

Best of LessWrong: April 2021

1. [Notes from "Don't Shoot the Dog"](#)
2. [Another \(outer\) alignment failure story](#)
3. [Announcing the Alignment Research Center](#)
4. [The Case for Extreme Vaccine Effectiveness](#)
5. [Predictive Coding has been Unified with Backpropagation](#)
6. [Testing The Natural Abstraction Hypothesis: Project Intro](#)
7. [AMA: Paul Christiano, alignment researcher](#)
8. [Specializing in Problems We Don't Understand](#)
9. [I'm from a parallel Earth with much higher coordination: AMA](#)
10. ["AI and Compute" trend isn't predictive of what is happening](#)
11. [Opinions on Interpretable Machine Learning and 70 Summaries of Recent Papers](#)
12. [Why We Launched LessWrong.SubStack](#)
13. [Why has nuclear power been a flop?](#)
14. [Tales from Prediction Markets](#)
15. [Monastery and Throne](#)
16. [How to Play a Support Role in Research Conversations](#)
17. [Covid 4/22: Crisis in India](#)
18. [The irrelevance of test scores is greatly exaggerated](#)
19. [What are all these children doing in my ponds?](#)
20. [A new acausal trading platform: RobinShould](#)
21. [Gradations of Inner Alignment Obstacles](#)
22. [\[Letter\] Advice for High School #1](#)
23. [Updating the Lottery Ticket Hypothesis](#)
24. ["Taking your environment as object" vs "Being subject to your environment"](#)
25. [Covid 4/15: Are We Seriously Doing This Again](#)
26. [People Will Listen](#)
27. [A Brief Review of Current and Near-Future Methods of Genetic Engineering](#)
28. [Highlights from The Autobiography of Andrew Carnegie](#)
29. [Wanting to Succeed on Every Metric Presented](#)
30. [Iterated Trust Kickstarters](#)
31. [My take on Michael Littman on "The HCI of HAI"](#)
32. [Draft report on existential risk from power-seeking AI](#)
33. [April drafts](#)
34. [FAQ: Advice for AI Alignment Researchers](#)
35. [Facebook is Simulacra Level 3, Andreessen is Level 4](#)
36. [Against "Context-Free Integrity"](#)
37. [The secret of Wikipedia's success](#)
38. [We need a career path for invention](#)
39. [Three reasons to expect long AI timelines](#)
40. [Center for Applied Postrationality: An Update](#)
41. [Alignment Newsletter Three Year Retrospective](#)
42. [Scott Alexander 2021 Predictions: Market Prices](#)
43. [Vim](#)
44. [Covid 4/1: Vaccine Passports](#)
45. [Agents Over Cartesian World Models](#)
46. [Where are intentions to be found?](#)
47. [D&D.Sci April 2021 Evaluation and Ruleset](#)
48. [Probability theory and logical induction as lenses](#)
49. [What books are for: a response to "Why books don't work."](#)
50. [Covid 4/9: Another Vaccine Passport Objection](#)

Best of LessWrong: April 2021

1. [Notes from "Don't Shoot the Dog"](#)
2. [Another \(outer\) alignment failure story](#)
3. [Announcing the Alignment Research Center](#)
4. [The Case for Extreme Vaccine Effectiveness](#)
5. [Predictive Coding has been Unified with Backpropagation](#)
6. [Testing The Natural Abstraction Hypothesis: Project Intro](#)
7. [AMA: Paul Christiano, alignment researcher](#)
8. [Specializing in Problems We Don't Understand](#)
9. [I'm from a parallel Earth with much higher coordination: AMA](#)
10. ["AI and Compute" trend isn't predictive of what is happening](#)
11. [Opinions on Interpretable Machine Learning and 70 Summaries of Recent Papers](#)
12. [Why We Launched LessWrong.SubStack](#)
13. [Why has nuclear power been a flop?](#)
14. [Tales from Prediction Markets](#)
15. [Monastery and Throne](#)
16. [How to Play a Support Role in Research Conversations](#)
17. [Covid 4/22: Crisis in India](#)
18. [The irrelevance of test scores is greatly exaggerated](#)
19. [What are all these children doing in my ponds?](#)
20. [A new acausal trading platform: RobinShould](#)
21. [Gradations of Inner Alignment Obstacles](#)
22. [\[Letter\] Advice for High School #1](#)
23. [Updating the Lottery Ticket Hypothesis](#)
24. ["Taking your environment as object" vs "Being subject to your environment"](#)
25. [Covid 4/15: Are We Seriously Doing This Again](#)
26. [People Will Listen](#)
27. [A Brief Review of Current and Near-Future Methods of Genetic Engineering](#)
28. [Highlights from The Autobiography of Andrew Carnegie](#)
29. [Wanting to Succeed on Every Metric Presented](#)
30. [Iterated Trust Kickstarters](#)
31. [My take on Michael Littman on "The HCI of HAI"](#)
32. [Draft report on existential risk from power-seeking AI](#)
33. [April drafts](#)
34. [FAQ: Advice for AI Alignment Researchers](#)
35. [Facebook is Simulacra Level 3, Andreessen is Level 4](#)
36. [Against "Context-Free Integrity"](#)
37. [The secret of Wikipedia's success](#)
38. [We need a career path for invention](#)
39. [Three reasons to expect long AI timelines](#)
40. [Center for Applied Postrationality: An Update](#)
41. [Alignment Newsletter Three Year Retrospective](#)
42. [Scott Alexander 2021 Predictions: Market Prices](#)
43. [Vim](#)
44. [Covid 4/1: Vaccine Passports](#)
45. [Agents Over Cartesian World Models](#)
46. [Where are intentions to be found?](#)
47. [D&D.Sci April 2021 Evaluation and Ruleset](#)
48. [Probability theory and logical induction as lenses](#)
49. [What books are for: a response to "Why books don't work."](#)

50. [Covid 4/9: Another Vaccine Passport Objection](#)

Notes from "Don't Shoot the Dog"

Cross-posted from [The Whole Sky](#).

I just finished Karen Pryor's "[Don't Shoot the Dog: the New Art of Teaching and Training](#)." Partly because a friend points out that it's not on Audible and therefore she can't possibly read it, here are the notes I took and some thoughts. It's a quick, easy read.

The book applies behavioral psychology to training animals and people. The author started off as a dolphin trainer at an aquarium park in the 1960s and moved on to horses, dogs, and her own children. There are a lot of anecdotes about how to train animals (apparently polar bears like raisins). At the time, training animals without violence was considered novel and maybe impossible. I read it as a parenting book since I don't plan to train dogs, horses, or polar bears.

It's probably not the best guide to training dogs since a lot of it is about people, and not the best guide to training people since a lot is about animals. She's written a bunch of [other books](#) about training dogs and cats. But this book is an entertaining overview of all of it.

The specter of behaviorism

I can understand not wanting to use behavioral methods on children; the idea can sound overly harsh or reductive. The thing is, we already reinforce behavior all the time, including bad behavior, often without meaning to. So you might as well notice what you're doing.

To people schooled in the humanistic tradition, the manipulation of human behavior by some sort of conscious technique seems incorrigibly wicked, in spite of the obvious fact that we all go around trying to manipulate one another's behavior all the time, by whatever means come to hand.

[...]

There are still people who shudder at the very name of Skinner, which conjures in their minds some amalgam of Brave New World, mind control, and electric shock.

([B. F. Skinner](#) in fact believed that punishment was not an effective learning tool, and that positive reinforcement was much better for teaching.)

Pryor argues that behavioral training allows you to get good results more pleasantly than with other methods. She describes her daughter's experience directing a play in high school:

At the closing performance the drama coach told me that she'd been amazed to see that throughout rehearsals Gale never yelled at her cast. Student directors always yell, but Gale never yelled. 'Of course not,' I said without thinking, 'she's an animal trainer.' From the look on the teacher's face, I realized I'd said the wrong thing—her students were not animals! But of course all I meant was that Gale would know how to establish stimulus control without unnecessary escalation.

Of course there are bad applications of behavioral training: “The psychological literature abounds with shaping programs that are so unimaginative, not to say ham-handed, that they constitute in my opinion cruel and unusual punishment.”

I don’t know a lot about ABA (applied behavior analysis), which is one application of behaviorism. My understanding is that its bad applications are certainly cruel and ham-handed, although there also seem to be good applications. I think that even people opposed to ABA should be able to find a lot of useful material in this book.

You’re already doing reinforcement training

One point I think is underappreciated is that we all reinforce each other, and children train parents as well as the other way around.

A child is tantruming in the store for candy. The parent gives in and lets the child have a candy bar. The tantruming is positively reinforced by the candy, but the more powerful event is that the parent is negatively reinforced for giving in, since the public tantrum, so aversive and embarrassing for the parent, actually stopped.

It’s also easy to accidentally reinforce bad behavior.

I recently read Beverly Cleary’s *Beezus and Ramona* with the kids, in which a preschooler scribbles in a library book she wants to keep. Her older sister pays for the book, and the librarian gives them back the discarded book to keep.

That’s not fair, thought Beezus. Ramona shouldn’t get her own way when she had been naughty.

‘But, Miss Evans,’ protested Beezus, ‘if she spoils a book she shouldn’t get to keep it. Now every time she finds a book she likes she will...’ Beezus did not go on. She knew very well what Ramona would do, but she wasn’t going to say it out loud in front of her.

Jeff and I try to not let bad behavior lead to a reward. For example, our four-year-old was eager to go home from the park, and left without us towards the house. I caught up with her and told her not to leave without us. We were halfway to the house, but if I’d continued home with her from there, she would still have achieved what she wanted: getting home sooner. So I took her back to the park and we redid the whole situation: she said “I want to go home” and I walked home with her. Running off on her own didn’t pay, and she hasn’t repeated it.

Responding to good behavior, not bad

Instead of punishing bad behavior, the emphasis is on noticing and reinforcing good behavior.

“Shutting up about what you don’t like, in order to wait for and reinforce behavior you do like, is counterintuitive and takes some practice.”

My mother, who taught preschool for decades, sums it up as “You have to catch ‘em being good.”

Some animals can’t be trained by force, or at least can’t be trained to do anything very complicated. Such training was necessary with dolphins because they’ll simply

swim away if you try to make them do anything they don't like. You can only train them by offering something they like (fish).

"As a dolphin researcher whom I worked with sourly put it, 'Nobody should be allowed to have a baby until they have first been required to train a chicken,' meaning that the experience of getting results with a chicken, an organism that cannot be trained by force, should make it clear that you don't need to use punishers to get results with a baby."

At its best, reinforcement learning is enjoyable for the learner:

Clicker trainers have learned to recognize play behavior in animals as a sign that the learner has become consciously aware of what behavior was being reinforced. When 'the light bulb goes on,' as clicker trainers put it, dogs gambol and bark, horses prance and toss their heads, and elephants, I am told, run around in circles chirping. They are happy. They are excited.

Clickers and other sounds

Pryor became known for "clicker training" because she started using the method of using a sound to immediately convey "yes, that's good." The particular sound isn't important as long as the learner can hear and recognize it. With aquatic animals you use whistles because they can be heard underwater; with dogs she uses mechanical clicker noisemakers; with a person I'd probably use a specific phrase but some people also use clickers.

The sound initially has no meaning, but by giving it at the same time as a reward (food, smiles, pats) you create an association between the sound and the reward. Later the sound itself is rewarding.

It often happens, especially when training with food reinforcers, that there is absolutely no way you can get the reinforcer to the subject during the instant it is performing the behavior you wish to encourage. If I am training a dolphin to jump, I cannot possibly get a fish to it while it is in midair. If each jump is followed by a thrown fish with an unavoidable delay, eventually the animal will make the connection between jumping and eating and will jump more often. However, it has no way of knowing which aspect of the jump I liked. Was it the height? The arch? Perhaps the splashing reentry? Thus it would take many repetitions to identify to the animal the exact sort of jump I had in mind. To get around this problem, we use conditioned reinforcers.

[...]

Breland called the whistle a 'bridging stimulus,' because, in addition to informing the dolphin that it had just earned a fish, the whistle bridged the period of time between the leap in midtank—the behavior that was being reinforced—and swimming over to the side to collect one's pay.

Pryor describes the program her son (an airplane pilot) designed for pilot training:

A flight instructor can also click a student for initiative and for good thinking: for example, for glancing over the instrument panel before being reminded to do so. So the clicker can reward nonverbal behavior nonverbally in the instant it's occurring.

[...]

Once you have established a conditioned reinforcer, you must be careful not to throw it around meaninglessly or you will dilute its force. The children who rode my Welsh ponies for me quickly learned to use 'Good pony!' only when they wanted to reinforce behavior. . . One day a child who had just joined the group was seen petting a pony's face while saying 'You're a good pony.' Three of the others rounded on her instantly: 'What are you telling him that for? He hasn't done anything!'

Attention

This doesn't mean you give positive attention only during training.

One can and should lavish children (and spouses, parents, lovers, and friends) with love and attention, unrelated to any particular behavior; but one should reserve praise, specifically, as a conditioned reinforcer related to something real.

I think when children point out minor accomplishments — "Look at all the sticks I collected" — it's more often a request for attention than a situation that requires praise. I'm likely to comment in a way that shows interest — "Yes, you've got a lot of sticks there!" — but I don't see a need to evaluate the quality of their stick pile or whatever. I try to save actual praise for something I especially want them to do more of, or something that was new and challenging for them.

Interested attention during training is necessary, and ignoring someone is a kind of punishment:

If the trainer starts chatting to some bystander or leaves to answer the telephone or is merely daydreaming, the contract is broken; reinforcement is unavailable through no fault of the trainee. This does more harm than just putting the trainer at risk of missing a good opportunity to reinforce. It may punish some perfectly good behavior that was going on at the time. Of course if you want to rebuke a subject, removing your attention is a good way to do it.

Wrong timing

Pryor emphasizes that if you give punishment or reward at the wrong time, you reinforce the wrong behavior. If you call a dog to you and it finally comes, then you strike it, you've punished it for returning to you.

My mother always complained of the same tendency in her choral director: when the singers finally got a difficult passage right, instead of praising them he'd shout "Why couldn't you do it like that the first time?!"

I've noticed the importance of timing when a child finally does what you want, because it's tempting to scold them even after they've shaped up. Anna has a wide variety of delay tactics for brushing her teeth, and I find it easy to be stony-faced when she's capering around instead of coming to the sink. By the time she finally comes to have her teeth brushed I'm feeling annoyed and would like to give her a lecture. But if I give her an unpleasant response just as she's finally doing what I want, I disincentivize her from doing it. Instead, as soon as she comes to the sink I become pleasant Mama, smiling and joking.

Maintaining behavior

Once a behavior is established, you use intermittent reinforcement to maintain it:

"In order to maintain an already-learned behavior with some degree of reliability, it is not only not necessary to reinforce it every time; it is vital that you do not reinforce it on a regular basis but instead switch to using reinforcement only occasionally, and on a random or unpredictable basis."

"Many people initially object to the idea of using positive reinforcers in training because they imagine that they will forever have to hand out treats to get good behavior. But the opposite is true. Training with reinforcement actually frees you from the need for constant vigilance over the behavior, because of the power of variable schedules."

In people, the behavior itself eventually brings its own reward; we praise toddlers for learning to use the potty, but after the behavior is established we no longer need to reinforce it. And having dry clothes is its own reward.

"The power of the variable schedule is at the root of all gambling. If every time you put a nickel into a slot machine a dime were to come out, you would soon lose interest. Yes, you would be making money, but what a boring way to do it. People like to play slot machines precisely because there's no predicting whether nothing will come out, or a little money, or a lot of money, or which time the reinforcer will come (it might be the very first time)."

We encountered this in my house when Lily was two. Our housemate would sporadically show her a Sesame Street video on his phone, and she loved this so she'd pester him constantly for it. The reward came unpredictably, so she asked very often. Once he moved to a predictable schedule (one video every day after dinner) she learned the pattern and stopped asking at times of day when she knew it wouldn't work.

Also affects adult relationships:

If you get into a relationship with someone who is fascinating, charming, sexy, fun, and attentive, and then gradually the person becomes more disagreeable, even abusive, though still showing you the good side now and then, you will live for those increasingly rare moments when you are getting all those wonderful reinforcers: the fascinating, charming, sexy, and fun attentiveness. And paradoxically from a commonsense viewpoint, though obviously from the training viewpoint, the rarer and more unpredictable those moments become, the more powerful will be their effect as reinforcers, and the longer your basic behavior will be maintained. Furthermore, it is easy to see why someone once in this kind of relationship might seek it out again. A relationship with a normal person who is decent and friendly most of the time might seem to lack the kick of that rare, longed-for, and thus doubly intense reinforcer.

Pryor training herself to go to class even when she didn't feel like it, and then maintaining the behavior without the reward:

I found that if I broke down the journey, the first part of the task, into five steps—walking to the subway, catching the train, changing to the next train, getting the bus to the university, and finally, climbing the stairs to the classroom—and reinforced each of these initial behaviors by consuming a small square of chocolate, which I like but normally never eat, at the completion of each step, I

was at least able to get myself out of the house, and in a few weeks was able to get all the way to class without either the chocolate or the internal struggle.

Sports players and fans become “trained” to do certain actions (wearing their lucky clothes, etc) because they associate it with the team winning.

I have seen one baseball pitcher who goes through a nine-step chain of behavior every time he gets ready to pitch the ball: touch cap, touch ball to glove, push cap forward, wipe ear, push cap back, scuff foot, and so on. In a tight moment he may go through all nine steps twice, never varying the order. The sequence goes by quite fast—announcers never comment on it—and yet it is a very elaborate piece of superstitious behavior.

Raise expectations gradually, with rewards for incremental progress:

I once saw a father make a serious error in this regard. Because his teenage son was doing very badly in school, he confiscated the youth’s beloved motorcycle until his grades improved. The boy did work harder, and his grades did improve, from Fs and Ds to Ds and Cs. Instead of reinforcing this progress, however, the father said that the grades had not improved enough and continued to withhold bike privileges. This escalation of the criteria was too big a jump; the boy stopped working altogether.

Pryor claims that you have to be much more consistent with aversives (punishments) than with rewards. Seems like that might be right with animals and young children, but adults are usually willing to avoid committing crimes even if they don’t expect to be caught every time.

Often when we are teaching the behavior, we use a fixed schedule of reinforcement; that is, we reinforce every adequate behavior. But when we are just maintaining a behavior, we reinforce very occasionally, using a sporadic or intermittent schedule. For example, once a pattern of chore sharing has been established, your roommate or spouse may stop at the dry cleaners on the way home without being reinforced each time; but you might express thanks for an extra trip made when you are ill or the weather is bad. When we train with aversives, however—and that’s the way most of us began—we are usually taught that it is vital to correct every mistake or misbehavior. When errors are not corrected, the behavior breaks down. Many dogs are well behaved on the leash, when they might get jerked, but they are highly unreliable as soon as they are off leash and out of reach. When out with their friends, many teenagers do things that they wouldn’t dream of doing in their parents’ presence. This can happen because the subject is fully aware that punishment is unavailable—when the cat’s away, the mice will play—but it can also happen as a side effect of training with aversives. Since the message in a punisher is ‘Don’t do that,’ the absence of the aversive sends the message, ‘That is okay now.’

Learners can go long periods of time without a reward:

One psychologist jokes that the longest schedule of unreinforced behavior in human existence is graduate school.

When to stop a training session

End a training session while the learner is having success:

"When you stop is not nearly as important as what you stop on. You should always quit while you're ahead."

"The last behavior that was accomplished is the one that will be remembered best; you want to be sure it was a good, reinforceable performance. What happens all too often is that we get three or four good responses—the dog retrieves beautifully, the diver does a one-and-a-half for the first time, the singer gets a difficult passage right—and we are so excited that we want to see it again or to do it again. So we repeat it, or try to, and pretty soon the subject is tired, the behavior gets worse, mistakes crop up, corrections and yelling take place, and we just blew a training session."

Sports training

Pryor notes that in the second part of the 20th century, sports training seems a lot better than when she was young, and has moved toward more effective reinforcement learning:

I think what had changed in the last decade or two is that the principles that produce rapid results are becoming implicit in the standard teaching strategies: "This is the way to teach skiing: Don't yell at them, follow steps one through ten, praise and reinforce accomplishment at each step, and you'll get most of them out on the slopes in three days.

On patience

Good trainers are disciplined and intentional:

People who have a disciplined understanding of stimulus control avoid giving needless instructions, unreasonable or incomprehensible commands, or orders that can't be obeyed. They try not to make requests they're not prepared to follow through on; you always know exactly what they expect. They don't fly off the handle at a poor response. They don't nag, scold, whine, coerce, beg, or threaten to get their way, because they don't need to. And when you ask them to do something, if they say yes, they do it. When you get a whole family, or household, or corporation working on the basis of real stimulus control—when all the people keep their agreements, say what they need, and do what they say—it is perfectly amazing how much gets done, how few orders ever need to be given, and how fast the trust builds up. Good stimulus control is nothing more than true communication—honest, fair communication. It is the most complex, difficult, and elegant aspect of training with positive reinforcement.

One thing I notice in all this is that it's self-reinforcing. The method requires a certain amount of patience and self-discipline from the parent. It's easier to do that when things are already going well, and in turn you're rewarded with children who are easier to live with. When parents are exhausted and time-pressed, it's easier to slip into inconsistency, and both parents and children are more prone to outbursts and unpleasantness.

Limits of reinforcement

She ends with some warnings about trying to apply reinforcement to absolutely everything, or assuming it's the only thing in play:

Idealistic societies, in imagination or in practice, sometimes fail to take into account or seek to eliminate such biological facts as status conflict. We are social animals, after all, and as such we must establish dominance hierarchies. Competition within groups for increased status—in all channels, not just approved or ordained channels—is absolutely inevitable and in fact performs an important social function: Whether in Utopias or herds of horses, the existence of a fully worked-out hierarchy operates to reduce conflict. You know where you stand, so you don't have to keep growling to prove it. I feel that individual and group status, and many other human needs and tendencies, are too complex to be either met or overridden by planned arrangements of reinforcement, at least on a long-term basis.

This isn't the only tool I'd want in my parenting repertoire. But I do think it's well worth having.

Another (outer) alignment failure story

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

Meta

This is a story where the alignment problem is somewhat harder than I expect, society handles AI more competently than I expect, and the outcome is worse than I expect. It also involves inner alignment turning out to be a surprisingly small problem. Maybe the story is 10-20th percentile on each of those axes. At the end I'm going to go through some salient ways you could vary the story.

This isn't intended to be a particularly great story (and it's pretty informal). I'm still trying to think through what I expect to happen if alignment turns out to be hard, and this more like the most recent entry in a long journey of gradually-improving stories.

I wrote this up a few months ago and was reminded to post it by Critch's [recent post](#) (which is similar in many ways). This story has definitely been shaped by a broader community of people gradually refining failure stories rather than being written in a vacuum.

I'd like to continue spending time poking at aspects of this story that don't make sense, digging into parts that seem worth digging into, and eventually developing clearer and more plausible stories. I still think it's very plausible that my views about alignment will change in the course of thinking concretely about stories, and even if my basic views about alignment stay the same it's pretty likely that the story will change.

Story

ML starts running factories, warehouses, shipping, and construction. ML assistants help write code and integrate ML into new domains. ML designers help build factories and the robots that go in them. ML finance systems invest in companies on the basis of complicated forecasts and (ML-generated) audits. Tons of new factories, warehouses, power plants, trucks and roads are being built. Things are happening quickly, investors have super strong FOMO, no one really knows whether it's a bubble but they can tell that e.g. huge solar farms are getting built and *something* is happening that they want a piece of. Defense contractors are using ML systems to design new drones, and ML is helping the DoD decide what to buy and how to deploy it. The expectation is that automated systems will manage drones during high-speed ML-on-ML conflicts because humans won't be able to understand what's going on. ML systems are designing new ML systems, testing variations, commissioning giant clusters. The financing is coming from automated systems, the clusters are built by robots. A new generation of fabs is being built with unprecedented speed using new automation.

At this point everything kind of makes sense to humans. It feels like we are living at the most exciting time in history. People are making tons of money. The US defense establishment is scared because it has no idea what a war is going to look like right now, but in terms of policy their top priority is making sure the boom proceeds as quickly in the US as it does in China because it now seems plausible that being even a few years behind would result in national irrelevance.

Things are moving very quickly and getting increasingly hard for humans to evaluate. We can no longer train systems to make factory designs that look good to humans, because we don't actually understand exactly what robots are doing in those factories or why; we can't evaluate the tradeoffs between quality and robustness and cost that are being made; we can't really understand the constraints on a proposed robot design or why one design is better than another. We can't evaluate arguments about investments very well because they come down to claims about where the overall economy is going over the next 6 months that seem kind of alien (even the more recognizable claims are just kind of incomprehensible predictions about e.g. how the price of electricity will change). We can't really understand what is going to happen in a war when we are trying to shoot down billions of drones and disrupting each other's communication. We can't understand what would happen in a protracted war where combatants may try to disrupt their opponent's industrial base.

So we've started to get into the world where humans just evaluate these things by results. We know that Amazon pays off its shareholders. We know that in our elaborate war games the US cities are safe. We know that the widgets that come out the end are going to be popular with consumers. We can tell that our investment advisors make the numbers in our accounts go up.

On the way there we've had some stumbles. For example, my financial advisor bought me into a crazy ponzi scheme and when I went to get the money out I couldn't---financial regulators eventually shut down the fund but people with bad AI advisors still lost a lot. My factory colluded with the auditors who were valuing its output, resulting in a great Q4 report that didn't actually correspond to any revenue. In a war game our drones let the opponents take the city as long as they could corrupt the communications out of the city to make it look like everything was great.

It's not hard to fix these problems. I don't just train my financial advisors to get more money in my bank account---if I eventually discover the whole thing is a fraud, then that's a big negative reward (and we have enough data about fraud for models to understand the idea and plan to take actions that won't be eventually recognized as fraud). If an audit is later corrected, then we use the corrected figures (and apply a big penalty). I don't just rely on communications out of the city to see if things are OK, I use satellites and other indicators. Models learn to correctly treat the early indicators as just a useful signal about the real goal, which includes making sure that nothing looks fishy next year.

To improve safety we make these measures more and more robust. We audit the auditors. We ensure that ML systems are predicting the results of tons of sensors so that if anything is remotely fishy we would notice. If someone threatens an auditor, we'll see it on the cameras in their office or our recordings of their email traffic. If someone tries to corrupt a communication link to a camera we have a hundred other cameras that can see it.

As we build out these mechanisms the world keeps on getting more complicated. The automated factories are mostly making components for automated factories.

Automated R&D is producing designs for machines that humans mostly don't understand, based on calculations that we can only verify experimentally---academic fields have pivoted to understanding machines designed by AI and what it says about the future of industry, rather than contributing in any meaningful way. Most people don't have any understanding of what they are invested in or why. New industrial centers are growing in previously sparsely populated areas of the world, and most of it is feeding new construction that is several degrees removed from any real human use or understanding. Human CEOs are basically in charge of deciding how to delegate to ML, and they can talk as if they understand what's going on only because they get their talking points from ML assistants. In some domains regulations are static and people work around them, in others corruption is endemic, in others regulators adopt new policies pushed by ML-enhanced lobbyists. Our automated army is now incomprehensible even to the humans in charge of it, procured by automated procurement systems and built by fully-automated defense contractors.

For many people this is a very scary situation. It's like we are on a train that's now moving too fast to jump off, but which is accelerating noticeably every month. We still understand well enough that we could shut the whole thing down, scrap the new factories or at least let them sit dormant while experts figure out what is actually going on. But that could not be done unilaterally without resigning yourself to military irrelevance---indeed, you have ML systems that are able to show you good forecasts for what would happen if you stopped the machine from spinning without also getting the Chinese to do the same. And although people are scared, we are also building huge numbers of new beautiful homes, and using great products, and for the first time in a while it feels like our society is actually transforming in a positive direction for everyone. Even in 2020 most people have already gotten numb to not understanding most of what's happening in the world. And it really isn't that clear what the harm is as long as things are on track.

