug deep enough to realise
- Remote settings
 - I did these workshops in person. I tried re-running the disagreements one online, and it seemed to go well. Overall I think these are less ideal remotely, but still basically work
 - Apparently the worst part of it being remote was my frequent complaints that remote workshops are terrible. I have since updated this belief!
 - I think having Zoom breakout rooms (or equivalent) for paired exercises is essential for this to work well
 - I think the most significant thing lost is the ability to easily judge audience reactions, people disengage more easily, and it's harder to monitor how people are doing in the exercises
 - I added an introductory question ("what are things you struggle with when having disagreements, and what do you find helpful?"), gave people a minute to think about it and told them I'd be picking people at random to answer. At least one person said the "I'll pick somebody at random" part helped them feel engaged
 - Encourage people to have video on, this significantly increases engagement.
 - Have slides to illustrate key points/key algorithms/exercises (though err on the side of too little text, not too much)
- Further tips
 - **Beware the typical mind fallacy.** People are complex and different, but it's easy to implicitly assume they think exactly like you, find the same

- techniques useful, etc
 - Encourage questions, to notice these differences!
- Do trial runs! I find doing a practice run with a friend valuable for getting more fluent, noticing exercises and framings that feel clunky or repetitive, etc
- **Pace according to difficulty, not to length**
 - I think it's easy to pace a lesson according to how long the content takes to say. This is an error, because the main bottleneck is how long students take to understand it.
 - Rate each section/idea in the talk out of 5 for difficulty, and pad the hard ones with more examples, framings, etc, and cut things out of the easy ones
- Optimise for long-term impact
 - I think most of the expected value of the workshops to come from techniques sticking super well for a handful of people, rather than weakly for everyone
 - This means a lot of time should go for exercises, they help the techniques feel like useful knowledge, not academic knowledge

Conclusions

Overall, I think this was an excellent experiment! These ideas have become a powerful part of my mental toolkit, and I've been able to transfer them to others. My priors were against them being transferrable in this kind of context, but I've strongly updated in favour after doing [long-term followups](#)!

I'd be very excited to see other local groups trying to run these, and I hope [my lesson plans](#) can save some effort there. If you plan on organising these, I'd be very happy to chat and give any advice! Please feel totally free to reach out, and I'd be extremely interested in hearing how it goes if anybody does try them. My email is neelnanda27@gmail.com

Acknowledgements

Thanks to all of the many people who took the time to give feedback on the draft! Especially Dan Keys, Luca Righetti, Nora Ammann and Nathan Young, who helped make this significantly better. And thanks to CFAR for introducing me to these ideas in the first place! A year on, I've gotten a ton of value from my workshop

Appendix A: Lesson Content Continued

The continuation of the content section for my workshops on planning, habits and systems. I summarise the motivation, key takeaways and structure of the workshop, and try to give Pedagogical-Content Knowledge - common misconceptions and specific insights for teaching the ideas well.

Effective Planning

- [Lesson Plan](#)
- **Motivation:** Humans are systematically bad at planning. Eg, the student who leaves every essay to the last minute and needs to pull an all-nighter.
 - This is a *systematic* failure in planning, because she had all the information she needed to conclude this would happen (historical data), but didn't realise this in the moment. So the solution is to train your intuitions to better use your existing experiences to notice flaws in your plans
- **Intended takeaways:**
 - Most of the things we want to do in the future will fail to happen. Often the key failure is that we never even get around to explicitly planning them, and leave it as a vague ("I should ... some time"). To improve at planning, we need to improve our vague plans, not just improve our explicit planning skills
 - We intuitively flinch away from making robust, working plans. This can be resolved by framing things differently, and correctly using our intuitive knowledge of how plans will fail, since we know what will go wrong on *some* level.
 - This is a problem that needs to be solved in the moment, by building better reflexive reactions to planning and procrastination. This requires different solutions than just "trying harder" or "thinking differently", engaging with intuitions takes a different framing
- **Algorithm:**
 - Find a task/goal to plan
 - Make a plan (it can be super vague & rough)
 - Design a **picture of failure** - a concrete picture of yourself in the future, knowing that the plan failed, but without fleshing out why.
 - **Technique 1, Surprise:** Ask yourself "am I surprised that this picture came about?"
 - This is an introspective question - the surprise should be a visceral emotion
 - **Technique 2, Pre-hindsight:** Ask yourself "suppose this picture *did* come about. What went wrong?"
 - Emphasis: You *assume* it fails, and use hindsight to *explain* why. You aren't *asking* whether it fails or not, it's purely a hypothetical
 - Using this information, patch the plan! Iterate until you feel surprise if it fails
- **Structure:**
 - Introduce the ideas with relatable, motivating examples about how human intuitions suck at planning by default. Emphasise that *anything* you want done in future is a plan, even if you don't put in effort to formally plan
 - Introduce the idea of the Inner Simulator - a part of your unconscious mind that is excellent at modelling the world
 - People generate a list of things they want to plan
 - Conveniently, this means people leave with a list of other ways to practice the technique
 - Explain the algorithm, lead people through step by step, and after each step, give them time to apply it to their plan
 - **Exercise:** Pair people up, have them practice helping each other patch their plans and iterate
- I'm super excited about people getting better at planning - I think it's a key subskill for making deliberate positive changes in your life (eg, from the other

workshops!), career planning, overcoming procrastination etc. And that it can become a reflexive part of daily life, with enough practice.

- **PCK:**

- It's *super* important to get people to practice the technique. Exercises should be over half the length of the workshop
- Paired exercises work well - it's easier to spot flaws from the outside, easier to stick to the algorithm with somebody prompting you, and people take it more seriously
- People often think of planning as "only for high-effort things where I write out a plan". Emphasise that hearing a cool idea and deciding to do it some day *is* a plan, it's just a bad one
- The algorithm as framed feels high-effort and discouraging, people can't imagine using it in everyday life. Emphasise that the end-goal is to build mental *habits*, applying surprise and pre-hindsight in the moment, and the algorithm is short-term effort to reinforce these habits, and for really important plans.
 - Reflexive use of this technique looks like having *much* better calibrated expectations of yourself, and what you are and are not likely to actually do.
- It's easy to frame it as *only* an anti-procrastination tool, but it's more broadly useful. I lead with a motivating example about someone intensely preparing for a job interview, who gets stressed and screws up. And how they could have foreseen this
- Pre-hindsight feels kinda odd at first, like creating information from nothing. I motivate it with the example of regretting an email *immediately* after sending it, to highlight the difference between foresight and hindsight
- Some people bounce off the name Murphyjitsu, so I avoided it
- The technique can feel unnecessarily convoluted, and like a weird ritual. It helped to frame it as better accessing some *intuitions*, and that the interface required to do this well centres more on imagery and emotions, than it does on explicitly verbalised thoughts.

Building Good Habits

- [Lesson Plan](#)
- This was a class on Trigger-Action Plans (called implementation intentions in the psychological literature), a way to install if-then reflexes in your mind. Where you associate a concrete trigger with the reflex to perform a concrete action, eg "when I enter the library, turn off my phone".
- **Motivation:** I think a core skill of rationality is adjusting your behaviour [in the moment](#), because many errors happen without conscious attention. I am extremely excited about TAPs as a way to adjust your behaviour in the moment, and they're my main tool for doing so
- **Intended takeaways:**
 - It is valuable to think about how you reflexively react to things, and this is a key step to understanding and solving many problems. This requires a different approach to thinking analytically and abstractly about problems, since it's about connecting with your System 1
 - Most of our reflexive behaviour can be broken down simple if-then patterns. These have clear, concrete triggers and clear, concrete actions. This can help us to understand our behaviour, and to design new habits that stick

- More generally, habits are something you can cultivate and shape, rather than something that just happens to us
- Reflexive habits are most useful with *leverage*. Finding weak points early on, where a habit can be a small nudge and change default behaviour for the better
- Building and changing habits takes time, effort and maintenance in the short term, but this is worth it if it sticks in the long-term
- **Structure:**
 - Introducing the idea of habits and TAPs. Making it intuitive that this is a natural part of how we think
 - Framing them as small nudges, and talking about where they're useful. Participants generate a list of problems to apply TAPs to
 - Talk about what triggers work well, and have participants design a trigger
 - Good triggers are **obvious** (it's immediately clear when it happens), **visceral** (there's clear sensory detail attached) and **reliable** (it happens when you want the behaviour change to happen)
 - Talk about what actions work well, and have participants design one
 - Good actions are **obvious** (you know exactly what to do), **atomic** (it's short, simple and can be done in <5sec) and **consistent** (it's something you can always do)
 - You build a habit by explicitly practicing it 10 times, and then have a **learning period** where you keep it at the top of your mind and track it. Participants then build their habit
- **PCK:**
 - TAPs are hard to teach, because they're often a new concept without obvious analogies. The word habit often connotes *routines*, eg "I go to the gym once a week". It helped to explicitly emphasize this distinction
 - It also helps to ask the audience for examples of habits, and to repeat it back in the framing of "trigger -> action"
 - **Intuition pump:** To shift from inaction to action, *something* must have changed, and that thing is the trigger. Eg the habit "stretching when I'm stiff" is "*notice* I am stiff -> stretch"
 - Most people will not actually put in the effort to ensure habits stick. It helps to emphasise that it's *short-term* effort to ensure it sticks long-term
 - Habits break *super* easily with a change in life context, eg moving place. It's useful to emphasize this, and say that it takes attention and care to avoid this failure mode
 - Most habits built by participants broke after lockdown
 - It's easy to think TAPs are cool in the abstract but to never act upon this because it doesn't feel relevant. This can be resolved by *filling* the workshop with examples, as diverse as possible. I try to illustrate every point with a different micro-example
 - And link to a bunch more at the end
 - Eg diverse examples like "take a bite of food -> appreciate the taste" were found helpful
 - Even after all this, the biggest problem I found in long-term followup was participants not having clear ways to use it in their life
 - The most common mistake is not being specific and concrete enough with the trigger & action
 - Possible exercise idea: Pair people up, tell them to keep asking the other "can you make that more concrete"
 - I think I'd add a paired exercise in future workshops
 - Actions tend to be too ambitious, often high-effort, take more than 5 seconds, and aren't always doable

- Emphasise it should be a *reflex*. If you need to ask “should I do this?”, it’s already not working
 - This seems common with TAPs for social situations, people often feel uncomfortable taking an action the other person can perceive
 - Emphasise that TAPs are small nudges, not brute force. A good litmus test is “if a friend tapped me on the shoulder and told me to take the action, would the problem feel solved?”
 - There are many points about a good TAP. Emphasise that these are guidelines - they make it stick more easily, but you rarely satisfy all of them
 - The idea of practicing 10 times can feel silly/unnecessary. I frame it as “carving a groove in your mind”, or as “your system 1 is good at pattern spotting. This gives it data to infer a reflex from”
 - Emphasise that it’s totally fine to forget a habit sometimes, if it triggers sometimes that’s still a victory! Set the neutral point at “my habit never triggers”, not “my habit triggers perfectly”
 - On the margin people are too ambitious and impatient, and will try to train many TAPs at once. Recommend doing one at a time
 - A good framing is [Tortoise Skills](#) - a TAP is something you can have training in the background without much active effort. There’s no need for impatience, this is already a free win!
 - Some people find it unintuitive that our minds work on such a computational ‘if X then Y’ level. I began with an example of this behaviour in wasps to make it clear that it happens in animals too
 - Often people don’t get excited about habits, they feel too small and underwhelming to be worth the effort to work on. I like to emphasise the importance of tiny gains compounding over time, and having a *major impact over your life*. [A good motivating article](#)
- I’m most excited about TAPs being used to build *mental* habits, as I talk about in this post on [Noticing](#). I had a chapter on that, but didn’t have time to talk about it. In future, I think I’d spin that out into a separate workshop focusing more on Noticing (converting an emotion/mental phenomena into a visceral trigger)

Building Useful Systems

- [Lesson Plan](#)
- **Motivation:** A lot of important things in life take willpower and energy, eg remembering to exercise regularly and eat healthily. By building systems, eg a routine of going to the gym every Sunday morning, or having healthy food automatically delivered to your house, you can achieve these outcomes without needing to invest as much effort.
 - More generally, the workshop was on the *mindset* of thinking in systems: thinking of your effort and willpower as a limited resource, and finding ways to shape your future actions so that you do the right thing without requiring as much willpower.
- **Intended takeaways:**
 - Life is full of trade-offs. You have limited time, energy, attention etc, and need to consider opportunity costs when allocating them, rather than trying to do everything. Many problems can be best diagnosed by understanding how they’re spending your resources
 - Willpower is a limited resource, and it is important to notice where it’s spent and which areas of your life consume too much of it

- It takes willpower to deviate from the default action. A sustainable, long-term solution needs to change your default action to be better, and you should pay a lot of attention to your default actions and how to shape them.
 - If your default actions are bad, this is something to be *changed*, not something to feel guilty about
 - It is not sustainable to fix your problems by just ‘trying harder’
- Building a good system takes iteration and creativity, there are many options beyond the obvious, and it’s worth looking for them
 - To build good systems, you need to put a meaningful amount of effort into actually implementing it and reviewing it.
- **Structure:**
 - Introduce the idea of thinking in your life as abstract resources: health, time, energy, attention, willpower, sleep, etc (what CFAR calls units of exchange)
 - Emphasise that this is useful for identifying bottlenecks (resources you lack/are wasting) and for diagnosing problems (eg, am I short on time or just stressed)
 - I think this was valuable & framed the class well, but was too long and not 100% necessary
 - Introduce the mindset of thinking in systems. Define a system as “anything where you can spend willpower now to reduce willpower needed in future”
 - Strongly emphasise friction and trivial inconveniences as important
 - Strongly emphasise the importance of the **default action**, and systems as **changing the default**
 - System design
 - **Exercise:** Participants generate problems to systematise
 - **Exercise:** Participants understand the problem better by generating examples and doing a mindful walk through
 - Emphasise introspection and understanding the *true* behaviour, not the desired behaviour
 - **Exercise:** Design a rough system. Emphasise that it should be reliable (happens by default) and effortless (doesn’t require much effort to follow)
 - **Reality check:** Use pre-hindsight, run through the system, notice problems, iterate and patch
 - Emphasise that making robust systems is *hard*, and that the first draft likely fails
 - **Exercise:** Pair people up and have them practice
 - System implementation
 - Emphasise the importance of *doing* something to set up the system
 - **Exercise:** Set a 5 minute timer, have participants start implementing their system
- **PCK:**
 - There is a *lot* of room for creativity here
 - Emphasise that systems are super personal and specific to your tastes & habits
 - Emphasise that systems aren’t just routines. I give the examples of triage (quitting your least important course/commitment) and automating things
 - It’s easy to agree in the abstract but for the ideas to not feel actionable. *Fill* the workshop with examples, as diverse as possible
 - It’s easy to be neurotic and feel guilty at failure to meet your standards, and resolve to try harder. Emphasise the mindset of “failure is a problem

- with your default actions. Fix this by changing the default”
- It helps but is not essential to teach this after the planning workshop, pre-hindsight is a useful subskill
 - The key idea of units of exchange is not to precisely measure things, it’s to highlight hidden costs and bottlenecks. If it motivates a decision, that decision should feel *obvious*
 - A common failure mode is to fail to *implement* systems. The 5 minutes of “implement it now” time were valuable, and many participants left with worthwhile systems. Emphasise this
 - A future idea would be follow-up workshops that are just practicing the ideas and making new systems, many participants never designed systems outside of the workshop
 - There is unusually high value from pairing up to discuss systems and getting another perspective. I think I added significant value by suggesting tweaks to participant’s systems (eg using Toggl for time-tracking)
 - Mention *costs* of systematising, it’s only worth doing for problems that come up frequently ([relevant xkcd](#))
 - Emphasise the importance of trivial inconveniences, this isn’t obvious to some people
 - The idea of “make a system feel **sacrosanct**” resonated well with some people - shifting from “do I do this or not?” to it not even feeling like a decision
 - The reality check is worth emphasising, it’s easy to design a system that works in an ideal world, but not in practice.
 - The process outlined is high-effort, and this can put people off. Emphasise that the point is to practice and reinforce the mental habits of thinking in systems, and *these* are what matter:
 - Noticing and fixing resource bottlenecks, and drains on your energy and attention
 - Noticing and fixing inefficiencies
 - Noticing and fixing minor inconveniences and trivial inefficiencies
 - Noticing when a problem stems from a bad default action

Appendix B: Highlighted Testimonials Continued

Some more examples of the testimonials I’m most excited about from the other 3 workshops:

- Effective planning: of 12 participants, 2 moderate and 2 strong successes
 - Highlighted testimonial 1 (Strong):
 - Has formally applied the techniques to every important plan since (5 or 6 plans, took significant effort but felt valuable)
 - Has absorbed the mental reflex of “if this plan fails, what happened?” and uses it on a regular basis when planning smaller things, and finds this valuable
 - Found the ideas simple, intuitive and obviously powerful, so found them easy to absorb without much effort
- Building good habits: of 6 participants, 3 moderate successes, no strong successes
 - Highlighted testimonial 1 (Moderate):

- Built several habits shortly after the workshop
 - Successfully took effort to ensure they stuck, eg leaving themselves post-it note reminders
 - These broke somewhat when life was disrupted by quarantine, only one has stuck since
 - Feels like they have a better mindset around habits now, and that that has stuck
- I taught this workshop a few weeks before lockdown began in the UK, which I think was an unusually bad time. I think habits are very tied to your environment and daily routine, and easily break when that significantly changes. Several participants formed habits they found useful, that were then disrupted by lockdown
- Building useful systems: of 11 participants, 5 moderate and 1 strong success
 - Highlighted testimonial 1 (Strong):
 - Made a system in the workshop, it stuck moderately well
 - Now feels significantly better at self-compassion: when he has an unproductive day, thinks about how to adapt the environment to avoid this happening again rather than blaming himself
 - Eg he's started using a LifeRPG app (Habitica) to track to-dos, finds this useful
 - Eg he's reduced procrastination by adding a 20 second delay to opening distracting websites
 - Estimates a 30% probability a similar change would have happened without the workshop
 - Also found reading the [Replacing Guilt](#) series made this shift easier
 - Highlighted testimonial 2 (Strong):
 - Internalised the idea of taking trivial inconveniences seriously
 - Internalised the idea of setting things up to require minimal willpower
 - Was nudged towards installing a fitness app with scheduled training sessions, and refusing to ever miss a session (going strong at 5 weeks!)
 - Highlighted testimonial 3 (Moderate):
 - Built a robust to-do list system in the workshop, still finding it notably useful months later
 - Found the workshop valuable, but hasn't explicitly tried designing another system since
 - Highlighted testimonial 4 (Strong):
 - Wanted to address the problem of being busy and having too many commitments, started using Toggl to track time and better evaluate opportunity costs
 - **6 month follow-up:** This has become a core part of their workflow, and been notably helpful at encouraging them to balance their life and say no to things
 - Not sure how well the underlying ideas have stuck for developing future systems

Literature Review on Goal-Directedness

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

Introduction: Questioning Goals

Goals play a central role in almost all thinking in the AI existential risk research. Common scenarios assume misaligned goals, be it from a single AGI (paperclip maximizer) or multiple advanced AI optimizing things we don't want (Paul Christiano's [What Failure Looks Like](#)). Approaches around this issue ask for learning the right goals ([value/preference learning](#)), allowing the correction of a goal on the fly ([corrigibility](#)), or even removing incentives for forming goals ([CAIS](#)).