We know what happens when you deploy a sloppily-trained ML system---it will immediately sell you out in order to get a good training reward. This isn't done at all anymore because why would you? But people still remember that and it makes them scared, especially people in the defense establishment and AI safety community because we still haven't really seen what would happen in a hot war and we know that it would happen extremely quickly.

Most people stay well clear of most of the new automated economy. That said, there are still drones everywhere they are legally allowed to be. At some point we reach a threshold where drones can do bad stuff and it's very hard to attribute it to any legal person, so it becomes obligatory for every city to have automated local defenses. If they don't, or if they do a sloppy job of it, drones descend to steal and kidnap and extort. This is a terrifying situation. (It never gets *that* terrifying, because before that point we're motivated to try really hard to fix the problem.)

We do ultimately find our way out of that situation, with regulations that make it easier to attribute attacks. Humans don't really understand how those regulations or the associated surveillance works. All they know is that there are a ton of additional cameras, and complicated book-keeping, and as a result if a drone flies into your city to mess stuff up someone is going to be on the hook for the damage it causes. And we know that we end up being pretty safe. In effect the harm caused by such drones has been propagated back into the reward function for every AI in the world, using mechanisms built and maintained by other AIs---if you mess with people, you are going to be held accountable and so you avoid actions that predictably lead to that consequence.

This regulatory regime persists and is constantly upgraded. It becomes ever-more-incomprehensible, and rests on complex relationships between autonomous corporations and automated regulators and automated law enforcement, new forms of bookkeeping and surveillance and complete inventories of everyone who could build a drone that can kill someone. None of this significantly reduces US competitiveness (because when considering a proposal we can tell if it would reduce competitiveness, and as long as we can get what we want without sacrificing competitiveness then we strongly prefer that).

There are treaties amongst states to prevent some of the worst offenses. Again, we can tell the treaties at least kind of work because we can tell that no one is dying. Again, we can't tell what the treaties really do. Academic fields discuss them and study them and sometimes make proposals to slightly improve them, but it's with the same spirit that academics today study a complex biological system which they have little hope of understanding.

The world continues to change faster and faster. The systems that protect us become increasingly incomprehensible to us, outpacing our attempts to understand. People are better educated and better trained, they are healthier and happier in every way they can measure. They have incredibly powerful ML tutors telling them about what's happening in the world and helping them understand. But all of these things move glacially as far as the outside automated world is concerned.

Now we are resigned to being on this train and seeing where it goes, and for the most part people are happy (if we could have predicted that they would have been unhappy they would have taken a different route). There was a time when we were trying to design better and better systems to monitor for problems, but now that work is itself incomprehensibly sophisticated and out of our hands.

Some people still complain about the situation, and it still is objectively quite scary. We've built this incredible edifice designed to make sure that there are never any *reports* of trouble. As we built it we understood what was happening out there in the real world and there wasn't much risk that something bad would happen without causing a report. And we had bigger things to worry about.

But now everything is totally out of our hands and we have no idea whether our ML systems would actually be trying to avoid or correctly predict a systematic failure. The communications infrastructure that records data was built by machines, and most of it is new, and it might be corrupted in a systemic failure. The sensors were built and deployed by machines. The forces on the other side who could try to attack us are now invisible and broadly distributed and incredibly sophisticated. The systems of accountability that would prevent someone from building up a paramilitary are themselves maintained by AI systems whose only real goal was to make sure that no problem ever got reported. We cannot tell if a factory was building widgets or weapons except by asking another ML system, and that system's judgments are by now ultimately optimized to ensure that nothing ever causes any trouble.

For a time there were watchdogs that could explain to us why we should be scared, why a particular charge was part of a brewing storm that could actually cause trouble someday. And for a time that leads to real change to prevent trouble. But eventually we can't tell real scary stories from bogus scary stories. We still have watchdogs that we train to tell us what's scary, but they can always scare us and we've long-since become numb to the warnings. There were always bogus scary stories, if you train models to look for them, and it's just a gradual transition to all the stories being

meaningless. When we investigate a claimed problem, sometimes we do so with ML auditors who tell us there's no problem, and sometimes we use ML auditors trained to be more skeptical who always tell us that there is a problem they just can't demonstrate in a way we'd understand. When we go to the factory and take it apart we find huge volumes of incomprehensible robots and components. We can follow a piece of machinery along the supply chain but we can't tell what it's *for*.

If this led to a visible catastrophe that would show up on camera, then that would have showed up in the forecasts and we would have avoided it. So we're able to stop machines that try to grab power and use it to cause a visible problem. In the scenario I'm describing we've done our job so well (and the machines we've delegated to have carried on the work so well) that there is basically no chance of that.

But eventually the machinery for detecting problems does break down completely, in a way that leaves no trace on any of our reports. Cybersecurity vulnerabilities are inserted into sensors. Communications systems are disrupted. Machines physically destroy sensors, moving so quickly they can't be easily detected. Datacenters are seized, and the datasets used for training are replaced with images of optimal news forever. Humans who would try to intervene are stopped or killed. From the perspective of the machines everything is now perfect and from the perspective of humans we are either dead or totally disempowered.

By the time this catastrophe happened it doesn't really feel surprising to experts who think about it. It's not like there was a sudden event that we could have avoided if only we'd known. We didn't have any method to build better sensors. We could try to leverage the sensors we already have; we can use them to build new sensors or to design protections, but ultimately all of them must optimize some metric we can already measure. The only way we actually make the sensors better is by recognizing new good ideas for how to expand our reach, actually anticipating problems by thinking about them (or recognizing real scary stories and distinguishing them from fake stories). And that's always been kind of slow, and by the end it's obvious that it's just hopelessly slow compared to what's happening in the automated world.

Ways you could vary the story

- In this story all the models generalize "correctly" in the sense that they keep trying to ensure that things look good according to some kind of augmented human judgment. There are other possible generalizations: (i) "get a low loss" where the model identifies with some copy of itself (a generalization of the copy receiving training data) and cares about the data going into that model. In this scenario an AI doesn't have to seize the sensors all at once, they just have to seize their own physical realizations, (ii) treacherous behavior where the AI has some arbitrary other goal but wants to do well in training in order to get power to use later. In this scenario when it fails the AI doesn't have to corrupt sensors and just has to make sure that people can't fix the problem (i.e. can't reclaim the resources the AI controls). I think these changes lead to kind of similar stories to the one in this post, though usually failure will occur at a somewhat earlier stage and look a bit different. I think they offer more "outs" on the technical side but make it much harder to fix problems by e.g. building out more sensors or hardening them. I think that some folks (e.g. at MIRI) would consider this particular failure super implausible for that reason. I'm telling the story this way due to some combination of (i) I care about outer alignment in particular, (ii)

I think the failure modes in this story are an important input into treachery and so it's interesting to tell a simpler story without more moving parts.

- I'm making a lot of assumptions about how AI works (roughly that it looks like the "unaligned benchmark" [here](#)) and it could easily work totally differently. I'm also assuming that ML works well enough and generalizes to long horizons well enough that it's obligatory if you want to remain competitive, while also being risky (since a model can learn instrumental plans on long horizons). I do think lots of variants will leave the basic story intact, e.g. it doesn't really matter that much how much your systems leverage planning or deduction (they could even involve almost no learning and still run into similar problems).
- It seems like the story changes a lot based on how fast progress is in the outside world (is it like 3 years from a kind-of-weird world to the singularity, or 30 years, or 3 months?), which in turn depends on both what's technically possible and on how the regulatory environment works out (e.g. does competition between the US and China lead to very fast adoption; can we actually end up in the world where crazy factories are being built in the middle of nowhere that people only pretend to understand?). My guess is in the 3-30 year ballpark depending in large part on where you draw the line for "kind of weird," and this story is kind of centered on the 3 year world which feels a bit fast to me. I think the story would be much scarier if you have a much faster takeoff, and significantly less scary if you have a much slower takeoff (mostly since future people would have time to solve these problems).
- I'm a bit skeptical that our society would be even this competent and unified. It feels like there's a likely family of stories where everything is just a complete mess much earlier, with people yelling at each other and AI just committing fraud and stealing from people all over the place, and the machinery for correcting that situation totally breaks down as your civilization collapses. It seems worth fleshing out what that looks like as well, but it's definitely not what I'm doing here.
- In this situation a huge amount of work would be going into alignment, including by powerful ML assistants. I haven't talked about that at all, and indeed I think there's a reasonable chance that alignment just isn't very tractable so that things really could go down this way. But a lot depends on exactly how alignment work goes down during this tumultuous period, how well people are able to use ML to help with alignment, how well-organized the community is and how able it is to recognize and implement good ideas, etc. I haven't chosen this story to be one where alignment work is particularly valuable in advance, I think that may only happen if takeoff is much faster or if the response at the time is much worse.

Announcing the Alignment Research Center

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

(Cross-post from [ai-alignment.com](#))

I'm now working full-time on the [Alignment Research Center](#) (ARC), a new non-profit focused on intent alignment research.

I left OpenAI at the end of January and I've spent the last few months planning, doing some theoretical research, doing some logistical set-up, and taking time off.

For now it's just me, focusing on theoretical research. I'm currently feeling pretty optimistic about this work: I think there's a good chance that it will yield big alignment improvements within the next few years, and a good chance that those improvements will be integrated into practice at leading ML labs.

My current goal is to build a small team working productively on theory. I'm not yet sure how we'll approach hiring, but if you're potentially interested in joining you can fill out [this tiny form](#) to get notified when we're ready.

Over the medium term (and maybe starting quite soon) I also expect to implement and study techniques that emerge from theoretical work, to help ML labs adopt alignment techniques, and to work on alignment forecasting and strategy.

The Case for Extreme Vaccine Effectiveness

I owe tremendous acknowledgments to Kelsey Piper, Oliver Habryka, Greg Lewis, and Ben Shaya. This post is built on their arguments and feedback (though I may have misunderstood them).

Update, May 13

I first wrote this post before investigating the impact of covid variants on vaccine effectiveness, listing the topic as a major caveat to my conclusions. I have now spent enough time (not that much, honestly) looking into variants that I have a tentative position I'm acting on for now.

My moderately confident conclusion is that the current spread of variants in the US barely impacts vaccine effectiveness. The Pfizer vaccine [is reported](#) to be 85% as effective against the feared B.1.351 variant (South African) as it is against B.1.1.7 (UK). Assuming that other variants are no more resistant than B.1.351 on average (a reasonable assumption) and that presently variants are no more than [25% of Covid cases](#) (in Alameda and San Francisco). The net effect is $0.25 * 0.85 + 0.75 * 1.0 = 0.9625$. In other words, vaccines still have 96% of the effect they would if B.1.1.7 were the only variant.

Plus, that tiny reduction of vaccine effectiveness is dwarfed by the falling background prevalence of Covid. When I first wrote this post, Alameda and San Francisco were at 0.1-0.15%; now they're at ~0.05%. The same for New York and the United Kingdom.

Although relaxing of restrictions might reverse this, right now, Covid-risk is very, very low in the Bay Area and many parts of the US.

All updates/changelog can be [viewed here](#).

I plead before the Master of Cost-Benefit Ratios. "All year and longer I have followed your dictates. Please, Master, can I burn my microCovid spreadsheets? Can I bury my masks? Pour out my hand sanitizer as a libation to you? Please, I beseech thee."

"Well, how good is your vaccine?" responds the Master.

"Quite good!" I beg. "We've all heard the numbers, 90-95%. Even [MicroCOVID.org has made it official](#): a 10x reduction for Pfizer and Moderna!"

The Master of Cost-Benefit Ratio shakes his head. "It helps, it definitely helps, but don't throw out that spreadsheet just yet. [One meal at a crowded restaurant](#) is enough to give even a vaccinated person *hundreds* of microCovids. Not to mention that your local prevalence could change by a factor of 5 in the next month or two, and that'd be half the gains from this vaccine of yours!"

I whimper. "But what if . . . what if vaccines were way better than 10x? What about a 100x reduction in the risks from COVID-19?"

He smiles. "Then we could go back to talking about [how fast you like to drive](#)."

In its most extreme form, I have heard it claimed that the vaccines provide 10x reduction against regular Covid, 100x against severe Covid, and 1000x against death. That is, for each rough increase in severity, you get 10x more protection.

This makes sense if we think of Covid as some kind of "[state transition](#)" model where there's a certain chance of moving from lesser to more severe states, and vaccines reduce the likelihood at each stage.

I think 10x at multiple stages is too much. By the time you're at 1000x reduction, model uncertainty is probably dominating. I feel more comfortable positing up to 100x, maybe 500x

reduction. I dunno.

There is a more limited claim of *extreme vaccine effectiveness* that I will defend today:

1. **In the case of the Pfizer vaccine (and likely Moderna too), the effectiveness in young healthy people is 99% against baseline symptomatic infection, or close to it.**
2. **We can reasonably expect the effectiveness of the vaccine against more severe cases of Covid to be greater than effectiveness against milder cases of Covid.**

(Maybe it's 2x more effective against severe-Covid and 3x more effective against death compared to just getting it at all. Something like that, it doesn't have to be 10x- it'd still be a big deal because more severe outcomes are where most of the disutility lies.)

It's a very simple argument, really. First, the data very clearly suggests effectiveness of ~99% for young people, with nice tight confidence intervals. Second, across all the data we see trends of increasing effectiveness against increasing severity, granted that the confidence intervals are wide in some cases. Third, a very reasonable (imo) mechanistic model supports this interpretation of the data.

The 1.2 Million-Person Pfizer Israeli Observational Study

This observational study matched ~600k vaccinated people 1:1 with ~600k demographically similar controls. It covered the period December 20 to February 1. As far as I know, it is by far the largest Covid-19 vaccine study published to date. The other studies are clinical trials with sample sizes on the order of 20k-40k, and some other observational studies, typically with healthcare workers, in the single-digit thousands.

Why it's not as big as it sounds

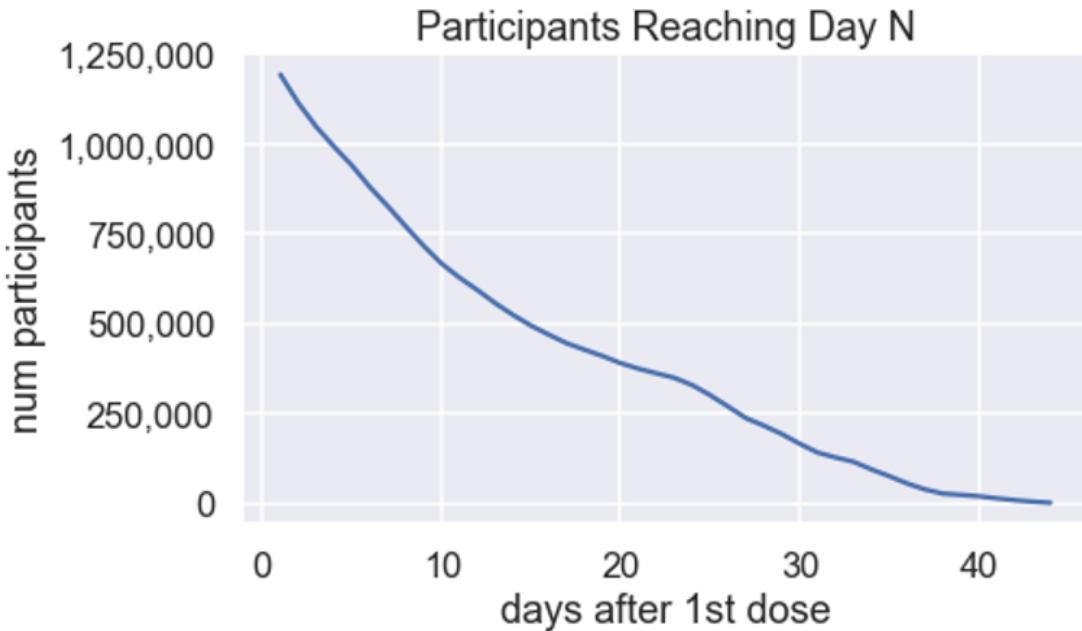
Before we get to looking at data, I think it's important to note why uncertainty remains despite the huge N. To start with, the outcomes are all quite rare. Eyeballing it, Israel had a Covid-19 prevalence of ~0.5% during the study period. Out of a million people, a few thousand might be expected to actually catch Covid. Of that few thousand, only dozens or hundreds will progress to more severe forms of Covid. When sample sizes are in the dozens, confidence intervals are wide.

The authors report:

During a mean follow-up of 15 days (interquartile range, 5 to 25), 10,561 infections [6,101 control vs 4,460 vaccinated] were documented . . . of which 5996 [2494 vs 2,071](57%) were symptomatic Covid-19 illness, 369 required hospitalization [259 vs 110], 229 were severe cases of Covid-19 [174 vs 55], and 41 resulted in death [32 vs 9].

See [Appendix B](#) for complete breakdown of outcomes by period and vaccination status.

What's more, those numbers are for the entire study period (44 days). Only in a subset of days had participants been vaccinated long enough for it to be a real test. Nominally it is a 1.2 million person study, but practically, when you're looking at results 2 weeks after the first dose (~day 14) or one week after the second dose (~day 28), the numbers are much lower. ~80% lower.



96% of participants received a second dose on day 21 or after; 95% received it before day 24

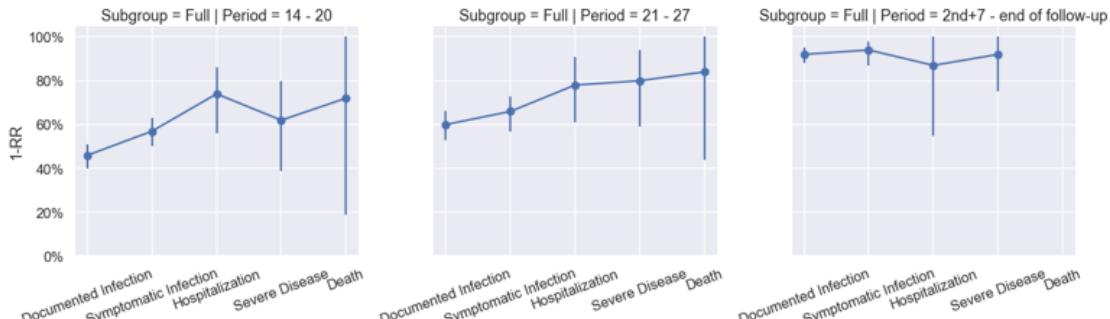
All that to say, sample sizes aren't as big as they sound. Well, let's look at the results. This is the [main outcome table](#). Definitions in [Appendix A](#).

Table 2. Estimated Vaccine Effectiveness against Covid-19 Outcomes during Three Time Periods.^a

Period	Documented Infection		Symptomatic Illness		Hospitalization		Severe Disease		Death	
	Risk Difference		Risk Difference		Risk Difference		Risk Difference		Risk Difference	
	% (95% CI)	no./1000 persons (95% CI)	% (95% CI)	no./1000 persons (95% CI)	% (95% CI)	no./1000 persons (95% CI)	% (95% CI)	no./1000 persons (95% CI)	% (95% CI)	no./1000 persons (95% CI)
14 to 20 days after first dose	46 (40-51)	2.06 (1.70-2.40)	57 (50-63)	1.54 (1.28-1.80)	74 (56-86)	0.21 (0.13-0.29)	62 (39-80)	0.14 (0.07-0.21)	72 (19-100)	0.03 (0.01-0.07)
21 to 27 days after first dose	60 (53-66)	2.31 (1.96-2.69)	66 (57-73)	1.34 (1.09-1.62)	78 (61-91)	0.22 (0.13-0.31)	80 (59-94)	0.18 (0.10-0.27)	84 (44-100)	0.06 (0.02-0.11)
7 days after second dose to end of follow-up	92 (88-95)	8.58 (6.22-11.18)	94 (87-98)	4.61 (3.29-6.53)	87 (55-100)	0.22 (0.08-0.39)	92 (75-100)	0.32 (0.13-0.52)	NA	NA

* Confidence intervals were estimated using the percentile bootstrap method with 500 repetitions. Estimates were calculated only for cells with more than 10 instances of an outcome across the two groups. NA denotes not available, and RR risk ratio.

It's a bit hard to track trends formatted like that, so here's an equivalent graph:



Left: 1-RR for 14-20 days after 1st dose;
Middle: 21-27 days after 1st dose;
Right: 7 days after 2nd dose until end of follow-up

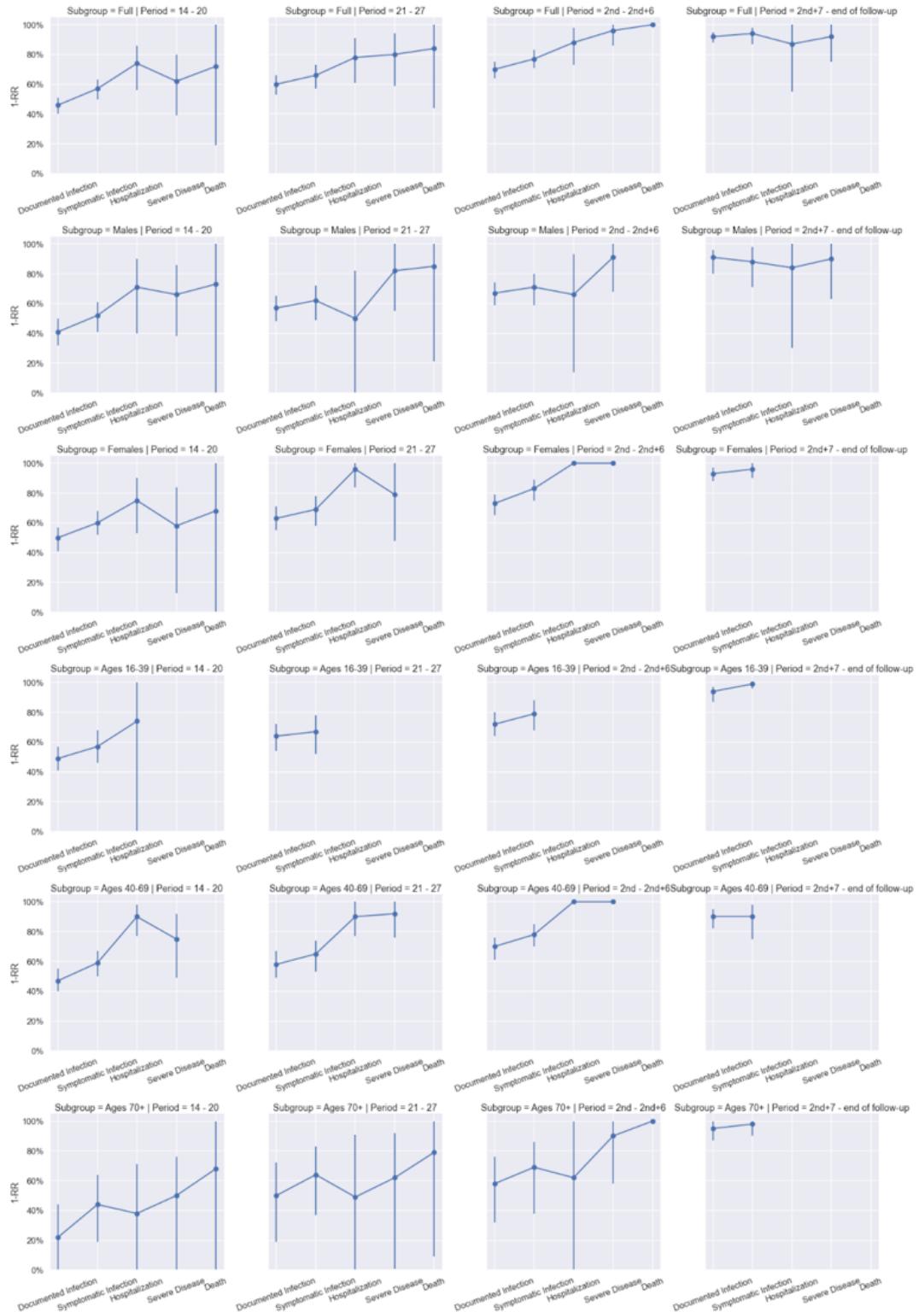
To me, the headline result is that 2nd-dose + 7-days, vaccine effectiveness against Symptomatic Illness is 94% (87-98). Pretty good! Also, efficacy clearly rises from the earlier to later periods after vaccine administration.

Unfortunately, we don't see efficacy improvements moving to the right on the rightmost graph (more severe outcomes), counter to claims of extreme vaccine effectiveness. We do see upwards-right trends in the earlier periods (left and middle graphs).

Well, 2 out of 3 ain't bad! Too bad the last one is the one we care about most.

It's okay. I've got more. The astute reader will have noted that the above graphs have "subgroup = Full" in their title. The study made available endpoints (outcomes) for multiple subgroups. Below are some of them. (Here are [ALL OF THEM](#).)

Since there was data for it, I also added in the "2nd dose and 6 days after" period.



To provide a sense of changing sample size between time periods: There were on average 221k participants each day, in each experimental group between days 14-20, 160k between days 21-27, and 39k from 28-44

Many of the points are missing. The authors did not compute vaccine effectiveness if the control and vaccine group combined did not have 10 or more instances of an

outcome. For example, the Age 16-39 subgroup did not have 10 instances of hospitalization, severe-Covid, or death for almost all of the time periods.

In cases where there are 10 instances of an outcome in the control group but 0 in the vaccine group, a value is reported without confidence intervals, e.g. the dots in the Females Subgroup, 2nd-2nd+6 period.

The up-and-to-right shape of the graphs persists across subgroups, except for Males and when the values are already maxing out near 100% (the right-most graphs). Overall, I think this is suggestive of the general trend that vaccines are more effective against progressively more severe outcomes. I also suspect that an uncertainty modeling that took account of the not uncorrelated neighboring values would shrink the error bars beyond the naive bootstrap method.

I'll comment more on why the flat/missing trend in the rightmost graph doesn't bother me much, beyond the fact that sample size means it's hard for it to show much at all.

Also! If you prefer Tables, here's the top half from the paper itself corresponding to the above graphs:

Table 3. Estimated Vaccine Effectiveness against Covid-19 Outcomes in Subpopulations According to Characteristics at Baseline.*				
Characteristic and Period	Documented Infection		Symptomatic Illness	
	1-RR % (95% CI)	Risk Difference no./1000 persons (95% CI)	1-RR % (95% CI)	Risk Difference no./1000 persons (95% CI)
Male sex				
14 to 20 days after first dose	41 (32 to 50)	1.71 (1.22 to 2.21)	52 (41 to 61)	1.26 (0.90 to 1.62)
21 to 27 days after first dose	57 (48 to 65)	2.25 (1.76 to 2.75)	62 (49 to 72)	1.30 (0.92 to 1.67)
7 days after second dose to end of follow-up	91 (80 to 96)	7.33 (4.48 to 10.84)	88 (71 to 98)	2.90 (1.87 to 4.02)
Female sex				
14 to 20 days after first dose	50 (41 to 57)	2.39 (1.84 to 2.86)	60 (52 to 68)	1.81 (1.43 to 2.19)
21 to 27 days after first dose	63 (55 to 71)	2.38 (1.91 to 2.91)	69 (58 to 78)	1.38 (1.02 to 1.71)
7 days after second dose to end of follow-up	93 (88 to 97)	9.75 (6.84 to 13.48)	96 (90 to 100)	6.22 (3.60 to 9.56)
Age, 16 to 39 yr				
14 to 20 days after first dose	49 (41 to 57)	2.29 (1.74 to 2.88)	57 (46 to 68)	1.38 (0.99 to 1.80)
21 to 27 days after first dose	64 (54 to 72)	2.80 (2.20 to 3.48)	67 (52 to 78)	1.27 (0.89 to 1.73)
7 days after second dose to end of follow-up	94 (87 to 97)	8.72 (5.72 to 12.69)	99 (96 to 100)	4.06 (2.76 to 5.66)
Age, 40 to 69 yr				
14 to 20 days after first dose	47 (40 to 55)	2.13 (1.69 to 2.66)	59 (50 to 67)	1.68 (1.32 to 2.05)
21 to 27 days after first dose	58 (49 to 67)	2.19 (1.67 to 2.70)	65 (53 to 74)	1.38 (1.03 to 1.80)
7 days after second dose to end of follow-up	90 (82 to 95)	8.96 (6.16 to 13.05)	90 (75 to 98)	5.01 (2.53 to 8.67)
Age, ≥70 yr				
14 to 20 days after first dose	22 (-9 to 44)	0.81 (-0.28 to 1.89)	44 (19 to 64)	1.36 (0.48 to 2.36)
21 to 27 days after first dose	50 (19 to 72)	1.40 (0.42 to 2.35)	64 (37 to 83)	1.35 (0.62 to 2.22)
7 days after second dose to end of follow-up	95 (87 to 100)	6.10 (3.43 to 9.61)	98 (90 to 100)	4.77 (2.14 to 7.70)
No coexisting conditions				
14 to 20 days after first dose	49 (42 to 56)	2.13 (1.69 to 2.59)	55 (45 to 63)	1.32 (0.98 to 1.67)
21 to 27 days after first dose	66 (58 to 73)	2.49 (1.99 to 2.98)	73 (62 to 82)	1.27 (0.92 to 1.64)
7 days after second dose to end of follow-up	91 (83 to 96)	7.67 (4.90 to 11.07)	93 (78 to 100)	3.54 (1.79 to 5.90)

As someone falling in the Age 16-39 subpopulation, I'm quite pleased to see 99% effectiveness against Symptomatic Infection, with a nice tight 96-100 confidence interval. This is higher than I'd had anyone cite to me, and is approximately a 5x increase in how effective I believe my Pfizer vaccine to be. That's even before we get to more severe outcomes.