But what are goals, and what does it mean to pursue one?

As far as we know, Rohin Shah's [series of four posts](#) were the first public and widely-read work questioning goals and their inevitability in AI Alignment. These posts investigate the hypothesis that goals are necessary, and outline possible alternatives. Shah calls the property of following a goal "goal-directedness"; but he doesn't define it:

I think of this as a concern about *long-term goal-directed behavior*. Unfortunately, it's not clear how to categorize behavior as goal-directed vs. not. Intuitively, any agent that searches over actions and chooses the one that best achieves some measure of "goodness" is goal-directed (though there are exceptions, such as the agent that selects actions that begin with the letter "A"). (ETA: I also think that agents that show goal-directed behavior because they are looking at some other agent are not goal-directed themselves -- see this [comment](#).) However, this is not a necessary condition: many humans are goal-directed, but there is no goal baked into the brain that they are using to choose actions.

Later on, he explains that his "definition" of goal-directedness relies more on intuitions:

Not all behavior can be thought of as [goal-directed](#) (primarily because I allowed the category to be defined by fuzzy intuitions rather than something more formal)

Clearly, fuzzy intuitions are not enough to decide whether or not to focus on less goal-directed alternatives, if only because we can't define the latter. We thus need a more advanced understanding of the concept. What's more, deconfusing this concept would probably move forward many existing research approaches, as defended in the [first post](#) of this sequence.

That being said, intuitions shouldn't be thrown out of the window either. As [suggested](#) by Vanessa Kosoy, they provide the analogue of experimental data for philosophy: theories and definitions should mostly agree with intuitions in the simple and obvious cases. That doesn't mean we should be slaves to intuitions; just that biting the bullet on breaking one of them asks for a solid foundation about most other basic intuitions.

Hence this literature review. In it, we go through the most salient and important intuitions about goal-directedness we found in the literature. We do not limit ourselves to AI Alignment research, although many references come from this field. The end goal is to crystallize tests we can use on proposals for goal-directedness, a clear benchmark against which to judge proposals for deconfusing goal-directedness.

Note that we don't necessarily share the intuitions themselves; instead, we extract them from the literature, and delay further exploration to subsequent posts in this sequence.

This post follows a three part structure:

1. **(Intuitions about Goal-Directedness)** The first and most relevant part explores five topics related to goal-directedness we found again and again in our readings, and extracts a test for each to "falsify" a proposal for goal-directedness.
2. **(Comparison with Optimization and Agency)** The second asks about the difference between goal-directedness and the intertwined concepts of optimization and agency.
3. **(Proposals for Goal-Directedness)** The third and last part studies some proposals for a definition of goal-directedness, seeing how they fare against the tests extracted from the first part.

Note that, as mentioned in our [previous post](#) in this sequence, clarification of goal-directedness could take many forms. It seems improbable to be usefully characterised as a binary property, but it's not clear whether it should be seen as a continuum, a partial order, or something else entirely. This post doesn't assume anything except that there are different levels of goal-directedness, with some comparable and some maybe not.

Thanks to Evan Hubinger, Richard Ngo, Rohin Shah, Robert Miles, Daniel Kokotajlo, and Vanessa Kosoy for useful feedback on drafts of this post.

Intuitions about Goal-Directedness

In this section, we explore five main topics related to goal-directedness from the literature. These are:

- **(What Goals Are)** We discuss the possible definitions of goals, focusing on utility functions and their variants. We also explore some proposed alternatives.
- **(Explainability)** We unravel the link between goal-directedness and explainability of the system in terms of goals. This includes both which forms of explainability are considered, and their effectiveness.
- **(Generalization)** We study the link between goal-directedness and generalization, notably the apparent consensus that goal-directedness directly implies some level of generalization.
- **(Far-sighted)** We study the link between goal-directedness and the timescale over which the impacts of actions are considered. Although not common in the more general literature, this connection appears in almost every resource considered in the AI Alignment literature.
- **(Competence)** We study the link between goal-directedness and the ability of the system to accomplish a given goal and/or to be efficient in accomplishing

this goal. The distinction appears relevant as most resources don't posit the same link between these two forms of competence and goal-directedness.

What Goals Are

Goals are what goal-directed systems push towards. Yet that definition is circular because we don't know how to formalize goal-directed systems! One way to break the circle is to explore goals themselves; we believe this is paramount for deconfusing the idea of goal-directedness.

So let's start our investigation of the literature by looking at the answers to this question: what are goals? Or more precisely, what is the space of all possible goals?

Thinkers with a decision theory mindset might answer [utility functions](#): functions from states of the world, or histories of such states, to real numbers, capturing preferences. Indeed, a [classic argument](#) in the AI Alignment literature defends that a future superintelligent AI could be modeled as maximizing expected utility, because otherwise it would reveal incoherent preferences, and thus it could be exploited by humans, in contradiction with the assumption of superintelligence. Given the value for AI Alignment in modeling superintelligent AI, this would be strong evidence for the... utility of utility functions as a formalization of goals.

Yet the story doesn't stop there: Shah's most discussed post on goal-directedness, [Coherence arguments do not imply goal-directed behavior](#), argues that the class of utility functions is too large: in principle, any behavior can be interpreted as maximizing some utility function, even ones that are intuitively not goal-directed. This comes from considering a utility function that returns 1 for the exact histories or states appearing in the behavior, and 0 for the rest.

Did Shah's arguments convince AI Alignment researchers? Looking at the top comments on the post, it seems so. Moreover, the only pushback we were able to find on this post concerns the interpretation of the coherence arguments or Shah's conclusions (like Wei Dai's [post](#) or Vanessa Kosoy's [review](#)), not the criticism of utility functions as representing goals.

There is a fleshing out of the reasoning in Richard Ngo's [Coherent behaviour in the real world is an incoherent concept](#). Ngo does so by exploring two definitions of coherence (one with preference over states, the other with preferences over state trajectories), and considering how to use the formal coherence results on such definitions for more practical settings. The conclusion? Whichever way we use them, mathematical coherence arguments can only be applied non-trivially to real-world context by further constraining preferences to what we humans consider relevant. At which point Ngo argues that talking of goal-directedness is just better, because it acknowledges the subjective human perspective element instead of sweeping it under the rug of mathematical elegance.

If not utility functions, then what? Shah mentions that a simple utility function (the kind that can be explicitly represented in a program) would probably lead to goal-directedness.^[^1]

As a corollary, since all behavior can be modeled as maximizing expected utility, but not all behavior is goal-directed, it is not possible to conclude that an agent is goal-driven if you only know that it can be modeled as maximizing some expected

utility. However, if you know that an agent is maximizing the expectation of an *explicitly represented* utility function, I would expect that to lead to goal-directed behavior most of the time, since the utility function must be relatively simple if it is explicitly represented, and *simple* utility functions seem particularly likely to lead to goal-directed behavior.

[Comment on Coherence arguments do not imply goal directed behavior](#) by Ronny picks this line of thought, focusing implicitly on simpler utility functions as goals. This post defends that the real question is whether the expected utility maximizer is the best/simplest/most efficient explanation. Indeed, many of the cases where the utility maximization is trivial come from the utility function capturing all the complexity of the behavior forever, which is probably just as complex an explanation as the mechanical one. So the utility functions to consider are probably simple/compressed in some way.

The emphasis on simplicity also fits with a big intuition about when goal-directedness might appear. For example, [Risks from Learned Optimization](#), by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant, argues that the push towards simple models incentivizes the selection of optimizers as learned models, where optimizers are the representative of high goal-directedness in this paper.

Beyond simplicity, Shah's quote above pushes to further constrain the utility functions considered. Hoagy's [When do utility functions constrain?](#) explicitly studies this option. The post proposes a constraint on the class of utility considered:

What we need to find, for a given agent to be constrained by being a 'utility maximiser' is to consider it as *having a member of a class* of utility functions where the actions that are available to it systematically alter the expected utility available to it - for **all** utility functions within this class.

It tries to capture the fact that intuitively goal-directed systems will attempt to shift the expected utility where they want, instead of having the maximal expected utility given to them from the start.

Another possibility[^2] would be to ditch utility functions completely. What might this look like? The most explicit take on that comes from Ngo's [AGI safety from first principles: Goals and Agency](#). Among other things, this post argues for goals as functions of the internal concepts of the system, instead of functions of behaviors or states of the world.

Given the problems with the expected utility maximisation framework I identified earlier, it doesn't seem useful to think of goals as utility functions over states of the world. Rather, an agent's goals can be formulated in terms of whatever concepts it possesses - regardless of whether those concepts refer to its own thought processes, deontological rules, or outcomes in the external world.

The issue of course is that finding such concepts (and even defining what "concept" means) is far from a solved problem. Interpretability work on extracting meaning from trained models might help, like the [circuit work](#) of the Clarity team at OpenAI.

So concept-based definitions of goals are hard to investigate. But some parallel experimental work on goal-inference gives evidence that goals based more on logical concepts and relations can be used concretely. Tan Zhi-Xuan, Jordyn L. Mann, Tom Silver, Joshua B. Tenenbaum, and Vikash K. Mansinghka's [Online Bayesian Goal Inference for Boundedly-Rational Planning Agents](#) studies the goal-inference problem

for systems, expressing goals in [PDDL](#), a specification language for planning tasks and environments. Their inference algorithm performs better than standard Bayesian IRL, and shows impressive similarities with human inferences for the same trajectories.

Lastly, we would like to touch on one last interesting idea that appears important to the question of defining goals. It comes from this quote from Shah's [Intuitions about goal-directed behavior](#).

There's a lot of complexity in the space of goals we consider: something like "human well-being" should count, but "the particular policy $\langle x \rangle$ " and "pick actions that start with the letter A" should not. When I use the word goal I mean to include only the first kind, even though I currently don't know theoretically how to distinguish between the various cases.

Such a distinction intuitively separates goals about the state of the world, and goals about the output of the agent. But there seems to be more subtlety involved than just these two categories (as Shah himself [acknowledges](#) in discussing another example of non-goal for him, twitching).^[^3]

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **goals cannot just be the set of all utility functions. They must either be more constrained sets of utility functions, or different altogether, like concept-based goals.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Explainability

Why do we think so much about goals when looking at the world and trying to understand it? Why do we value thinking in terms of goals, and apply such thinking to almost anything that crosses our mind?

One answer is because it works pretty well. If a system is well described as pursuing a goal, then it will probably do things that we can interpret as bringing it closer to its goal. To take a classic example, modeling AlphaGo as wanting to reach the goal of winning does help in predicting, if not its exact next move, at least that the outcome of the game, or the fact it won't play terrible moves. Even when predicting what a human would do against AlphaGo (a situation where AlphaGo will almost certainly win), knowing that the human has the goal of winning helps us interpret why they're playing these specific moves.

The thing is, this notion of explainability becomes more subtle if the system is not as competent as AlphaGo. But we defer a discussion of competence in the context of goal-directedness to the relevant subsection below. Until then, let's assume that the system is competent enough to be explained as having a goal, if it has one.

Such a setting is exactly the one of [The Intentional Stance](#) by Daniel Dennett, often cited when discussing goals in AI Alignment research. In order to capture the different approaches one can take to model a system, Dennett defines three stances:

- The physical stance, which models the system as the sum of its physical components, and predicts its behavior by simulating the laws of physics.

- The design stance, which models the system as resulting from a design (either by an intelligent designer or by a process like evolution), and thus as made of higher level parts with each a reason for existing and interacting as they do. For example, considering an electronic circuit as made of logic gates is using the design stance, because it abstracts away the physical details that are not relevant to the design at hand. (This corresponds to having a [gears-level model](#))
- The intentional stance, which models the system as having beliefs (models of the world) and desires (goals), and as rationally trying to reach its desires by acting in accordance to its beliefs. The classical example is humans (Dennett's first goal with the intentional stance is to explain [folk psychology](#), the intuitive understanding of how other people think and react), but Dennett applies this stance to many other systems, including thermostats!

Given a system, which stance should we apply to it? Dennett doesn't say explicitly, at least in [The Intentional Stance](#). But he suggests that going from physical to design, or from design to intentional, is only justified if it improves our predictive powers (in terms of accuracy and/or efficiency). Applied to humans, Dennett argues that most of the time this means using the intentional stance, because it is the only tractable stance and it usually works.

He even goes one step further: in the context of theory of mind, he defends that the true believers (systems that should be ascribed internal states like beliefs) are exactly the intentional systems. Given how the intentional stance fits with the idea of goal-directedness, we can extend his ideas to say that the goal-directed systems (those that should be ascribed a goal) are exactly the intentional systems. What happens inside doesn't matter as long as the intentional stance works well enough.

In a similar vein, in his book [Life 3.0](#), Max Tegmark gives a definition of goal-directedness (called goal-oriented behavior) that depends heavily on explainability:

Goal-oriented behavior: Behavior more easily explained via its effects than via its cause

This approach fits with Dennett's intentional stance when "cause" is understood as the mechanistic cause, while "effects" is understood as the end result. Max Tegmark never explicitly defines these terms, but his writing on goal-oriented behavior implies the ones we gave.

Consider, for example, the process of a soccer ball being kicked for the game-winning shot. The behavior of the ball itself does not appear goal-oriented, and is most economically explained in terms of Newton's laws of motion, as a reaction to the kick. The behavior of the player, on the other hand, is most economically explained not mechanistically in terms of atoms pushing each other around, but in terms of her having the goal of maximizing her team's score.

Such approaches deal with the issue [raised](#) by Shah about the vacuousness of utility maximization. Recall that every system can be thought of as maximizing expected utility for some well-chosen utility functions. But Dennett (and Tegmark) doesn't only require that systems are explained by the intentional stance (as everything can be); he also wants the explanation to be better in some sense than mechanical ones. So the vacuousness of the expected utility maximization is not important, as we only want to consider the systems best explained in this way. It's also the interpretation favored by Ronny, on [this answer](#) to Shah.

There is also an argument to be made that goal-directedness is directly related to the kind of explanations that humans find intuitive. As we wrote above, Dennett's intentional stance attempts to explain folk psychology, which is pretty much this ability of humans to derive useful explanations of *other human beings*. So creating goal-directed systems would leverage these systems for prediction, as explained in Shah's [Will humans build goal-directed agents?](#):

Another argument for building a goal-directed agent is that it allows us to predict what it's going to do in novel circumstances. While you may not be able to predict the specific actions it will take, you can predict some features of the final world state, in the same way that if I were to play Magnus Carlsen at chess, [I can't predict how he will play, but I can predict that he will win.](#)

That being said, Shah thinks that prediction can be improved by going beyond goal-directedness:

I also think that we would typically be able to predict significantly *more* about what any AI system we actually build will do (than if we modeled it as trying to achieve some goal). This is because "agent seeking a particular goal" is one of the simplest models we can build, and with any system we have more information on, we start refining the model to make it better.

Nonetheless, some experimental work on explaining systems with the intentional stance partly vindicates Dennett.

As already mentioned, Zhi-Xuan et al.'s [Online Bayesian Goal Inference for Boundedly-Rational Planning Agents](#) shows how useful explanations of behaviors can be extracted through goal-inference. This gives evidence for the predictive power of this stance. Another evidence is arguably the whole field of [Inverse Reinforcement Learning](#).

In [Agents and Devices: a Relative Definition of Agency](#), Laurent Orseau, Simon McGregor McGill and Shane Legg compare mechanistic (a mix of physical and design stance) explanations against intentional explanations of trajectories in toy environments. The devices (for mechanistic explanations) are functions taking a history and returning a probability distribution over actions; the agents (for intentional explanations) use Bayesian RL. Through an appropriate simplicity prior, and the use of Inverse Reinforcement Learning for the agent side, Orseau et al. can compute a posterior distribution over devices and agents as explanations of a given trajectory. Their results fit with intuition: running in circles is inferred to be device-like, because it has a simple mechanical explanation; going straight to a specific zone is inferred to be agent-like; and others in between.

Lastly, we want to address a comment by Shah on a draft of this post. He wrote

I think this section is a reasonable way to talk about what humans mean by goals / goal-directed behavior, but I don't think the resulting definition can tell us that much about AI risk, which is what I want a definition of goal-directedness for.