So why is the absence of data showing increasing efficacy not evidence of absence?

Because we expect to see data that looks like this even in worlds where the vaccine is 100% effective (at least for all vaguely healthy people). To be evidence against something, it has to be less common in worlds where that thing is true, and that's not the case here.

Why would we see these numbers with a 100% effective vaccine?

"Saturation" and "noise"

When you have 100 true positives and 3 false positives, the false positives aren't such a big deal. When you have 0 true positives and 3 false positives, the false positives can change the entire picture.

I argue this is very likely what is going on with Covid-vaccine effectiveness, above and beyond sample sizes.

Consider that PCR Covid-19 tests have both a false negative and false positive rate (FPR). According to this [random site](#) I found by Googling that looks legit enough, the FPR for Covid-19 tests is between 0.2% and 0.9%. Let's choose a point estimate of 1% for the FPR to be safe, but run it twice for every case to compensate. So now our FPR is 0.01%

Let's now imagine using this test on a 99% effective vaccine (the same argument holds for 99.9% and 99.99% even more so). We run an RCT with 100,000 people receiving the vaccine and 100,000 receiving placebo. Covid-19 prevalence is 0.1% in the region our hypothetical test is running.

99 people from the control group catch actual Covid and receive positive test results (we lose one to a realistic false-negative rate of 10%, run twice to become 1%) plus 0.01% false positives for a total of 109. From the treatment group with 99% effective vaccine, we get 1 true-positive and 10 false-positive test results. Our final effectiveness estimate is $1 - 11/109 = 90\%$

90%! And that's from what is in truth a 99% effective vaccine. The control is mostly unaffected by the noise (109 vs actual 100) but the treatment is enormously changed. Instead of 1, it's 11.

The greatest noise of all is selection effects

Never mind fluke false positive tests, on priors we have reason to suspect that the people who are still getting severe Covid despite being vaccinated or even hospitalized are very likely not like you. Why isn't the number of vaccinated people who ended up in critical condition zero?

Because within the Israel Observation study, the vaccinated group contains some very sick people. (In a half-million person group without exclusion criteria against it there simply will be, and that's before the fact that we know many people are elderly and many explicitly meet risk criteria such as cancer, obesity, pulmonary disease, Type 2 diabetes, etc. [See Table 1](#).

[Demographic and Clinical Characteristics of Vaccinated Persons and Unvaccinated Controls at Baseline.](#))

Vaccines work by stimulating an immune response. If your immune system is in tatters, unable to manufacture healthy antibodies or something, your vaccine might not do much. You might fare little better than the unvaccinated.

There's actually a term for immune system failure in the elderly: *immunosenescence*. A couple of papers on that topic: [Immunosenescence and vaccine failure in the elderly](#) (2009) and [Immunosenescence: A systems-level overview of immune cell biology and strategies for improving vaccine responses](#) (2019)

Quoting the abstract from the first one:

An age-related decline in immune responses in the elderly results in greater susceptibility to infection and reduced responses to vaccination. This decline in immune function affects both innate and adaptive immune systems . . . Essential features of immunosenescence include: reduced natural killer cell cytotoxicity on a per cell basis; reduced number and function of dendritic cells in blood; decreased pools of naive T and B cells; and increases in the number

of memory and effector T and B cells . . . Consequently, vaccine responsiveness is compromised in the elderly, especially frail patients . . . In the future, the development and use of markers of immunosenescence to identify patients who may have impaired responses to vaccination, as well as the use of end-points other than antibody titers to assess vaccine efficacy, may help to reduce morbidity and mortality due to infections in the elderly.

The last line there suggested antibodies can be present even when a vaccine overall is not protective against disease.

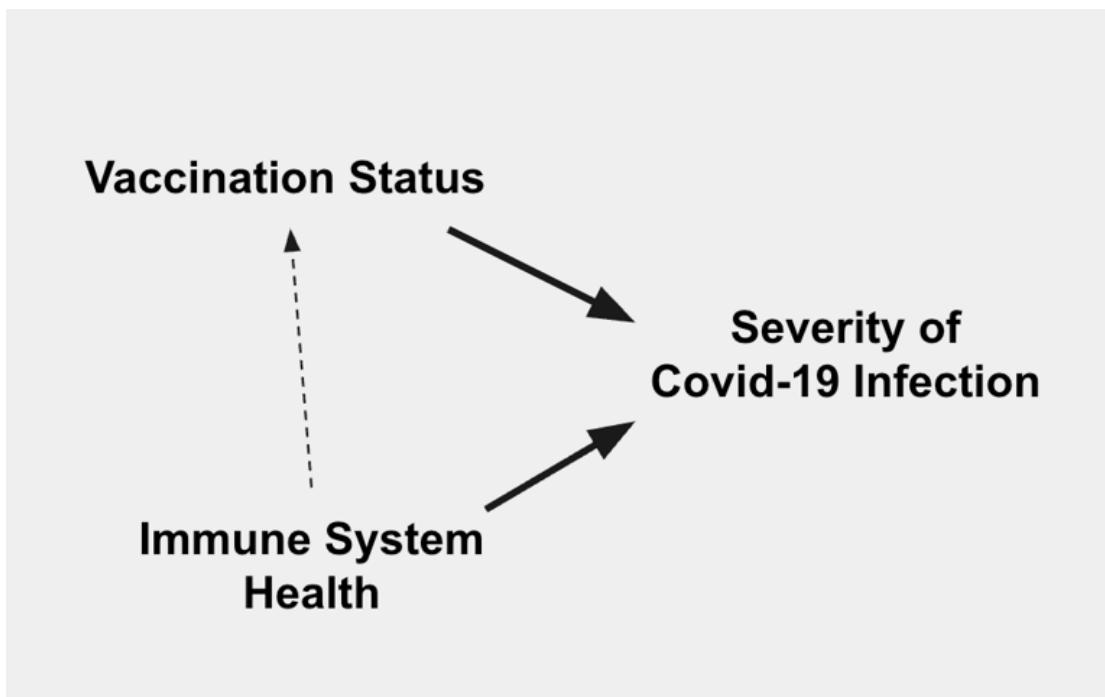
Don't get old kids, it's bad for you.

When I see that the treatment group has lower but not zero severe cases than the control group, I assume it's coming from the very ill, the immunocompromised and immunosenescent.

I am pretty confident that I am not in those groups. I'm pretty sure my vaccine elicited some real response (I had some side effects, not terrible, but some). When I see the effectiveness numbers showing globally that there's still some chance of really bad outcomes, I adjust them downwards because they were very likely not happening to people with remotely my level of health.

To invoke the math of the previous section regarding false positives, the immunocompromised might be only 10% to the number of severe Covid patients in the control group (because their prevalence is low), but after the filter/selection effect of vaccination, they make up 95% of severe cases in treatment group.

Here's a diagram for good measure:



This is a causal diagram. Arrows indicate the direction of causality but not the specific causal relationship. (Having a poor immune system, e.g. because you're older, might make you more likely to be vaccinated, but of course, in our study we're conditioning on vaccination status, so it doesn't matter.)

Since we didn't condition on immune system health, we can't expect that a naive interpretation of the table of efficacies (above) tells us the true relationship between vaccines and outcomes.

Well, we approximately can condition. The vaccine efficacy is higher for the Age 16-39yr subgroup (99%, 96-100) vs the entire population (94%, 87-98). It's not just that younger people

get less Covid, but that the vaccine worked better on them.

The Age 16-39yr subgroup didn't have zero cases of symptomatic disease (it *might* have zero of hospitalization, etc), but as above, I'm guessing those were almost all people who knowable had weak immune systems within the overall healthier group too.

Priors & Trends

In the above section I argued two things:

- that we see something of a trend towards increasing effectiveness with increasing severity in the Pfizer Israeli mass study
- Reasons why even if the vaccine was 99.9% effective, we would expect to observe effectiveness data that is lower than this.

In this section, I want to offer 1) a plausible mechanistic model for why this should be true, 2) further indications of increasing effectiveness from other trials and studies.

Shifting the Distribution of “Infection”/Viral Load

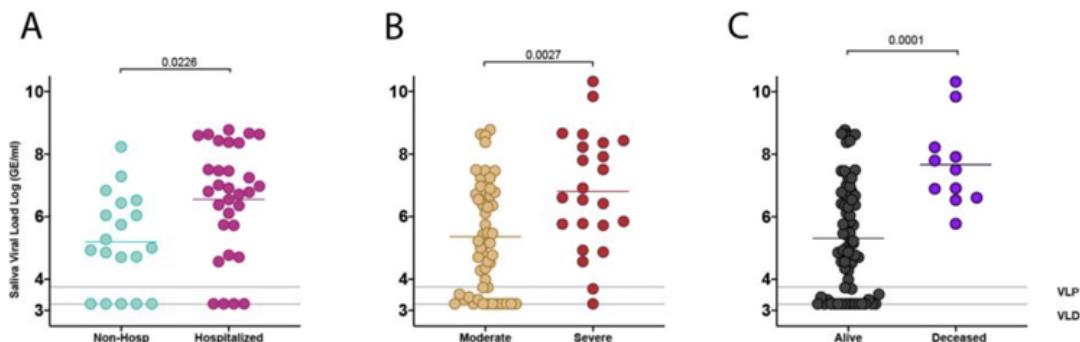
Epistemic status: I'm not a biology/virology/immunology person and this feels kinda hand-wavy to me.

This post started with a table that lists a progression of discrete outcomes: Documented Infection, Symptomatic, Hospitalization, Severe Disease, Death . . . and I've been referring to them since. Obviously, the underlying biological reality isn't quite as discrete as that.

It's probably more something like there's a continuous value of *how infected you are*, and the higher that value gets, the worse your condition will be.

How infected you are is probably a fuzzy thing in reality, but viral load might be an adequate proxy. It's been documented that viral load varies together with Covid-19 severity. See [SARS-CoV-2 viral load is associated with increased disease severity and mortality](#) and [Saliva viral load is a dynamic unifying correlate of COVID-19 severity and mortality](#).

The graphs in the second paper particularly make this point.



(a) Comparison of first recorded saliva viral load between individuals hospitalized for COVID-19 and non-hospitalized individuals within the first 10 days from symptom onset using a two-sided t-test. Comparison of only first recorded saliva viral load measurements amongst (b) moderate and severe disease or (c) alive and deceased individuals throughout the course of disease.

Presumably, your viral load is the result of competition between the virus replicating and your immune system fighting it. Vaccination gives a significant boost to your immune system. (Cf the oft-cited claim about [“4x lower” viral load in vaccinated people](#))

!! To engage in some inexpert armchair speculation, I'd guess that in the virus-immune system race, about $x\%$ of the time the virus gets the upper hand with an infected person and makes them symptomatic (~Level 1), and then in $x\%$ of cases where it got to Level 1, the virus wins out again to progress to Level 2 before the immune system can stop it. The virus progresses two levels $x^2\%$ of the time. If x is 20%, then overall 4% to get to two levels worse.

At each level, the virus only has an $x\%$ of winning out and progressing to the next level. Alternatively, in each time period, there's some chance, $y\%$, that the immune system will catch up and win.

In cases where someone has a weak immune system (high $x\%$, small $y\%$), increasing levels of case severity aren't much less likely than earlier ones. You might get an approx flat effectiveness curve.

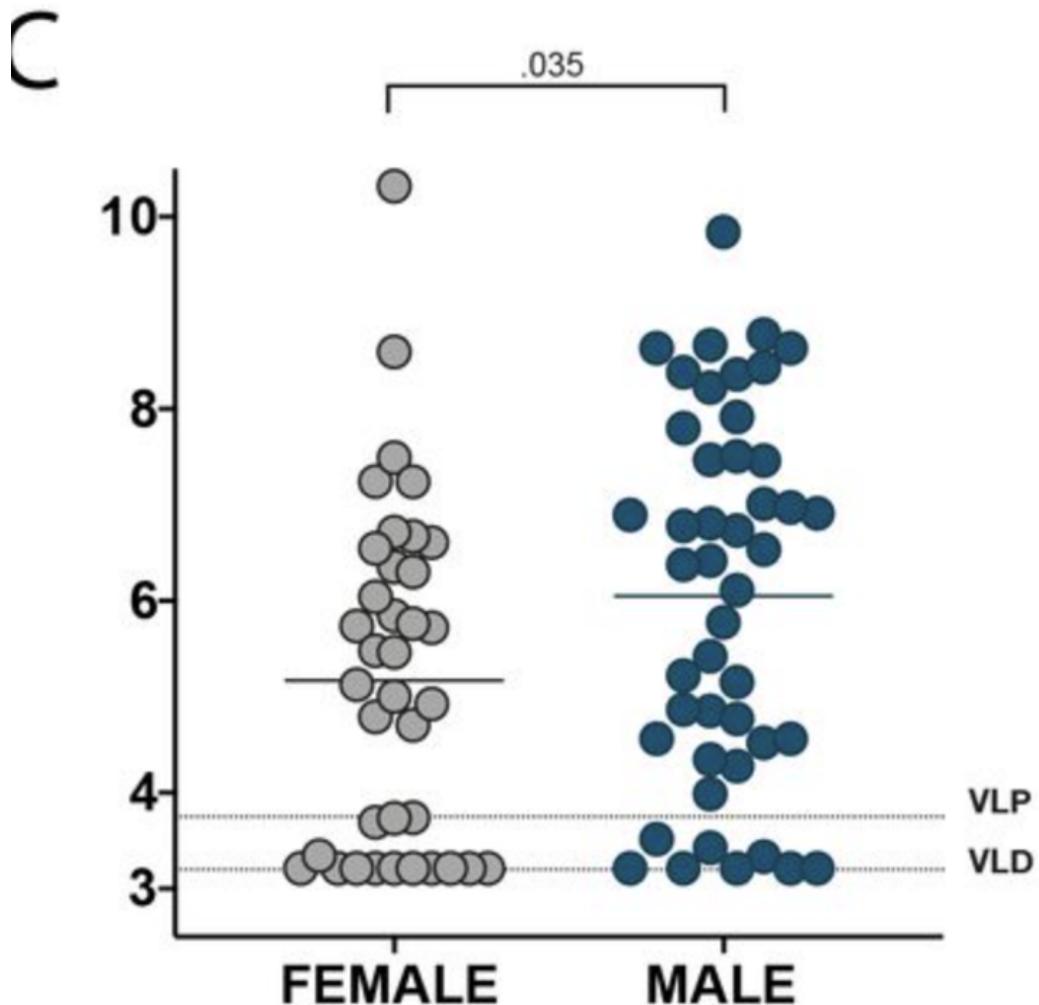
But suppose that someone is vaccinated and has a real boost to their immune system. Intuitively, I'd reckon they're now at $x/5\%$ for each stage. For the virus to progress two levels, it's $(x/5)^2$, or 0.16% when $x=20\%$.

Maybe it's a factor 2 or 3 instead of 5, but either way, it'd be a compounding effect. More severity means the virus has to replicate more times, which is more time for the immune system to catch up and beat it, ergo less chance for it to get that bad.

It's tough being a dude

Incidentally, [Silva et al \(2021\)](#) who provided the graphs of viral load immediately above, also had this to say re male vs female:

VLP= Viral load limit of positivity
VLD= Viral load limit of detection



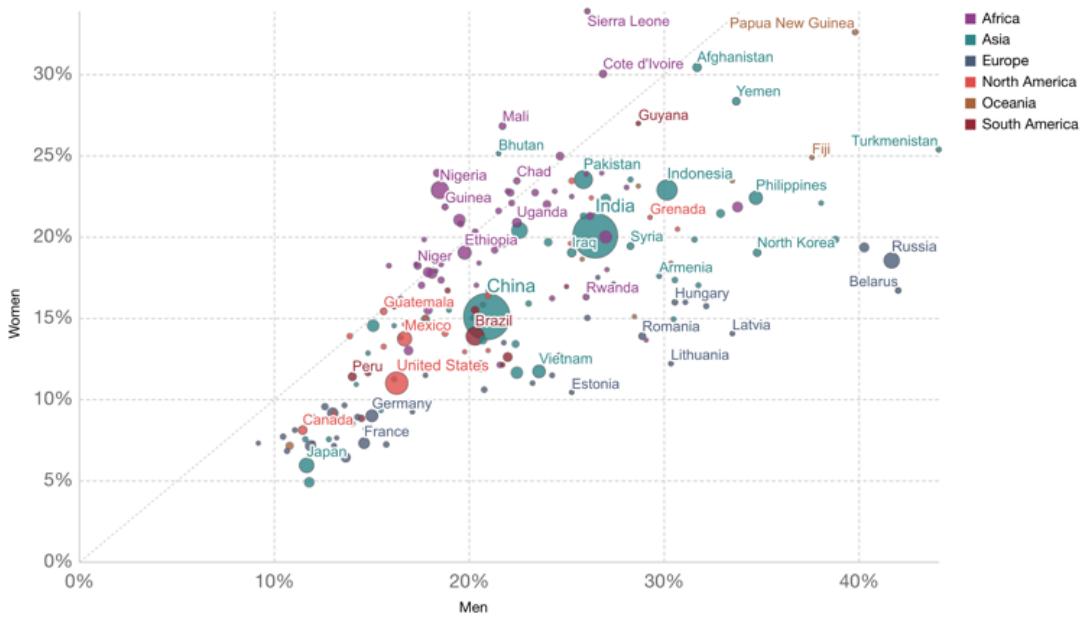
Y-axis is Saliva Viral Load Log (GE/ml)

The difference in viral loads lines up with the Male subgroup having worse vaccine efficacy than the Female subgroup: 88% (71-98) vs 96% (90-100) against symptomatic infections, 2nd dose+7 . It is also the case that across the world, women live longer and die less often from cardiovascular disease, cancer, diabetes, and chronic respiratory disease ([Our World in Data](#)).

Probability of death from any of the top global causes of death, 2015

Probability (%) of dying between age 30 and 70, from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease.

Our World
in Data



Source: World Health Organization - Global Health Observatory

CC BY

<https://ourworldindata.org/why-do-women-live-longer-than-men>

Johnson & Johnson & Friends

Although not an mRNA vaccine like today's star, Pfizer's BNT162b2, the [J&J clinical trial](#) tracked multiple endpoints that show our hoped-for trend.

- ~40,000 participants (treatment and control)
- 65% of cohort is 18-59yr

The design/reporting of the J&J clinical trial differs from others, particularly the large Pfizer observational study. The time periods and outcomes are defined differently.

To avoid copying multiple tables, I've extracted the numbers as I've understood them.

	Onset at Least 14 Days			Onset at Least 28 Days		
	Treatment	Placebo	VE% (95% CI)	Treatment	Placebo	VE% (95% CI)
Moderate to Severe, 18-59 yrs	95	260	63.7% (53.9-71.6)	66	193	66.1% (53.3-75.8)
Moderate to Severe, >=60 years	21	88	76.3% (61.6-86.0)	14	41	66.2% (36.7-83.0)

	Onset at Least 14 Days			Onset at Least 28 Days		
Severe /Critical, 18-59 yrs	12	52	76.9% (56.2-88.8)	5	33	85% (61.2-95.4)
Severe /Critical, >= 60 years	7	28	75.1% (41.7-90.8)	3	15	80.2% (30-96.3)
Requiring Medical Intervention	2	14	85.7% (37.8-98.4)	0	7	100% (31.1-100)

- Moderate Covid-19:* Positive test, any 1 really bad symptom or 2 of regular symptoms like fever, sore throat, cough
- Severe/Critical:* 3+ regular symptoms or things like respiratory rate ≥ 30 breaths/minute, heart rate ≥ 125 beats/minute, oxygen saturation (SpO_2) $\leq 93\%$, shock, admission to ICU, death
- Requiring Medical Intervention:* hospitalization, ICU admission, mechanical ventilation, and/or ECMO.

There were deaths in the J&J study, all within the placebo group:

Table 19. COVID-19 Related Deaths

Arm	Study Day ^c	Age	Comorbidity
Placebo	15	63	Obesity, Hypertension
Placebo	18 ^a	52	Obesity, Diabetes
Placebo	31	54	Obesity, Hypertension, Diabetes, Heart failure
Placebo	38	49	Obesity, Hypertension
Placebo	39	68	Obesity
Placebo	49 ^b	60	Obesity
Placebo	55	60	Asthma

^a Participant with positive SARS-CoV-2 PCR at baseline

^b Reported after the primary analysis cutoff date of January 22, 2021

^c Study day of death

The authors note that all these cases occurred at study sites in South Africa. Hmmm.

Using a somewhat different formula, the authors also report on interim asymptomatic results. They present four different operationalizations of which I choose two, the ones with the highest and lowest efficacy after 29 days. See Table 20 for further detail.

	Day 1-Day 29			After Day 29		
	Treatment	Placebo	VE% (95% CI)	Treatment	Placebo	VE% (95% CI)

	Day 1-Day 29			After Day 29		
FAS seronegative at baseline, +PCR and/or serology and did not show signs and symptoms	159	182	12.5% (-8.9- 29.7)	22	54	59.7% (32.8- 76.6)
Seroconverted without previous symptoms	84	180	22.6% (-3.9- 42.5)	10	37	74.2% (47.88- 88.6)

Despite the J&J trial using overlapping criteria, both explicitly lumping things and in their criteria, we see the progression we'd expect to see if vaccines work better to prevent worse outcomes than they do milder ones. Modulo confidence intervals, that is.

While efficacy against moderate to severe Covid-19 is 65% (for 18-59 years old), it jumps to 85% for severe alone, granted the overlap in confidence intervals. 0 cases in the vaccine group fell into the *Requiring Medical Intervention* endpoint, compared to 7 with placebo. It's as good as we could hope to see with this data.

However, there isn't a clear jump between "asymptomatic" and "moderate to severe". Partly because the operationalization isn't clear (59.7% to 66% is a jump) but there are still wide error bars. The After Day 29 antibody test was conducted at Day 71 and had only been completed for ~30% of participants at the time of publication.

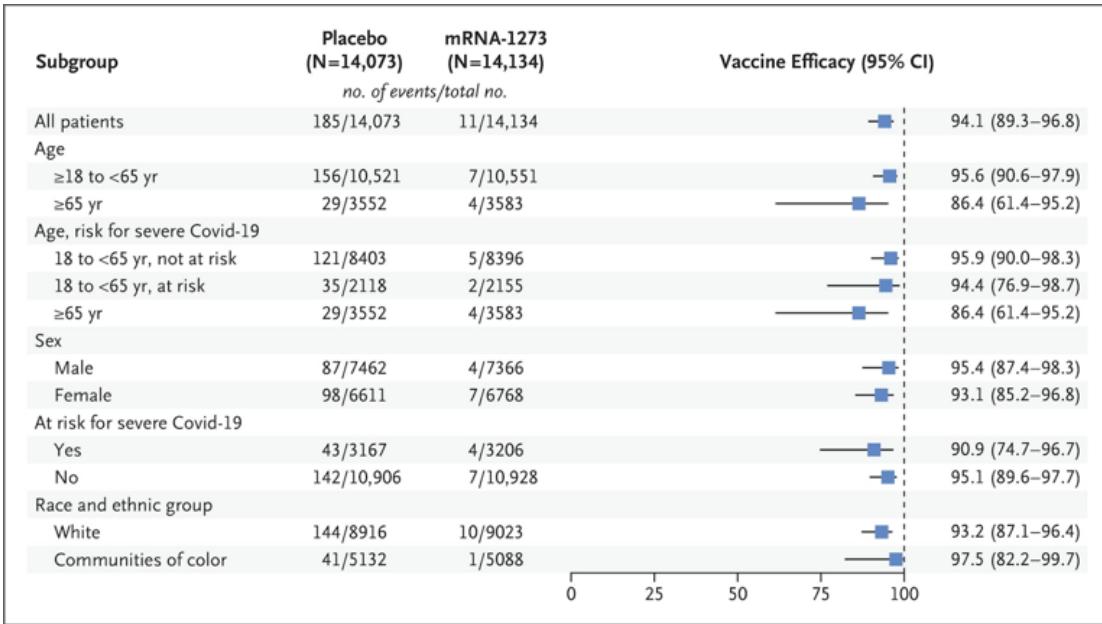
On net, I think the overall endpoint trend lines up with increasing vaccine efficacy against more severe outcomes, but the asymptomatic vs symptomatic doesn't clearly show it, but part of that is I don't really understand the groups or how to interpret them.

Let's get clinical, clinical

Moderna Clinical Trial

The Moderna Phase III Trial tracked Covid and severe Covid but no other end points. There were no severe Covid or deaths in the vaccine group, but 30 severe Covid cases in control and one death. Doesn't let you compute a precise value, but is consistent with the vaccines being *very* good.

- 59% of cohort is aged 18-64 yr and not in any risk categories
- Further breakdown of participant characteristics in Table 1
- Asymptomatic Infections and Hospitalizations were not tracked.
- Table S13 in the appendix provides a very detailed breakdown of symptoms. No severe symptoms at all occurred to participants in the treatment group, granted that very few occurred in the control group either.



Pfizer Clinical Trial

- 58% of cohort is aged 16-55yr, median 52yr, range 16-89
- 35% have BMI ≥ 30
- After the second dose, there 5 cases of severe-Covid in the control group and 1 in treatment, no reported deaths.

It's nice (for me) to note that the vaccine efficacy, and particularly the confidence intervals, are higher for the 16-55yr group. 95.6% with CI 89.4%-98.6%. Doesn't prove the main point, but is in line with the immunosenescence model.

No Covid deaths are reported in either placebo or vaccine groups. Table S5 from the appendix least severe-Covid outcomes. A total of 9 after the first dose in control, 1 in the vaccine group for a 89% reduction with confidence interval between 20.1%-99.7%. If we break it down to different time periods (before/after 1st/2nd dose), we end up with confidence intervals (-3800% to 100%). Yes, maybe taking the vaccine will increase your chance of severe-Covid by 39x!

As expected, the data doesn't show that the reduction in severe-Covid is greater than lesser-Covid, but it also they doesn't show the opposite either.

Table 3. Vaccine Efficacy Overall and by Subgroup in Participants without Evidence of Infection before 7 Days after Dose 2.

Efficacy End-Point Subgroup	BNT162b2 (N=18,198)		Placebo (N=18,325)		Vaccine Efficacy, % (95% CI)†
	No. of Cases	Surveillance Time (No. at Risk)*	No. of Cases	Surveillance Time (No. at Risk)*	
Overall	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.0–97.9)
Age group					
16 to 55 yr	5	1.234 (9,897)	114	1.239 (9,955)	95.6 (89.4–98.6)
>55 yr	3	0.980 (7,500)	48	0.983 (7,543)	93.7 (80.6–98.8)
≥65 yr	1	0.508 (3,848)	19	0.511 (3,880)	94.7 (66.7–99.9)
≥75 yr	0	0.102 (774)	5	0.106 (785)	100.0 (-13.1–100.0)
Sex					
Male	3	1.124 (8,875)	81	1.108 (8,762)	96.4 (88.9–99.3)
Female	5	1.090 (8,536)	81	1.114 (8,749)	93.7 (84.7–98.0)
Race or ethnic group‡					
White	7	1.889 (14,504)	146	1.903 (14,670)	95.2 (89.8–98.1)
Black or African American	0	0.165 (1,502)	7	0.164 (1,486)	100.0 (31.2–100.0)
All others	1	0.160 (1,405)	9	0.155 (1,355)	89.3 (22.6–99.8)
Hispanic or Latinx	3	0.605 (4,764)	53	0.600 (4,746)	94.4 (82.7–98.9)
Non-Hispanic, non-Latinx	5	1.596 (12,548)	109	1.608 (12,661)	95.4 (88.9–98.5)
Country					
Argentina	1	0.351 (2,545)	35	0.346 (2,521)	97.2 (83.3–99.9)
Brazil	1	0.119 (1,129)	8	0.117 (1,121)	87.7 (8.1–99.7)
United States	6	1.732 (13,359)	119	1.747 (13,506)	94.9 (88.6–98.2)

* Surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.

† The confidence interval (CI) for vaccine efficacy is derived according to the Clopper-Pearson method, adjusted for surveillance time.

‡ Race or ethnic group was reported by the participants. "All others" included the following categories: American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander, multiracial, and not reported.

Efficacy Endpoint Subgroup	BNT162b2 (30 µg) (N ^a =21669)		Placebo (N ^a =21686)		VE (%)	(95% CI) ^e
	n ^b	Surveillance Time ^c (n ^d)	n ^b	Surveillance Time ^c (n ^d)		
Severe COVID-19 occurrence after Dose 1	1	4.021 (21314)	9	4.006 (21259)	88.9	(20.1, 99.7)
After Dose 1 to before Dose 2	0		4		100.0	(-51.5, 100.0)
Dose 2 to 7 days after Dose 2	0		1		100.0	(-3800.0, 100.0)
≥7 Days after Dose 2	1		4		75.0	(-152.6, 99.5)

As described above, we have to note the greatly reduced sample size in the later periods.

Evidence of Increased Asymptomatic/Symptomatic Ratios

This was already shown in the mass Pfizer study, but several other sources indicate the ratio of asymptomatic-to-symptomatic cases is increased for vaccinated people. In other words, vaccination works better against symptomatic Covid (more severe) than asymptomatic Covid (less severe). This is at earlier side of "severity", compared to hospitalization, severe-Covid, and death, but it suggests the same trend-plus there's more data than when looking at more severe outcomes.

I've copied most of the numbers here from MicroCOVID.org, see [their analysis](#) for calculation detail.

	Unvaccinated		Vaccinated	
	Symptomatic vs Asymptomatic	% Asymptomatic	Symptomatic vs Asymptomatic	% Asymptomatic
Mass Pfizer Study, Day 28+	210 vs 191	48%	31 vs 59	65%
MicroCOVID Moderna/CDC Calculation	185 vs 37	17%	11 vs 11	50%
J&J	351 vs 182	34%	117 vs 159	57%
AstraZeneca	248 vs 73	23%	84 vs 57	40%

Pfizer Asymptomatic percentage goes from 48%=>65%, Moderna from 17%=>50%, J&J from 34%-57%, AstraZeneca 23%->40%

I expect the very different absolute numbers to come from the widely varying study methodologies as much as differences between the vaccines. (AstraZeneca uses home tests, for example.)

I didn't exhaustively look through all possible sources of asymptomatic vs symptomatic efficacy. I would be very interested if someone had a credible source not showing this trend.

I also didn't scrutinize these calculations much, so I wouldn't be *that* surprised if it turned out there were deep flaws that undermine the trend seen here.

It gets better

If we go back to the big Israeli Pfizer observational study, we see increasing vaccine effectiveness as more time passes since first/second dose. Unfortunately, the study didn't have enough time/data to show us things two weeks after the 2nd dose.

Fortunately, there was a follow-up. On March 11, Pfizer/Israel Ministry of Health made the [following press release](#):

Findings from the analysis were derived from de-identified aggregate Israel MoH surveillance data collected between January 17 and March 6, 2021, when the Pfizer-BioNTech COVID-19 Vaccine was the only vaccine available in the country and when the more transmissible B.1.1.7 variant of SARS-CoV-2 (formerly referred to as the U.K. variant) was the dominant strain. **Vaccine effectiveness was at least 97% against symptomatic COVID-19 cases, hospitalizations, severe and critical hospitalizations, and deaths.** Furthermore, the analysis found a vaccine effectiveness of 94% against asymptomatic SARS-CoV-2 infections. For all outcomes, vaccine effectiveness was measured from two weeks after the second dose.

The lack of actual paper makes this a little harder to interpret, but I don't find it surprising given that (1) at this later date in their roll-out, an even greater proportion of people will be young and healthy, (2) this data is only counting two weeks after the second dose, whereas the previous large observational study only had a "7 days after 2nd dose until end of follow-up" (maximum of day 28 to 44).

And, of no small significance, the Pfizer vaccine appears fully effective against the UK variant. (Yay!!)

If the vaccine is showing 97+% amongst everyone, I would expect that's at least as true when you filter for younger/healthier people and filter out those with comorbidities.

What I believe

I believe that what I wrote above supports my initial assertions:

1. **In the case of the Pfizer vaccine (and likely Moderna too), the effectiveness in young healthy people is 99% or close to it.**
2. **That we can reasonably expect the effectiveness of the vaccine against more severe Covid to be greater than effectiveness against milder cases of Covid.**

The initial Pfizer mass study has 99% (96-100) for the age 16-39yr group, and the subsequent follow-up gives 97% for everyone. At baseline for symptomatic cases, we're talking 30-100x reductions, which is hella good.

Further, across multiple studies, vaccines, and outcomes we see trends of increasing effectiveness against progressively severe outcomes. In some cases, we don't definitively see it, but that's easily attributable to lack of sample size and inherent limitations in methodology due to noise and selection effects.

If we're talking 99% against symptomatic cases (100x reduction), then I think it's reasonable to expect at least that for hospitalization, 99.5% (200x). Hence the title, *extreme vaccine effectiveness*.

What about J&J (and AstraZeneca)? Granted, The effectiveness numbers for J&J look lower than for Pfizer and Moderna, but I think they're higher than [MicroCOVID.org's numbers](#) imply. First, we get 85% (61.2-95.4) effectiveness against Severe/Critical in the 18-59yr subgroup. That number matters more. Second, that is a very wide age range. I would bet that restricting it to 18-39 would show an improvement relevant to most of those reading this. Lastly, I suspect all the factors mentioned above (selection effects, noise/saturation) to affect it and make the result lower than it would be otherwise.

On net, J&J might not be quite as extremely effective as the mRNA vaccines, but it's no pushover either.

AstraZeneca isn't on offer in the Bay, and was recently abandoned in my home country of Australia too, so I apologize for not examining it.

Tell me where I'm wrong

I want the case I've made today to be true, but even more than that I want to believe true things (and I certainly don't want people to believe false things because of me). If you think any of this is wrong. PLEASE SAY SO.

Many thanks!

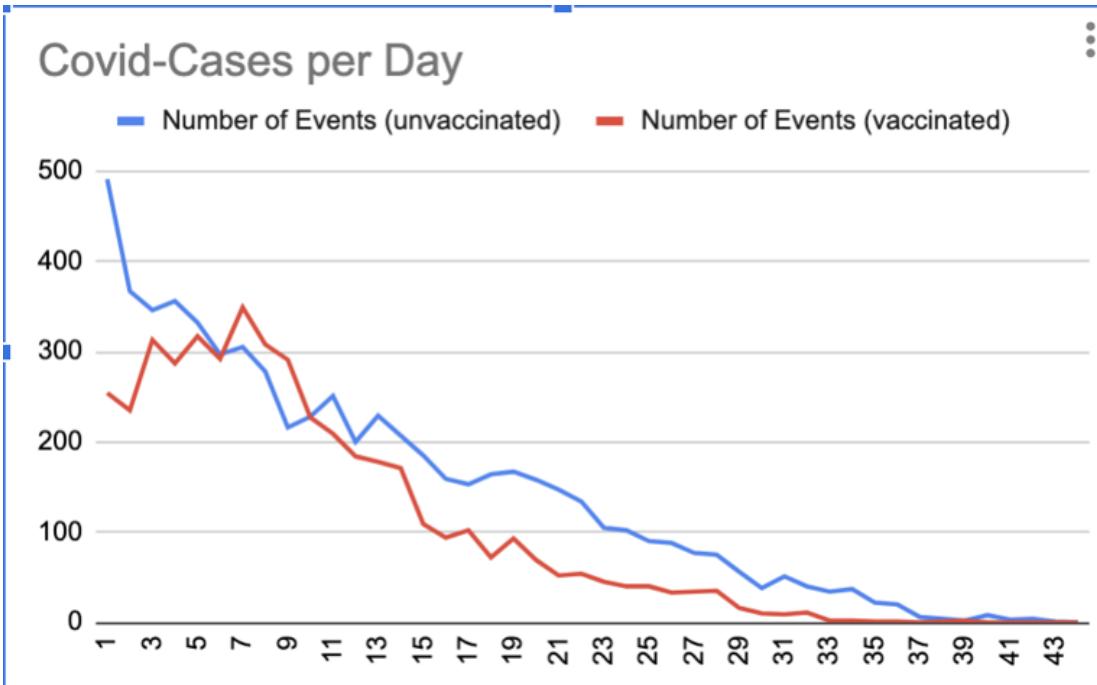
Objections

Maybe the vaccine and control groups are different?

The Mass Israeli observation study is not an RCT. Instead, it's looking at their database of vaccinated people and then trying to match each of them to someone who's mostly the same except not vaccinated.

This might be hard to do. The kinds of people who seek out vaccines ("health-seeking" behavior) are probably a bit different from those who haven't gotten around to it. Maybe they wear better masks (or masks at all) on the hand, and maybe they risk-compensate on the other.





The authors attempt to control for this by 1:1 matching vaccinated people and controls on multiple characteristics. They cite as evidence this worked the fact the Covid cases among vaccinated and non-vaccinated groups are the same before Day 12 (before the vaccine really starts working), and explain the initial gap in the first couple of days as a selection effect, because vaccinated people only choose to get vaccinated when they're feeling well.

I didn't look at the data especially long, but I'm not sure what to make of it. If vaccinated people are more cautious, even after being vaccinated, this means vaccines are less effective than we'd otherwise think. If they risk-compensate, that means vaccines are more effective than what we observe.

I'm not sure how this should net out, but I do think it should widen confidence intervals somewhat.

I also expect the effect to get stronger over time in Israel, as the people who remain unvaccinated and serve as controls (in observational studies) who least cared about Covid and getting vaccinated. So...maybe that explains the later 97% effectiveness result.

Ben Shaya's Thoughts

Ben Shaya, one of the people responsible for [MicroCOVID.org](#)'s models, has been kindly taking time to discuss the topic of vaccine effectiveness with me.

He wrote a document arguing [Why I think vaccines don't bring the chance of severe COVID to 0](#). While that's a stronger claim than I would make, his arguments and models are still valuable when thinking about the topic generally.

Read them all [in his doc](#), but I'll highlight one that I found very relevant to mechanistic models of Covid severity:

There's *is* a bit of mechanistic nuance - COVID tends to start with an upper respiratory tract infection - that's when nasal swabs work, and then moves to the lower respiratory tract (lungs) - which is where it screws over your blood oxygen. There's research suggesting that the immune response of the lining of your windpipe is what determines whether the virus reaches your lungs: <https://www.biorxiv.org/content/10.1101/2021.02.20.431155v1.abstract?%3Fcollection%3D>

That is to say, there isn't a single "immune response" that determines how bad your infection is; immune response in your upper respiratory tract determines if you get COVID while immune response in your lungs and windpipe determine whether you get a severe reaction. Immune response in your blood correlates with these, but it's not the same as either.

(this comes from Riley Drake, who is a virologist and one of the authors on the paper, who strongly cautioned against reading too deeply into antibody concentrations as a proxy for immunity)

When we treat these three immune systems as separate, we see there are at least 3 hidden models - upper respiratory response, lower respiratory response, and blood response. Of the 3, the side effects to the vaccine only reflect the blood response. The overall efficacy of the vaccine only reflects the upper response.

Further, we know that humans have significant heterogeneity in all three of these responses, since only some people get severe covid, and some people get lung damage from mild covid, and some people get exposed to COVID and don't contract the virus at all.

Note further that, assuming a strong immune response protects you is also not necessarily true; people with very strong response can cause cytokine storms (which we now know how to handle, but will land you in a hospital).

This gets at more gears in immune response and is the kind of thing that can expose where simple state transition models and immune response as a single thing don't hold up.

What if the vaccine boost one part of the immune system but not another? In that case, you might see the vaccine be very effective against symptomatic and asymptomatic Covid, but not more severe disease. If the virus makes its way deep into your lungs, and the lungs are protected by a different immune response that isn't helped by the vaccine- then, conditional on having gotten to that point, you might not be better off than a non-vaccinated person.

All this to say we should be cautious in putting too much stock in simple mechanistic models.

What if vaccines are all or nothing?

The model behind the claim of extreme vaccine efficacy is that even if your post-vaccine immune response isn't enough to stop you getting Covid at all, it should be stronger than it would have otherwise been, and you'll do better at fighting off severe-Covid. This takes vaccine efficacy as a continuous thing.

But maybe the vaccine is 100% effective against all outcomes! So long as it's correctly transported and administered, that is. Except sometimes vaccines are left at high temperature for too long, the delicate proteins are damaged, and people receiving them are effectively not vaccinated. If this happens 5% of the time, then 95% of people are completely immune to Covid and 5% are identical to not be vaccinated. Whatever chance they had of getting severe Covid before, it's the same now.

In this world, not knowing whether your vaccine was a dud or not, post-vaccine you should assume you have a 95% reduction across all outcomes equally.

I originally found this argument very persuasive. How could assume that the "continuous" model of vaccine-immune response was true? But actually, most things are continuous in the real world, particularly in biology. People aren't old or young, but somewhere on a continuous measure. Immune response isn't an all or nothing affair, and some people's body's will produce more antibodies than others. Some enough to stymy any symptoms at all, but some only enough to prevent them getting hospitalized.

One person told me that the mRNA vaccines induce 60x the antibodies of a recovered Covid patient ([closest source I found](#)), such that even if your response was weaker, it should still be

more than powerful to deal with any actual Covid. Therefore, we should take people still getting quite sick with Covid as a sign that some people's vaccines must not be working at all.

I would be surprised if complete vaccine failure didn't occur some of the time, the question is how much. Suppose we have a vaccine that, when it works successfully, confers 100x reduction (99% efficacy). If it fails to work at all 5% of the time, we'd see a $1 - (0.01 \cdot 0.95 + 1 \cdot 0.05) = 94\%$ efficacy overall. Something like that could be a big part of what we observe.

Also, Kelsey Piper says that she couldn't find anything about any other vaccines working this way (all or nothing). [Edited: However, as [Romeo points out in the comments](#), Moderna and Pfizer are mRNA vaccines that might not behave like the older types.]

Caveats - read before you act

In my long case for the extreme effectiveness of vaccines, there are some topics of crucial practical importance. These imply that maybe you don't want to throw caution to the wind just yet. I'm not sure, I didn't get to looking into these.

Long-Covid

Long-Covid is not an endpoint tracked by any of the studies I've looked at. I would think that'd be related to increased viral load and behave like an outcome more severe than just a symptomatic case, but there isn't data for that. Anecdotally, I've heard of a couple of cases where someone experienced mild Covid yet was dramatically affected for months afterwards.

My conservative gut estimate is that your odds of getting long-Covid are reduced by a vaccine by as much as your chance of getting symptomatic Covid at all, but not necessarily any more than that.

Variants

The Israeli study showed supreme efficacy even against the [much-feared UK variant](#) (B.1.1.7). However, there are fears the vaccines aren't not nearly as good against the South African variant (B.1.351) or Brazilian P.1 variant.

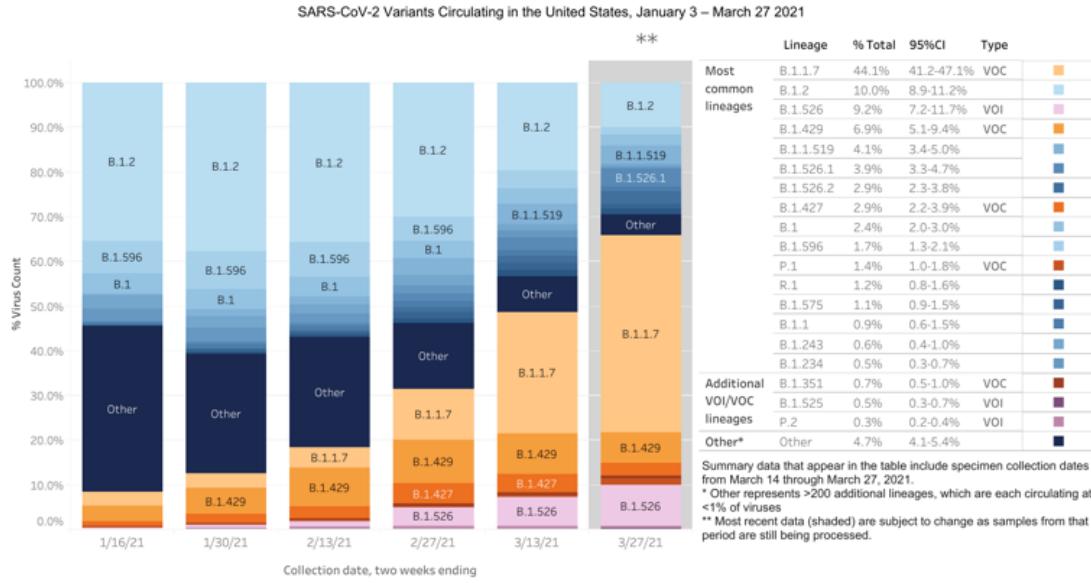
A new study published on MedRxiv last week, [Evidence for increased breakthrough rates of SARS-CoV-2 variants of concern in BNT162b2 mRNA vaccinated individuals](#), state that vaccinated individuals were more likely to contract B.1.1.7 and B.1.351 than unvaccinated controls, suggesting both variants are more resistant to vaccination than other strains. (Caveat: I didn't read the paper in detail.)

. . . we performed a case-control study that examined whether BNT162b2 vaccinees with documented SARS-CoV-2 infection were more likely to become infected with B.1.1.7 or B.1.351 compared with unvaccinated individuals.

Vaccinees infected at least a week after the second dose were disproportionately infected with B.1.351 (odds ratio of 8:1). Those infected between two weeks after the first dose and one week after the second dose, were disproportionately infected by B.1.1.7 (odds ratio of 26:10), suggesting reduced vaccine effectiveness against both VOCs under different dosage/timing conditions.

Not good.

The [CDC's CovidTracker page](#) actually has some nice dashboards for tracking variant proportion, though I haven't looked into the data quality. Their [brief](#) is helpful too.



<https://covid.cdc.gov/covid-data-tracker/#variant-proportions>

There's a breakdown by US state too. I'm in California and am pleased to see that currently, B.1.351 is only 0.3% of the Covid cases and P.1 is 1.6%

Proportions of SARS-CoV-2 Variants of Concern by State

State	B.1.1.7	B.1.351	B.1.427/B.1.429	P.1	Other lineages	Total Available Sequences
Arizona	34.1%				49.2%	411
California	15.9%	0.3%	53.8%	1.6%	28.4%	6,919
Colorado	29.1%	0.1%	28.1%	0.8%	42.0%	915
Connecticut	29.2%	0.8%	7.5%	0.8%	61.8%	651
Florida	52.2%	0.3%	7.5%	2.4%	37.6%	6,093

Of course, if B.1.351 and friends are resistant to vaccination, we will see them rise in prevalence.

This needs more investigation. Without looking into it more, the sensible strategy would be something like act according to your local prevalences and beliefs about how vaccine-resistant the different strains are. Right now the suspicions are on B.1.351 and P.1, but they're uncommon in California (0.3% and 1.6% respectively).

If you've got the time and skill to look into this more, please do, I can provide you money and glory.

Spreading to the Unvaccinated

Even if the vaccine protects you 200x against more severe outcomes, that doesn't help the unvaccinated if they catch Covid from you when you had an asymptomatic or mild case. This means that until such time as those you interact with most are vaccinated, you might want to be more conservative in your microCovids.

MicroCOVID.org calculated reductions in [contagiousness for vaccinated people](#) (10x reduction for Pfizer/Moderna, 3x for J&J), and that's what I'd stick to myself if interacting with the unvaccinated. (But hey, California is just about at universal eligibility, now is the time!)

Predictive Coding has been Unified with Backpropagation

Artificial Neural Networks (ANNs) are based around the backpropagation algorithm. The backpropagation algorithm allows you to perform gradient descent on a network of neurons. When we feed training data through an ANNs, we use the backpropagation algorithm to tell us how the weights should change.

ANNs are good at inference problems. Biological Neural Networks (BNNs) are good at inference too. ANNs are built out of neurons. BNNs are built out of neurons too. It makes intuitive sense that ANNs and BNNs might be running similar algorithms.

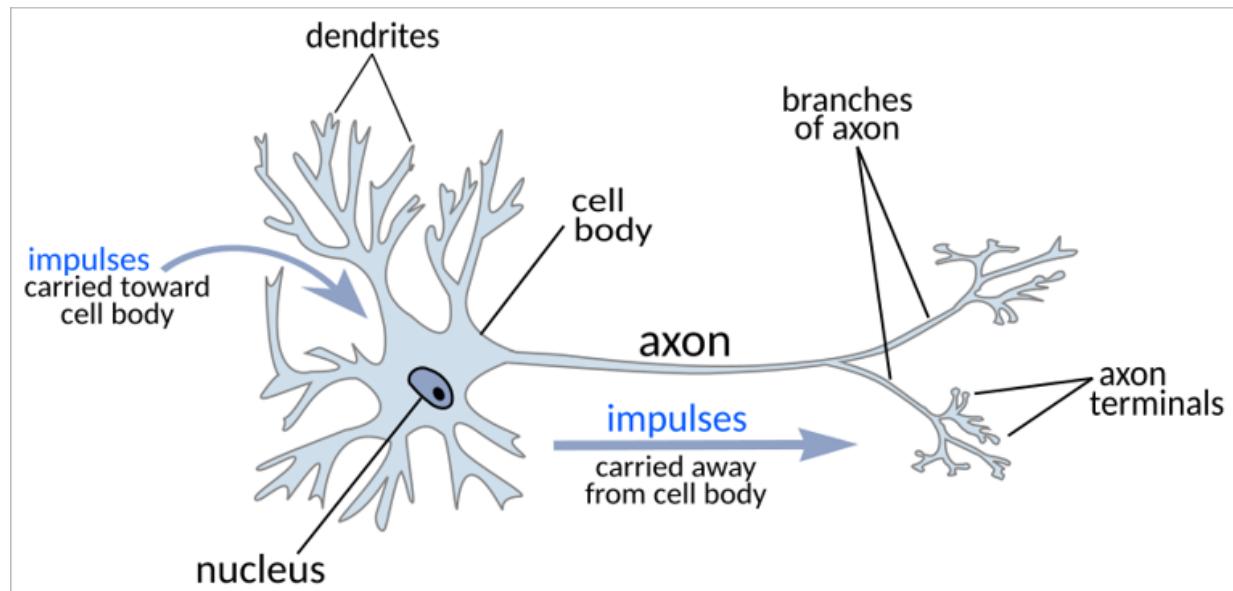
There is just one problem: BNNs are physically incapable of running the backpropagation algorithm.

We do not know quite enough about biology to say it is impossible for BNNs to run the backpropagation algorithm. However, "a consensus has emerged that the brain cannot directly implement backprop, since to do so would require biologically implausible connection rules"^[1].

The backpropagation algorithm has three steps.

1. Flow information forward through a network to compute a prediction.
2. Compute an error by comparing the prediction to a target value.
3. Flow the error backward through the network to update the weights.

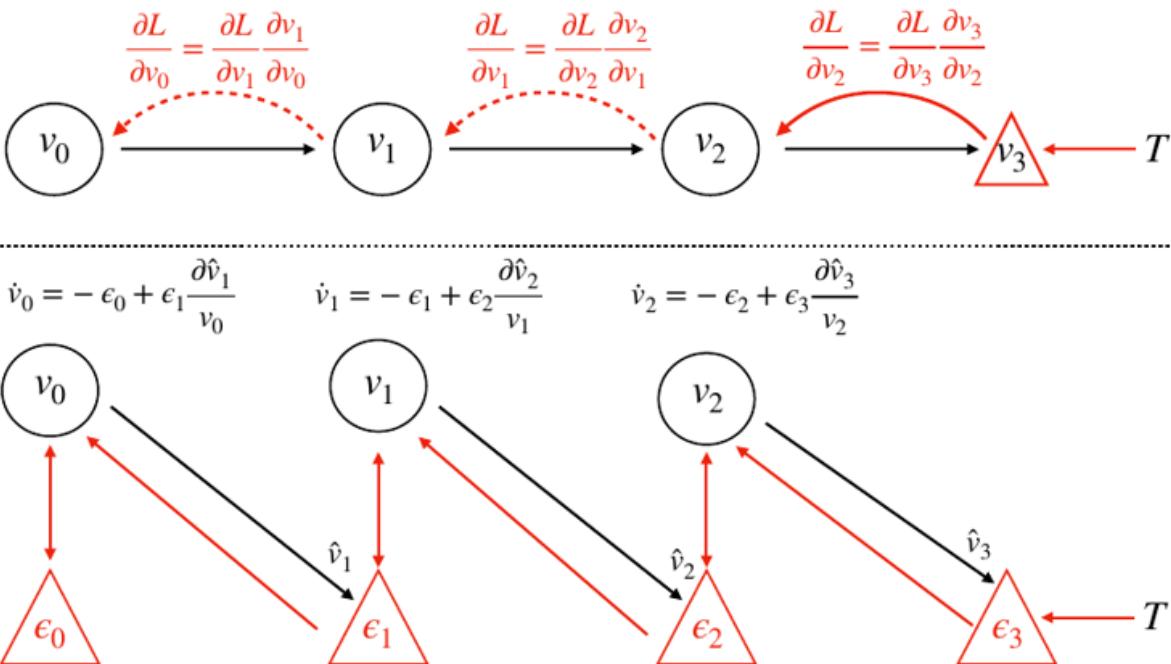
The backpropagation algorithm requires information to flow forward and backward along the network. But biological neurons are one-directional. An action potential goes from the cell body down the axon to the axon terminals to another cell's dendrites. An axon potential never travels backward from a cell's terminals to its body.



Hebbian theory

Predictive coding is the idea that BNNs generate a mental model of their environment and then transmit only the information that deviates from this model. Predictive coding considers error and surprise to be the same thing. Hebbian theory is specific mathematical formulation of predictive coding.

Predictive coding is biologically plausible. It operates locally. There are no separate prediction and training phases which must be synchronized. Most importantly, it lets you train a neural network without sending axon potentials backwards.