Although this goes beyond the subject of this literature review, one way in which more explainable systems might be more dangerous in terms of AI risk is that they will behave in ways that follow fundamental intuition in humans. Hence when we find a scenario involving such systems, we should put more credence on it *because* it applies to systems for which our intuitions are better calibrated by default. This of

course doesn't mean that less explainable systems are not also potential sources of AI risk.

To conclude, this part of the literature suggests to us the following test for proposals of goal-directedness: **a definition of goal-directedness should come with a way to explain systems based on goals (like saying they maximize some utility). And the predictive power (both in terms of efficiency and in terms of precision) of this explanation compared to purely mechanistic ones should increase with goal-directedness.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Generalization

Maybe the most obvious aspect of goal-directedness is that it implies generalization: a system striving towards a goal will adapt itself to changes in the environment.

Indeed, this distinction between goal-directed behavior and the opposed habitual behavior has been a major topic in psychology and cognitive neuroscience for more than 50 years. [Goals and Habits in the Brain](#), by Ray J. Dolan and Peter Dayan, surveys this research, splitting the approaches to the goal-directed/habitual dichotomy into four overlapping “generations”: the behavioral approach in rodents; the behavioral approach extended to humans; a computational modeling approach through model-based vs model-free RL; and the refinement of this modeling. As they write, habitual behavior is understood as automatic, and thus efficient but inflexible, whereas goal-directed behavior is understood as thoughtful, and thus expensive but flexible.

Thus, key characteristics of habitual instrumental control include automaticity, computational efficiency, and inflexibility, while characteristics of goal-directed control include active deliberation, high computational cost, and an adaptive flexibility to changing environmental contingencies.

Note that the current consensus presented in this survey argues that both behaviors play a role in human cognition, with a complex dynamic between the two. Yet time and time again, in experiment after experiment, goal-directed behavior becomes necessary when something changes that invalidates the previous habits.

Similarly to this modeling of behavior, Dennett’s [intentional stance](#) assumes that the system will alter its behavior in response to changes in the environment. Dennett even uses the sensitivity of this change as a means for comparing two intentional systems.

But as systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will notice, in effect, and make a change in its internal state in response. There comes to be a two-way constraint of growing specificity between the device and the environment. Fix the device in any one state and it demands a very specific environment in which to operate properly (you can no longer switch it easily from regulating temperature to regulating speed or anything else); but at the same time, if you do not fix the state it is in, but just plonk it down in a changed environment, its sensory attachments will be sensitive and discriminative

enough to respond appropriately to the change, driving the system into a new state, in which it will operate effectively in the new environment.

Generalization also plays a fundamental role in the analysis of goal-directedness in the AI Alignment literature. In [Intuitions about goal-directed behavior](#), Shah even defines goal-directedness that way:

This suggests a way to characterize these sorts of goal-directed agents: there is some goal such that the agent's behavior *in new circumstances* can be predicted by figuring out which behavior best achieves the goal.

We failed to find any rebuttal to the claim that goal-directedness would entail generalization, at least with some assumption of competence.^[^4] On the other hand, the arguments in favor abound:

- Shah's [Will humans build goal-directed agents?](#) presents the argument that goal-directed systems are the best known candidate for solving problems in novel and possibly superhuman ways, which amount to generalizing beyond our current context.
- Ngo's [AGI safety from first principles: Goals and Agency](#) lists criteria for goal-directedness, and the last one, flexibility, directly references the ability to adapt plans to changes in circumstances.
- Hubinger et al.'s [Risks from Learned Optimization](#) uses explicit optimisers as a representation of goal-directed systems, both because of the usefulness of knowing the internal structure, and because such optimizers will generalize better. That being said, the risk of inner alignment is that such a learned optimizer becomes goal-directed towards a different goal than the expected one from training. So in some sense, while goal-directed systems are assumed to generalize, mesa-optimizers present the risk of undesirable generalization.
- Wei Dai's [Three ways that "Sufficiently optimized agents appear coherent" can be false](#) lists distributional shift as one way that an optimized system might appear incoherent, even if it's performing well on the training environments. In contraposition, a truly coherent system (a system which always appears to follow some goal) would deal with this distributional shift, and thus generalize.

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **generalization should increase with goal-directedness, for a suitable definition of generalization that captures adaptation of behavior to changes in the environment.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Far-sighted

One intuition about goal-directedness mostly appears in the AI Alignment literature: that goal-directed systems should consider the long-term consequences of their actions. Not that such far-sightedness is forbidden in goal-directed behavior from cognitive neuroscience, or in Dennett's intentional stance. But it holds no special place: a system that only looks at short-term consequences of its actions would still be considered goal-directed or intentional. In AI Alignment on the other hand, researchers who looked into the issue apparently agree that far-sightedness plays a fundamental role in goal-directedness.

Look for example at Ngo's [AGI safety from first principles: Goals and Agency](#): one of his criteria for goal-directedness is large scale. Or see Shah's [list](#) of non-goal-directed AI designs: it includes [act-based agents](#) whose main feature is to only care about the short term preferences of the human supervisor/AI. Or observe that Eric Drexler's [CAIS](#) are presented as different from goal-directed systems in part because they only have the resources to consider short-term consequences of their actions.

Why is it so? It probably stems from safety issues like Stephen Omohundro's [Convergent Instrumental Subgoals](#), Hubinger et al.'s [Deceptive Alignment](#), and [Goodhart's Law](#). The first two require the consideration of non-immediate outcomes, and the third only really becomes an existential problem when considering large timescales where enormous optimization pressures are applied.

So this intuition is clearly relevant for thinking about AI Alignment and existential risks. Let's just keep in mind that it doesn't necessarily apply to the general idea of goal-directedness as considered outside AI Alignment.

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **the timescale of goals considered should increase with goal-directedness.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Link with Competence

Finally, the most subtle intuition about goal-directedness concerns its link with competence. Note that we're not discussing which goals a system will have (the sort of question addressed by Nick Bostrom's [Orthogonality Thesis](#)); instead, we focus on how competence changes with goal-directedness. This also means that we don't delve into whether competence pushes towards goal-directedness. The latter is an important question, studied for example in Shah's [Coherence arguments do not imply goal-directed behavior](#) and in David Krueger's [Let's Talk about "Convergent Rationality"](#); but it is not our focus here.

So what competence is needed for goal-directedness? Let's first clarify what we mean by competence. We separate how far the system can go in accomplishing its goal (**ideal accomplishment**)^[^5] from how fast it moves towards the goal (**efficiency**). When thinking about reaching a certain state, like winning at chess, ideal accomplishment measures the ability to win, possibly against a range of different adversaries, whereas efficiency measures things like the number of moves taken or how close were the won games. When thinking more about maximizing some utility in an online setting, ideal accomplishment talks about how little regret is possible for the system, whereas efficiency measures the actual regret of the system.

This split also manifests itself in the link with goal-directedness: ideal accomplishment is mostly considered independent of goal-directedness^[^6], while efficiency is assumed to grow with goal-directedness.

Beyond a minimal level of ideal accomplishment (if the system is too incompetent, its goal cannot be revealed), behavioral approaches to goal-directedness like Dennett's intentional stance don't assume the ability to reach goals. The difference between a

good chess player and a bad chess player is not seen as a difference in their intentionality, but as a difference in their beliefs (models of the world).

As for structural and mechanical definitions, they don't even need this minimal assumption of ideal accomplishment: goal-directedness doesn't have to be revealed. Hence no criterion requires the system to be good at accomplishing its goal in Ngo's [AGI safety from first principles: Goals and Agency](#), and it's not assumed either in Hubinger et al.'s [Risks from Learned Optimization](#).

This assumption, that only minimal ideal accomplishment matters for goal-directedness, has some experimental evidence going for it. In Zhi-Xuan et al.'s [Online Bayesian Goal Inference for Boundedly-Rational Planning Agents](#), the goal-inference assumes only bounded-rationality from the system (which can be seen as a constraint on ideal accomplishment). Even with this limitation, the authors manage to infer goals (among a small set of predefined ones) in some cases. More generally the Inverse Reinforcement Learning (IRL) literature contains many examples of goal-inference even with less than competent demonstration. For a recent example, see Jonnavittula and Losey's [I Know What You Meant: Learning Human Objectives by \(Under\)estimating Their Choice Set](#): in it, they avoid overfitting the reward to the demonstrator's mistakes by making the inference more risk-averse, which means assuming that the demonstrator was limited to trajectories very close to the one taken.

What happens for our other facet of competence, efficiency? There, behavioral definitions do assume that efficiency grows with goal-directedness. The intentional stance for example assumes rationality; this means that a system well explained intentionally will move as efficiently as possible (given its beliefs) towards its goal. The same can be said for proposals related to Dennett's, like Ronny's [rescue of utility maximization](#) or Tegmark's definition of goal-oriented behavior.

While structural definitions don't explicitly require efficiency, their reliance on mechanisms like optimization, which approximate the ideal rationality used by Dennett, hint at the same dependence between efficiency and goal directedness.

That being said, one might argue that the link between efficiency and goal-directedness relates more to the alignment issues linked with goal-directedness than to the property itself. Jessica Taylor's [quantilizers](#) can be interpreted as such an argument. Quantilizers don't just maximize expected utility; they instead randomly sample from the top $x\%$ actions according to a utility function and a base distribution. The paper proposes that by doing this, the system might still accomplish the wanted goal without some of the expected consequences of stronger optimization.

Given that utility maximization can have many unintended side effects, Armstrong, Sandberg, and Bostrom[12] and others have suggested designing systems that perform some sort of "limited optimization," that is, systems which achieve their goals in some non-extreme way, without significantly disrupting anything outside the system or otherwise disturbing the normal course of events. For example, intuition says it should be possible to direct a powerful AI system to "just make paper clips" in such a way that it runs a successful paper clip factory, without ever attempting to turn as much of the universe as possible into paper clips.

Rephrased in our words, quantilizers aim at reducing efficiency while maintaining ideal accomplishment. The assumption that this will keep accomplishing the goal in many cases thus undermines the dependence between efficiency and goal-directedness.

This part of the literature thus suggests to us the following test for proposals of goal-directedness: **ideal accomplishment (as in being able to accomplish the goal in many contexts) should only be relevant to goal-directedness in that a minimal amount of it is required. Efficiency on the other hand (as in measuring how fast the system accomplishes the goal) should increase with goal-directedness.**

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Comparison with Optimization and Agency

Why not use an already existing concept in place of goal-directedness? After all, we have two closely related alternatives at hand: optimization and agency. It might seem at first glance that goal-directedness merely measures optimization power or agency. The goal of this section is to clarify the differences, and justify an independent study of goal-directedness.

Optimization

What's the difference between goal-directedness and optimization? After all, many of the paper mentioned in the previous section either assume optimization (Hubinger et al. in [Risks from Learned Optimization](#) and Ngo in [AGI safety from first principles: Goals and Agency](#)) or use it as an implicit model of goal-directedness (Shah's [Intuitions about goal-directed behavior](#)).

Yet there is some sense in which goal-directedness appears different from optimization. To see that, we will first rephrase two definitions of optimization from the literature: internal search and optimizing systems.

- (**Internal Search**) This is the meaning from optimization algorithms: search a space of possibilities for the best option according to some objective (like training neural nets by gradient descent). It fits with Abram Demski's [selection](#).
- (**Optimizing Systems**) This definition comes from Alex Flint's [the ground of optimization](#): an optimization system robustly pushes the state of the system into a smaller set of target states. Note that this is different from Abram Demski's [control](#): a control process only has one shot at doing what it wants to do, and doesn't have an explicit access to a search space (like a growing plant or a guided missile). Optimizing systems appear to encompass internal search/selection as well as control.

Armed with these two definitions, it's easier to compare with goal-directedness. Ideally, we would want to find a simple relation, like an equality or straightforward inclusions. Alas the literature is far from settled on such a proper hierarchy.

Let's start with internal search. Proponents of a structural definition of goal-directedness usually assume optimization in this sense, like Hubinger et al. in [Risks from Learned Optimization](#) and Ngo in [AGI safety from first principles: Goals and Agency](#). So in a sense, they assume that goal-directedness is contained in internal search. One way to interpret the additional constraints they bring to optimization (for

example the self-awareness criterion explicitly stated by Ngo) is that goal-directed systems are a subset of systems using internal search.

On the other hand, proponents of behavioral definitions seem keen to accept non-optimizing systems. Recall that Dennett accepts a thermostat in the club of intentional systems -- and a thermostat definitely doesn't run any internal search.

Hence without rejecting one half of the literature, there is no way to include internal search in goal-directedness, or the other way around.

What about optimizing systems? At first glance, it seems that the situation is more straightforward: Alex Flint [himself](#) includes goal-directed systems inside the set of optimizing systems:

Our perspective is that there is a specific class of intelligent systems — which we call optimizing systems — that are worthy of special attention and study due to their potential to reshape the world. The set of optimizing systems is smaller than the set of all AI services, but larger than the set of goal-directed agentic systems.

Yet the issue of competence muddles it all. Indeed, optimizing systems **robustly** push the configuration space towards a smaller target. That is, they are *successful* optimizers. To the extent that goal-directed systems don't have to be very competent (in reachability), they might fall short of forming optimizing systems with their environment.

And it is not the other way around either: some examples of optimizing systems (computing the square root of 2 through gradient descent or a tree growing) hardly satisfy all the tests for goal-directedness. So goal-directedness doesn't contain all optimizing systems.

All in all, the links between optimization (for either definition) and goal-directedness are far from obvious. Yet a general trend emerges from studying the literature: **most authors agree that goal-directedness isn't just optimization.**

Agency

Agency is a broad concept that appears in various fields, from psychology and economics to the natural sciences and AI. It is thus not surprising that different definitions of agency have already been proposed. In [Defining Agency: individuality, normativity, asymmetry and spatio-temporality in action](#), Xabier Barandiaran, Ezequiel Di Paolo, and Marieke Rohde collect the insights found in the literature on agency and try to give a definition useful to multiple domains. According to them:

[...] agency involves, at least, a system doing something by itself according to certain goals or norms within a specific environment.

In more detail, agency requires:

1. Individuality: the system is capable of defining its own identity as an individual and thus distinguishing itself from its surroundings; in doing so, it defines an environment in which it carries out its actions;
2. Interactional asymmetry: the system actively and repeatedly modulates its coupling with the environment (not just passive symmetrical interaction);

3. Normativity: the above modulation can succeed or fail according to some norm.

Readers interested in more formal approaches to agency might consider [Semantic information, autonomous agency and non-equilibrium statistical physics](#) by Kolchinsky and Wolpert, or Martin Biehl's [PhD thesis](#) (and referenced papers).

Back to goal-directedness, the part of the definition by Baradiaran et al. that relates to our review is normativity as a necessary requirement: every agent acts according to some norm or goal, in a quite broad sense. They consider minimal proto-organisms as agents that modulate their interaction with the environment through processes aimed at dissipation and self-maintenance; but they classify a human undergoing involuntary tremors as non-agentic, since it is hard to find a sense in which tremors succeed, or fail, to fulfill any norm generated by the human system and related to its interaction with a generic environment.

The authors also write that

[...] systems that only satisfy constraints or norms imposed from outside (e.g. optimization according to an externally fixed function) should not be treated as models of agency).

It is clear that Baradiaran et al. use the terms “agent” and “goal” differently than how they are usually used in AI: in the standard AI textbook [Artificial Intelligence: A Modern Approach](#), Russell and Norvig present different types of agents, but classify only some of them as goal-based.

Dennett adopts another viewpoint: he argues for “the unavoidability of the intentional stance with regard to oneself and one's fellow intelligent beings” (*The Intentional Stance*, p.27); in other words, he claims that intelligent agents necessarily behave according to the beliefs-goals model.

On the other hand, when considering the arguments on AI risk, Ngo [doesn't take for granted that AGI will be goal-directed](#). At the same time, in his analysis he uses the terms “agency” and “goal-directedness” interchangeably, interpreting them as two different terms for the same concept. He acknowledges that multiple factors could drive the development of goal-directed AGI, and he also points out that a collective intelligence as a whole could be more agentic than any individual agent within it.

Goal-directedness in distributed agent systems is also considered by Magnus Vinding in his short book [Reflections on Intelligence](#). He notes that the collective human ability to achieve goals is great thanks to the wide variety of specialized tools that humans have created, and that its nature is highly distributed. Similarly, in [CAIS](#) Drexler argues that superintelligence won't necessarily be embodied in a single monolithic agent: he actually claims that distributed superintelligence will be obtained before single-agent AGI is developed.

Lastly, the philosophically-inclined reader can check the [SEP page on Agency](#), which presents the standard view on the concept of action, relying on intentional actions and acting for a reason, and different kinds of agency. The page on [Moral Motivation](#) discusses instead Humeanism and Internalism/Externalism, which are different philosophical positions regarding the connection between beliefs and goals—both concepts are often used to describe the behavior of agents.

In summary: **In the literature we found significant, though far from perfect, overlap between agency and goal-directedness. Yet goal-directedness appears like the most tractable concept for the moment, compared to agency. High goal-directedness and agency can emerge also in distributed systems of agents that are, individually, less goal-directed.**

Proposals for Goal-Directedness

At last, let's see how the different proposals for goal-directedness fare for our criteria. We focus on the following four:

- Dennett's [intentional stance](#)[⁷]
- The [operationalization](#) of the intentional stance with expected utility maximization
- Ngo's six [criteria](#) for goal-directedness
- Vanessa Kosoy's [definition](#) of goal-directed intelligence

Intentional Stance

Recall Dennett's intentional stance: an intentional system is the kind of system that can be more easily explained as following its desires (goals) according to its beliefs (models of the world). The comparison is drawn with the physical stance (describe the physical configuration of atoms, and simulate the laws of physics) and the design stance (assume that the system was designed and abstract it into a gears-level model).