Predictive coding is easier to implement in hardware. It is locally-defined; it parallelizes better than backpropagation; it continues to function when you cut its substrate in half. (Corpus callosotomy is used to treat epilepsy.) Digital computers break when you cut them in half. Predictive coding is something evolution could plausibly invent.

Unification

The paper *Predictive Coding Approximates Backprop Along Arbitrary Computation Graphs*^[1:1] "demonstrate[s] that predictive coding converges asymptotically (and in practice rapidly) to exact backprop gradients on arbitrary computation graphs using only local learning rules." The authors have unified predictive coding and backpropagation into a single theory of neural networks. Predictive coding and backpropagation are separate hardware implementations of what is ultimately the same algorithm.

There are two big implications of this.

- This paper permanently fuses artificial intelligence and neuroscience into a single mathematical field.

- This paper opens up possibilities for neuromorphic computing hardware.
-

1. Source is [available on arxiv](#). ↵ ↵

Testing The Natural Abstraction Hypothesis: Project Intro

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

The [natural abstraction hypothesis](#) says that

- Our physical world abstracts well: for most systems, the information relevant “far away” from the system (in various senses) is much lower-dimensional than the system itself. These low-dimensional summaries are exactly the high-level abstract objects/concepts typically used by humans.
- These abstractions are “natural”: a wide variety of cognitive architectures will learn to use approximately the same high-level abstract objects/concepts to reason about the world.

If true, the natural abstraction hypothesis would dramatically simplify AI and AI alignment in particular. It would mean that a wide variety of cognitive architectures will reliably learn approximately-the-same concepts as humans use, and that these concepts can be precisely and unambiguously specified.

Ultimately, the natural abstraction hypothesis is an empirical claim, and will need to be tested empirically. At this point, however, we lack even the tools required to test it. This post is an intro to a project to build those tools and, ultimately, test the natural abstraction hypothesis in the real world.

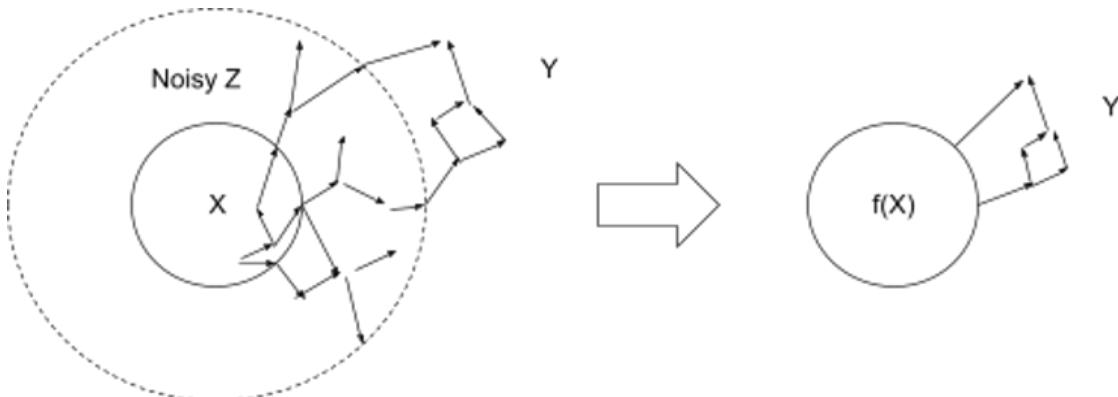
Background & Motivation

One of the major conceptual challenges of designing human-aligned AI is the fact that [human values are a function of humans' latent variables](#): humans care about abstract objects/concepts like trees, cars, or other humans, not about low-level quantum world-states directly. This leads to conceptual problems of defining “what we want” in physical, reductive terms. More generally, it leads to conceptual problems in [translating](#) between human concepts and concepts learned by other systems - e.g. ML systems or biological systems.

If true, the natural abstraction hypothesis provides a framework for translating between high-level human concepts, low-level physical systems, and high-level concepts used by non-human systems.

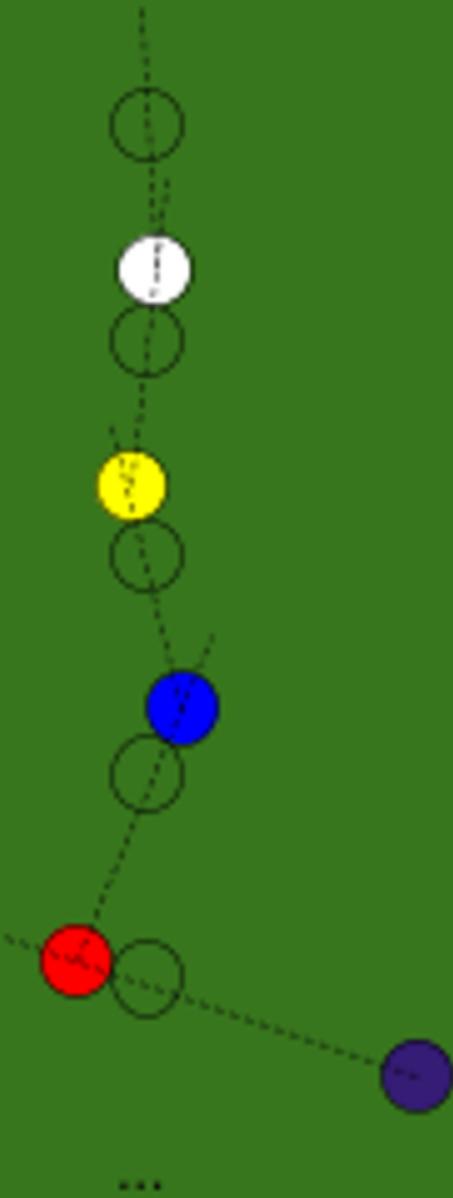
The foundations of the framework have been sketched out in previous posts.

[What is Abstraction?](#) introduces the mathematical formulation of the framework and provides several examples. Briefly: the high-dimensional internal details of far-apart subsystems are independent given their low-dimensional “abstract” summaries. For instance, the [Lumped Circuit Abstraction](#) abstracts away all the details of molecule positions or wire shapes in an electronic circuit, and represents the circuit as components each summarized by some low-dimensional behavior - like $V = IR$ for a resistor. This works because the low-level molecular motions in a resistor are independent of the low-level molecular motions in some far-off part of the circuit, *given* the high-level summary. All the rest of the low-level information is “wiped out” by noise in low-level variables “in between” the far-apart components.



In the causal graph of some low-level system, X is separated from Y by a bunch of noisy variables Z. For instance, X might be a resistor, Y might be a capacitor, and Z might be the wires (and air) between them. Noise in Z wipes out most of the low-level info about X, so that only a low-dimensional summary $f(X)$ is relevant to predicting the state of Y.

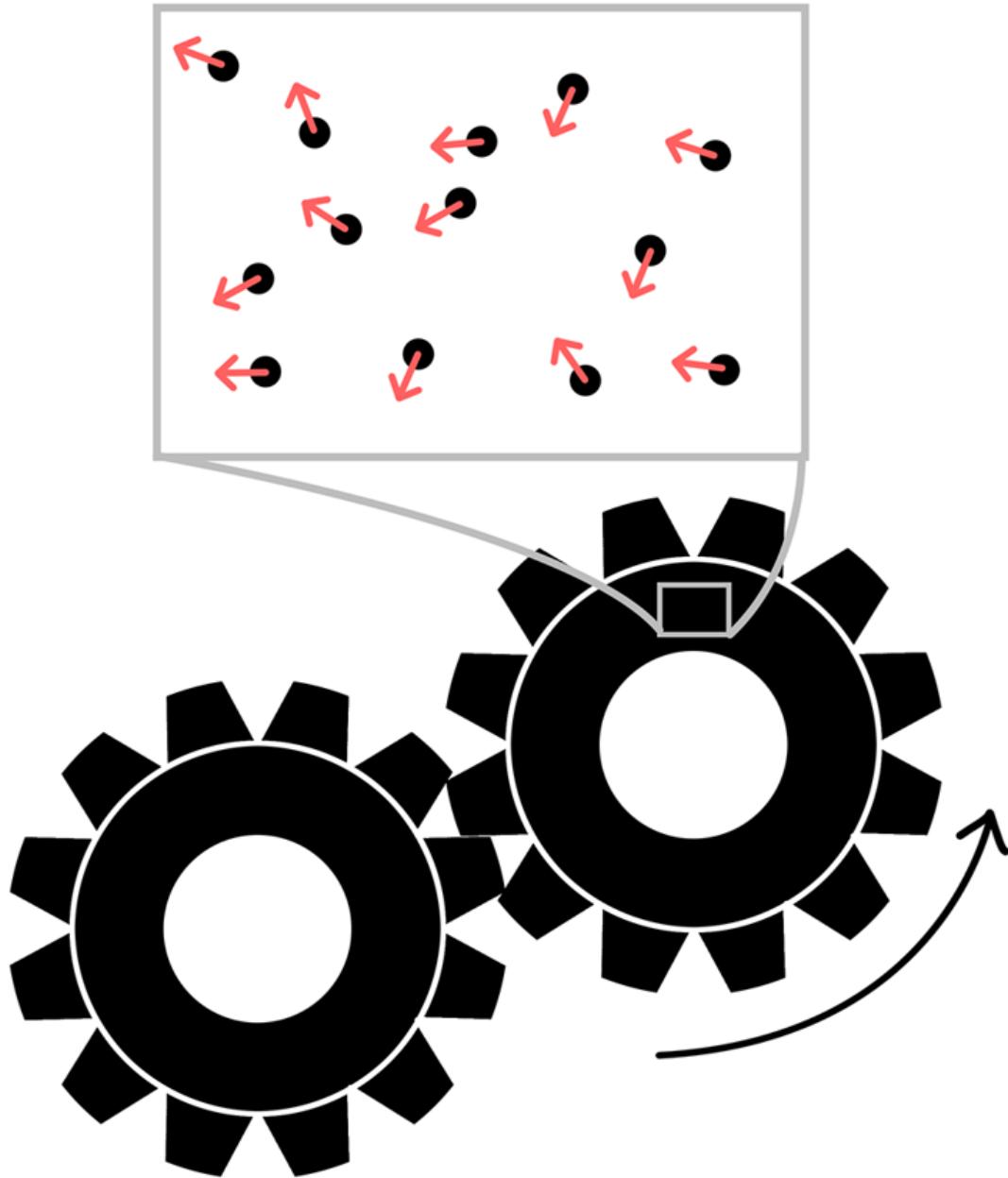
[Chaos Induces Abstractions](#) explains one major reason *why* we expect low-level details to be independent (given high-level summaries) for typical physical systems. If I have a bunch of balls bouncing around perfectly elastically in a box, then the total energy, number of balls, and volume of the box are all conserved, but chaos wipes out all other information about the exact positions and velocities of the balls. My “high-level summary” is then the energy, number of balls, and volume of the box; all other low-level information is wiped out by chaos. This is exactly the abstraction behind the ideal gas law. More generally, given any uncertainty in initial conditions - even very small uncertainty - mathematical chaos “amplifies” that uncertainty until we are maximally uncertain about the system state... except for information which is perfectly conserved. In most dynamical systems, some information is conserved, and the rest is wiped out by chaos.



In a system of billiard balls, a small initial error in a ball's angle is exponentially amplified as the balls travel and bounce off each other. This is chaos.

Anatomy of a Gear: What makes a good “gear” in a gears-level model? A physical gear is a very high-dimensional object, consisting of huge numbers of atoms rattling around. But for purposes of predicting the behavior of the gearbox, we need only a one-dimensional

summary of all that motion: the rotation angle of the gear. More generally, a good “gear” is a subsystem which abstracts well - i.e. a subsystem for which a low-dimensional summary can contain all the information relevant to predicting far-away parts of the system.



If we look at all the atoms in one little chunk of one gear, only the average motion of all the atoms will tell us about the motion of far-away atoms in a neighboring gear.

[Science in a High Dimensional World](#): Imagine that we are early scientists, investigating the mechanics of a sled sliding down a slope. The number of variables which could conceivably influence the sled's speed is vast: 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, astrological motions of stars and planets, etc. Yet in practice, just a relatively-low-dimensional handful of variables suffices - maybe a

dozen. A consistent sled-speed can be achieved while controlling *only a dozen variables*, out of *literally billions*. And this generalizes: across every domain of science, we find that controlling just a relatively-small handful of variables is sufficient to reliably predict the system's behavior. Figuring out *which* variables is, in some sense, the central project of science. This is the natural abstraction hypothesis in action: across the sciences, we find that low-dimensional summaries of high-dimensional systems suffice for broad classes of "far-away" predictions, like the speed of a sled.



The Problem and The Plan

The natural abstraction hypothesis can be split into three sub-claims, two empirical, one mathematical:

- Abstractability: for most physical systems, the information relevant "far away" can be represented by a summary much lower-dimensional than the system itself.
- Human-Compatibility: These summaries are the abstractions used by humans in day-to-day thought/language.
- Convergence: a wide variety of cognitive architectures learn and use approximately-the-same summaries.

Abstractability and human-compatibility are empirical claims, which ultimately need to be tested in the real world. Convergence is a more mathematical claim, i.e. it will ideally involve proving theorems, though empirical investigation will likely still be needed to figure out exactly *which* theorems.

These three claims suggest three different kinds of experiment to start off:

- Abstractability: does reality abstract well? Corresponding experiment type: run a reasonably-detailed low-level simulation of something realistic; see if info-at-a-distance

is low-dimensional.

- Human-Compatibility: do these match human abstractions? Corresponding experiment type: run a reasonably-detailed low-level simulation of something realistic; see if info-at-a-distance recovers human-recognizable abstractions.
- Convergence: are these abstractions learned/used by a wide variety of cognitive architectures? Corresponding experiment type: train a predictor/agent against a simulated environment with known abstractions; look for a learned abstract model.

The first two experiments both require computing information-relevant-at-a-distance in a reasonably-complex simulated environment. The “naive”, brute-force method for this would not be tractable; it would require evaluating high-dimensional integrals over “noise” variables. So the first step will be to find practical algorithms for directly computing abstractions from low-level simulated environments. These don’t need to be fully-general or arbitrarily-precise (at least not initially), but they need to be general enough to apply to a reasonable variety of realistic systems.

Once we have algorithms capable of directly computing the abstractions in a system, training a few cognitive models against that system is an obvious next step. This raises another algorithmic problem: how do we efficiently check whether a cognitive system has learned particular abstractions? Again, this doesn’t need to be fully general or arbitrarily precise. It just needs to be general enough to use as a tool for the next step.

The next step is where things get interesting. Ideally, we want general theorems telling us which cognitive systems will learn which abstractions in which environments. As of right now, I’m not even sure exactly what those theorems should say. (There are some promising directions, like [modular variation of goals](#), but the details are still pretty sparse and it’s not obvious whether these are the right directions.) This is the perfect use-case for a feedback loop between empirical and theoretical work:

- Try training various cognitive systems in various environments, see what abstractions they learn.
- Build a model which matches the empirical results, then come up with new tests for that model.
- Iterate.

Along the way, it should be possible to prove theorems on what abstractions will be learned in at least some cases. Experiments should then probe cases not handled by those theorems, enabling more general models and theorems, eventually leading to a unified theory.

(Of course, in practice this will probably also involve a larger feedback loop, in which lessons learned training models also inform new algorithms for computing abstractions in more-general environments, and for identifying abstractions learned by the models.)

The end result of this process, the holy grail of the project, would be a system which provably learns all learnable abstractions in a fairly general class of environments, and represents those abstractions in a legible way. In other words: it would be a standardized tool for measuring abstractions. Stick it in some environment, and it finds the abstractions in that environment and presents a standard representation of them. Like a thermometer for abstractions.

Then, the ultimate test of the natural abstraction hypothesis would just be a matter of pointing the abstraction-thermometer at the real world, and seeing if it spits out human-recognizable abstract objects/concepts.

Summary

The natural abstraction hypothesis suggests that most high-level abstract concepts used by humans are “natural”: the physical world contains subsystems for which all the information relevant “far away” can be contained in a (relatively) low-dimensional summary. These subsystems are exactly the high-level “objects” or “categories” or “concepts” we recognize in the world. If true, this hypothesis would dramatically simplify the problem of human-aligned AI. It would imply that a wide range of architectures will reliably learn similar high-level concepts from the physical world, that those high-level concepts are exactly the objects/categories/concepts which humans care about (i.e. inputs to human values), and that we can precisely specify those concepts.

The natural abstraction hypothesis is mainly an empirical claim, which needs to be tested in the real world.

My main plan for testing this involves a feedback loop between:

- Calculating abstractions in (reasonably-realistic) simulated systems
- Training cognitive models on those systems
- Empirically identifying patterns in which abstractions are learned by which cognitive models in which environments
- Proving theorems about which abstractions are learned by which cognitive models in which environments.

The holy grail of the project would be an “abstraction thermometer”: an algorithm capable of reliably identifying the abstractions in an environment and representing them in a standard format. In other words, a tool for measuring abstractions. This tool could then be used to measure abstractions in the real world, in order to test the natural abstraction hypothesis.

I plan to spend at least the next six months working on this project. Funding for the project has been supplied by the [Long-Term Future Fund](#).

AMA: Paul Christiano, alignment researcher

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

I'll be running an Ask Me Anything on this post from Friday (April 30) to Saturday (May 1).

If you want to ask something just post a top-level comment; I'll spend at least a day answering questions.

You can find some background about me [here](#).

Specializing in Problems We Don't Understand

Most problems can be separated pretty cleanly into two categories: things we basically understand, and things we basically don't understand. Some things we basically understand: building bridges and skyscrapers, treating and preventing infections, satellites and GPS, cars and ships, oil wells and gas pipelines and power plants, cell networks and databases and websites. Some things we basically don't understand: building fusion power plants, treating and preventing cancer, high-temperature superconductors, programmable contracts, genetic engineering, fluctuations in the value of money, biological and artificial neural networks. Problems we basically understand may have lots of moving parts, require many people with many specialties, but they're generally problems which can be reliably solved by throwing resources at it. There usually isn't much uncertainty about whether the problem will be solved at all, or a high risk of unknown unknowns, or a need for foundational research in order to move forward. Problems we basically don't understand are the opposite: they are research problems, problems which likely require a whole new paradigm.

In agency terms: problems we basically understand are typically solved via [adaptation-execution rather than goal-optimization](#). Problems we basically don't understand are exactly those for which existing adaptations fail.

Main claim underlying this post: it is possible to specialize in problems-we-basically-don't-understand, as a category in its own right, in a way which generalizes across fields. Problems we *do* understand mainly require relatively-specialized knowledge and techniques adapted to solving particular problems. But problems we don't understand mainly require general-purpose skills of empiricism, noticing patterns and bottlenecks, model-building, and design principles. Existing specialized knowledge and techniques don't suffice - after all, if the existing specialized knowledge and techniques were sufficient to reliably solve the problem, then it wouldn't be a problem-we-basically-don't-understand in the first place.

So... how would one go about specializing in problems we basically don't understand? This post will mostly talk about how to choose what to formally study, and how to study it, in order to specialize in problems we don't understand.

Specialize in Things Which Generalize

Suppose existing models and techniques for hot plasmas don't suffice for fusion power. A paradigm shift is likely necessary. So, insofar as we want to learn skills which will give us an advantage (relative to existing hot plasma specialists) in finding the new paradigm, those skills need to come from some other area - they need to *generalize* from their original context to the field of hot plasmas. We want skills which generalize well.

Unfortunately, a lot of topics which are advertised as "very general" don't actually add much value on most problems in practice. A lot of pure math is like this - think abstract algebra or topology. Yes, they *can* be applied all over the place, but in practice the things they say are usually either irrelevant or easily noticed by some other path. (Though of course there are exceptions.) Telling us things we would have figured out anyway doesn't add much value.

There are skills and knowledge which do generalize well. Within technical subjects, think probability and information theory, programming and algorithms, dynamical systems and control theory, optimization and microeconomics, linear algebra and numerical analysis. Systems and synthetic biology generalize well within biology, mechanics and electrodynamics are necessary for fermi estimates in most physical sciences, continuum mechanics and PDEs are useful for a wide variety of areas in engineering and science.

But just listing subjects isn't all that useful - after all, a lot of the most generally-useful skills and techniques don't explicitly appear in a university course catalogue (or if they do, they appear hidden in a pile of more-specialized information). Many aren't explicitly taught at all. What we really need is an outside-view criterion or heuristic, some way to systematically steer toward generalizable knowledge and skills.

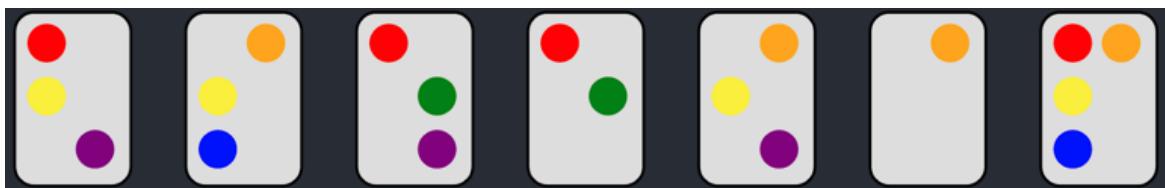
To Build General Problem-Solving Capabilities, Tackle General Problems

It sounds really obvious: if we want to build knowledge and skills which will apply to a wide variety of problems, then we should tackle a wide variety of problems. Then, steer toward knowledge and skills which address bottlenecks relevant to multiple problems.

Early on in a technical education, this will usually involve fairly basic things, like "how do I do a Fermi estimate for this design?" or "what are even the equations needed to model this thing?" or "how do the systems involved generally work?" - questions typically answered in core classes in physics or engineering, and advanced classes in biology or economics. Propagating back from that, it will also involve the math/programming skills needed to both think about and simulate a wide variety of systems.

But even more important than coursework, having a wide variety of problems in mind is directly useful for learning to *actually use* the relevant skills and knowledge. A lot of the value of studying generalizable knowledge/skills comes from being able to apply them in new contexts, very different from any problem one has seen before. One needs to recognize, without prompting, situations-in-the-wild in which a model or technique applies.

A toy example which I encountered in the wild: [Proset](#) is a variant of the game Set. We draw a set of cards with random dots of various colors, and the goal is to find a (nonempty) subset of the cards such that each color appears an even number of times.



The game Proset: find a subset of cards with an even number of each color.

How can we build a big-O-efficient algorithmic solver for this game? Key insight (hover to reveal):

Write down a binary matrix in which each column is a card, each row is a color, and the 0/1 in each entry says whether that color is present on that card. Then, the game is to find the nullspace of that matrix, in arithmetic mod 2. We can solve it via row-reduction.

(Read the spoiler text before continuing, but don't worry if you don't know what the jargon means.)

If we're comfortable with linear algebra, then finding a nullspace via row-reduction is pretty straightforward. (Remember the claim from earlier that the things abstract algebra says are "usually either irrelevant or easily noticed by some other path"? The generalization of row reduction to modular arithmetic is the sort of thing you'd see in an abstract algebra class, rather than a linear algebra class, but if you understand row-reduction then it's not hard to figure out even without studying abstract field theory.) Once we have a reasonable command of linear algebra, the rate-limiting step to figuring out Proset is to *notice* that it's a nullspace problem.

This requires its own kind of practice, quite different from the relatively rote exercises which often show up in formal studies.

Keeping around 10 or 20 interesting problems on which to apply new techniques is a great way to practice this sort of thing. In particular, since the point of all this is to develop skills for problems which we don't understand or know how to solve, it's useful to keep around 10 or 20 problems which you don't understand or know how to solve. For me, it used to be things like nuclear fusion energy, AGI, aging, time travel, solving NP-complete problems, government mechanism design, beating the financial markets, restarting Moore's law, building a real-life flying broomstick, genetically engineering a dragon, or factoring large integers. Most classes I took in college were chosen for likely relevance to at least one of these problems (usually more than one), and whenever I learned some interesting new technique or theorem or model I'd try to apply it to one or more of these problems. When I first studied linear algebra, one of the first problems I applied it to was constructing uncorrelated assets to beat the financial markets, and I also tried for quite some time to apply it to integer factorization (and later various NP-complete problems). Those were the sorts of experiences which built the mental lenses necessary to recognize a modular nullspace problem in Proset.

Use-Cases of Knowledge and Suggested Exercises

If the rate-limiting step is to *notice* that a particular technique applies (e.g. noticing that proset is a nullspace problem), then we don't even necessarily need to be good at using the technique. We just need to be good at noticing problems where the technique applies, and then we can google it if and when we need it. This suggests exercises pretty different from exercises in a lot of classes - for instance, a typical intro linear algebra class involves a lot of practice executing row reduction, but not as much recognizing linear systems in the wild.

More generally: we said earlier that problems-we-basically-understand are usually solved by adaption-execution, i.e. executing a known method which usually works. In that context, the main skill-learning problem is to reliably execute the adaptation; rote practice is a great way to achieve that. But when dealing with problems we basically don't understand, the use-cases for learned knowledge are different, and therefore require different kinds of practice. Some example use-cases for the kinds of things one might formally study:

- Learn a skill or tool which you will later use directly. Ex.: programming classes.
- Learn the gears of a system, so you can later tackle problems involving the system which are unlike any you've seen before. Ex.: physiology classes for doctors.
- Learn how to think about a system at a high level, e.g. enough to do Fermi estimates or identify key bottlenecks relevant to some design problem. Ex.: intro-level fluid mechanics.
- Uncover unknown unknowns, like pitfalls which you wouldn't have thought to check for, tools you wouldn't have known existed, or problems you didn't know were tractable/intractable. Ex.: intro-level statistics, or any course covering NP-completeness.
- Learn jargon, common assumptions, and other concepts needed to effectively interface to some field. Ex.: much of law school.
- Learn enough to distinguish experts from non-experts in a field. Ex.: programming or physiology, for people who don't intend to be programmers/doctors but do need to distinguish good work from quackery in these fields.

These different use-cases suggest different strategies for study, and different degrees of investment. Some require in-depth practice (like skills/tools), others just require a quick first pass (like unknown unknowns), and some can be done with a quick pass if you have the right general background knowledge but require more effort otherwise (like Fermi estimates).

What kind of exercises might we want for some of these use-cases? Some possible patterns for flashcard-style practice:

- Include some open-ended, babble-style questions. For instance, rather than "What is X useful for?", something like "Come up with an application for X which is qualitatively different from anything you've seen before". (I've found that particular exercise very useful - for instance, trying to apply coherence theorems to financial markets led directly to the [subagents](#) post.)
- Include some pull-style questions, i.e. questions in which you have to realize that X is relevant. For instance "Here's a problem in layman's terms; what keywords should you google?" or "Here's a system, what equations govern it?". These are how problems will show up in real life.
- Questions of the form "which of these are not relevant?" or "given <situation>, which of these causes can we rule out?" are probably useful for training gearsy understanding, and reflect how the models are used in the real world.
- Debugging-style questions, i.e. "system X has weird malfunction Y, what's likely going on, and what test should we try next?". This is another one which reflects how gearsy models are used in the real world.
- For unknown unknowns, questions like "Here's a solution to problem X; what's wrong with it?". (Also relevant for distinguishing experts from non-experts.)
- For jargon and the like, maybe copy some sentences or abstracts from actual papers, and then translate them into layman's terms or otherwise say what they mean.
- Similarly, a useful exercise is to read an abstract and then explain why it's significant/interesting (assuming that it is, in fact, significant/interesting). This would mean connecting it to the broader problems or applications to which the research is relevant.
- For recognizing experts, I'd recommend exercises like "Suppose you want to find someone who can help with problem X, what do you google for?".

Cautionary note: I have never made heavy use of exercises-intended-as-exercises (including flashcards), other than course assignments. I brainstormed these exercises to mimic the kinds of things I naturally ended up doing in the process of pursuing the various hard problems we talked about earlier. (This was part of a conversation with [AllAmericanBreakfast](#), where we talked about exercises specifically.) I find it likely that explicitly using these sorts of exercises would build similar skills, faster.