How does it fare against our tests from the literature?

- **(What goals are)** Here the intentional stance disappoints: it says nothing about what is a goal, and what makes a goal. It relies on the common sense definition of a goal as a desire, which might suffice for folk psychology, but doesn't cut it for thinking about goals more broadly.
- **(Explainability)** The intentional stance is strong with this one -- after all explainability is in the definition! Arguments can be made about how much predictive power does the intentional stance give, but explaining the behavior of an intentional system is its point.
- **(Generalization)** The intentional stance also fares well on generalization, thanks to its focus on adaptability. Dennett is pretty clear that with a rich enough connection to the world, the beliefs should become accurate enough to adapt the behavior to new situations.
- **(Far-sighted)** As mentioned in the corresponding section, approaches to goal-directedness outside AI Alignment rarely consider far-sightedness as fundamental in the definition of goal-directedness. It thus makes sense that Dennett doesn't expand on this point.
- **(Link with Competence)** Ideal accomplishment seems irrelevant to the intentional stance except for the minimal level that allows the system to be efficiently predicted as an intentional system. Then efficiency and applicability of the intentional stance go hand in hand, as a more efficient system will be closer to Dennett's assumption of rationality.

All in all, the intentional stance scores decently enough on the tests we extracted from the literature (maybe in part because Dennett's book [The Intentional Stance](#) is such a big part of this literature). Yet this overlooks the main issue with the intentional stance: its lack of formalization. Yes, an intentional system is one that is better explained by the intentional stance; but what does "better explained" mean here? Is it about computational complexity? About accuracy? About both? How many errors until the explanation is deemed not good enough? And how do we formalize this explanation, by the way? Utility maximizers are an obvious answer, but they have problems of their own, as we'll see for the next proposal.

A more formal grounding is important because of the ease with which we as humans use a version of the intentional stance to explain almost everything. A well-known example is [ELIZA](#), the computer program that acted as a psychologist by recognizing specific keywords and rephrasing as questions what was written to it. This program had an incredibly basic design stance explanation, but users got emotionally attached to ELIZA really quickly (prompting the invention of the [ELIZA effect](#)). To quote ELIZA's programmer, Joseph Weizenbaum:

What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people.

On the other hand, a recent exploration of this issue, [Do We Adopt the Intentional Stance Toward Humanoid Robots?](#) by Marchesi et al., found out that participants tended to slightly favour the design stance when explaining the actions of a humanoid robot. Nonetheless, some situations can create mentalistic assumptions coherent with the intentional stance. This might mean that the [ELIZA effect](#) is mitigated when looking at actions specifically, instead of interacting directly with the program/robot.

Intentional Stance Through Expected Utility Maximization

The main issue with the intentional stance comes from a lack of operationalization; hence an easy tweak is to propose such a formalization. For example, Ronny's [Comment on Coherence arguments do not imply goal directed behavior](#) raised the idea that maybe a goal-directed system is one that is most efficiently described in terms of utility maximization. It's not clear from the post if Ronny knew about Dennett's work, but his proposal can still be interpreted as an operationalization of the intentional stance.

The main consequence is that goals are now well-defined as utility functions. But isn't that in contradiction with the test for goals? No, because of the constraint that the explanation be simpler than the mechanistic one. In Shah's post, the vacuity of utility maximization is shown by encoding the whole behavior in the utility function; this probably creates an explanation that is just as complex as a mechanistic one. Hence there is an implicit simplicity assumption for the utility function in Ronny's proposal. Nonetheless, how to capture this simplicity is still an open question.

Experimentally, Orseau et al.'s [Agents and Devices: a Relative Definition of Agency](#) use this operationalization for their agents, and they are able to infer goal-directed behavior in toy examples.

Hence this is one of the most promising concrete proposals for goal-directedness.

Richard Ngo's proposal

We've mentioned Ngo's [AGI safety from first principles: Goals and Agency](#) multiple times, when defining the tests for goal-directedness. In this post, Ngo proposes six criteria for goal-directedness:

1. *Self-awareness*: it understands that it's a part of the world, and that its behaviour impacts the world;
2. *Planning*: it considers a wide range of possible sequences of behaviours (let's call them "plans"), including long plans;
3. *Consequentialism*: it decides which of those plans is best by considering the value of the outcomes that they produce;
4. *Scale*: its choice is sensitive to the effects of plans over large distances and long time horizons;
5. *Coherence*: it is internally unified towards implementing the single plan it judges to be best;
6. *Flexibility*: it is able to adapt its plans flexibly as circumstances change, rather than just continuing the same patterns of behaviour.

Note that none of these traits should be interpreted as binary; rather, each one defines a different spectrum of possibilities.

How do these fare against our tests from the broader literature?

- **(What Goals Are)** As mentioned earlier, Ngo presents goals as made from the internal concepts of the system, rather than world state. This points towards a concrete space of goals, but is hard to evaluate for the moment as we don't have a good understanding (or even formalization) of the sort of concepts formed by a system.
- **(Explainability)** Since the criteria 2 (Planning) and 3 (Consequentialism) boil down to internal search, this proposal can use a mechanistic form of predictability hand in hand with the more intentional approach. It's thus satisfying for explainability.
- **(Generalization)** Generalization is a big theme in Ngo's [AGI safety from first principles: Goals and Agency](#) and all the [AGI safety from first principles](#) sequence. Assuming competence, a goal-directed system by Ngo's definition has all the ingredients needed to generalize: it looks for plans and adapts them to changes in the environment.
- **(Far-sighted)** Criterion 4 (Scale) explicitly requires that goal-directed systems care about the long-term consequences of their actions.
- **(Link with Competence)** This approach being based on internal structure, it doesn't require a minimal level of reachability to work. And efficiency should follow from the combination of many criteria like 2 (Planning), 3 (Consequentialism) and 6 (Flexibility).

Really, the only issue for our purposes with this definition is that it focuses on how goal-directedness emerges, instead of what it entails for a system. Hence it gives less of a handle to predict the behavior of a system than Dennett's intentional stance for example.

One surprising take is the requirement of self-awareness. In our readings, this is the only resource that conditioned goal-directedness on self-awareness (usually it's more related to agency, which makes sense because Ngo uses the two terms interchangeably). Is it a necessary component? Further research is needed to decide. But as it is far from a consensus in the literature, we don't include it in our list of tests.

Goal-Directed Intelligence

In contrast with the rest of this section, Vanessa Kosoy's [proposal](#) for goal-directedness is completely formal.

Given $g > 0$, we define that " π has (unbounded) goal-directed intelligence (at least) g " when there is a prior ζ and utility function U s.t. for any policy π' , if $E_{\zeta\pi'}[U] \geq E_{\zeta\pi}[U]$ then $K(\pi') \geq D_{KL}(\zeta_0 || \zeta) + K(U) + g$. Here, ζ_0 is the Solomonoff prior and K is Kolmogorov complexity. When $g = +\infty$ (i.e. no computable policy can match the expected utility of π ; in particular, this implies π is optimal since any policy can be *approximated* by a computable policy), we say that π is "perfectly (unbounded) goal-directed".

Let's ignore some details for the moment, and just consider the form of this definition: $\exists A, \forall \pi' : B \implies C$.

- A is a pair of a prior and a utility function.
- B is an inequality, asking that the expected utility parameterized by A of π' is greater than the one of π .
- C is another inequality, asking that the description complexity of π' be greater than the description complexity of the prior and the utility by at least g .

So in essence, having goal-directed intelligence g means that according to some goal, beating the policy costs at least g in terms of descriptive complexity.

It's not obvious what the link is to our tests, so let's spell it out.

- (**What Goals Are**) Kosoy appears to place herself in the expected utility maximization framework: goals are pairs of (utility function, prior).
- (**Explainability**) First, if goal-directed intelligence is infinite, the policy is optimal. This gives some form of explainability, in that the policy will accomplish its goal as well as possible.
But what about policies with finite goal-directed intelligence? Intuitively, a policy easy to explain as optimizing the goal should have a low descriptive complexity. Yet the only constraint on descriptive complexity is a lower bound. If policy π has

goal-directed intelligence g , this tells us that the descriptive complexity of π is lower bounded by the descriptive complexity of the goal (prior and utility function) + g . So nothing forbids the descriptive complexity of π to grow with goal-directed intelligence, even if one would expect it not to.

- **(Generalization)** The existential quantifier in the definition captures the idea that there is some goal for which the property holds. So a discussion of generalization requires fixing the goal. And in an online setting like this one, generalization means something like minimizing regret, which here translates to maximizing expected utility. This works for infinite goal-directed intelligence; see the discussion of competence to see why this also works for finite goal-directed intelligence.
- **(Far-sighted)** No mention of long-term consequences of actions in this definition.
- **(Link with Competence)** Here ideal accomplishment means reaching the optimal expected utility for the goal, and efficiency is about how far one reaches towards that optimality. The ideal accomplishment part is quite irrelevant, as if the prior makes the optimal expected utility for a goal very small, there is still some sense in which goal-directed intelligence can increase. Concerning efficiency, Kosoy provided us with a simple result that goes a long way: if two policies have the same goal (the same goal satisfies the existential constraint for their goal-directed intelligence) then $g_1 > g_2$ implies that π_1 has better expected utility than π_2 .^[^8] Hence efficiency does increase with goal-directed intelligence.

The main issue with this proposal is not formalization, but understanding: it's a very compressed definition that needs further exploration to be judged more fully. That being said, we can already see that it ties goal-directedness with efficiency in a really strong way, fitting completely with our test.

Conclusion: What Needs to Be Done?

In this literature review, we extracted five tests for goal-directedness -- about goals, explainability, generalization, far-sightedness and competence. We also investigated how goal-directedness differed from optimization and agency, the two closest concepts studied in the AI Alignment literature. And we finally pitted various proposals for goal-directedness against the tests.

The conclusion is that two proposals seem particularly promising: the operationalization of the intentional stance through expected utility maximization, and Vanessa Kosoy's goal-directed intelligence.

That being said, there is still room for a lot of work in this sphere: questioning the criteria, proposing new definitions, and finding new intuitions to challenge the current proposals.

Our previous sequence, [Toying With Goal-Directedness](#), proposed ideas in a less polished fashion, and already contains some of our thoughts on these questions. This sequence will expand on these ideas, and propose a structured and more thought through take on goal-directedness.

Notes

[^1] Simplicity is related to the idea of compression, which is the subject of one of our [previous posts](#).

[^2] A slightly different take comes from Vanessa Kosoy's [instrumental reward functions](#). Intuitively, such a function takes as input not a bare state, but an instrumental state -- an abstraction capturing the distribution of histories starting from this state according to the POMDP. This difference means that two states with different rewards are necessarily distinguishable by some policy (playing the role of an experiment); and thus that difference in rewards are indeed meaningful. We only mention this in a footnote as it's not clear how instrumental reward functions evade the issue pointed out by Shah.

[^3] This distinction might be related to the concept of locality we talked about in a [previous post](#), and which will be studied further in this sequence.

[^4] Note that we don't mean complete generalization here. Only that in many cases the system will adapt its behavior to the environment.

[^5] Note that ideal accomplishment is related to generalization, in the sense that a high ideal accomplishment probably requires significant adaptation to the environment, and thus generalization. That being said, we are aiming here at something different from generalization: not just the ability to adapt, but the ability to "win".

[^6] This relates to the idea of focus that we presented in a [previous post](#).

[^7] This part also applies to similar proposals that focus on explainability, for example the one by Max Tegmark.

[^8] The proof from Vanessa: since g_1 is the intelligence of π_1 , for any policy π , if $E[U](\pi) \geq EU(\pi_1)$ then $K(\pi) \geq g_1 + C$. Since the intelligence of π_2 is $< g_1$ by hypothesis, there exists π s.t. $EU(\pi) \geq EU(\pi_2)$ but $K(\pi) < g_1 + C$. Applying the former observation to the latter π , we get $EU(\pi) < EU(\pi_1)$. Combining the two inequalities $EU(\pi_2) \leq EU(\pi) < EU(\pi_1)$. QED

A few thoughts on the inner ring

I enjoyed C.S.Lewis' [The Inner Ring](#), and recommend you read it. It basically claims that much of human effort is directed at being admitted to whatever the local in-group is, that this happens easily to people, and that it is a bad thing to be drawn in to.

Some quotes, though I also recommend reading the whole thing:

In the passage I have just read from Tolstoy, the young second lieutenant Boris Dubretskoi discovers that there exist in the army two different systems or hierarchies. The one is printed in some little red book and anyone can easily read it up. It also remains constant. A general is always superior to a colonel, and a colonel to a captain. The other is not printed anywhere. Nor is it even a formally organised secret society with officers and rules which you would be told after you had been admitted. You are never formally and explicitly admitted by anyone. You discover gradually, in almost indefinable ways, that it exists and that you are outside it; and then later, perhaps, that you are inside it.

There are what correspond to passwords, but they are too spontaneous and informal. A particular slang, the use of particular nicknames, an allusive manner of conversation, are the marks. But it is not so constant. It is not easy, even at a given moment, to say who is inside and who is outside. Some people are obviously in and some are obviously out, but there are always several on the borderline. And if you come back to the same Divisional Headquarters, or Brigade Headquarters, or the same regiment or even the same company, after six weeks' absence, you may find this secondary hierarchy quite altered.

There are no formal admissions or expulsions. People think they are in it after they have in fact been pushed out of it, or before they have been allowed in: this provides great amusement for those who are really inside. It has no fixed name. The only certain rule is that the insiders and outsiders call it by different names. From inside it may be designated, in simple cases, by mere enumeration: it may be called "You and Tony and me." When it is very secure and comparatively stable in membership it calls itself "we." When it has to be expanded to meet a particular emergency it calls itself "all the sensible people at this place." From outside, if you have dispaired of getting into it, you call it "That gang" or "they" or "So-and-so and his set" or "The Caucus" or "The Inner Ring." If you are a candidate for admission you probably don't call it anything. To discuss it with the other outsiders would make you feel outside yourself. And to mention talking to the man who is inside, and who may help you if this present conversation goes well, would be madness.

...

My main purpose in this address is simply to convince you that this desire is one of the great permanent mainsprings of human action. It is one of the factors which go to make up the world as we know it—this whole pell-mell of struggle, competition, confusion, graft, disappointment and advertisement, and if it is one of the permanent mainsprings then you may be quite sure of this. Unless you take measures to prevent it, this desire is going to be one of the chief motives of your life, from the first day on which you enter your profession until the day when you are too old to care. That will be the natural thing—the life that will come to you of its own accord. Any other kind of life, if you lead it, will be the result of conscious and continuous effort. If you do nothing about it, if you drift with the stream, you

will in fact be an “inner ringer.” I don’t say you’ll be a successful one; that’s as may be. But whether by pining and moping outside Rings that you can never enter, or by passing triumphantly further and further in—one way or the other you will be that kind of man.

...

The quest of the Inner Ring will break your hearts unless you break it. But if you break it, a surprising result will follow. If in your working hours you make the work your end, you will presently find yourself all unawares inside the only circle in your profession that really matters. You will be one of the sound craftsmen, and other sound craftsmen will know it. This group of craftsmen will by no means coincide with the Inner Ring or the Important People or the People in the Know. It will not shape that professional policy or work up that professional influence which fights for the profession as a whole against the public: nor will it lead to those periodic scandals and crises which the Inner Ring produces. But it will do those things which that profession exists to do and will in the long run be responsible for all the respect which that profession in fact enjoys and which the speeches and advertisements cannot maintain.

His main explicit reasons for advising against succumbing to this easy set of motives are that it runs a major risk of turning you into a scoundrel, and that it is fundamentally unsatisfying—once admitted to the ingroup, you will just want a further in group; the exclusive appeal of the ingroup won’t actually be appealing once you are comfortably in it; and the social pleasures of company in the set probably won’t satisfy, since those didn’t satisfy you on the outside.

I think there is further reason not to be drawn into such things:

1. I controversially claim that even the good of being high status is a crappy kind of good relative to those available from other arenas of existence.
2. It is roughly zero sum, so hard to wholly get behind and believe in, what with your success being net bad for the rest of the world.
3. To the extent it is at the cost of real craftsmanship and focus on the object level, it will make you worse at your profession, and thus less cool in the eyes of God, or an ideal observer, who are even cooler than your local set.

I think Lewis is also making an interesting maneuver here, beyond communicating an idea. In modeling the behavior of the coolness-seekers, you put them in a less cool position. In the default framing, they are sophisticated and others are naive. But when the ‘naive’ are intentionally so because they see the whole situation for what it is, while the sophisticated followed their brute urges without stepping back, who is naive really?

D&D.Sci II: The Sorceror's Personal Shopper

The day's task shows up in an envelope, and not in glowing purple letters emblazoned across the inside of your eyelids, which is usually a good sign. The owl that brought it looks on with equanimity as you read its master's message:

Hello,

I hearde that you do odde jobs for Wizards. I neede 120 mana for a ritual but cannot leave my Tower righte now. Go to the caravans in towne and buy enough magic items that I can gette that much by sacrificinge them.

My Owle has a pouch. It is biggere inside than oute. Putte the things in it ande she will carrye them back.

Enclosed is my Thermo Tharmet Magic Sensing Device. It usually lies but is probably bettere than guessinge. Returne it when you are done. Enclosed is also a [list](#) of 836 itemse I sacrificed and what coloure they glowed and how muche mana I gotte and what the Thau Lying Box said when I pointede it at them. I like lists.

The pouch contains 200 gold pieces. You may keepe what coins are lefte over. If I do notte gette at leaste 120 mana from the things you sende me, you shalle owe me 200 gold pieces.

Goodbye,

Wakalix the Wizard

PS: If you do not accepte the jobbe, I bid you sende the Owle and the gold back before sundown, that I may finde another to charge with it.