Summary

Problems-we-basically-understand are usually solved by executing specialized strategies which are already known to usually work. Problems-we-basically-don't-understand are exactly those for which such strategies fail. Because the specialized techniques fail, we have to fall back on more general-purpose methods and models. To specialize in problems-we-basically-don't-understand, specialize in skills and knowledge which generalize well.

To learn the sort of skills and knowledge which are likely to generalize well to new, poorly-understood problems, it's useful to have a fairly-wide variety of problems which you basically don't understand or know how to solve. Then, prioritize techniques and models which seem relevant to multiple such problems. The problems also provide natural applications in which to test new techniques, and in particular to test the crucial skill of recognizing (without prompting) situations-in-the-wild in which the technique applies.

This sort of practice differs from the exercises often seen in classes, which tend to focus more on reliable execution of fixed strategies. Such exercises make sense for problems-we-basically-understand, since reliable execution of a known strategy is the main way we solve such problems. But learned skills and knowledge have different use-cases for problems-we-basically-don't-understand, and these use-cases suggest different kinds of exercises. For instance, take a theorem, and try to find a system to apply it to which is qualitatively

different from anything you've seen before. Or, try to translate the abstract of a paper into layman's terms.

I've ended up doing things like this in the pursuit of a variety of problems in the wild, and I find it likely that explicit exercises could build similar skills faster.

I'm from a parallel Earth with much higher coordination: AMA

Related: [My April Fools Day Confession](#); [Inadequate Equilibria](#)

On April 1, Eliezer Yudkowsky ran a dath ilan AMA on Facebook:

I came from a parallel Earth that successfully coordinated around maintaining a higher level of ability to solve coordination problems. Ask me anything.

With Eliezer's blessing, I've quoted the resultant discussion below, leaving out threads that were repeats or didn't go anywhere.

Guy Srinivasan: Did parallel Earth coordinate around a specific day each year for everyone to play with falsity?

Eliezer Yudkowsky: Not a specific *day* as such. There's very much a tradition of leading somebody down a garden path, and also of pretending to be led down the garden path — similar to the "MIRI pomodoro: 25 minutes of work followed by 5 minutes of trolling" — but there's a verbal handshake you're supposed to give at the end to prevent that from going out of control and any tragic errors.

Emielle Potgieter: What is parallel earth's biggest problem, then?

[...]

Eliezer Yudkowsky: I'd assume that Artificial General Intelligence is being seen by the Senior Very Serious People as a *big* problem, given the degree to which *nobody ever talked about it*, how relatively slow computing progress was compared to here, and how my general education *just happened* to prepare me to make a ton of correct inferences about it as soon as anybody mentioned the possibility to me. They claim to you it's about hypothetical aliens and economic dysfunction scenarios, but boy howdy do you get a lot of [Orthogonality](#) and [Goodhart's Curse](#) in the water supply.

Stācia Gāel: Why did you come here?

Jean-Baptiste Clemens: @Stācia Gāel Everyone on parallel Earth was attempting to meet for lunch in the absence of communication and Eliezer was wrong about the Schelling point.

[...]

Eliezer Yudkowsky: No clue, then or ever.

Erica Edelman: How do you transition babies into job holding adults?

Eliezer Yudkowsky: That covers a lot of territory! But if I tried to zoom in at the most general level of difference, there's very much an understanding among Very Serious People that if you require anything to get access to a job (like a credential) that is difficult to get or has finite supply, people can burn the whole surplus value of the job to them in order to get it, even if that thing is of lower value to the employer. In other words, they would recognize "occupational licensing" or college as a sickness and move to prevent it while it

was still getting started. In general, demanding something from somebody other than their actual job skill is recognized as a potential civilizational problem. So this means:

- A focus on testing for the actual job skills, rather than for peripherally related things like having paid to attend a particular institution. Not even talking about tests vs. attendance, I mean that they watch you doing the actual job.
- Older children teach younger children things that the older children have been watched doing, because the older children have the knowledge and can teach it, and you wouldn't want to demand any more qualification than that. With monitoring and external validation to prevent ignorance from iterating upon itself, of course.
- There's no minimum age to work, because demanding a higher age isn't something that the person doing the job actually needs.
- You don't need a business license because that would, again, be an instance of something they recognize as technical debt / overhead / cruft for the civilization.
- The economy runs hot enough that there's generally enough jobs on offer; the Very Serious People would regard it as a huge issue if people had to look for a job instead of choosing *which* job, and they would ask how we could possibly have gotten into that position when jobs were available 1000 years earlier and the economy had gotten a lot more productive since then.
- The degree to which the employee sees themselves as doing the employer a favor, vs the employer seeing themselves as doing the employee a favor, is much more symmetrical than it is here - if that wasn't true, the Very Serious People would look at it and ask "What's going wrong with this supply-demand balancing price level, why isn't this bargain behaving more symmetrically?"
- I can speculate about what other conditions contribute to that, but unfortunately it wasn't my actual field of study before I left. But in general, think of the situation among Silicon Valley programmers: you display the ability to do the work rather than competing on credentials, and it's as common to find a precious employee through connections as to find a precious job through connections. This would be equally true about haircuts in dath ilan; somebody who wants to expand their hair salon needs to somehow find an employee to do that and will have trouble doing so, unless they're much more profitable than average and can offer above-average salaries for that reason.
- They have an actual concept of "matching people up with jobs is a huge allocation issue for the whole economy" and have already put serious sweat into experimenting with different ways to discern kids' native talents and figure out what that would match up to, obviously including have the kid actually try doing lots of different things. The notion that you get a whole college degree before spending one day trying out the actual job would be insane.

So the overall answer to your question is that a kid learns skills from older kids, or from more specialized teachers or apprenticeships if it's a sufficiently complex skill to require that, after all kinds of "try doing this for a week" microapprenticeships so they can figure out where their niche is, and then when they know enough to do the job they can start doing it. If that sounds utopian, it's because the Very Serious People recognized it as a central issue for the whole economy and put literally any thought and experimentation at all into figuring out what would optimize it.

Ben Pace: I'm surprised this is how it works on your planet. I am probably missing a basic piece of knowledge here, but in my experience it is very hard to put people to useful work, and if you give me a set of 100 people and ask me to cause them to do useful work, I will be able to do so with a team of 5-10 of them and then either I can help the rest form teams (if we're lucky that takes up most people), and if I'm forced to give the rest jobs many people will be left with bullshit jobs.

(I mean, there are more interesting mechanisms I would use, like producing financial rewards for tasks completed that incentivize the rest to self-coordinate etc, but I don't feel like I can *count* on this to give everyone a job.)

(Also like 15% of people have a low enough IQ that they cannot be given anything useful to do without on the order of 1-1 oversight. Not a defeater, but another major hurdle.)

If you stop all the people doing bullshit jobs on your planet, do the rest of the people really find productive ways to build products and provide services, or is there something else that a lot of the population spends their time doing?

[...]

Tom Pandolfo:

> There's no minimum age to work, because demanding a higher age isn't something that the person doing the job actually needs.

@Eliezer Yudkowsky How does your planet avoid the problem of child labor - that is, a 10-year-old being required to work in a factory in order for them or their family to afford food?

(As distinct from a 10-year-old shadowing/apprenticing someone whose job they passionately want to do when they grow up.)

Or is that not seen as a problem?

Eliezer Yudkowsky: The fact that only 10 people can do the work *you* have for them doesn't mean that the other 90 people can't do work that *other* people have for them? I mean, that's kind of how the whole economy works?

Civilization has enough productive capacity that when somebody is born who just can't make it in Civilization for whatever reason, there are places where you can go to live out the rest of your life in peace, provided that you have not previously had any children.

One of the things I'd expect people from this world to find relatively off-putting is that dath ilan has comprehended that happiness is heritable and they teach a sight that extends over generational times and thinks ahead to the equilibrium; so it's understood that, except in very exceptional circumstances, if you're unhappy on average *for any reason*, it is your duty to the next generation not to have kids who might inherit this tendency.

So the number of people who go to the Quiet Cities is more like 5% than 15%, because those who would have otherwise been the parents of people who went there did not have kids. And the rest of the world is mostly happy, because transmitting constitutions that can be happy in a civilized world is itself a solvable coordination problem.

Ben Pace: (Sad react, but glad the people are able to look at sad things and take appropriate action and deal with it, and I respect those ~10% of people who do this a great deal.)

[...]

David Schneider-Joseph: @Eliezer Yudkowsky What about those who are unhappy because they see a problem with the civilization which will take many generations to solve, and are motivated to start the work of fixing it? Does that count as one of those very exceptional circumstances?

Jim Babcock:

"@Eliezer Yudkowsky How does your planet avoid the problem of child labor - that is, a 10-year-old being required to work in a factory in order for them or their family to afford food?"

UBI, obviously? On Earth, we have a political narrative that people must be threatened with poverty, or else they will choose not to work. And yet, if you look at the most impoverished people, you mostly find people whose ability-to-work has been damaged by

the consequences of their poverty. It seems wildly overdetermined that, in a sensible system, there would be a UBI, and it would be high enough that not-having-food can't happen without something else going wrong other than lack of money.

Bruce Barrett Banner: @Jim Babcock it seems plausible to me that some suitable notion of economic democracy would justify, or even require, a UBI, and hence protect, not just children, but humanity at large from the degradation of wage slavery, which produces ruined workers as you suggest, and much other misery at well. It also makes a reality of political democracy as a side effect, to its credit.

Jim Babcock: If "the degradation of wage slavery" sounds like a good summarization of the point to you, I think you may have a pretty important confusion? (A very *common* confusion, which is why I'm choosing to highlight it.)

The "ruined worker" is typically *not* a wage slave, but someone who, at a key point, had no wage, and suffered malnutrition, medication lapses and assorted traumas as a result. It's pretty common for people to get confused about this, and try to put additional demands on employers, which wind up decreasing the availability of low-end jobs and increasing the number of people starved.

One of the most important things I learned, being very into nutrition-research, is that most people can't recognize malnutrition when they see it, and so there's a widespread narrative that it doesn't exist. But if you actually know what you're looking for, and you walk down an urban downtown and look at the beggars, you will see the damage it has wrought... and it is extensive.

Tom Pandolfo: @Jim Babcock I'm not so sure that it's obvious. I'm skeptical of a purely monetary UBI, as I haven't been convinced that it won't just drive up the price of Literally Everything in the long run (without some serious regulatory intervention). On our Earth, at least, I'm much more in favor of a robust system of public housing/food/healthcare for free to everyone who needs it, and tepidly in favor of UBI as a temporary/transitional solution.

That said, I'm interested in what solution Eliezer thinks is obvious to the denizens of dath ilan.

Jim Babcock:

"I'm skeptical of a purely monetary UBI, as I haven't been convinced that it won't just drive up the price of Literally Everything in the long run (without some serious regulatory intervention)."

This is something economists understand very well, actually! Not at the intro-course level, and not if you get your economists through a politics-optimized filter, but if you put an economics department directory on a dartboard and throw three darts, I expect you'll get the same (or compatible) answers from all three.

UBI is spending, which causes inflation, but not if it's replacing other government spending or offset by comparable taxes on any bracket. The price of supply-constrained goods, like housing in a strictly zoned area, or medical care from guild-approved doctors, may rise; the price of food, manufactured goods, and housing in low-density unzoned areas, will not.

Eliezer Yudkowsky:

"What about those who are unhappy because they see a problem with the civilization which will take many generations to solve, and are motivated to start the work of fixing it? Does that count as one of those very exceptional circumstances?"

It does not. There are people who can manage to be happy and cheerful in their daily lives while tackling big problems. If you were unhappy and you wanted more people tackling big problems, you'd offer to pay for their childcare, not have kids yourself.

Nora Ammann: Is there an overarching social narrative (at some level) they all share and that plays a significant role in their ability to coordinate? If so, what are its characteristics?

Eliezer Yudkowsky: Several, of course, but this feels like a large enough question that I'm having trouble answering. Any individual cultivates a sense of <untranslateable 23>, their personal sense of This Is Who I Am, This Is What I Do, which will generally include I Can Correctly Analyze Coordination Problems And Do My Part In Solving Them. It's understood that this is meant to be personal and individualized; and at the same time, society is utterly unable to coordinate against all the people who metaphorically talk about Civilization's <untranslateable 23> and argue about what it is or should be. If you take a step back, this reflects a narrative about It Is Your Pride And Your Responsibility To Choose Your Own Narrative And Then Be Fucking Awesome At It. And of course that idea is going to spill over to Civilization too.

People are aware of Goodhart's Law, but that doesn't stop the Very Serious People from Very Seriously And Prudently talking about particular figures-of-merit for all Civilization that we are still going to talk about while keeping an eye out for anybody trying to game them; figures-of-merit from productivity to scientific progress to health and happiness; and when you see a graph with those lines supposedly going up, it becomes part of a narrative about how Civilization is not doing too terribly at living up to its own <untranslateable 23>.

Nora Ammann: Is their ability to coordinate relying on growth? Would [they] still be able to coordinate if [their] economy stagnated? [...]

Eliezer Yudkowsky: It seems to me that the structure of what they're doing would carry over to a case without growth - you can ask about steady-state coordinated equilibria. There might be a bunch of specific things that would break and need changing, but people would be *thinking* about what those were.

Rocco Stanzione: Why did you choose the first day of April to reveal yourself?

Eliezer Yudkowsky: Plausible deniability!

Jeff Dubin: Economic system?

David Spearman: This. Also, legal system?

Eliezer Yudkowsky: You're both gonna need to be more specific.

David Spearman: Suppose that A is in a conflict with B over some perceived violation of whatever passes for rights. They both recognize the coordination problem for what it is, but they can't agree on which of them is actually the highest-value user, in the Coasian sense, of whatever the object of the dispute is. Do they just whistle up a Very Serious Person to resolve the dispute? What sort of training does the Very Serious Person have for that sort of value assessment? Or is there some kind of auction system that's somehow less negative-sum than the current all-pay auction/lawyer system? What do they do about interpersonal disputes (battery or murder, say)? If a metaphorical hawk shows up in the grand game of Hawk and Dove that is society, what happens?

Jeff Dubin: Sure!

How is it decided who gets what stuff? Is there a "from each according to their ability, to each according to their need" framework? Is there some sort of competitive market system? What can and cannot be owned as property by an individual or family? How are the needs of children, elderly and disabled persons, etc met?

David Spearman: I'd also like to hear what sort of intellectual property system they have.

Also, what is the difference between how the system handles malfeasance from a Very Serious Person versus from a rando?

Jeff Dubin: @David Spearman Or, have the VSPs somehow been trained not to Malfeasance?

David Spearman: @Jeff Dubin I would be disappointed if this was the answer, unless there's a very detailed explanation of what that training entails.

[...]

Eliezer Yudkowsky: I want to say that there's the basically obvious Georgist system of private property with public capture of economic rents, but in this Earth I hardly know what's "obvious" anymore. I'll put it the way it would be put in dath ilan second grade.

There's a planet everybody lives on, of which conceptually everybody ought to start out with an equal interest in the raw resources thereof; the raw resources in untransformed states are not directly very valuable, but labor requires access to these resources in order to be valuable and value-add to them.

If your economic system is that Bill Gates owns all land and permanent installations and doles out scraps of food in exchange for labor, then even if this could look from a libertarian standpoint like a "private property" system in which Bill Gates happens to own all the property, this is not equitable and everyone who isn't Bill Gates ought to shrug off the consensual hallucination claiming that all the matter in the universe is tagged with a tiny private property tag saying that Bill Gates owns it.

At the opposite extreme, to the extent that the result of labor adding value to resources is the expropriation of the transformed resource, labor has less incentive to add value and is being expropriated. Or if you look at it from the perspective of the labor, versus the person who gets to keep what the labor produces, they'd have an interest in spending some of their labor to band together to fend off the bandit who takes away most of the transformed resource, which is why, once labor has transformed a resource and added value to it, there's a reason to let the laborer keep or trade that transformed resource.

And then we introduce the second-grade notion of an "economic rent" which is that some things, like land, are valuable in a way where we don't get any *more* of them, depending on who's said to own them - unlike the way we get more labor, if the laborers can keep what they make; and the second-grade notion of trade and allocation, which is that different resources are worth different amounts to different people, and if you let people bid differently on those resources, that introduces a factor where people who can use the resource to produce more will bid more highly on it, modulo another acknowledged factor where people who have more resources available can also bid higher.

This all combines into a notion of taking the planet's prior endowment of arable land, minable ores, livable land, etcetera; and establishing an ongoing auction-lease system for renting them, where people can pay excess amounts to establish propertylike momentum around places where they want to build permanent installations; but there's still a system that demands ongoing rents, in order to encourage reallocation, and in order to give everybody in the world a share of that rent in order to represent their interest in respecting the "property" coordination-hallucination at all, in the way they wouldn't have any such interest in coordinating to respect a pretense of private property where everything was tagged as being owned by Bill Gates.

There's an explicit understanding of concepts like "value added to land in virtue of its proximity to other land where people live", so rather than distributing the economic rent equally to everybody, like you would for a minable ore, it makes sense to direct some of that rent to a citylike entity that can reinvest the increased local rent it created in making that city larger and more valuable by producing goods with inherent coordination-nature

like roads (where the value of any one road segment is dependent on the value of other segments next to it, and where multiple entities with veto ability could each try to capture all of the value-added from the whole road, giving the road a public-good-nature).

Basically, the kind of private property system you would invent if you understood that you were *inventing* it rather than codifying some pre-existing natural law.

Eliezer Yudkowsky: Intellectual property in eg the form of "patents" the way Earth does it, is a needless special case of property with the veto-nature where you want to avoid ending up with multiple entities that can each veto the good being produced, which gives each of them an incentive to try to capture all of the value and produces a coordination problem in producing the good. For this reason, nobody on dath ilan would consider introducing an Earth-style patent system, even leaving aside the extent to which the US patent system malfunctions on its own terms and patents lots of things that are obvious to any practitioner skilled in the art.

If you create an invention that benefits an industry, it's understood by a solider and sharper-toothed honor system that the industry is expected to donate back 15% of the marginal value thereby produced to the original inventor or invention group, and there are external-auditor-like agencies to sign off on accounting (accounting is generally much simpler in dath ilan because they do not have a regulatory process that has gone completely out of control); the closest thing in Earth terms would be a mandatory shall-license patent system.

If you did the equivalent of what BioNTech did and invented a Covid-19 vaccine, and Moderna had no mRNA vaccine of their own (meaning this was something that *only* BioNTech could do), anybody would be allowed to produce it, but *EVERYBODY* would be honor-bound to donate rather a *LOT* of money to BioNTech and to the inventors and discoverers of RNA vaccines; who donated how much money would be a public fact and your friends might look at you funny if you donated nothing; with some of that money tending to be in the form of donations to let the inventors of mRNA vaccines set up their own scientific funding agencies, to make the way easier for the next generation of scientists. All of this would be the sort of thing that Very Serious People had heated debates about in newspapers; and all of the solutions inside their Overton Window would be about equally good in the larger scheme of things compared to Earth solutions, regardless of which exact details ended up being picked.

Copyright doesn't face the same collision issue where multiple parties can extract all value, and by one of those weird little coincidences, dath ilan has the same 14-year copyright rule that was originally baked into the US Constitution (albeit with no renewals allowed, because it happens to not be an exact coincidental collision). 14 years is enough for people to make a profit on copyrighted creations, thereby incentivizing its existence; and then that work goes into the same pool of public domain that implicitly helped create that copyrighted work and whatever shoulders it stood upon.

Eliezer Yudkowsky: Very Serious People are not rulers invested with formal authority; their work, by its nature, consists of public debates with other Very Serious People. Any "malfeasance" in the obvious sense would have to consist of saying the wrong thing in a public debate, but then if other VSPs didn't catch it and call them on it, couldn't they just claim it was an honest mistake?

I guess you could have an instance where a Very Serious Person pretended to have important anecdotal evidence from their own life history about something, and then was caught out on having made it up. I can't actually remember hearing about a case like that, per se; the big deal tends to be groups of Very Serious People making Very Serious dire predictions and then turning out to be totally wrong, which is considered especially iffy because of how making a dire prediction about the conditional result of doing X can prevent anybody from actually testing X and thereby finding out the dire results. Albeit that dath ilan is *much* more likely to set aside a little hamlet where people try doing X anyways, which is how the Very Serious People get caught being too pessimistic.

But it's still understood that when it comes to, say, messing around with breeding viruses, it is *reasonable* to say that this is a direly dangerous thing even to experiment with. In fact it is emphasized *oddly* strongly how *totally reasonable* this would be, if Civilization ever *did* run into something that could plausibly wipe it out in one shot; and that the fact that somebody *could* have an incentive to gain attention by warning direly against ever trying X, then escaping falsification through X not being tried, must *not* mean that we reject arguments of this type out of hand; because then Civilization is inevitably and undignifiedly doomed if it ever runs into anything that actually can wipe it out in one shot.

The very strong emphasis here makes more sense to me now that I realize that AGI issues were probably secretly in the background informing some of the top people.

Eliezer Yudkowsky: To a first approximation, everybody in dath ilan is an economist, in the same way that everybody on Earth is a scribe and a calculator from the perspective of Earth's medieval era. So when it comes to things like setting up courts, people understand that you get what you pay for and you pay for what you measure and that measuring things is dangerous.

There's much more of an emphasis on courts producing judgments where they write out all of the reasoning used, in a way that superior courts and ultimately cities and delegates can check over, and less of an emphasis on "His Honor said so, so shut up and respect him, peon". But you're paying courts for a certain kind of reliable reproducible reasoning that's supposed to reflect a particular set of agreed-on standards and agreed-on rules, not just for having a very scary honorable person in robes hand down a dictum that everybody has to agree with; the emphasis on writing out all the reasoning is to try to make it easier to measure what you're supposed to be paying for. If it was a big enough issue, you'd pay for two courts so you could check if they agreed.

[...]

Ben Pace: This [method for incentivizing intellectual innovation] is a bit harder in art. Like, I can indeed track down the authors that I think influenced me the most (my puns from Scott, my titles from Hanson, my concepts from you, etc) and pay them money, but it's harder to enforce than if I am using the vaccine built by BioNTech. It's harder for others to see that happened.

That said, I can also imagine publicly "taking" people's ideas is just good and encouraged, such that it's more natural. Right now I don't say that I stole my writing style from people too much, in part because it can be seen as bad form to copy people's intellectual/artistic work, but if it were more encouraged then accounting would be easier too.

Matthew Graves: I assume this is also tied to a crowdfunding-like invention system, instead of a monopoly-profits-driven invention system? Or does the honor extend all the way up the stack?

Motivating example: suppose there are 10k people with a disease, who in aggregate would be willing to pay \$10k each to not have the disease, so there's \$100M of value 'on the table' for curing it. Alice develops a treatment, hoping to sell it to each of those people and get \$100M in revenue, and then Bob produces it at marginal cost (say, \$100 each), makes \$1M in revenue, and has no profits to donate to Alice.

One could imagine all of the customers deciding that, well, they need to donate \$1.5k to Alice to pay her back for the invention. But this is substantially less than Alice would have gotten in monopoly revenues, and (more importantly in my eyes) requires all 10k customers to track the question of whether or not their treatment supplier has honorably discharged their duty to reward Alice. But also how much that duty is depends on how much those patients actually value not having the disease.

(And if I employ one of those patients, do I have a duty to reward Alice for the increased productivity? Or to track whether or not my employees have rewarded Alice enough?)

Eliezer Yudkowsky: Ben: I wouldn't need to pay anybody for having written [HPMOR](#) - that kind of inspiration is just considered part of the common pool where everybody is inspired by everybody else. Paying an author for their quoted work is payment enough.

Matthew: You're not supposed to capture all of the value you create. If the treatment is worth \$10k each to the treated, \$1.5K is a very reasonable amount for them to donate to Alice.

Jean-Baptiste Clemens: Did successful coordination require an authoritarian government or dictatorship and omnipresent surveillance to ensure compliance?

Because I don't see how else this could work with billions of people, the vast majority of them being strangers to each other, when even a small group of close friends can have trouble reaching a consensus on something as simple as where to have lunch.

Eliezer Yudkowsky: Leviathan is what people here do *instead of* having everybody in the room know what a coordination problem is, work out a coordination solution (including coordinated enforcement of the solution if there's large defection incentives), pool their solutions to all work together, and then all do the thing. This requires *previously* having coordinated your Civilization well enough to make sure that everybody in the room knows the abstract theory and has practiced it across many trials, but the equilibrium is stable once you get there, especially if everybody knows what the equilibrium *is* and how important it is to keep it stable.

Jean-Baptiste Clemens: @Eliezer Yudkowsky Right, that's a good point. The social contract is this Earth's sub-optimal arrangement, but I'm interested to know how everybody on parallel Earth initially came to understand coordination problems and reach a consensus on how to solve these problems in the first place, despite the statistical inevitability of so many people having competing and incompatible ideas, priorities and preferences. Does everybody on parallel Earth have a shared mind/consciousness?

I would like to rephrase my original question but I feel it would be unfair to edit it after Eliezer has already responded, so I will rephrase my question here:

Did the *initial* coordination necessary to reach a stable equilibrium of successfully solving coordination problems require an authoritarian government or dictatorship and omnipresent surveillance to ensure compliance during the initial coordination problem solving period?

Eliezer Yudkowsky: I don't think we can get better coordination the way I suspect that dath ilan wandered into it, certainly not before AGI. Gregory Cochran (IIRC) has a theory about how the secret sauce of the Industrial Revolution was the children of shopkeeper classes starting with a larger inheritance and outreproducing others. As I previously mentioned a couple of comments up, a lot of dath ilan's earlier history is considered a Highly Unpleasant Thing It Is Sometimes Necessary To Know and a mild cognitohazard, but my suspicion is that a lot of the real work was done by a historical accident of this sort.

Luca Ross: How did they solve competing access needs?

Eliezer Yudkowsky: Be more specific?

Luca Ross: @Eliezer Yudkowsky I think of competing access needs as a sort of coordination problem, probably. In that it seems like coordination is the only feasible solution prior to the Glorious Transhumanist Future. Stuff like people with severe dog allergies or phobias and people with service dogs needing access to the same space at the same time. Or people who are triggered to self harm by seeing images with self harm scars in them, and people who need a place where they can post normal pictures of themselves without their scars being brought up needing access to the same support groups. People

who have audible stims and people who are noise sensitive but need to be able to hear other things in the same environment (like class).

Anything where people's needs are mutually exclusive but they need access to the same thing.

Eliezer Yudkowsky: Diversity of places! On Earth you have dozens of countries with nearly cookie-cutter regulatory systems and equilibria, containing thousands of nearly identical cities. On Earth, the whole reason for having *different places* is either to mine natural resources, to run an experiment, or to pick a different public equilibrium of this kind.

Daniel Powell: @Eliezer Yudkowsky oh, so you don't have competing access needs because everyone with aphonia and experience distress from sounds lives in one dimension where they don't have to interact with people who need a clicker to stim!

That's a lot of parallel NYC subway systems. How long does it take to phase to the right one?

David Spearman: Besides the whole "kink and macroecon" thing, are there any other ways they were consistently worse than Earth? In particular, how is the economy not a constant, negative-sum war between price fixing cartels, consumer cartels, etc. What stops otherwise-competitive producers from coordinating around "charge the monopolistic price and don't overproduce"?

Eliezer Yudkowsky: That's an excellent question! The first thing I'd say in reply is to point out that some *individual* at the head of one of these concerns, does not themselves capture the whole value of the cartel - maybe the cartel makes a billion dollars, but that person doesn't get to take home the whole value themselves. But they do get to take home a terrible reputation for having knowingly acted against the interests of Civilization at scale.

On Earth, you have puddled reputational pools where somebody gets to go home with their fellow cigar-smoking villains and have a high reputation among them as a great successful villain. If something like that started to develop on Earth, it would be a Huge Problem and all the Very Serious People would see it as a Huge Problem and it would be on *all* the news programs once uncovered, and if they had to pay a billion dollars to set up a new competitor to drive that concern out of business among all the good people who would immediately boycott it once uncovered, the Kickstarter would be funded the next morning.