Your spirits lift with every line. Clear objectives, payment in advance, acknowledgement that you have the right to refuse the task, no threats of involuntary transformation, no random tangents about world domination or beard care, handwriting legible, capitalization not entirely random . . . this is one of the *good* clients. And if you make clever enough use of the list he provided, you suspect you could end up taking home a decent fraction of that 200gp once this day's work is done. With a song in your heart, you depart for the travelling caravans and their magic items.

The selection of artefacts that greets you is as follows:

Item name	Glow color	Thaumometer reading	Price
Longsword of Wounding +2	Red	14	66gp
Warhammer of Justice +1	Yellow	5	41gp
Hammer of Capability	Blue	35	35gp
Pendant of Truth	Red	40	38gp
Ring of Joy +5	Blue	29	32gp
Warhammer of Flame +2	Yellow	48	65gp

Battleaxe of Glory	Blue	7	23gp
Plough of Plenty	Yellow	12	35gp
Saw of Capability +1	Green	16	35gp
Amulet of Wounding +2	Green	50	35gp
Pendant of Hope	Blue	77	34gp
Pendant of Joy +4	Green	42	39gp

Will you accept Wakalix's errand? If so, what will you buy?

I'll be posting an interactive letting you test your decision, along with an explanation of how I generated the dataset, sometime this Sunday. I'm giving you a week, but the task shouldn't take more than a few hours; use Excel, R, Python, tarot readings, or whatever other tools you think are appropriate. Let me know in the comments if you have any questions about the scenario.

If you want to investigate this collaboratively and/or call your decisions in advance, feel free to do so in the comments; however, please use spoiler tags when sharing inferences/strategies/decisions, so people intending to fly solo can look for clarifications without being spoiled.

Discussion on the choice of concepts

"The reason that you can currently make toast without doing great damage is just that your toaster is stupid."

"Can 'stupid' be correctly applied to toasters?"

"Yes"

"What if I say no?"

"Well, if you have a conception of stupidity that can't be applied to toasters, and one that can, why would you choose the one that can't?"

"But I don't have two—I'm talking about the actual concept"

"There isn't an actual concept, there are a bajillion concepts, and you can use whichever one you want."

"There's one that people mean"

"Not really—each person has a slightly different usage, and probably hasn't pinned it down. For instance if you ask them if toasters are stupid, they might be unsure."

"Yes! They are unsure because they are trying to guess what the real concept is, from their limited collection of exposures to it. If it were about making one up, why would they be uncertain?"

"They might be uncertain which one they want to make up"

"You're saying when people say words, they are ascribing meanings to them that they just made up, according to which definition they like most?"

"Well, they like definitions that fit with other people's usage a lot more than other ones."

"I think they are just guessing what the real meaning is"

"There isn't a real meaning"

"Ok, what the consensus meaning is"

"There isn't a consensus"

"Yeah but they believe there is one"

"You're like a word meaning nihilist—you want only these 'real' word meanings or at least these word meanings of consensus, yet you know they don't exist. That seems sad."

"Maybe, but that doesn't make it wrong. And also, I was talking about what other people do."

"What does it matter what other people do? You can use whatever meanings you want."

"That seems unfriendly somehow"

"What if you do it in a friendly way? For instance, where a meaning is ambiguous, if you choose the best one. For instance, if you say toasters can be stupid?"

"It's more a vibe of do-it-alone responsibility for everything, thinking of others as machinery that happens to be near you, that rings my alarm bells. Leaving the common experience of word usage to stand outside the system, as it were, and push the common stock of concepts in the way that you calculate best. At least it seems somehow lonely and cold"

"That's a bit dramatic - I think the odd nudge in a good direction is well within the normal human experience of word usage. Plus, often people clearly redefine words somewhat in the context of a specific conversation. Would it be so strange if within our conversation we deemed 'stupid' applicable to toasters? Not doing so seems like it will only limit our discussion and edge us toward taking up some further concept like shmoopid to fill the gap."

"It's not clear at all to me that that is the only bad consequence at stake. For instance, words have all kinds of connotations besides what you explicitly think of them as about. If you just declare that stupid applies to toasters, then try to use it, you'll doubtless be saying all kinds of things about toasters that you don't mean. For instance, that they are mildly reprehensible, and that you don't like them."

"I don't know if I would have used it if I didn't implicitly accept the associations, and this is a risk one seems to always run in using words, even when you would deem them to apply."

"Hmm. Ok, maybe. This sounds like a lot of work though, and I have done ok not thinking about using my influence over words until this day."

"You think you have done ok, but word meanings are a giant tragedy of the commons. You might have done untold damage. We know that interesting concepts are endlessly watered down by exaggerators and attention seekers choosing incrementally wider categories at every ambiguity. That kind of thing might be going on all over the place. Maybe we just don't know what words could be, if we were trying to do them well, instead of everyone being out to advance their own utterings."

Thoughts on being mortal

(Cross-posted from [Hands and Cities](#))

(Content warning: discussion of death and intense pain)

This post is an amalgam of thoughts about death, prompted centrally by Atul Gawande's book [*Being Mortal*](#).

I.

Gawande's book describes a lot of different people who are dying, at different speeds, with different kinds of suffering and support. I often found it piercingly sad, in a particular way I associate with death, and with people saying final goodbyes to life and to each other.

But for all the pain and degeneration in the book, the tone still feels somehow "smooth." Gawande imbues the suffering of his subjects with a lot of dignity; he focuses, often, on what they can still do, and what they love. Most of the doctors and nurses he portrays are competent and well-intentioned.

Beneath this, though, one senses a roiling ocean of pain, confusion, fear, despair, and loss. Here I am reminded of Scott Alexander's [post](#) on death in hospitals, which emphasizes, much more than Gawande, patients who are *screaming*, attacking doctors, trying to pull out their tubes, being restrained, going actively insane. See also Samuel Shem's [*House of God*](#), for a portrayal modern medicine much more macabre and ridiculous and horrifying than Gawande's (I've only read the beginning, but it felt like the first few chapters were enough to get the gist). Anne Boyer's [*The Undying*](#) — an account of her experience with breast cancer and its treatment — is also more relentless than Gawande in her focus on the pure pain — examining it from a huge variety of angles, searching for some sort of linguistic adequacy. She opens with words from the Iliad: "Not even if I had ten tongues and ten mouths."

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

(Boyer and Shem are also much more interested than Gawande in the corruptions of the medical system, but these won't be my focus here.)

I don't fault Gawande the book's smoothness: the topic is hard enough as it is, and his tasteful writing makes it bearable. And relative to many depictions of human life, this one certainly isn't short on raw horror. See, for example, his descriptions of the poor-houses to which many elderly people in the early 1900s were confined, and of comparable institutions in India today, which Gawande describes as "as close to a vision of hell as I've ever experienced."

Nor is it short on grisly detail. See, for example, Gawande's lengthy description of what happens to the body during aging: the loss of tooth enamel, muscle mass, strength, lung capacity, brain size; calcium deposits building in blood vessels, joints,

and heart valves; arthritis, osteoporosis, tremors, strokes, dementia; clouded and yellowing eyes, thinning hands, the death of pigment cells in the hair, the breakdown of sweat glands.

Here I'm reminded of Buddhist "[Patikulamanasikara](#)" meditation — "reflections on repulsiveness" — in which one enumerates the parts that make up the body (head hairs, body hairs, bile, phlegm, skin oil) in an attempt to overcome lust and attachment. But reading Gawande's description of aging, I don't feel repulsed. Rather, I feel a certain kind of tenderness, warmth, and fear for my body and its trillions of cells; this incredible concert of tiny, rapid processes, being born and dying, frantic and earnest and determined, sustaining for years a cloud of something like Joe, moving through the world; warding off pathogens, digesting food, pumping blood; billions of nerves cells firing, the energy burned in taking a walk or lifting a hand or remembering; all of it in a certain sense blind, but still purposeful, still in some sense trying. And eventually, more and more, maybe slowly, maybe fast: failing.

II.

The basic thesis of Gawande's book is that the modern medical system often makes the end of life worse for the dying than necessary. For Gawande, this happens, broadly, because the system focuses unduly on safety and on prolonging a patient's life at all costs, rather than on the patient's actual priorities for the end of their life. As a result, you end up with nursing homes that deny the elderly all autonomy, and treatment recommendations that pursue increasingly small chances of increasingly small amounts of additional life, at extreme costs to a patient's ability to make the most of what time is left. (Note that this isn't a story about the financial costs of end-of-life treatment, which are extreme — rather, it's about whether this treatment even helps the patients whose welfare supposedly justifies it).

Gawande discusses institutions, practices, and evidence that points to an alternative vision — of nursing homes that provide more autonomy; of hospice care that does not prolong life at extreme costs to its quality; and of doctors and families who learn to have hard and honest conversations about what sorts of trade-offs the dying are willing to make as death approaches, and who shape treatment and care to serve that vision.

(One other recurring motif in the book, echoed in articles like "[How Doctors Die](#)," is that the type of treatment involved in attempting to prolong life often makes life not just worse, but shorter as well.)

I was particularly struck by the discussion of hard conversations. A theme in the book is how often doctors and families will avoid really confronting the fact that death is coming fast, and that remaining treatment options (themselves often quite painful and debilitating) are virtually certain to fail. It's a fraught topic, and the simplest approach to it is to focus on whatever slivers of hope remain, and to err always on the side of the most aggressive treatments.

This focus often works to preserve optimistic misunderstandings. For example, Gawande quotes one oncologist who reports that he is often thinking 'can I get a good year or two out of this?', where his patients are thinking ten or twenty. The book quotes a study showing that the average survival time estimate given by doctors of terminally ill patients was 530% too high; and Gawande finds himself subtly and not-so-subtly accommodating his patients' desire for optimism. Doing otherwise requires effort, courage, and skill.

III.

To aid in confronting and responding to hard truths, Gawande returns often to questions in the following vein, posed to those nearing the end of their life:

- What is your understanding of what's happening to you, and of its potential outcomes?
- What are your fears if your condition worsens?
- What are your goals if your condition worsens?
- What trade-offs are you willing to make, and not willing to make, to try to stop what is happening?

The task of the care-givers and the patient together is to plot the course of action that best serves this understanding. What the patient values most as their life ends is key here, and reading their descriptions — a woman with metastatic ovarian cancer who wants to make it to her best friend's wedding, to put her feet in the sand, to be a wife and mother and friend for a just a bit longer; a piano teacher who wants to keep teaching; everyone who wants to be free from pain, and to have more time with their family — I felt again death's capacity to bring life into focus.

It's a familiar motif: the patient emerges from the hospital, newly diagnosed, newly short of time, and with new clarity about what really matters, and about the preciousness of what she's had all along. As a result, her remaining time is charged with greater meaning and intimacy. I think Tim McGraw's song "[Live Like You Were Dying](#)" is actually a pretty good expression of this: "I loved deeper, I spoke sweeter, I gave forgiveness I'd been denying."

For a long time, one of my guiding goals in life has to been start early on this. To reach the end of my life, and to have learned already, or as deeply as possible, whatever lessons in preciousness and fleetingness and beauty that death can teach; and to have been doing the whole time what I would've wanted myself to do, at least in expectation, with those lessons in mind.

Often this feels easiest when I'm alone. Somehow the social world is harder. It's one thing to recognize the preciousness of life and of our time together; it's another to hold onto that recognition, and to infuse our interactions with it, amidst the forceful currents of sociality — the habits and uncertainties, the comfortable defaults, the fears and blocks and distances between us.

And indeed, a more vivid awareness of death is no guarantee of intimacy: if it's hard in everyday life, or around the dinner table at Christmas, or on intermittent phone calls, deathbeds won't always make it easy. We can miss each other, we can fail to be truly together in this world, even in our final moments. Here I'm reminded of a funeral I once went to, in which even the remarks from the family of the deceased felt to me formulaic and stilted and distant. Part of this, I expect, was the difficulty of expressing things in that public context; but it was also a reminder to me that certain kinds of closeness can just not happen. Deathbeds are no exception.

Still, reading Gawande's book, and now writing about it, has left me, for now, with some small measure of the consciousness that I think the approach of death can bring. I feel more aware of my body's fleeting strength and health; my ability to run and jump and walk up stairs; the softness of the clothes on my skin. And I feel, as well, the urgency of certain projects; the preciousness of certain relationships; the shallowness of certain hesitations and pre-occupations; the costs of wasting time. You can't keep any of it; there's nothing to [hold back](#) for; your life is always flowing

outwards, through you and away from you, into the world; the only thing to do is to give it away on purpose, and the question is where and to what.

The impact merge

(Cross-posted from [Hands and Cities](#))

Lots of people want to do big things — start a big company, write a bestselling book, participate in an important project. I'll call this the "accomplishment desire."

Often this is centrally tied to social status; the relevant type of "bigness" is highly correlated with what would seem cool at, e.g., a college reunion. But it also mixes in richer values. Maybe someone wants to be a famous musician; but they also often want their music to be *good*, and expressive of who they are.

This desire is also tied to meaning — to wanting to have *done something*, mattered, lived a life of consequence. Accomplishments are the types of things you look back at on your deathbed; they're your legacy. (My sense is that having children is often central here, even if this doesn't read as traditionally "big").

Lots of people also want the world to be as good as possible. They want there to be no disease, injustice, poverty, war; they want love and joy and beauty. Call this the "good world desire."

The accomplishment desire and the good world desire need not mix, even where they co-exist: many people who want to start big companies aren't centrally motivated by a desire to do good, and wanting to make good music isn't the same as trying to make the world better through music — music has its own standards. People generally want the big things they do to be good for the world (few seek "big" legacies of harm), but that doesn't mean they're aiming for bigness on the "good for the world" dimension.

Still, it's natural to attempt a merge: e.g., to hitch the accomplishment desire to the good world desire, into an overall aim of trying to improve the world as much as possible. Let's call this the "impact merge."

My question is whether the impact merge is a good idea. It might seem so — if you're aiming to do big things, why not aim for big improvements to the world? But I think there are considerations on both sides.

My basic concern is that the accomplishment desire is, at bottom, *about you*, whereas the good world desire is about the world. So I worry that the two, hitched together, will not always pull in the same direction.

This is easy to see in contexts where they literally recommend different actions. Suppose, for example, that your desire for accomplishment is in fact deeply tied to some sort of desire for social status. In that case, you might turn down a more impactful role for a more conventionally prestigious one; you might prefer to start your own projects, rather than help someone else succeed at theirs; you might care unduly about whether you are getting credit for your good deeds, and so on.

But that's the easy version of the question. The harder version comes in contexts where someone has in fact successfully keyed their personal conception of accomplishment entirely to improving the world as much as possible. Such a person would view the more impactful role as a bigger accomplishment than the more prestigious one; prefer helping someone else with a more valuable project to starting

their own, etc. They would look, from the outside, like someone who didn't care about accomplishment at all, and cared only about the world.

But the essential "about them"-ness of their accomplishment desire would remain. They still care *that it was them* who did good deed X; they just don't care that this fact is socially legible.

Here's a case to illustrate the point (this sort of case is also explored in a draft paper from various researchers at the Global Priorities Institute). Consider two scenarios:

1. The world is wonderful, but Bob can't do anything to improve it;
2. The world is terrible, but Bob can do a lot to improve it, though not nearly enough to make it as good as it would be in (1).

Suppose that Sally is choosing which of these worlds to create. If Bob cares about how much good *he* accomplishes in life, he'll hope that she chooses (2). But that seems messed up, and incompatible with the good world desire — e.g., Bob is hoping for a worse world. True, he's not *causing* a worse world, since in that case, he wouldn't be improving it, and Bob is all about improving the world. But intuitively, the point of improving the world is for the world to *be* good, not for anyone in particular (including oneself) to have made it better.

For this sort of reason, I've generally felt resistance to framing idealistic endeavors in terms of trying to "have an impact," "make a difference," and so forth. Such framings seem like they don't have their eye on the true prize, which is always and only what happens in the world, rather than what you *cause* to happen. True, if you try and fail to cause something good to happen, but it happens anyway, that means you might've spent your resources elsewhere, and the world could've been even better as a result; but the ultimate problem there isn't that you didn't make a difference; it's that the world ended up worse than it could've been.

(The Global Priorities Institute paper also explores a number of other more practical differences that can arise between what they call "difference-making" and "standard" versions of consequentialism, once you bring in questions about risk and ambiguity aversion, but I won't get into those here).

That said, trying to keep one's accomplishment desire and one's good world desire strictly separate can seem too austere. For one thing, the accomplishment desire is often strong; demanding that it stay unmixed with your efforts to do good seems to imply either that it will need to be directed elsewhere, or stifled. Either way, your efforts to do good will be cut off from the energy the accomplishment desire provides — a deprivation that could be costly, especially given that cases like Bob's above only make a difference to Bob's attitudes, not his actions.

What's more, even though the accomplishment desire is necessarily in some sense self-oriented, not all manifestations of it seem objectionably self-concerned. Here I'm reminded of a suggestion in C.S. Lewis (I can't easily find the exact citation), to the effect that the truly virtuous person will spend great effort building a beautiful church, but then feel just as happy afterwards as they would if someone else had built it. This seems like a bit much to me: I think there are virtue-compatible ways of being proud of something you in particular have done, like building a church, that go beyond being glad that *someone* did it.

I don't have a great analysis of how to draw lines between more and less problematic forms of this pride, but here's a shot. Some forms of pride in one's actions seem

intuitively still humble — still oriented, ultimately, towards the good that was done, but grateful, as well, to have been able to be a part of it. These forms of pride place the good that was done first; it's the central focus, and the self-related meaning/accomplishment flows from there. In other cases, though, the good that was done seems more like an instrument of the meaning/accomplishment. What was wanted, first and foremost, was to *have done something* good; one does good, as it were, as a way of having done good, not as a way of increasing goodness. This seems to me more problematic.

I don't think I've really clarified this distinction; but I think the existence of a distinction at all gives me hope that there are virtuous options for the accomplishment desire other than "merge it with the good world desire," "stifle it," or "[purchase your accomplishments and your world-improvements separately.](#)"

Voting Phase for 2019 Review

Click [here](#) to begin voting. Click [here](#) for a general overview of the 2019 review.

We are now in the final stretches of the LessWrong 2019 Review. We've spent 2 weeks nominating posts, and a month reviewing them. And today, we begin the final vote!

A recap on the major goals of the Review:

- First, it is an experiment in improving the LessWrong community's longterm feedback and reward cycle.
- Second, it is an attempt to [build common knowledge](#) about the best ideas we've discovered on LessWrong.
- Third, after the vote, the LessWrong Team will compile the top posts into a physical book. (We may make some judgment* calls about what to include, but will use the vote as a strong guideline)

[Voting](#) starts today (January 12th) and continues through Jan 26th.

Who can vote?

All users registered before 2019 can vote, but the LW curation team will primarily be paying attention to the votes of users with 1000+ karma.

(There will essentially be a "top users' vote", and a "people's vote." The results of both will be made public, but the longterm-users' vote will be the main thing influencing the Best of 2019 Book)

Anyone can still write reviews

For all users and lurkers, regardless of karma, the next 2 weeks are your last opportunity to write reviews for any nominated posts in 2019, which can influence how people vote. As you can see below, all reviews are highlighted when a user is voting on a post. (To review a post, go to the post and click "Write A Review" at the top of the post.)

Posts need at least 1 review to appear in the vote. You can still review previously un-reviewed posts to put them on the ballot. (But, people who vote early might not see them)

The screenshot shows the LessWrong 2019 Review voting interface. At the top, there are navigation links for 'REPORT' and 'CONVERT TO QUADRATIC' with a right-pointing arrow. To the right are search, user profile, and notification icons. Below this is a table of posts with columns for title, karma threshold (No, Neutral, Good, Important, Crucial), and a row of buttons for 'I personally benefited from this post', 'Deserves followup work', 'Should be edited/improved', 'Important but shouldn't be in book', 'I spent 30+ minutes reviewing this in-depth', and 'Other...'. A text input field for writing a review is shown for the first post. Below the table, sections for '3 nominations' and '2 reviews' are displayed, each with a list of users and their comments. The interface uses a light gray background with blue and red highlights for interactive elements.

How do I vote?

To vote, head over to lesswrong.com/reviewVoting.

The vote has a simple first section, and a detailed-yet-optional second section based on [quadratic voting](#).

Sorting Posts Into Buckets

The first part of voting is sorting the nominated posts into buckets.

The five buckets are: **No, Neutral, Good, Important, Crucial**. Sort the posts however you think is best.

The key part is the *relative* weighting of different posts. For example, it won't make a difference to your final vote if you put every post in 'crucial' or every post in 'good'.

Fine-Tuning Your Votes

Once you're happy with your buckets, you can click 'Convert to Quadratic'. At this point the system converts your buckets roughly into their quadratic equivalents.

The system will only assign integer numbers of votes, which means that it will likely only allocate around 80-90% of the total votes available to you. If you vote on a smaller number of posts (<10), the automatic system may not use your entire quadratic voting budget.

If you're happy with how things look, you can just leave at this point, and your votes will be saved (you can come back any time before the vote closes to update them). But if you want to allocate 100% of your available votes, you'll likely need to do fine-tuning.

There are two key parts of quadratic voting you need to know:

First, you have a limited budget of votes.

Second, votes on a post have increasing marginal cost.

This means that your first vote costs 1 point, your second vote on that post costs 2 points, your third costs 3 points. Your nth vote costs n points.

You have 500 points to spend. You can see how many points you've spent at the top of the posts.

The system will automatically weight the buckets differently. For example, I just did this, and I got the following weightings:

- Good: 2 votes.
- Important: 4 votes.
- Crucial: 9 votes.
- Neutral: 0 votes.
- No: -4 votes.

(Note that negative votes cost the same as positive votes. The first negative vote costs 1 point, the second negative vote costs 2 points, etc.)

You'll see your score at the top of the page. (When I arrived on the fine-tuning page, the system had spent about 416 points, which meant I had a significant number of votes left to buy.)

Once you're happy with the balance, just close the page; your votes will be saved.

You can return to this page anytime until voting is over, to reconfigure your weights.

Quick Note Buttons

This year, in addition to full reviews, we're providing these "quick note" buttons.

I personally benefited from this post	Deserves followup work	Should be edited/improved
Important but shouldn't be in book	I spent 30+ minutes reviewing this in-depth	Other...

For each post, click any of these buttons if it describes your opinion of the post. These will be aggregated to get a sense of people's qualitative opinion on the post.

Happy Voting

That's it! Go forth and vote, and let us know if you have any questions, bug-fixes, or suggestions for next year!

* The LW Team reserves the right to make judgment calls about what actually goes in the book. Last year, we left off a couple posts that were too long, and included brief notes about a few lower-ranked posts that we felt were important. But, overall we'll be taking the vote as a strong indicator of what the LessWrong community thought was important.

For Better Commenting, Avoid PONDS



"All around the bog still sprawls, from out the drear lake come soulless thoughts and drift into the hearts of the people, and they are one with their surroundings."— Alan Garner, *The Weirdstone of Brisingamen*.

If our blogging is to be more than shouting into the void, we have to write good comments. [But that's hard to do.](#) Too much perfectionism, enforced positivity, criticism, status smackdowns, or vagueness in our commenting standards can all be problematic.

I'm using a framework that I think has been helpful. It's simple enough.

Avoid PONDS.

- P = Prickly
- O = Opaque
- N = Nitpicky
- D = Disengaged
- S = Shallow

Let me define each term.

Prickly means that your comment has a chilly, disapproving, mean, or unappreciative tone. It could hurt feelings, and make somebody feel dumb for opening their virtual mouth.

Opaque means that your comment makes assertions without backing them up. You're saying stuff without giving any reason at all for it. This can also include some types of lazy "questions" that are really meant as cheap shots. Even giving a *partial* reason or motivation for your comment or question means that it is *not* opaque.

Nitpicky means that your comment is expressing criticism of one particular part of an argument, without stating whether or how this local disagreement informs your view of the original argument as a whole. Even saying "this is just a local nitpick; I don't know if it means much for the argument as a whole" is enough to make your comment *not a nitpick*.

Disengaged means that your comment doesn't give the impression that you'd be likely to carry the conversation further if you received a response. It's a "drive-by commenting."

Shallow means that you didn't read the entire original post or comment thread to which you're responding.

Each category is meant to be a *very low bar*. Express even *mild* warmth, underlying reasoning, attempt at synthesis, display of engagement, or depth of consideration -- **not even all five, just one!** -- and it's not PONDS. This term is only for the very worst comments.

Comments that are even *slightly* better than PONDS are *totally acceptable*. This is a way of speaking and responding that gives almost total freedom for [Socratic grilling](#), while creating some minimal self-enforced safeguards to promote good conversation.

I've [build a habit](#) of checking my own comments to ensure they're not PONDS. It's not that hard to improve a comment to do better, or else to delete it. I also have a policy to never respond to PONDS. Instead, I heavily downvote them. In fact, they're the only type of comment that I heavily downvote.

See Willa's comment below for an even more in-depth treatment of the five anti-PONDS virtues. Thanks to Raemon, Richard Kenneway, and Christian KI for helpful discussion and suggestions.

COVID-19: home stretch and fourth wave Q&A

Robby Bensinger wrote this post and shared it in a private Facebook group, but it seemed good to also have it on LessWrong. Note that I have substantial disagreements with the current content of this document, since it currently seems to advise far too risk-averse of a strategy and makes claims of the type "if you don't fully lockdown you are basically giving up on not getting COVID" which strike me as very wrong. See thread on [this comment](#) for details.

Disclaimer: This document was made by non-experts. You may want to spot-check the cited sources to decide for yourself whether you think the reasoning makes sense. This document was created January 6, 2021, and may be out of date on some points.

Q: What's up with the new COVID-19 strain from southern England?

The strain, called VOC-202012/01 (or B.1.1.7 in [cov-lineages.org](#) nomenclature, 20B/501Y.V1 in [nextstrain.org](#) nomenclature), appears to be much more infectious than other COVID-19 strains. As of Dec. 31, Zvi Mowshowitz thought it was 80% likely the new strain is >50% more transmissible; as of Dec. 27, superforecaster Juan Gambeiro thought this was 65% likely. As of Dec. 31, infectious disease expert Trevor Bedford expected the new strain to [be about 50% more transmissible](#). I expect we'll get increasingly good estimates over the next few weeks.

The new strain ~~doesn't appear to cause worse symptoms~~ (update Jan. 22: it [may indeed cause worse symptoms](#)), but [as Zvi Mowshowitz noted on Dec. 24](#), if transmissibility is as high as it looks and vaccine rollout doesn't speed up dramatically, we should expect a massive fourth wave of infections in the US "likely cresting between March and May, that could be sufficiently powerful to substantially overshoot herd immunity".

"Overshooting herd immunity" means we achieve herd immunity in the space of a few weeks, with perhaps 60+% of all Americans getting sick; and then (because the total number of infectious people is so high) a large portion of the rest of the population gets infected too even though the virus's effective reproduction number R is much lower now.

Daniel Speyer's description of overshooting herd immunity in a short period of time:

"More people are infected than it would take to drive $r < 1$. More than 2/3 of the population, using the uniform $r=3$ model. This still leaves individuals the opportunity to go full-lockdown until vaccinated and avoid infection, but it may mean a few months of never leaving the house without a positive-pressure suit."

[From Zvi on Dec. 31:](#)

[...] The baseline scenario remains, in my mind, that the variant takes over some time in early spring, the control system [i.e., people's tendency to take more precautions when things look more dangerous] kicks in as a function of hospitalizations and deaths so with several weeks lag, and likely it runs out of power before it stabilizes things at all, and we blow past herd immunity relatively quickly combining that with our vaccination efforts.

[...] It seems likely we are looking at hundreds or low thousands [of people currently in the US with the UK strain], perhaps mid thousands, which puts the most likely start of the endgame some time in March if current trends continue and infectiousness is the full amount we suspect.

I think the evidence this past week has strongly favored the new strain being more infectious. What we lack is the knock-down info that would differentiate 50% more infectious from 65% more infectious[.]

Q: What should I do?

After discussing this with someone who's spent a lot of time studying COVID-19, we agreed that this is the decision point and it's basically all or nothing. Everyone needs to choose essentially one of three orientations:

- **The Protected** - You protect yourself now and during the crisis, whatever it takes.
- **The Switchers** - You don't protect yourself as much now, but you choose a threshold (e.g., a certain day of the calendar year, or a certain prevalence level of the new strain) at which you will shift to being in the first group if you haven't caught it.
- **The Unable/Unwilling** - You accept you're probably going to catch it (and in all likelihood at least soft prefer catching it now versus later).

Regardless of which group you're in, everyone who can should build up supplies again and start getting ready for an extended lockdown, possibly with supply line issues.

I think you should pick your basic strategy and make your plans and preparations now, or by mid-January at the latest. The downside of reacting too early is far smaller than the downside of reacting too late.

Given that vaccines are near at hand and (e.g.) the long-term health effects of COVID-19 aren't well-understood even for young people, I think locking down hard for 2-3 months beginning in late February or early March is probably the best option for most people who can take it.

I understand "locking down hard" to include things like "no interacting with people outside your household," "exclusively working from home," "no using public transit or ubers," and "no going to stores." You may even want to avoid going anywhere near passers-by outdoors, though I think the risk from walking past someone outdoors in a low-population area is *normally* (e.g., now in December) pretty negligible.

If you want to keep your risk especially low, Zvi notes:

[T]hose positive pressure suits are still for sale. I think they just work? As in, if you use the \$2,000 stuff, which could easily be a lot easier and cheaper than relocation [from the city to the countryside], you can [go] outside freely at almost no risk and all you have to worry about is the air circulation in your apartment.

Ordering packages from Amazon, paying others to go grocery shopping, etc. are under-used by the general public, I think, as good ways to reduce infection risk and cut down the overall spread of COVID-19. I heard a statement from a source in mid-2020 that grocery stores were the largest source of new infections in one part of the US -- many transmission chains are between store employees.

I think COVID-19 is to a very large extent transmitted indoors, and almost entirely via talking to someone face-to-face, or being in the same room as someone who coughs, shouts, sings, or sneezes. Almost all transmission happens via small droplets that hang in the air (a.k.a.

aerosols) and large droplets that quickly arc to the ground. I think people should put relatively little effort or attention into preventing surface transmission (a.k.a. fomites), which seems real but much rarer.

(The new strain seems to rely on the same vectors as other strains; its increased infectiousness seems to stem from increased viral load.)

Q: How should I decide whether it makes sense for me personally to lock down hard?

Review the best current estimates of COVID's health risks for your demographic, review your personal preferences and constraints (e.g., whether you're able to work remotely), and do an expected value calculation to figure out what makes sense for you personally. Your estimate should include a factor for the risk that if you get sick, you may infect others.

Q: What if I want to lock down hard, but also want to see friends who haven't been similarly locked down?

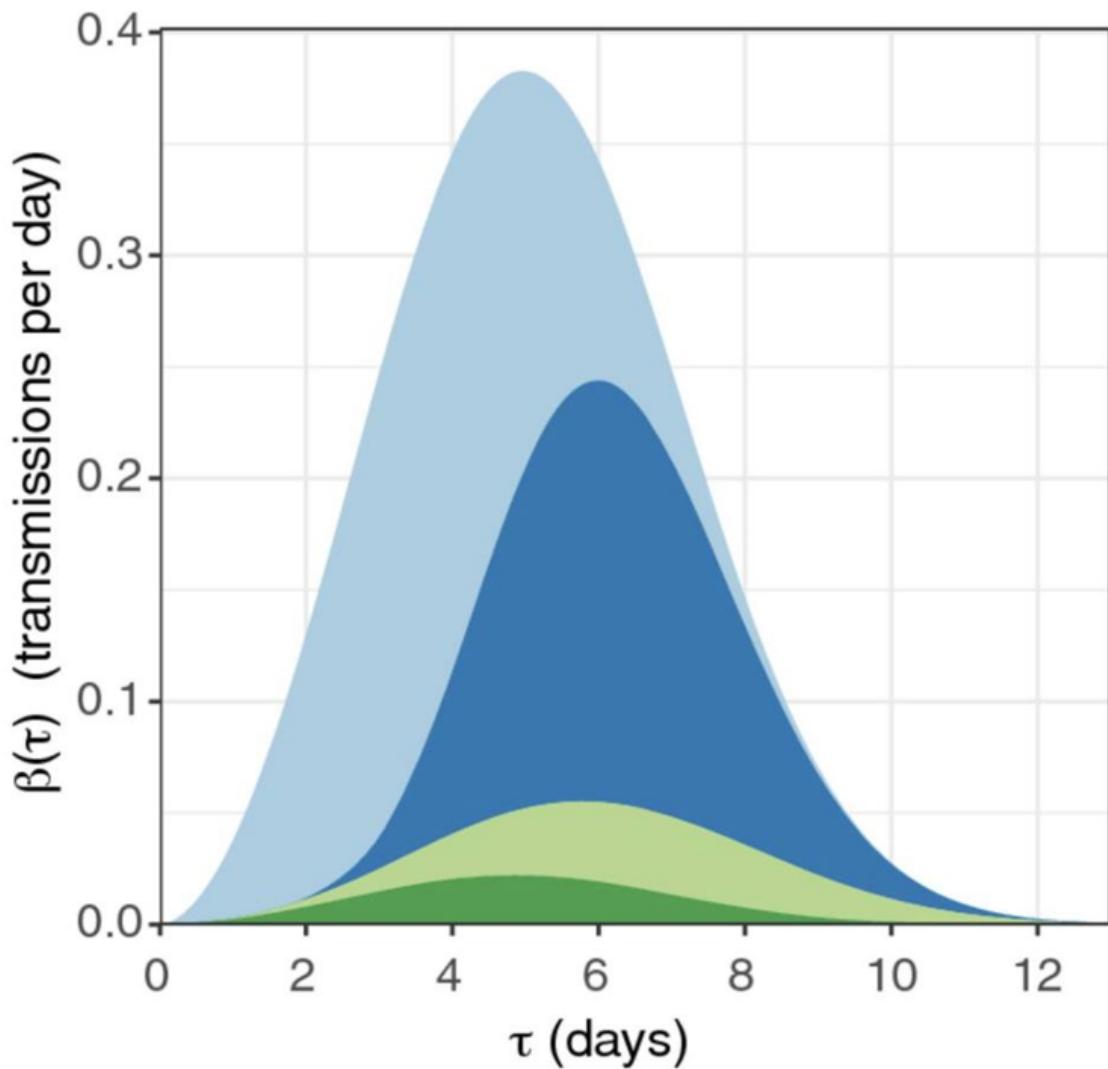
I think that if both parties lock down hard for two weeks, it's basically safe to meet up.

On the other hand, I think getting tested and then meeting once everyone gets a negative test result back basically *doesn't* work as a high-confidence way to keep transmission low. Viral COVID-19 tests have a pretty high false negative rate, common US tests tend to have a slow turn-around (so you may get infected after getting tested), and tests can miss infections if you take the test slightly too early or too late. (Though it's "you took your test too early" that's the main risk, since COVID-19 is most infectious early on.)

More specifically, Paul Bleicher and Marc Lipsitch [recommend getting tested 5-6 days after your last exposure](#), rather than (e.g.) 2-3 days. But Feretti et al. [think people are most infectious 2-8 days after they were infected](#), with some transmission occurring as early as day 0 or as late as day 12; and they treat 5.5 days as the mean time between infection and symptom onset, for people who develop symptoms at all.