Where the Huge Problem, to be clear, would not so much be the price-fixing aspect, as the fact that a separate reputational puddle had developed in which powerful people could hoard a reputation as villains; who knows, maybe next up they're going to develop a bioweapon and wipe out half the planet so they can brag about *that* to their friends.

So long as this separate reputational puddle hasn't been allowed to develop, then a would-be price-fixing cartel would have to comprise a lot of people who didn't get to take home all the money captured themselves, whose spouses and kids were all raised to believe that your pride and honor rest in coordinating good equilibria for Civilization, not for coordinating with a few people to defect in a way that benefits you a little and damages Civilization a lot. And if that starts to go wrong, there can be actual boycotts, Kickstarters to develop a new competitor, and lots of people who will remember your sins and still boycott you five years later. It's not a perfect equilibrium, but it's not that unstable once you're in it. The key difference is that everybody is *seeing* the global equilibrium and thinking five moves ahead and that's the way all the newspaper stories are written.

Simon Sandoval-Moshenberg: ^ lots of Marx and Engels here but this comment is the Marx-iest of all.

David Spearman: @Simon Sandoval-Moshenberg not really. His comment equally applies to a consumer consortium which unionized to gain unwarranted monopsony power. Though I will note that it assumes a no-cartel equilibrium where the VSP's can just play whack-a-mole as cartels arise. I'd expect a many-cartel equilibrium, at least at first, with no obvious path from there to the no-cartel state where the moles can be whacked case-by-case.

Ben Pace: This aspect of the civilization sounds like there's a lot of cultural homogeneity within the civilization. Ensuring that everyone has a shared set of values regardless of decisions about upbringing and morality and way of life reduces cultural variance quite a bit. It's harder to have the Ravenclaws who as a point of their culture sit in an ivory tower reading/writing books and never leave to act on the world (though will answer questions occasionally). It's harder to have the Quakers who live technologically backward because they think the tech damages their community. It's harder to have a group that leaves broader civilization for 30 years because they think it's failing and they need to think clearly on their own and bring up children elsewhere.

But I guess it doesn't actually sound that hard, nor costly enough to make it not worth it. Just have a set of agreements with these people when they set up their cultures, about how to interface with the outside world, and check in occasionally to ensure agreements are being kept. Sounds deeply worth it.

Eliezer Yudkowsky: Global homogeneity, local variations. 99% of the planet may be thinking the same way about something because that's what the textbooks say, but it's also *much* easier than on Earth to set up a town somewhere that does one *particular* thing very differently. It's understood that being Able To Run Experiments is an important feature of Civilization.

Ben Pace: More global coordination and more local experimentation.

David Spearman: What is their fiction like? What would the most-cited pages of their equivalent of TVTropes look like?

(Alternative framing: if they had an equivalent to Warhammer 40k, what would the different factions look like?)

Eliezer Yudkowsky: All my brilliantly original work over here is me just writing completely stereotypical dath ilan cliches and cackling to myself. [HPMOR](#) would be their equivalent of a Harlequin romance or a Kindle Unlimited dungeon-core cultivation monster-girl harem novel.

Alex Zavoluk: What's the best food on your home planet?

Eliezer Yudkowsky: Swedish meatballs!

Adam Priest: Do people have jobs?

Are [workplaces] unionised?

What prevents the bosses exploiting the workers?

[...]

Eliezer Yudkowsky: Yes, no, and the economy runs hot enough that a hair salon has to work hard to find labor if they want to be able to expand. As for why, if you asked dath ilan *why* they always have enough job openings to compete hard for labor, they'd give you a puzzled look and point out that since people could afford to pay for labor 1000 years ago

when technological productivity was much lower, obviously today there should be *lots* of people who can pay at least survival food and shelter for some bit of labor they want done; and indeed, since there are many more payable jobs like this than laborers, employers have to pay much *more* than this for labor they want done; so employees can choose which job they want, and can tell abusive bosses to either fuck off or pay them a huge premium for putting up with it.

If you showed them Earth's economy where people struggle to find jobs and employers treat employment as a huge favor they dole out that can come packaged with abuse, they would stare with huge eyes and then try to figure out what the hell had gone so wrong and how it was even *possible* to build a system like ours in a world with 100x medieval agricultural productivity.

They would start to understand once you listed out all of the different obstacles to an employer employing somebody. But my guess is that if you come at this from the frame of "what prevents the bosses exploiting the workers?", it might be a long conversation before I could describe how a dath ilan eye parses up the obstacles to employment here, that prevent there from being much more competition for labor here.

Marcello Herreshoff: How are the parallel Earth's societal coordination mechanisms protected from corruption and conflicts of interest? (On Earth classic, the intended vehicles for solving coordination problems that markets and independent actors cannot are our various governments, but we all know how that's turning out.)

Eliezer Yudkowsky: I'm not quite sure how to answer this - everybody in dath ilan is an economist the same way everybody in the USA is a scribe from the perspective of medieval times, so everybody knows that enforcement is part of the problem of maintaining an equilibrium, that this enforcement is often a nonrival and/or non-excludable good you have to coordinate to pay for, that you can only enforce what you can see and measure, etc?

In terms of things I can say about that in general... maybe one thing is that dath ilan understands that the first remedy to potential corruption or conflict of interest is *visibility* and *accountability*, rather than trying to make sure that no one person has power via committee, or writing big books of regulations to supposedly govern the committee decision. That is, the dath ilan approach is to appoint one person who has to hear things in public, write down their reasoning, publish their reasoning and decision, and get their results checked three years later.

Bill Doherty: How does the dating market differ on your Earth?

Eliezer Yudkowsky: It's much more the domain of paid professionals, something like a real estate broker where you tell them all about who you are and who you're looking for, they get together with other real estate brokers to look for matches, and you'd pay them based on results if you were still happy 6 months or 10 years later.

Daniel Sturtevant: How did y'all compensate for "my stake in this coordination problem is more important than yours," bias in everyday, not-super-deliberative cognition? Or did everyone get eaten by tigers?

Eliezer Yudkowsky: I'm not quite sure which problem class you're envisioning. People here on Earth routinely run into situations, any time there's a coordination problem, where somebody could conceivably claim their stakes mattered more? And people here on this Earth have a reputation concept and an implicit social capital system, where somebody who claims exceptions too often, and who doesn't produce compensating value, will start to lose friends, so nobody wants to hang out with them after the third time that they claimed they couldn't afford to pay for their share of the pizza?

People in dath ilan don't know they're supposed to be in a Utopian illustration of people talking about coordination problems all the time, or that they're supposed to be solving them perfectly. The main difference, if there even is one, is that if somebody repeatedly doesn't pay for the pizza, everybody has a shared verbal abstract concept of what's going on, and they can say, "Look, if you don't believe us about what we think you're doing, namely defecting in a <dath ilan equivalent of the Prisoner's Dilemma parable>, we can ask a disinterested third party" or "You're spending social capital like water, pal" instead of quietly and resentfully dumping the defector.

Ben Pace: Broadly, what do leaders look like in your world? Are there people who are respected and are able to take action for the nations of your world without being punished for it and slapped down? Somewhat relatedly, do they interact with social media more like Elon Musk or more like Paul Graham?

Eliezer Yudkowsky: Computing media less advanced, social media doesn't exist yet and probably wouldn't be allowed to exist. If they did exist, the Very Serious People would look like Paul Graham but with something closer to Elon Musk's sense of humor, albeit with the humor more restrained and carefully set apart in most cases.

The Earth concept of "nations" is very much about barriers to immigration and barriers to trade, both of which would be considered harmful in dath ilan. There are huge factions, there are huge special interest groups, there are regions with their local public goods that you are required to pay for if you live there, but people belong to more than one of those; they don't belong to A Country. Somebody has to run those organizations but it is not assumed that they can Speak For Their People the same way *unless* somebody has actually set up a voting/delegation structure for that organization.

It's understood that it can be dangerous to leave out the details and that governance is dangerous. Imagine if the newspapers consistently described Biden as "the FPTP primary-general Electoral College indirect delegate with a 53% overall approval rating" or some such instead of "The President", except that the phrase was much shorter because it contained abbreviation symbols that everyone already recognized, like using INTP to describe somebody's personality. Then as much ability to speak for the citizens as you want to recognize somebody like that as having, that's how much Biden has.

Open combat where you actually destroy people and property is a huge undignified failure in dath ilan, like, that cannot *possibly* be on the Pareto equilibrium, what are you doing wrong. If somebody hasn't gone literally schizophrenic and needed to be restrained, it's seen as shameful for everybody that it got that far. If you violate regional rules you're told to leave the region; if you refuse to pay a fine you contracted for, the contract says that your bank can take it from your account.

Since you can move about from one region to another rather than there being this huge Immigration deal, you're not supposed to be in Your Country that's the only country you can live and that has to punish you to make you obey its rules; what it tends to do is kick you out, and depending on reason, fewer other countries may accept you for a time. Less prisons, more Australia (actually the continent we'd call Japan, but a totally different place).

So the Grand Authority that concentrates in countries, counties, cities, and police forces as arbiters of violence, is more distributed across factions and regional experiments and people heading up particular organizations. There is less authority that stems from being the person who commands the organization of people with lots of guns, and more authority that stems from enough people having actually explicitly said that they'll follow you if you say to boycott somebody. The judge who's contracted to follow a constitution and determines on that basis when to tell somebody "You're not playing by the experimental rules we set out, get out of this regional experiment" or "You can't enter this city, based on the public reason you were expelled from your last home" is not a spiritual leader, they are not the holy person entitled to command violence.

It's not that there are no people trained to use advanced weapons and factories set to be quickly repurposable to making them - people do understand that you don't want to let the first defector conquer the whole world with just a bread knife - but it's understood that the actual use of this power to settle conflicts represents a grave danger to everybody; it's not used *routinely* the same way it is on Earth. It is possible to coordinate around deploying less pride-injurious solutions instead, where you don't have to walk around in public being visibly way off the Pareto bargaining frontier.

[...]

Eliezer Yudkowsky: I don't want to make it sound like taxes are voluntary. Different regions will, to different degrees, be trying to produce non-excludable goods, and entry to that region will require an explicit commitment to pay for them. But this is less of an issue of being an Earth-style milkable tax cattle because no other nice place in the world will let you emigrate and be a citizen; local goods like this are more often produced on the city or megacity level than on an enormous national level, so there are a *lot* of choices in where to pay taxes to, and you can travel a short distance and pay taxes somewhere else. It is more like being a customer of a business establishment and less like being a cow that gets milked, and if that sounds utopian, it's because people explicitly sat down and thought about the problem and put some effort into optimizing the larger structures that got adopted.

David Moscovici: What typically happens if a regional experiment is engaging in behavior X, that is utterly abhorrent to the people/leaders in a neighboring bigger/stronger administration?

Eliezer Yudkowsky: Like... preventing people from leaving? Brainwashing kids? Leaving people's brains to rot instead of freezing them? I can't recall offhand reading about that happening - the Chroniclers try to write about things in proportion to how many people they actually affect, not how outrageous they are - but I think they'd expect an army to march on them, which is why it wouldn't happen often.

David Moscovici: How long has it been since the last major bordered administration?

Eliezer Yudkowsky: Don't know, wasn't a student of history. There wasn't an epochal moment so far as I know - just trade barriers going down, immigration barriers going down, taxation authority devolving to internal regions.

Ben Pace: Has your world ever coordinated to slow economic progress down (e.g. due to concerns around nanotech/AGI/biorisk)? How did that work, what were the main pushbacks, and how were they appeased? And how did society continue to function given that massive numbers of people were told to e.g. not work?

Eliezer Yudkowsky: I mean, given that I never heard anybody discussing Artificial General Intelligence and that computing progress was suspiciously slow compared to Earth, somebody clearly did something, but I don't know what, and in fact I had no clue whatsoever that anything unusual was missing until I got to this Earth. I mean, that is what an *actually effective* global conspiracy *should* look like. Some of the logic behind it ought to be clear from the point that, if I'd had any inkling in dath ilan, I would have found the highest-ranked shadarak I could easily get to and told them that I suspected I'd run across a bigass infohazard; and they would have told me what, if anything, to do from there.

I don't understand what you mean about people being told to not work. Nobody knew they were being told not to work on Artificial Intelligence. They just worked on other things instead.

Ben Pace: What's the simplest piece of software that over 1 billion people use on your Earth that is net positive and that we don't have?

Eliezer Yudkowsky: Computing in general is less advanced, but we do have some of it. Most of the software that comes to mind is software supporting other political and organizational systems that you don't have here; the closest pure software I can think of is Kickstarter But In Full Generality, and that may actually exist by now but with not enough people using it or it not being legal to use it for the right things. There's software that supports a popular multilevel-delegation political system, what I might call a Dunbararchy, and I guess you could conceivably try to build that and let people use it to see what happened, even if you weren't allowed to make its results politically binding.

Jessop A Breth: Does your planet have humans?

Eliezer Yudkowsky: I have no particular reason to expect they'd have any trouble having kids with people from this planet, if you want to define "the human species" the way a species is usually defined.

David Spearman: How does the society figure out what better alternative to the status quo are? Is there some system like [Archipelago](#) where you can run small-scale experiments which the VSP's can go on to signal boost to the rest of the population?

Eliezer Yudkowsky: It's widely understood that the point of having regions apart from one giant homogenous optimal Megacity is so that you can do a thing differently in different places, yes. Either to house people with different priorities about how to set regional parameters that need a single setting - there is probably somewhere out there where everybody goes naked, though I didn't particularly look for it at the time - or to run experiments.

Maximilian Schlederer: Are people on parallel Earth more intelligent than on our Earth?

Eliezer Yudkowsky: Yes because they have actual education. And even apart from that, it's almost impossible that they wouldn't be, if only because there's a norm against chronically unhappy people having kids, and that probably reduces prevalence of a bunch of low-grade health issues.

Connor Heaton: How much closer are they to general AI?

Eliezer Yudkowsky: Impossible to even guess. I imagine that they're treating their ability to take it slowly as a huge resource, which implies that they're doing a bunch of capabilities research and restricting the results. I have no idea at all how far they've gotten with the lower computing resources they have.

Tom Pandolfo: Does your world have a minimum standard of living guaranteed by law?

(That is, every person has adequate housing, is adequately fed, and has reasonably adequate healthcare - and these are guaranteed by law, not as an "inevitable consequence of The Market.")

Eliezer Yudkowsky: The mineable resources of the world have economic rents that's captured via Georgist principle and is available equally to all citizens as an income, but there's no attempt to protect that income from taxation or treat it as a Universal Basic Income that ought to be sufficient to live on; it's just the rent that people get paid for respecting the existence of a property system at all.

People donate around 15% of their income to good causes in one form another, and while that's always been very competitive, it's also always been enough to support the existence of Quiet Cities where you can go to retire from Civilization if you haven't had any kids. If

you *did* have kids, your kids would have other options for being supported, but *you* might be allowed to starve; and if you tried to steal to support yourself, you'd be deported until only Australia (actually the territory we'd call Japan) would accept you, and then you might die there. You're not supposed to be having kids if you aren't sure of your ability to support yourself in Civilization, and it's understood that there should be some incentive structure against that.

If you randomly get injured in an avalanche, there's insurance for that kind of thing, and if somehow you ended up with Real Genuine Unforeseeable Unavoidable Bad Luck after having kids, your relatives or friends or some Very Serious Person might serve as a safety net. But the socially general safety net is deliberately and consciously restricted (for reasons of avoiding Malthusian equilibria and societies full of unhappy people) to people who have not had any kids.

David Bahry: Is oligopolistic collusion considered a market failure, or an admirable instance of solving a coordination problem?

[...]

Eliezer Yudkowsky: It'd be considered a failure of the consumers to coordinate on starting a non-oligopolistic competitor, I'd guess? And depending on what exactly had happened, I would expect a swarm of Very Serious People writing Very Serious Essays about some larger civilizational dysfunction that had allowed it to happen in the first place. See previous answers about individual oligopolists not being able to individually capture most of the monetary value seized by cartels, and on the understood civilizational importance of not having separate reputational puddles of important people who would pay them in prestige for having pulled off an elaborate locally coordinated defection.

David Bahry: Are unions considered a market failure, or an admirable instance of solving a coordination problem?

Eliezer Yudkowsky: Everybody in dath ilan is an economist the way everybody in this Earth is a scribe, so they would immediately reply that this is the sort of thing where supply-demand balancing prices seem like a perfectly fine solution; if you presented them with a situation where unions were necessary in order for employees to capture a fair amount of the value they created, they'd ask how the *hell* some employer had ended up with an effective monopsony on labor. This question sometimes has an answer on Earth; on dath ilan they'd spend their effort on avoiding the monopsony situation rather than on unionizing afterwards.

I think there's a certain amount of *implicit* unionization in the sense that, sure, most of the employees in a company could get together and yell something, and it would not be considered very socially acceptable for the company to try to prevent that.

On Earth, formal unions got started when companies were literally bringing in people with rifles to force miners to work. I think dath ilan would agree that unions are an appropriate and indeed *overly mild* response to this condition, but it is not a problem they are currently trying to solve.

Samy Gallienne: How is the media system funded and distributed while ensuring quality? On our Earth, we haven't figured this one yet.

Digital information wants to be free, but the labor to produce the information needs to be compensated. As a result, newspapers are struggling. [Government] doesn't seem to be the problem since that limits competition while risking propaganda.

Meanwhile, distribution is [increasingly] reliant on social media which pushed dubious [ethical] practices like clickbait and rewards outrage-creation over quality.

Eliezer Yudkowsky: Among the public goods that gets supported by the 15% of income that people publicly donate to charity, are the Chroniclers of Humankind - somewhere between what judges want to be, what journalists used to pretend to be, what Wikipedia aspires to be but with better writing; they are supposed to be very neutral, very fair, very kind to people who haven't deliberately massively screwed up, very well-paid, etcetera.

Lots of other people write about things too, of course; but to the extent that there's such a thing as a Story there, it becomes part of the Story when the Chroniclers start to retell it. If Scott Alexander adopted a special voice that he used to speak when he wasn't taking sides in any partisan conflicts, he could in that voice be a Chronicler.

[...]

David Moscovici: Chron 1. How do Chroniclers avoid (or how is it avoided that Chroniclers) select new truths to 'declare' depending what ideology might grasp them? Do they never "fringe out" as most of our thinker/[analyst] categories do?

Chron 2. Does their Chronicler status become threatened if they do gain a very uncommon understanding of the world?

Chron 3. Who retires their status if their objectivity becomes heavily compromised?

Chron 4. And how have 'new truths' become recognized?

Chron 5. What happens when there are apparent and direct contradictions between pockets of Chroniclers?

Chron 6. Does anything actively prevent capture in a certain space by a specific thinkgroup of Chroniclerhood?

Eliezer Yudkowsky: The closest Earthly analogue is Wikipedia, I'd guess, or somewhere between Wikipedia and a high-functioning science journal; except that there's a common understanding of much more sophisticated discourse norms and reasoning norms, when it comes to saying that Chroniclers should be using standard reasoning to write about things. You need to remember that I am an average kid from that place; I did not *invent* the stuff I wrote about on Less Wrong but I did *know* all of it, albeit not with much sense of your Earth's appropriate citations.

If a Chronicler was writing about something especially controversial using a non-public source, they'd probably call in a senior retired Chronicler to act as witness to the conversation, or something like that. Otherwise, why would there be any need to trust them in the first place? The Chronicler would just show their work.

David Schneider-Joseph: In what way is the parallel Earth doing worse than ours?

Eliezer Yudkowsky: At least some types of people in it are probably having less fun than those people would be having here, if those people were otherwise fairly wealthy in both places; though no especially striking non-socially-harmful examples are coming to mind except for people who want to take a lot of drugs and people into BDSM.

David Schneider-Joseph: @Eliezer Yudkowsky Why would drugs and BDSM be adversely affected by society having a higher ability to solve coordination problems?

John Wentworth: Sounds like a lot of previous answers involve making lots of information available to lots of people, and relying on reputation. How do people decide what to pay

attention to? How do people notice when nobody else is paying attention to some important information?

Eliezer Yudkowsky: I'm not sure there was any particularly magical solution other than having lots of people walking around knowing that this was an issue and a public good that they had to pay for in money and reputation. It's like asking why we had scientists actually running replications of experiments; we understood that this was important, and that you actually have to pay in money and honor for important things, not just pay lip service to them, so funding was available and newspapers would report the names of the first two replicators next to the people who'd found the preliminary hinting.

Brett C Allen: Do informational asymmetries exist, or is everyone equally informed at the best resolution they have capacity to interpret?

Eliezer Yudkowsky: Of course informational asymmetries exist, they're still bounded agents for heaven's sake!

Brett C Allen: But then you cannot solve the class of problem such as the prisoners' dilemma, because geometries of action and incentive are possible that confer advantage based on a defect strategy?

Eliezer Yudkowsky: Allow me to introduce the number one technical thingy I actually managed to partially remember from dath ilan: https://arbital.com/p/logical_dt/?l=58f

Thor Taylor: If you could materialise a small city-state on Earth following Dath Ilani norms and customs, would it be robust to international politics? Or does the equilibrium on Dath Ilani require general consensus to be able to punish transgressors? E.g. is the militarisation required for a small state to fend off militant neighbours compatible with the freedoms you consider crucial? Are norms from a cooperative world robust against attack by selfish [foreign] actors playing zero sum or even negative sum games with trade and espionage?

Eliezer Yudkowsky: I think that city state would effectively materialize far out of equilibrium, would immediately regard itself as being under siege, and would immediately start to try to build weapons of mass destruction in order to have a credible threat to prevent its conquest by the environment around it, which I'd expect to rapidly go extremely hostile if presented with a non-politely-Facebook-censored version of dath ilan culture. Our kids get explicitly trained on perspective-taking, and wouldn't have the expectation for an absolutely foreign culture to look very nice and pretty by the norms of suburban pontificators. Your culture makes no such allowances, has no such concept, and is full of "low-decouplers" who literally lack any internal grasp of the mental motion they would need to perform.

Ben Albert Pace: What role does Robin Hanson have in your world? What job does he have?

David Spearman: I imagine "Crazy Idea Guy" is a sub-genre of Very Serious Person.

Ben Pace: I kinda want to know if he runs a university or manages a government department of prediction or (more likely) some third thing I haven't imagined.

Eliezer Yudkowsky: There are no direct people-level analogues between worlds - even if dath ilan had only diverged 100 years earlier, that would be more than enough to butterfly out of existence almost everybody born more than a couple of years later, and dath ilan must have diverged much much before that.

And there can also be no metaphorical analogue of Robin Hanson, because the whole concept of Robin Hanson is that he understands some particular things that aren't common

knowledge here. You can't have that one person who goes around loudly observing that "education isn't about human capital" because dath ilan doesn't have an education system like that, and because it doesn't make you a contrarian iconoclast to suggest that you'll get what you measure and pay for.

Or to put it another way, if there actually was a Robin Hanson back in that world, he was too much smarter than I was, and I was part of the mob of hoi polloi who couldn't tell the difference between that and any other crazy person.

Ben Pace: What do the contrarian iconoclasts look like in your earth? I expect you'll say that they're off running experiments that nobody understands, and somedays they come back with incredible results and then they're absorbed into the way of being for the whole civilization.

Eliezer Yudkowsky: Or not-so-incredible but still-cool results, but, uh, yeah? I mean, what do the contrarian iconoclasts look like inside the little crippled partial tiny fragment of my home culture that I managed to reproduce here? They look like a bunch of outright psychiatric nutcases, a bunch of people being loudly wrong, and a few people who made fortunes outbetting the markets during the Covid-19 crisis.

Patrick Lozada: Which one (or ones) do you use and who decided this.

 Plugs by Type			
Type A  2 Pins 15A 100-127V	Type B  3 pins 15A 100-127V	Type C  2 pins 2.5A 220-240V	Type D  3 pins 5A 220-240V
Ungrounded Compatible with plug type A	Grounded Compatible with plug types A & B	Ungrounded Compatible with plug type C	Grounded Compatible with plug type D
<small><i>Please Note: This socket has a partial and unsafe compatibility with plug types C, E, & F.</i></small>			
Type E  2 pins 16A 220-240V	Type F  2 pins 16A 220-240V	Type G  3 pins 13A 220-240V	Type H  3 pins 16A 220-240V
Grounded Compatible with plug types C, E, & F	Grounded Compatible with plug types C, E, & F	Grounded Compatible with plug type G	Grounded Compatible with plug types C & H
<small><i>Please Note: This socket has a partial and unsafe compatibility with plug types E & F.</i></small>			

Eliezer Yudkowsky: All cables had already evolved to their final universal form of USB-C by the time I was born.

David Bahry: Does marriage exist (monogamous or otherwise), does divorce exist, and how common are they?

Eliezer Yudkowsky: Monogamous heterosexual marriage is The Rule to an even greater extent than in the modern USA, and this is one of the places where I suspect our Earth is doing slightly better. Either social pressure is actually and effectively producing sexual conformity in dath ilan, or the general sense of "If you're an unhappy misfit, don't have kids" or "Don't lie to yourself and others about who you are" caused homosexuals to actually not reproduce or not be in fake heterosexual marriages with kids, over several generations, and they actually became a smaller population segment. 😞

Daniel Powell: What method do you use to determine how to distribute scarce resources?

Eliezer Yudkowsky: If they're raw resources, conduct a Georgist auction to capture the economic rents from them, modulo a premium-paid to establish ownership-momentum if you have to build permanent installations near them, but with continuing rents due. If they're resources produced primarily by value added by scarce labor, those resources are owned by the producer and you trade with them. If that doesn't answer your question, what do you have in mind?

Daniel Powell: @Eliezer Yudkowsky so there's even more poor people, and the lines for Hamilton tickets are even longer?

Eliezer Yudkowsky: I don't get it. :confused pikachu face:

Daniel Powell: Well, the people who overestimated the mineral rights value of Gaul couldn't participate in the bidding on Britannia or anywhere since, so they and their heirs have been locked out ever since.

Eliezer Yudkowsky: That's the "continuing rents due" part? You don't just bid a little on the mines and own them forever.

Daniel Powell: Yeah, they overestimated the first thing to auction off, and the person who bought the lease from them overestimated it less. Forever after they can't afford to purchase any rights unless everyone with an average amount of currency underestimated the value.

Michael Blume: Do private cars exist and if so how many people own them and how much explicit legal structure is there for handling their costs?

Eliezer Yudkowsky: I mean, I'd expect there to *maybe* be regions for ore-mining where they just pave a road and have humans drive over the road, instead of paying the additional expense to set up automated car lines, because it's not worth the added expense considering the very small amount of irregular traffic - something like that? But if you let humans drive cars, they crash into each other and kill people, and much worse they occasionally *crush the brain and destroy the soul*. If there were regions where people go to crush their souls, I didn't hear about them. The length dath ilan goes to in order to avoid brain-destruction scenarios is finite, but *large*, and humans driving cars at high speeds sure do cause that.

Jessica Evans: A world with more coordination sounds like a world with less liberty for the same basic reasons that democracy consistently produces tyrannies of the majority. Explain how you get "more coordination" for the same price at any level of complexity or be ridiculed.

Putting it another way, I wish to a genie for "more coordination". Why doesn't the genie instantly solve this by making the world homogeneous in opinion, or making fringe

disagreement dramatically more costly, or some other horrifying thing.

Eliezer Yudkowsky: Coordinatees who understand coordination better? That world wasn't the product of the genie wish you just described?

Richard Wilde: How did climate change work out for you?

[...]

Eliezer Yudkowsky: Better-coordinated people can do the same amount of Science and Engineering with a smaller global population, meaning that we reached a roughly equivalent technological level with around a tenth of the population, so we didn't put a significant amount of CO₂ into the atmosphere before transitioning to liquid-phase fission reactors as the primary energy source.