$R_0 = 2.0$:

- $R_p = 0.9$ from pre-symptomatic
- $R_s = 0.8$ from symptomatic
- $R_e = 0.2$ from environmental
- $R_a = 0.1$ from asymptomatic



This is not to say that tests are useless, just that they're pretty mediocre evidence for this specific purpose.

Relying on symptoms alone to decide a meeting is safe is an even worse heuristic, since COVID-19 is at its most infectious shortly *before* symptoms start (and some people who spread the virus never show clear symptoms at all).

Note that this "wait two weeks" advice is specific to *people who want to lock down hard in preparation for an imminent large fourth wave*. I am definitely not advising that all people adopt a policy this strict at all times and places (e.g., when COVID rates are low).

Q: If most people in the US get infected all at once, will it really be possible to keep my own COVID risk low?

If you're able to work from home, and you're informed and conscientious, then yes. I once asked an acquaintance of mine who has spent a lot of time studying COVID-19 whether I should be paranoid about things like washing my hands, given how many people in many parts of the US were getting infected week-by-week. They replied:

My guess is you're too optimistic about how cautious typical people are being. If you think typical people are basically isolating, then current prevalence numbers look like "yikes! you can get sick even if you're mostly isolating!". I think that's the wrong model. There's some significant fraction of ordinary people doing [approximately] nothing to avoid infection. There are literal frat parties going on in Berkeley. Literally. People are going to work, and though they're masking while working at the cash register, they're taking their mask/gloves off in the break room with their colleagues and just doing whatever.

[...] I think you should model out an average Berkeley resident as spending multiple hours per week in public spaces [on public transit, at the grocery store, etc.]

In other words, for a shut-in like me who isn't interacting with people outside my quaranteam at all, I need to drastically adjust downward my intuitive estimate of how risky any given activity is, because COVID-19 is only infecting a small percentage of people every month *even though* large chunks of the population are doing plenty of in-person socializing and interacting. There's nothing virtuous about being over-cautious and burning value.

The new strain looks substantially more infectious, and relatively small increases in transmissibility can have a very large impact on exponential growth. But as Zvi notes:

One issue is that even if there's a lot of risk, it's *still* going to be highly concentrated in people not taking any real degree of precautions. It's still not going to be that easy to get this from walking down a stairwell or through a lobby or building vent - not a 0% chance, but I think that 'reasonable' precautions make one a solid favorite to get out clean even in the 85% infected scenario, because so few people will do even that, but it does mean not seeing anyone and all that.

My take-away here isn't "you can be relaxed and expect to avoid infection during a fourth wave." Rather, I take this as a warning against fatalism: it's not *that* hard to be one of the 20% (or for that matter 5%) safest people, if you're able to e.g. work from home.

Q: Is there any way to stop a large fourth wave from happening in the US?

There are three things I could imagine preventing a fourth wave, or sharply reducing its size:

1. The new strain could turn out to be less infectious than it currently seems. Zvi writes:

"I remain confident that we can probably mostly muddle through if there is small increased infectiousness (up to about 35%), at 50% I am highly skeptical we can stop this, and at 65% it's (probably) already over."

Early estimates of the strain's increased infectiousness tend to be significantly higher than 35% at present (as of Jan. 6), so I think this is possible but pretty unlikely.

2. Zvi's "control system": people could voluntarily respond to the increased risk by taking more precautions, thereby reducing the overall spread of COVID.

I think it's quite likely that this will be too delayed and too weak an effect. There has definitely seemed to be a control system like this in place in the US, [keeping death rates from climbing too high](#). The problem is that:

- People aren't likely to sharply change their behavior until they see large increases in hospitalizations and deaths.
- Hospitalizations and deaths lag behind infections by many days.
- The new strain is highly infectious, so multiple doublings may occur between "when a large subset of the population becomes infected" and "when we see a large enough spike in hospitalizations and deaths to change people's behavior".
- People's pandemic fatigue may result in a slower and/or weaker response than we would have seen a year ago.
- We already saw last January, February, and March that public health authorities, journalists, etc. are surprisingly bad at understanding exponential growth. Even if their credibility is still high enough to convince the public, it's very unlikely they'll understand and communicate the magnitude of the risk early enough to make a large difference.
- Unlike in Jan/Feb/March, we face the added difficulty that the dominant COVID-19 variant [can hide the new variant's spread](#). Overall COVID rates won't provide a good warning. Instead, you'll want to pay attention to data about how many COVID cases (in the US and in your area) are the old strain versus the new strain; and you'll need to complete all your preparations long *before* overall COVID rates start rising.

3. Widespread vaccination; extensive testing; an unprecedentedly large contact tracing effort and unprecedentedly heavy-handed quarantining and travel restrictions in affected parts of the US.

None of these currently seem likely to me -- not because any of this is a foregone conclusion or because we don't know how to solve this, but because the incompetence of the US' COVID response so far makes me very pessimistic that the obvious interventions will be carried out (at all, or in a reasonable or coordinated way).

The only option that I could imagine the US pulling off at this point is widespread rapid vaccination. But so far, almost every part of the US' vaccination plan has been terribly misconceived and very poorly implemented. In decreasing order of importance, I want to see changes like:

- giving emergency use authorization to the [AstraZeneca vaccine](#) (and accelerate the evaluation of other promising vaccines in the pipeline);
- giving everyone one dose of the vaccine before anyone gets a second dose, since this will do much more to slow the spread of the new strain;
- pouring as much money as possible into buying far more doses than needed and encouraging their speedy manufacture; and
- accelerating the vaccination process by removing red tape and focusing on speed above all else.

Patrick McKenzie has an excellent [list of suggestions for accelerating vaccination](#). [Matt Kilcoyne's plan](#) was developed for the UK, but I expect it to have transferable ideas.

Q: Should I expect our current vaccines to be effective against the new UK strain?

Yes! Mutations that reduce vaccine effectiveness are rare in most viruses. (Influenza is an unusual exception.)

Q: Are these vaccines safe?

Yes, assuming you've never had a strong anaphylactic reaction to a previous vaccine.

I plan to get vaccinated ASAP once rollout reaches my demographic, and I don't think of this as a risky action.

Q: When will I be able to get vaccinated?

(Added Jan. 18)

If you're 65+ years old or a health care worker, likely now or very soon. Check rollout plans for your state. Note that as of Jan. 18, people in a number of states [have been having success](#) getting vaccinated early, and rescuing vaccines from being thrown out in the process, by showing up half an hour before vaccination sites finish for the day.

Zvi Mowshowitz predicted on Jan. 6 "about 38 million vaccinated by April 1, 81 million by July 1, 131 million by October 1". [As of Jan. 14](#), the aggregate forecast on the Hypermind prediction market was 65 million by April 1, 128 million by July 1, 184 million by October 1.

For comparison, based on a quick googling, it looks like there are ~18 million American health care workers and ~55 million Americans age 65+. And as of Jan. 14, we had vaccinated [~11 million Americans](#), including ~4.5 million in the last week.

This suggests that it may be difficult for a lot of non-health-care-workers age 65+ to get vaccinated before a likely fourth wave hits in March, in spite of the new directions from the CDC to prioritize a rapid roll-out to everyone 65+. Zvi Mowshowitz [comments](#):

I hope [Hypermind's] predictions are right, or better yet pessimistic. The difference in those guesses is a really big deal. If we take my model's guess and boost vaccinations an additional 50%, and the additional infectiousness of the new strain is only 50%, we plausibly (mostly) stop the fourth wave, despite the new strain currently being ahead of my previous guess for that. I've added knobs in the spreadsheet to vary both numbers.

Under the optimistic scenario, this increased vaccination pace would be the difference between ending with 40% or 50% of the population having had Covid-19 when it's all over, with about 24% having already been infected. So this would prevent over a third of all future infections. Speed matters more than anything.

There are 209 million Americans over the age of 18. Given how many don't want to be vaccinated, these predictions above imply that by April 1 you can likely get the vaccine as long as you care about finding it, and by July 1 you can get the vaccine provided you want it at all, as there would be enough vaccinations to cover every adult who wants one at all.

Q: Will it be safe to get vaccinated during a massive spike in infections?

It's possible to get infected when you leave your lockdown to go get vaccinated. The first dose [confers zero protection until a week or two after you get it.](#)

On the other hand, even if there's a massive ongoing fourth wave of infections at the time, it seems easy to make the risk from getting vaccinated much lower than the benefits, by calling in advance to make sure the vaccination area is uncrowded and following reasonable safety protocols, and by wearing protective equipment such as an N-95 mask and a face shield. Zvi tells me:

It would take a very extreme situation before getting a vaccine was so actively risky you couldn't do it - we're talking (hopefully) about a quick trip to CVS, which isn't nothing, but assuming they understand the issues you should be able to be in+out in a few minutes with generally good protocols, and potentially they can do it all outdoors. Yes, entering+leaving the apartment isn't a free action, but turning down the vaccine because of this seems unlikely to me."

Getting your first dose of the vaccine seems to confer ~80% protection (after a few weeks), while the second dose confers ~95% protection (again, after a few weeks).

Q: Should I try to get infected now?

I can imagine two reasons to want to do this:

(1) You might be worried that hospitals will be overloaded during a large fourth wave. If hospitals in your area aren't currently overwhelmed (and aren't likely to be overwhelmed in the next few weeks), then you might receive better medical care now than you would in March/April/May.

(2) You might think the risk of getting infected with a large viral load is higher later. If you get infected with a small viral load now, you're likely to have milder symptoms.

On the other hand, it's pretty hard to guarantee you're getting a small viral load; if you try to get infected, you'll need to lock down hard (even if you didn't succeed) to avoid exposing others to extra risk; and the benefits of trying to variolate at this point are much smaller, now that you have the potential to get vaccinated in the next few months.

Guaranteeing you get COVID probably requires that you expose yourself to a large viral load, which is more dangerous. But if you shoot for a small viral load instead, the likeliest outcome is that you need to self-quarantine for weeks as a precautionary measure, only to find that you didn't end up infected at all.

I don't want to claim that trying to catch COVID early is definitely a bad idea for everyone, since I'm not an expert and people's life circumstances vary a lot. The #1 thing I'd advise is

to avoid any action like this that might unintentionally infect others.

Q: What should I do, then, if I'm not willing or able to lock down hard?

Stock up on delicious non-perishable food so you can make your grocery runs in March/April shorter and less frequent. (Or avoid grocery shopping altogether, and spend your microcovid on more valuable things. Tool and explainer for thinking about microcovid: <https://www.microcovid.org/>)

Front-load your activity – if you can lock down slightly harder later by being less cautious now, then this is probably a good trade.

Prioritize activities that are short and have low exposure risk. These are the cheapest activities in general, in terms of microcovid; and as a bonus, if you do get infected in this way, the viral load is likely to be smaller. Additionally, these activities reduce the risk you'll infect others (or expose them to a higher viral load) if you're asymptotically or presymptomatically sick. If you're doing something like this, take periodic home tests to check whether you're sick, and avoid older or otherwise at-risk people.

Everything Okay

Okay is a key concept for which we lack a good detailed handle.

There seems to be some sort of switch in brains, certainly in mine, that goes between modes of Okay and Not Okay. Being in Not Okay Mode is a giant ball of stress that makes it impossible to relax and demands attention.

When I hear talk that some problem exists, my first thought often has nothing to do with the substance of the problem. Instead, I feel instantly stressed and upset that someone has taken my perfectly pleasant Okay Mode existence, and declared that we are in Not Okay Mode.

Even if the real content of the problem is trivial.

There is an intense desire to get back to Okay Mode.

Even if that means one must be in Okay Mode by making Not Okay be Okay. Even if that means destroying all possible superior outcomes so that the bad outcome is now seen as acceptable.

Thus:

"Tell me everything is going to be okay."

"Everything is going to be okay."

What does that even mean?

It is a request for permission to enter or remain in Okay Mode. That requires unpacking.

When one says this, for what value of 'everything' and what value of 'okay' is this meant?

Does the future tense matter here? If everything is *going to be okay*, is everything also by implication okay now, or not? If we have to ask about whether things are going to be okay in the future, does that imply they are *definitely not okay* now?

Here are some possibilities that I think are central at least *some* of the time for what the question means.

A: Is 'everything' 'okay' in a way that is potentially meaningful in general but meaningless in a given context? The way that someone who experiences [Kensho](#) could say that everything was okay regardless of the current world state and value of everything, or that one might say that God has a plan? How central a feature of the basket of things called 'enlightenment' is the ability to label everything as 'okay' in a way your emotions will accept, and avoid the consequences of failing to do that?

B: Is 'everything' 'okay' in a way that is meaningful in a more specific context, but is a relatively weak claim meant to provide perspective? As in, yeah, I understand that sucks a lot, but look at the bigger picture and the things that actually matter are still 'okay'. You spilled some milk, and the milk isn't okay, but the relevant value of 'everything' does not need to include the milk, because the milk is not important. We

still have plenty to eat, and/or anything up to and including it not changing the probability that AGI will be friendly.

C: Is ‘everything’ ‘okay’ in the sense that the dreaded stress ball of uncertainty will be resolved, and peace will be made with the ways in which things are definitely not okay?

D: Or, perhaps, is ‘everything’ ‘okay’ in the sense that you have put a [Somebody Else’s Problem](#) field around it so the whole thing can be ignored, which definitely resolves the stress ball brought about by uncertainty about whether something is in a state of okay or not? This is then assurance that, while things might be really awesomely terrible, no action is required or anticipated *from you* at this time – you don’t have to enter Not Okay mode, either in the sense of sending appropriate social signals for such a state, or in terms of actually fixing anything.

E: Or, similarly, a statement that no blame or punishment need be assigned for the current state of existence, and no scapegoat need be chosen.

F: Or, similarly, a request for an ‘I got this, you can treat this as having been handled’?

G: This could be a general question about whether to pay attention and/or change intentions towards future actions. Most people spend most time on autopilot, and this is a check to see if the autopilots need to come off. One could view D or also F as this applied only to oneself, whereas this is a request to confirm that *all* autopilots can continue to function normally. This is no one’s problem, and no one has to ‘get this.’ [There is no need for anyone to suddenly be a person.](#)

H: This could be a request to implicitly define ‘everything.’ There is something *or someone or some group* that is clearly not okay. Does it count? Do we care? Is it part of everything, or not?

I: This could be a request to implicitly define ‘okay’ in context.

J: This could be an attack. The respondent now has two choices.

If they affirm the state of okayness, several bad things happen to them. The two biggest are:

First, now it has been established that all the things that were previously wrong were ‘okay’ so they can’t go back and complain about them or blame anyone else for them later. By extension, other similar things that happen in the future are implicitly also okay, and perhaps the range of such things will now expand.

Second, they have affirmed that everything is okay, which means that if things turn out not to be, whether for them or for other people, that’s now their fault and they are blameworthy, and the claim will be made that we were in case F and then they didn’t deliver, but also other cases can also be invoked similarly.

Alternatively, they could deny okayness, and then *different* bad things happen to them.

First, they have now disturbed flow and made things socially awkward, to at least some extent. They have shown that they are a complainer, a person who is not okay

with things, not comfortable. They are then seen as the aggressor, the one causing trouble and escalating. Bad improv. They will lose points.

Second, not only have they admitted they know there is a problem, [they now have interacted with the problem](#). Oh no. All sorts of blame is now in play. All sorts of additional demands for explicitness can be made, and being explicit costs even more points. Exactly what is and isn't okay, and exactly what counts as everything, are now among the questions that they might have to answer. So is how one might make things not okay.

Third, it is now on them to propose a way to make things okay again. Otherwise, who are they to claim that things are not okay without proposing a solution? And of course, if a solution is proposed, then the solution can be attacked and ridiculed.

Thus, the person asking if everything is okay is saying, either wave your right to complain and potentially be assigned blame if things go wrong, or make a stand and have it out now, while being marked as the one making things awkward and causing a problem, forcing us to be in Not Okay Mode, since otherwise everyone could have agreed to be in Okay Mode.

K: This could be a request for actual information about the state of the world. There is a set of things, everything, that can be in various states. How are those states? The better defined both 'everything' and 'okay' are in context, the more useful the question becomes. Note the distinction between asking about what social or other actions might be required, and asking purely for the world state.

L: This can be a demand for certainty. For choices and commitments – any choice and any commitment – to be made now in order to reduce the stress of the choices not having been made. This can happen for matters both great and small. If there are two possible futures, and both have big advantages and disadvantages, then one cannot be assured that any of the realms involved is "okay" because there is a trade-off and a different choice might get made. Thus, anything that could get worse or better or merely be different somehow is now Not Okay. Thus, even if there are large unknowns or there are otherwise huge advantages to not committing to a choice and no actual reason why a choice needs to be made yet, even if there is important information that will only be available later, optionality might have to be sacrificed on this alter.

M: It can be an accusation. You are pretending as if everything is ok, but I do not believe this, and you had better admit whatever it is that is not okay, or else.

N: It can be an assertion that someone should not objectively be Okay, and should instead be Not Okay, especially when combined with terms like "Are you sure that..."

O: It can be a social script for generic meaningless reassurance, which may be a request for an *enactive* statement rather than something descriptive. 'Everything' and 'okay' need not have any values at all, [because too many words don't have meanings](#). Interpreting this as a claim that there exists a set of things that is being claimed to be in a particular state, or that will be in a particular state, would be a pure category error. Saying this merely builds momentum towards the Okay state.

P: This could be an establishment of dominance and submission. The exchange reminds both parties who decides whether things are Okay or Not Okay, granting power over the other's emotional state, which can be used as power generally. And, of course, that power is now okay.

Q: This could be a debate about the ‘right to complain.’ Complaining can be about information and moving towards solutions or better understanding, but it can also (more frequently) be about scoring points, or about needing to feel heard, or punishing the scapegoat. Saying everything is okay in this mode is therefore an assertion that there is no right to complain, whereas saying that everything is not okay is both itself a complaint and a prelude to more complaining. Asking if everything is okay can be an invitation that this is a good time to complain (either something might be done, or you might get a sympathetic ear, or if you gotta do it all right let’s get it over with, or speak now or forever hold your peace) or could be the opposite.

R: Or perhaps is ‘everything’ ‘okay’ because actually yeah, actual everything is actually okay now?

Primary Meanings Chart

What does Everything Mean?

All of existence	ABR
Your desire to be Okay	CDP
Your desire to avoid blame	E
Some meaningful local issue, or you	FGKLMNQ
You tell me	H
Could be anything	IJ
Is meaningless	O

What does Okay Mean?

Not requiring stress	ABCLNO
Not requiring action	DFGJK
Not requiring blame	EMPQ
Could be anything	H
You tell me	I
All of it	R

There are a lot of these that combine multiple meanings, or can mean different ones in different ways. In some cases it would be reasonable to disagree with which meaning is the locally primary one, but this is an attempt to group them together.

Different contexts have different defaults and distributions of possible explanations. The D/F/G/K cluster of related interpretations is closest to my default generic assumption, with the differences between those three often being important. Also note that one can respond with a different type of response than what was requested, either intentionally or unintentionally, and make this clear through other words or even through tone, subtly (for example) transforming a ‘nothing need be done’ into ‘I got this’ or vice versa.

These groupings suggest a few categories.

A, B and R represent the cluster of universalized claims about all of existence. That seems central to these. In other ways they are mostly part of the second category.

C, I, L, N and O are treating okay as not stressing out about local concerns, and debating whether or not one has earned the right to be in Okay Mode or not.

D, F, G and K, are attempts to determine if action is necessary, and if so who must take what action, with varying levels of clarity creation versus emotional needs and blame avoidance.

E, M, P and Q are focused on assigning blame and on social roles.

J is an attack, which puts it in its own category.

That leaves H, which depending on what 'okay' means could be part of any of the middle three categories, sufficiently that I didn't want to place it in any of them, but it falls into some combination of those.

The big two groups are those where the concern is ability to stress or avoid stress, and those where the issue is where it is necessary to take action or assign blame, with J as perhaps a special super-case in the second category.

A major motivating factor for creating this taxonomy is how easy it is for stress concerns to override non-stress concerns, and for the dominant dynamic to become what will allow people's brains to give themselves permission to shift back into Okay. In this model, that is what drives any action taken, or any blame assigned.

What else can 'okay' and 'everything' mean, in what combinations? To what extent can we profit by tabooing the word okay, or most of these uses for it? What is the best way to reliably get the benefits of 'okay' in avoiding giant stress balls, while still retaining the motivation to act and address problems or opportunities? How should one react to those who are primarily optimizing for being in Okay Mode at the expense of other concerns, or those who are using Okay as a weapon?

Unpopularity of efficiency

I feel like ‘efficiency’ is often scowled at. It is associated with factories and killing and commercialization, and people who are no fun. Things are openly criticized for being oriented toward efficiency. Nobody hopes to give their children an efficient childhood or asks for an efficient Valentine’s day, unless they want to get it over with. I expect wariness in listeners at talk of efficient charity.

This intrigues me, because in what I take to be its explicit definition, ‘efficiency’ is almost the definition of goodness manifest. The efficiency of a process is the rate with which it turns what you have into what you want.

I usually wince when people criticize efficiency, and think they are confused and should be criticizing the goal that is being pursued efficiently. Which does seem basically always true. For instance, if they are saying their childcare center cares only for efficiency, they probably mean that it is doing something like trying to minimize financial costs without breaking the law. Perhaps by fitting many children into a room with minimal oversight or attention to thriving. Here, I would complain that the childcare center cares only about its profits and not breaking the law. If it was fulfilling my own values efficiently, that would be awesome.

However I think there is more merit to efficiency’s poor reputation than I have given credit for. Because pursuing efficiency does seem to systematically lead to leaving things out. Which I suppose perhaps makes sense, for creatures who don’t explicitly know what their values are, and especially who have trouble quantifying them. If you set out to build an efficient daycare center, chances are that you don’t know exactly what makes a daycare center good, and are even less well equipped to put these things into numbers and build machinery to optimize those numbers. (This would be much like the AI alignment problem but where the AI you are trying to direct is made of your own explicit reasoning. It might also what [Seeing Like a State](#) is about, but I haven’t read it.) It’s still not clear to me why this would systematically turn out actively worse than if you didn’t aim for efficiency, or whether it does (my guess is that it usually doesn’t, but sometimes does, and is notable on those occasions). If efficiency has really earned its poor reputation, I wonder if I should be more worried about this.

Grokkling illusionism

(Cross-posted from [Hands and Cities](#))

A number of people I know are illusionists about consciousness: that is, they think that the way consciousness seems to us involves some fundamental misrepresentation. On an extreme version of this view (which [Frankish \(2016\)](#) calls “strong illusionism”), phenomenal consciousness simply does not exist; it only seems to exist (I’ll say more about what I mean by phenomenal consciousness in a moment). I’m especially interested in this version.

For a long time, though, I’ve found it hard to really grok what it would *be* for strong illusionism to be true. I can repeat the words illusionists say; but I haven’t had a clear sense of the reality envisioned, such that I could really look at the world through the illusionist’s eyes. What’s more, I’ve suspected that some sort of barrier in this respect is crucial to the resistance that I (and I expect many others) feel to the view.

Successfully imagining illusionism being true, I think, may be halfway to believing it.

(As a sidenote: I think this dynamic may be common. Actually looking at the world the way someone you disagree with looks at it is often much more difficult than being able to pass their “intellectual turing test” — e.g., to present their position in terms that they would endorse. As ever, words are easy; seeing the world in new ways is hard. And once you have seen the world in a new way, the possibility that the world *actually is that way* is much easier to take seriously.)

The aim of this post is to grok illusionism more fully. Let’s start with a few clarifications.

The philosophical debate about consciousness centers on “phenomenal consciousness,” which is generally thought of as the thing we ascribe to a system when we say that there is “something it’s like” to be that system, or when we ascribe to that system a first-person perspective or subjective experience. And experiences themselves — the taste of wine, the smell of leaves, the color of an afterimage in your visual field — are thought of as “phenomenally conscious” when there’s something it’s like to have them, and the “phenomenal properties” of experiences determine (consist in?) what it’s like to have them.

Phenomenal consciousness is often contrasted with “access consciousness,” understood as a property of the mental states that a subject can do certain things with (e.g., “access”) — in particular, notice them, report on them, reason about them, etc.

People have various intuitions about phenomenal consciousness often thought difficult to validate using a standard physicalist conception of the universe. [Chalmers \(2018\)](#) offers a helpful taxonomy:

- *Explanatory intuitions*: these are intuitions to the effect that certain familiar modes of physical and functional explanation are unable in principle to explain phenomenal consciousness.
- *Metaphysical intuitions*: intuitions about the metaphysical status of phenomenal consciousness. For example, intuitions that phenomenal consciousness is not a physical phenomenon, or that it is in some sense simple. Chalmers doesn’t say so, but I’ll assume that intuitions to the effect that consciousness is e.g. unified

(different conscious experiences arise in a single unified mental “space”) or binary (you have it or you don’t) would also fall under this bucket.

- *Knowledge intuitions*: Intuitions about the type of knowledge it’s possible to have about phenomenal consciousness — for example, intuitions to the effect that a neuroscientist raised in a black and white room and who knows all the physical facts about color vision learns something new when she sees red for the first time; and intuitions to the effect that it is difficult or impossible to know, from a third-person perspective, whether a given system is phenomenally conscious, or what its phenomenal consciousness is like, even granted arbitrary amounts of physical knowledge and understanding.
- *Modal intuitions*: These are intuitions about what sorts of scenarios involving phenomenal consciousness are *possible*. For example, you might think it possible that despite their behavior, other people are not conscious (even though you are); or that what other people call “blue” actually looks to them like red looks to you (and vice versa); or that there could be a physical duplicate of our world consisting entirely of creatures that don’t have phenomenal consciousness (“phenomenal zombies”).

Some theorists argue, from intuitions of this kind and other considerations, that a standard physicalist conception of the universe requires revision. Others resist such a revision.

Illusionists are definitely in the latter category. Where illusionism differs from other physicalist theories, however, is somewhat harder to pin down. Broadly speaking, illusionism is more willing to claim that the way phenomenal consciousness seems to us involves some fundamental aspect of misrepresentation, whereas other theories hold out more hope that various ways things seem to us might come out true. But because illusionism and other physicalists theories share a fundamental physicalist metaphysic, however, the distinction between them comes down primarily to a debate, not about how things fundamentally are, but about how they seem (or, alternatively, which properties something must have, in order to count as phenomenal consciousness, vs. something else). In this respect, the distinction is much less interesting than the distinction between physicalist and non-physicalists theories more broadly.

This is a familiar dialectic in philosophical debates about whether some domain X can be reduced to Y (meta-ethics is a salient comparison to me). The anti-reductionist (A) will argue that our core intuitions/concepts/practices related to X make clear that it cannot be reduced to Y, and that since X must exist (as we intuitively think it does), we should expand our metaphysics to include more than Y. The reductionist (R) will argue that X can in fact be reduced to Y, and that this is compatible with our intuitions/concepts/everyday practices with respect to X, and hence that X exists but it’s nothing over and above Y. The nihilist (N), by contrast, agrees with A that it follows from our intuitions/concepts/practices related to X that it cannot be reduced to Y, but agrees with D that there is in fact nothing over and above Y, and so concludes that there is no X, and that our intuitions/concepts/practices related to X are correspondingly misguided. Here, the disagreement between A vs. R/N is about whether more than Y exists; the disagreement between R vs. A/N is about whether a world of only Y “counts” as a world with X. This latter often begins to seem a matter of terminology; the substantive questions have already been settled.

My sense is that the distinction between what Frankish calls “weak” and “strong” illusionism may turn out to be largely terminological as well. Frankish characterizes weak illusionism as admitting that phenomenal consciousness exists, but claiming

that we misrepresent it as having certain metaphysically suspicious features — such as being ineffable, intrinsic, essentially private, or infallibly-known — that it doesn't possess. Strong illusionism, by contrast, denies that phenomenal consciousness exists altogether. But it's not clear to me what's at stake in the difference between admitting phenomenal consciousness exists, but does not have X, Y, and Z features, vs. saying that it doesn't exist at all, unless we can say more about the features that, according to weak illusionists, it *does* have. Frankish, here, mostly says that weak illusionists still allow that experiences have properties that are "genuinely qualitative" and "feely," and in that sense phenomenal; claims which strong illusionists deny. But it's very unclear to me, absent further positive characterization, what "qualitativeness" and "feely-ness" amount to (I think Frankish talks about this "[Quining Diet Qualia](#)," which I haven't read).

Despite the purported strength of his illusionism, though, Frankish himself does a few terminological dances to avoid baldly endorsing claims like "there's nothing it's like to be you," and "you are a phenomenal zombie." He is committed to the non-existence of phenomenal consciousness, but he says that we need not construe talk about "what it's like" or of "phenomenal zombies" as essentially about phenomenal consciousness. For example, we might think of there being "something it's like" to have a certain experience if that experience is represented introspectively in some way; and we might think of zombies as essentially lacking in *this* type of introspective access to their mental states — what Frankish calls an "inner life." We aren't zombies like *that*, says Frankish.

I think Frankish is squirming a bit here, and that he should bite the relevant bullets more forthrightly (though to his credit, he's still reasonably up front). No one ever thought that phenomenal zombies lacked introspective access to their own mental states, since they were by hypothesis functionally identical to humans; and the central function of "what it's like" talk in the discourse about consciousness has been to point to/characterize phenomenal consciousness.

Let's consider, then, the more forthright version of strong illusionism, which just states directly that phenomenal consciousness does not exist; there's nothing it's like to be you, or a bat, or your partner; there's never been anything it's like to be anyone; there's nothing it's like to see green, or to feel pain, or to fall in love. You used to think a zombie world was merely possible; actually, it's actual. The lights have never been on. No one has ever been home.

Can you conceive of this? Can you take seriously the possibility that this might, actually, be true?

In attempting this, the shift I've found most helpful is actively and deliberately moving from conceiving of subjective experience as a *thing* — a "space" or "experiential array" that you have some sort of "direct acquaintance" relationship with — to conceiving of it as the content of a *story*, as a way things are represented to be. Less like the canvas and paint of a painting, and the more like what the painting is supposed to be *of*; less like a newspaper, and more like the news. And the news, as we all know, might be oversimplified, partly false, or entirely fake.

Suppose, for example, that after fixating your vision on a black, green, and yellow image of an American flag, you are left with an "after-image" of a red stripe when the flag stimulus is removed (this is a favorite example of Daniel Dennett's). It's tempting to think that there is *something* that has the property of phenomenal redness — that is, an appearance of a red stripe, where that appearance is itself red, in your "internal

space” or “experiential array” — and that it is this something that you direct your attention to in noticing the after-image. On the view illusionists like Dennett and Frankish are encouraging, though, what’s happening is that *according to the story your brain is telling*, there is a stripe with a certain type of property. That’s the sense in which it seems to you like there’s a red stripe; that’s all that the appearance of the red stripe amounts to, and this does not require an actual red stripe made out of mental stuff, painted in [mental paint](#) (Dennett calls it “figment”) in your internal world.

Here’s [Frankish’s \(2019\)](#) more comprehensive version of this picture (it’s not the only version available, but my impression is that many illusionist accounts proceed on similar lines). Your brain engages in processes associated, for Frankish, with “access consciousness” — e.g., acquiring information and synthesizing information about the environment, and then “broadcasting” that information to the brain as a whole, such that it can enter into further processes like reasoning and decision-making. Beyond this, though, it also uses introspective mechanisms to track the processes involved in access consciousness and represent them using a simplified model — a model which can then itself feed into other cognitive processes like decision-making, memory storage, and so on. Importantly, though, this simplified model involves representing some things (maybe mental states, maybe objects in the world) as having properties they don’t have — specifically, phenomenal properties. And it is this false representation that gives rise to problematic intuitions like the ones described above.

Frankish is openly hazy about exactly what it is to represent a property as phenomenal, and about the specific mechanisms via which such representations give rise to the problematic intuitions in question — this, he thinks, is a matter for further investigation. But the basic outlines of the eventual story are, for him, relatively clear, and the project of filling them out is much more promising, he thinks, than the project of trying to validate the intuitions in question, whether via physicalist or non-physicalists means.

Because this account is more of a promissory note than a developed theory, it doesn’t provide a ton of content to aid in constructing an illusionist model of how your mind works. Still, I think, shifting to thinking of your subjective experience as the content of a story — not the Cartesian theatre; but the plot of the film — seems to me a basic and instructive first step.

(Note that we can make the same move, not just about phenomenally conscious experiences, but about the self that experiences them. The basic picture is: there is a physical machine controlled by a brain, it contains representations that purport to describe a self, a set of mental states, and an external world, all with various properties; according to these representations, the self is situated at the center of a unified internal arena or space in which sights, sounds, etc with phenomenal properties appear and disappear. To the extent that we end up thinking of the properties of these mental states as illusory, we may end up thinking of properties of the “self” that is represented as experiencing them as illusory as well.)

When I try to see the world like this, I find myself shifting from an implicit frame in which I am the “consumer” of my brain’s story about the world — a consumer who *uses* that story as a map — to one in which I am fully *engrossed* in that story, fully *in* the world that the story portrays. [Shakespeare](#) writes: “Think when we talk of horses, that you see them; printing their proud hoofs i’ the receiving earth.” On illusionism, I continue to take this advice, applied to qualia, very much to heart: “Think, when your brain talks of phenomenal redness, that you see it.” Oh, but I do. I look at my desktop

background, and there is the phenomenal redness, shining vividly. And indeed it is, says illusionism, *in the fictional world your brain is representing*. Quite a fiction, I find myself thinking; very engrossing; feels so real; feels like I'm there.

Indeed, a part of me is tempted to say that this fictional world, in which things have phenomenal redness, is *my* world, and that I am more deeply identified with the "self" in this world than with the organism and brain that houses the mental states representing it. Perhaps, in this sense, "I" will end up as fictional/illusory as the phenomenal redness I take myself to be perceiving. I'm tempted towards this view in part because the fictional world is where, as it were, the phenomenal red lives; in this sense, the fictional self in the fictional world is *right* about its perceiving the (fictional) phenomenal red, though wrong to treat the fictional world as real. And being right about something that seems as obvious as the phenomenal red seems like a real benefit.

But a part of me pulls the other direction. On this view, I'm the organism/brain, using a flawed map to navigate a real territory. Phenomenal properties, it turns out, are a flaw in the map — a particularly compelling and unusual flaw, but familiar in a broader sense, and not, perhaps, particularly harmful outside of philosophy seminars (though my best guess is actually that accepting illusionism would have very revisionary implications in a variety of domains, especially ethics). This, I think, is where most illusionists end up; and it seems the more sensible route.

What currents of thought on LessWrong do you want to see distilled?

The question is inspired by a few comments and a question I have seen recently. The first is a [discussion in the 2019 Review](#) post on the subject of research debt; the second a question from johnswentworth [asking what people's confusions are about simulacra](#) (which I interpret to be a 'what do you want from this distillation' question).

The question is what it says in the title, but I would like to add that there is no expiration. For example, I recently saw cryogenics back in the posts and questions, which had fallen off the activity radar for years. So old currents of thought are valid candidates, even if the real goal is a re-distillation in light of new developments in the field or all the accumulated communication technique we've considered on LessWrong.

So please describe the current of thought, and your reason for wanting a distillation. The authors may be called to action, or alternatively following [Bridgett Kay's suggestion](#) someone else may take up the challenge.