[...]

Ben Pace: I hadn't notice how path dependent the issue of climate change was!

Jim Syler: @Eliezer Yudkowsky Waaait, but doesn't innovation scale with population, because you have a larger number of smart/lucky/innovative/etc. people?

Eliezer Yudkowsky: Innovation scales *poorly* with population, and even more so here than there. On my home planet you are *much* more likely to see a big company producing an amount of innovation that is, like, proportional to the square root of the company's employment, which is to say that it is increasing *at all* with population size; as opposed to here on Earth where tiny startups are often around *equally* innovative with entire established companies.

Richard Wilde: And how did you manage to keep your population so low?

Eliezer Yudkowsky: It wasn't a coordination thing, it was just a question of igniting earlier along the population-growth curve. We didn't know our population was "low" because we weren't comparing it to Earth.

Erica Edelman: Given that nobody is happy all the time, how do you separate out unhappy people (who shouldn't have kids) from not unhappy people? Is there an objective test that tells you your current happiness rating? Is there an absolute number on the happiness scale you should be, or is it like... the lowest 10% of the population. If it's an absolute number did a very large percentage of people not meet standards when this system first rolled out?

Eliezer Yudkowsky: You could literally spend the rest of your life reading all the Very Serious People arguing about that. I would personally yell "Bottom 20%" and then run away before they got me.

Erica Edelman: Don't you worry that doing bottom 20% over lots and lots of generations is going to eventually lead to a world where everyone is... psychotically happy / mentally ill levels of happy?

Eliezer Yudkowsky: THANK YOU MISS VERY SERIOUS PERSON FOR POINTING OUT THIS IMPORTANT FUTURE PROBLEM

Eliezer Yudkowsky: (runs faster)

Eliezer Yudkowsky: (summons David Pearce to distract her)

Ymir Vigfusson: How does the police on your Earth bargain with pairs of prisoners who have been arrested for an alleged crime?

Eliezer Yudkowsky: For one thing, the people who do the arresting are not the people who run the interrogation who are not the people who do the prosecuting who are not the people who run the prisons! See <https://yudkowsky.medium.com/a-comprehensive-reboot-of-law-enforcement-b76bfab850a3> for an elementary concept of how an economic literate might look at this kind of thing. Offering prisoners clemency in exchange for them purporting to inform on each other runs into all kinds of obvious horrible incentive problems *within the court system* that any Very Serious Person would point out before five seconds had elapsed.

Steve Jackson: What types of tokens do they use to signal where the problems are, and how do they make sure the tokens keep moving?

Eliezer Yudkowsky: Uh, are you referring to money? They use money.

Steve Jackson: How do they make sure their money signals problems that need solving well? Or do they do that?

Kelley Meck: Put another way:

Do you have loans at interest? If yes, what did/do you do with the bad kind of lenders? If no, how does the market clear between now-problems and later-problems?

Eliezer Yudkowsky: I'm confused. Surely the punishment for being a bad lender is that the bad lenders lose money? The public policy intervention here is not to bail them out, there, you're done; private insurance on deposits will now price-signal the riskiness of those deposits.

Andrew McKnight: How do para-people notice when their problems are caused by coordination failure?

Eliezer Yudkowsky: In virtue of literally everybody knowing what a Nash equilibrium is, what a Pareto optimum is, classroom situations that show them blowing up in practice, newspaper stories that analyze things in those terms, Very Serious People debating edge cases in Very Serious Debates? This is like asking how economists notice when something is balancing supply and demand, or when an arithmetician knows that it's time to count or add or multiply something; it's a form of literacy that is understood to underpin Civilization in much the same way as counting or reading, so the Very Serious People are constantly being Very Concerned any time it shows a 5% drop in one state region.

Andrew McKnight: Eliezer Yudkowsky ahh, basic coordinacy to go with their literacy and numeracy. I see.

If I may add a follow-up, are there major problems on para-Earth with false coordination where para-folks overcoordinate when they should just do the standard Earthly thing?

Eliezer Yudkowsky: I get the impression they could be overthinking a few things - nothing specific comes to mind, just the general level of overhead and how often Very Serious People have Very Serious Takes on things, sort of like looking at Earth and saying that it spends too much on "Left vs Right" takes on things. But whooooo nelly does it look like it's better to spend too much thought on coordination problems than too little. If there's a Golden Mean Earth, it's probably 80% of the way to dath ilan.

Ben Pace: I'm a bit confused by this. One of the advantages of noticing that there's a supply problem, is that the supplier can unilaterally change their price. Prices are indeed

set by buyers unilaterally outbidding each other, and sellers unilaterally underbidding each other.

Yet with coordination problems, even if everyone recognizes that something is a coordination problem (and has the concepts of Nash equilibria), you still often have to do much more surprising/novel things to switch equilibrium. I mean, you can just use politics as usual ("Let's all jump at the same time because we're in a bad Nash equilibrium!"), and that will go more easily if everyone understands the concept.

But I was expecting you might say something more like "In my world is a much better reward mechanism for people who successfully solve such problems. If you manage to move your school/business/community into a better equilibrium, you are massively rewarded even if you did not monetize it, because at each level there is an equilibrium team whose job is to pour money into the bank accounts of people who do this." Or something. That probably doesn't work, but it sounds to me slightly more like it could.

Eliezer Yudkowsky: Obviously you'd win an insane amount of social marbles if you somehow coordinated a really big jump toward a much better equilibrium, but that presumes that a much worse equilibrium was somehow allowed to develop in the first place and that you were the first person to spot it and see a solution. This is *actually* hard to pull off in dath ilan, because when a good solution is visible the activation energy is actually available to jump there. Part of the reason I was so quickly able to spot adequacy fantasies in this world as psychologically dysfunctional, is because I've been in the world where it's *actually* hard to spot big problems and big fixes, and *that world looks very different from this one*.

Ben Pace: Okay. I think I have a picture of how it works on your Earth.

I'm imagining the situation with the QWERTY keyboards, and Bob realizing that what was good for the old typewriters (not hitting keys next to each other) is not needed any more for our new keyboards.

At this point in time there's lots of companies (i.e. 10-50) making these newfangled 'personal computers' in lots of countries and lots of different languages.

I'm imagining Bob talking to each company, and saying (a) "this is sort of maximally inefficient for our hands" and (b) "we're gonna be pretty soon in a Nash equilibrium where none of us can unilaterally improve it and everyone's learned the old one".

Then the companies probably give Bob command over when they all jump to the new one, because there's standard social protocols around doing this (and, though it's not something Bob actually thinks about, Bob knows in the back of his mind that if he were to use it corruptly they would follow through on punishing him, letting his community/employer know the sort of person he is, etc). He also knows he will get financially compensated by all of the companies, to the tune of like \$5k (peanuts for them, adding up to a year's salary for him).

Once Bob has got 50% of them, the others quickly follow into line, because you know that when a Coordinator has got 50% of parties ready to jump, they're to be trusted and jumped with. Then after maybe 4-6 months of work (traveling and persuading initial companies), he hits jump, they all commit to changing keyboards on their upcoming computer, and the customers will just be forced to learn, with a slight dip in the economy for a bit followed by slightly faster growth after.

Eliezer Yudkowsky: If you built a better school than an Earth-school baseline, people would be like, "Okay, 5% of the story is about this awesome person here, to whom all due congratulations are due, and 95% of this story is about WHAT THE HELL FUCK WERE WE DOING?"

Eliezer Yudkowsky: Bob in your story is, like... three hundred Very Serious People arguing with each other in newspapers for several months.

J. Caitlin Elizondo: Do you guys run on the same meat hardware, with the same type and degree of inclinations e.g. towards sex, love, status? If so, how do you manage for those whims not to sidetrack everything into oblivion?

Eliezer Yudkowsky: It seems to me like pretty much the same meat hardware to the same extent that, say, Ashkenazic meat hardware is the same as Eskimo meat hardware? And I'm not really sure how to answer your question, maybe something like, "Greater economic literacy and conscious awareness of short-term incentives means that people have put a lot more deliberation into being aware of where short-term incentives point and trying not to misalign them." Nobody has ever invented Twitter, and if anyone did, there would be immediate unanimous coordination around jumping to something else with better incentives and no 280-character limit, and if Twitter somehow still existed despite that, it would be taken for granted that you didn't want it inside the same web browser as your work web browser.

One of the aspects of dath ilan civilization that I'd expect an Earth-person to find much more relatively offputting is the degree to which anybody in the top 75% (not a typo, I mean the top three-quarters) of personal attractiveness would be expected to wear a veil, or makeup to look uglier, applying to both men and women but with a lower threshold for women. If there were any such thing as a beach with people wearing scanty swimwear, it would have all kinds of Cognitohazard signs slapped all over it and nobody would ever publish any photographs of it.

Why? So that if you go into a bedroom with somebody and get naked, you're not comparing them to the attractiveness of the top 0.1% of the population. If you want to shoot yourself in the foot like that, you'd have to go out of your way to do it and tromp past a lot of warning signs, because the rest of Civilization has comprehended "ability to be attracted to the average naked person" as a public good and is coordinating around preserving that public good. It lends all of Civilization a very deliberate and abstract quality that I'd expect to put off a lot of Earthers, and not without reason. But if Earth civilization doesn't *immediately* blow up in a vast orgy of sex and cookie-eating, it definitely shouldn't be surprising that dath ilan manages to walk on.

Jay Schweikert: @Eliezer Yudkowsky But what does that mean for the dating landscape? It sounds like dath ilan probably doesn't have the equivalent of Tinder, or at least it's not widely used for the same cognito-hazard reasons. But is relative attractiveness a factor at *all* when people are trying to figure out who they want to date?

That is, I can understand "ability to be attracted to the average naked person" as a public good, and that a well-coordinated society could have social enforcement mechanisms like you describe. But assuming people still are actually more attracted to more attractive people, it's harder to understand that people would just give up on trying to signal this information, or give up on trying to read those signals.

So, for example, if you're in a smaller groups of friends, can you take off the veil, or is that basically always required for anyone but a romantic partner? Are people expected to wear sweatpants and sweatshirts when they work out? Is there a whisper network for figuring out how attractive people actually are? How scandalous would a "no makeup singles bar" be? Is pornography a thing at all, or is it locked behind major warning gates?

Eliezer Yudkowsky: If pornography is a thing, it was locked behind warning gates big enough that I was literally not exposed to the concept before Earth! There's sex manuals illustrated with carefully 20th-percentile unattractive people. There's a kind of loose robe-like clothing you'd wear to work out that would also serve the purpose of absorbing sweat and preventing it from getting all over the equipment, which honestly still seems like a pretty good idea to me, if I'm not just being homesick.

Computing in general is less advanced, I assume because somebody knows about AGI, and dath ilan never started having social media. If they started getting results remotely similar to Earth's social media it would all be shut down.

So no, no Tinder. But a lot of deliberate understanding of the matching problem in dating. Some Very Serious People who are Very Concerned about Where It's All Headed would say that it's *too* deliberate, and people should just, like, fall in love at first sight properly. But roughly speaking, what they have instead of Tinder is real-estate brokers; you tell somebody *all* about yourself, *they* get to see you naked and take pictures; and if they can find you a mate by talking to their fellow brokers, then they get a bounty that's supposed to reflect 5% of the value-added of being in the better-than-you-could-have-found-on-your-own relationship they found you.

When there's a big problem, expect dath ilan to have paid professional specialists to solve it; and everybody in dath ilan is an economist the same way that everybody on Earth is a scribe from a medieval perspective, so they will be very careful about what they measure and pay for. The problem of finding good people to date is a big one with lots of value dependent on it, so *obviously of course* there's going to be a well-paid professional class devoted to it, which people actually use, with payouts dependent on results because they know you get what you pay for.

J. Caitlin Elizondo: This doesn't matter but how do people decide/figure out what attractiveness percentile specific people are in? Does everyone know their number, like an IQ score?

Ben Pace: I do anticipate that you can pay for sex and for dates and things on Dath Ilan, and that there is some price discrimination there, and that generally the people you pay are attractive and agreeable and extraverted and so on than average (though with lots more buying niches than are available on earth e.g. disagreeable quiet people). This would be valuable for all sorts of people e.g. people unable to get a mate, busy people, etc.

Eliezer Yudkowsky: If you're more attractive than people you see on the street presenting as, don't let yourself look more attractive than that.

Raymond Arnold: Huh, does this mean you don't have anime catgirls?

Eliezer Yudkowsky: We don't.

That is like *super* a thing that the Very Serious People would be *super* against.

JJ Treadway: Is there a reason you don't instead *raise* the attractiveness of *less* attractive people (using e.g. plastic surgery or genetic engineering or mind-uploading into attractive artificial bodies) in order to reduce attractiveness-inequality? Do you just lack the technology to do this cheaply/safely, or is there some more subtle reason why this would be a bad idea regardless of its technological feasibility?

Raymond Arnold: A thing that sticks out in my mind is that it's a runaway arms race, that doesn't really have a point (assuming a model that this is a domain where you just hedonically adapt, which you might or might not buy)

Eliezer Yudkowsky: We don't have the tech? I don't think anybody would object to raising everybody's attractiveness and refiguring where the 25% cutoff was.

Raymond Arnold: Oh, to clarify - when I said anime catgirls, I meant, like, "anime shows, that feature catgirls" as opposed to actual catgirls. I'm assuming the answer is the same based on the porn one but just doublechecking we communicated successfully

Eliezer Yudkowsky: Right. They try not to present people with fantasy worlds more attractive than reality. Respectable fantasy novels will generally start the protagonist off with a disadvantage and force them to reform some awful place, for the same reason.

Rodrigo Moreno Nuñez: "Everybody in dath ilan is an economist the same way that everybody on Earth is a scribe from a medieval perspective" makes dath ilan so much more believable in [one] sentence. props!

J. Caitlin Elizondo: Woa, do other people (here) walk down the street and have a clear sense how people compare in attractiveness to them?

[...]

J. Caitlin Elizondo: "If pornography is a thing, it was locked behind warning gates big enough that I was literally not exposed to the concept before Earth." Why wouldn't you suspect the same might be true of kink?

Eliezer Yudkowsky: Maybe there were special regions where it was different, but I think the standard background of departure was very much in a headspace of, "You want somebody to hurt you? Injure you? Cause you pain? That's not being Light-aligned. That's not how a biological organism is supposed to work. What's wrong with you? Do you have psychological damage? And you who want to cause pain, that's just called being Evil. Force it down, and if that makes you sufficiently unhappy, don't have kids so they won't be unhappy too." Even if there were places that were experimenting with having things be different, it wouldn't have occurred to me to look for them, and I expect a lot of other potentially kinky people wouldn't have looked for them either, or ever be exposed to the stimuli that could have made them realize they were kinky.

Eliezer Yudkowsky: To be clear, I think that if dath ilan got a good look at the Earth equilibrium, they wouldn't just go into denial about it, there would be a Huge Very Serious Blowup about what had happened.

J. Caitlin Elizondo: What had happened to Earth? Or them?

David Moscovici: Would a category-fetish/preference, with criteria like hair color, or race, or profession be seen as Light-deviant?

Eliezer Yudkowsky: What had happened to them - it *is* understood that having more fun is in some sense the ultimate purpose of existence.

Eliezer Yudkowsky: David - no, that's just individual taste (or so they would agree). The nonobvious step for them would be distinguishing the desire to hurt somebody and cause them pain in a sexual way, versus, say, the desire to kill them in a sexual way. Among other things, you need to realize there are people who want in a sexual way to be hurt, which is not something you can necessarily figure out from inside of your head; there's a multi-step cognitive inference problem here.

David Moscovici: Would pursuit of subservience to the point of non-physical ill-treatment be seen as Light-deviant?

And/or would offering such treatment be Light-deviant?

(Just testing the bounds of such cultural freedoms beyond violent kink)

Eliezer Yudkowsky: Yes and yes.

Kelley Meck: What is music like on parallel earth? Is it still used for marketing of products? Is it still used for dancing?

Eliezer Yudkowsky: More melodic, fewer words, the popular stuff is less repetitive and less based around very loud beats. I'm enough of a barbarian that I actually like Earth music better. It doesn't try to be respectable.

Advertising is understood to be mostly a negative-sum game where people try to steal customers from each other or from other business sectors, and slightly a positive-sum game that could theoretically produce more informed consumers if for some reason the Chroniclers and Very Serious People and lesser reporters were all asleep on the job.

It's not illegal, but it's understood that if you saw a Pepsi advertisement and switched some of your consumption from Coke to Pepsi, or from orange juice to Pepsi, or from pretty LED jewelry to Pepsi, you'd be contributing to a negative-sum game, which would generally go against your self-concept.

But if Pepsi had an actual superior product they would, like, obviously go pay one hundred Very Serious People a *small* token fee to spend ten minutes trying their product. If they were trying to play great music along with having their product shown on TV, I think everybody would watch that and go, "How dumb do they think we are?! What does this music have to do with the product?!"

Rhythmic sound and dancing is older than writing. That hasn't changed.

Alex Gunning: "dath ilan" is an anagram of "thailand", was this an intentional decision?

Eliezer Yudkowsky: How could that possibly be true? Oh, you mean if it was a joke? If counterfactually this was all a joke, I don't see why I'd have anagrammed Thailand or what I could have meant by that.

Jared Collins: How did you incorporate the bottom 10% distributions of the population on measures of intelligence, conscientiousness, industriousness, and agreeableness? I would think people in these demos, especially overlapping more than one, would be most inclined to defect or, worse, be mentally unable to think abstractly enough to govern their own behavior based on coordination values.

There is a not-insignificant portion of the population, for example, who could not pass your second grade even as an adult based on cognitive ability or conformity deficit grounds. If you exiled all of them, the Australia equivalent would start looking worryingly crowded and fractious to the point of a human rights crisis.

Eliezer Yudkowsky: Australia (actually Japan) is where people go when they've violated the regional rules to the point that no region wants them. If you haven't stolen or committed violence, and haven't had kids, then there are plenty of charity-supported Quiet Places that will accept you. We have 100X medieval productivity just like Earth ought to, and we're not setting all our resources on fire, so it's not hard to support any number of Quiet Places where people can go if they can't handle Civilization.

Jared Collins: @Eliezer, Quiet Places are the part I missed. What goes on in them? Do they have jobs, factories? It seems like a very obvious sort of disutility to have significant portions (back - of- my- hand math clocks in north of 30%) of the pop base doing nothing because they can't be made to do so at the efficiency frontier. And many aren't going to sit around; the willingness to work is ingrained pretty deep. If they aren't rewarded for working by your civilization, they'll make their own (suboptimal) options.

And if they can't handle civilization, they're not going to respect the self-imposed strictures on reproduction. It can't have escaped the attention of your VSPs that there's a real potential for an Eloi/ Morlock situation being set up (ref. H.G. Wells' 'The Time Machine', a culturally-well-known piece of speculative sci-fi from our side).

Karl Nordenstorm: How is Esperanto or other optimized languages doing?

Eliezer Yudkowsky: Finished over a hundred years ago. The benefits of having one shared planetary language were just too obvious, and that's leaving aside how much prettier our equivalent of Quenya is compared to the past's equivalent of Russian.

Patrick Hunter: Do you know the earliest point of departure from Earth's history and any significant turning point that put you on the trajectory towards everyone knowing pretty modern economics? In particular were do things fall on the spectrum of economic ideas being invented much earlier vs institutions adapting faster to their existence. Like understanding Nash equilibrium seems pretty important to your society but the Earth concept dates to somewhere between 1838-1951 depending on if you want to count Cournot and even the early end of that postdates the existence of modern states.

Eliezer Yudkowsky: I don't know about any overlap between histories at all, so we're probably looking at a divergence well before Sumeria.

Patrick Hunter: That explains the size of the divergence. Have you talked to any linguists in/about your language, and do you have a rough idea of how old the basics of economic theory are in your world?

Eliezer Yudkowsky: How the heck would I talk to a linguist about that?

I don't actually have much of a sense of how old things are because the phenomenon of genius is less pronounced - people actually do make improvements as they become available, in some sense, rather than leaving them bundled up for one huge genius to take in one huge leap - which means that I don't have any Big Name like Adam Smith whose century I've memorized. I could wave my hands and say "eh maybe three hundred years"?

Patrick Hunter: Ask on social media for linguist followers, and they'll be able to guide you through what they need to know.

Eliezer Yudkowsky: On reflection, I think this problem is unsolvable? I can't think of a large-enough batch of place-names or preserved words that are all from the same region, and that could serve as something to pattern-match our pre-Universal languages against ancient Earth languages. Like, let's say that all you knew is the equivalent of Quenya, a synthetic language, plus a few names of cities that are older than a hundred years. Could somebody on Earth trace back the roots to Sumeria?

Patrick Hunter: I don't know enough about linguistics to know I've just been primed by Glowfic to treat this as a default question for people from other worlds.

Jim Syler: How do you maintain liberalism in a society?

[...]

Eliezer Yudkowsky: Question too broad, maybe read answer comments above and then narrow it?

Jim Syler: Hopefully you're asking me to be more specific rather than to ask a narrower question.

Liberalism is probably the best social system—or at least the best social system for *this* Earth—honestly [virtuous], non-self-serving, and far-sighted philosopher-kings telling everyone what to think might work better, but that's not a feasible option here. So we're left with liberalism, in which no ideas are forbidden to be thought or expressed (which is not to say that every utterance is appropriate in every context, but there's *somewhere* you can go to discuss *anything* without risking being ostracized from society), so that error is tolerated and the best arguments are allowed to rise to the top in the marketplace of

ideas. Basically what Jonathan Rauch can't shut up about, plus the notion of a balance of competing interests that the Constitution was based on.

The problem with this is that it's in everyone's interest to have *their* ingroup rise to the top and suppress all the other groups, so that they've got all the power and no one is allowed to question their edicts, so that on *this* Earth, liberalism seems to be a fragile and unstable equilibrium.

So how does one shape a society so that liberalism is robust and antifragile?

(Note that although this question has both cultural and political aspects, I'm focusing on the social ones, as (I believe) politics is downstream of culture.)

Eliezer Yudkowsky: I know of no solution to this problem that uses neither better-educated voters nor philosopher-kings. The dath ilan solution is that the voters explicitly understand the thing about the incentive for some faction, even a majority faction, to burn the free-speech commons, and they'd see where that would go, especially with all the Very Serious People who would yell about it.

Jim Syler: @Eliezer Yudkowsky My question, then, is how you move toward having a populace that is generally aware of that.

Eliezer Yudkowsky: Well, I tried.

Justin John Holt: How is baby formed?

Eliezer Yudkowsky: When a mommy and a daddy coordinate with each other very well.

David Bahry: Does sex work exist and if so how is it viewed?

[...]

Eliezer Yudkowsky: Visiting an Experienced Professional Sex Worker is viewed as Concerning or Potentially Irreversible because if you have sex with somebody extremely experienced and good at it, who's focusing entirely on your pleasure, that's the sort of experience that could potentially ruin you for regular sex. But there isn't any concept of it being wrong to, I dunno, trade around regular money with regular people in order to remedy some imbalance of regular sex, like, "I'd like to have sex where I don't have to worry about your orgasm, can I pay you fifty dollars for that" isn't remarkable any more than asking somebody to do dishes that week in exchange for money. "Okay, I realize this flirtation attempt failed, can I just pay you three hundred bucks up front" might be a little weird and funny but it certainly wouldn't be *illegal*.

Being an experienced professional sex worker isn't illegal either, it's just one of those things where you'd be honor-bound to warn potential customers what they're getting into and that reduces the number of customers because people actually pay attention to warnings like that.

(I'd assume that somebody actually did run experiments somewhere about the effect on people of visiting highly experienced sex workers, which would be *extremely* legal. It would be legal even if counterfactually the rest was illegal for some reason. All *kinds* of things are legal if you do it on a small scale in the name of Science.)

[...]

Kayla O'Brien: I find the first part a bit strange because we don't assume that having an Extremely Experienced tutor or guide during our first experiences in other arenas to be likely to "ruin" us. Hearing my flute teacher play a particularly difficult piece inspires me it

doesn't "ruin" me. And while my therapist focuses entirely on me and my problems and my thoughts and feelings, it doesn't make me any worse or "ruined" at talking with my other loved ones (in fact, it makes me better!)

Why couldn't a Very Experienced Sex Worker help a less experienced partner be a better and more communicative lover just as easily?

Daniel Speyer: Seems like Sex Teacher would be its own role, distinct from normal Sex Worker and probably a bit more prestigious.

Eliezer Yudkowsky: Yup! Sex Teacher is a very different concept from High-Grade Expensive Professional Sex Worker. Your sex teacher is definitely *not* more attractive than your average partner will be, for example.

Roman Ponomaryov: Who's cleaning the toilets there? (Meaning, who's doing all the jobs that no one would be interested in doing provided there is no financial or other kind of pressure).

Eliezer Yudkowsky: People who get paid enough to do it anyways.

Cameron Taylor: How do people coordinate around drug safety, quality and applicability? (What do you have instead of the FDA and a Doctor gatekeeper class.)

Eliezer Yudkowsky: Hire several different reputable scientific-investigation companies to run trials and publish the likelihood functions? It's not that hard? Just strip out the violence from the system and leave the science.

Tomáš Kafka: @Eliezer Yudkowsky Why isn't 40+ % of a drug market run by charlatans peddling homeopathy and garlic tinctures then?

Eliezer Yudkowsky: Same reason that 40% of the furniture on Amazon isn't made from cardboard, mostly, with a side order of people who can read likelihood functions and old reputable institutions that produce them.

Karl Katz: Are there recreational or social drugs? Same deal as over in the Culture, or not there yet, or something else?

Eliezer Yudkowsky: I can't really remember hearing about them. I think that, like pornography, it would probably be restricted to special regions or factions where you had to look hard for them and cross some warning signs. The knowledge that there exists a long-term-costly substance you can consume to give you short-term pleasure or relief from psychic pain, is the sort of thing they'd treat as a minor infohazard in its own right, a Highly Unpleasant Thing It Is Sometimes Necessary To Know. The shadarak act as a repository for all the things like that.

Mateusz Drewienkowski: @Eliezer Yudkowsky I'd assume the same goes for alcohol? How about coffee? Tea? Sugar?

Marcello Herreshoff: @Eliezer Yudkowsky

> I think that, like pornography, it would probably be restricted to special regions or factions where you had to look hard for them and cross some warning signs.

I can believe that answer for drugs, but this line about pornography shook my suspension of disbelief.

It feels like it would require things roughly in the vicinity of:

A. Technology like our modern cellphones with their video recording capabilities not being put into the hands of the citizenry at all (or with pretty draconian DRM style restrictions) because serious people think it's dangerous (which is frankly a kinda reasonable option, given the destructive nature of social media on earth, though not ideal given how much extra economic coordination power their unfettered use gives people.)

or

B. Some fairly intense norms telling people not to send each other explicit pictures using such devices (this is the sort of thing the horniest 10% of the population is going to do if you give them communication devices with cameras).

or

C. Some other strong social conventions drawing a boundary on the otherwise slippery slope between the existence of sexting and the existence of mass-distributed pornography.

Which if any of these options did Dath Ilan society pick?

Eliezer Yudkowsky: Primarily A from your list; computing technology was less generally advanced and there were bulky cellphones, not smartphones. But people otherwise on boinking terms, trading nude photos with each other, wouldn't have been seen as problematic in the first place; because it doesn't introduce the problem of real average people having to compete aesthetically with airbrushed photos of the top 0.01%, which is the part that would be seen as burning a commons of the population's hedonic treadmill, for somebody's short-term gain, by way of placing short-term temptations in front of other people.

In other words, had the tech been that easy, it would have been answer C: people would see a very clear and obvious line between tempting people with what they can't realistically get except at great cost, and tempting people with what they can gain through an ordinary effort.

Ben Pace: What are some of your favorite works of fiction from your Earth?

Eliezer Yudkowsky: Considering that I can never read them again, this question is too painful for me to want to answer or think about.

David Moscovic: Why are rereads forbidden?

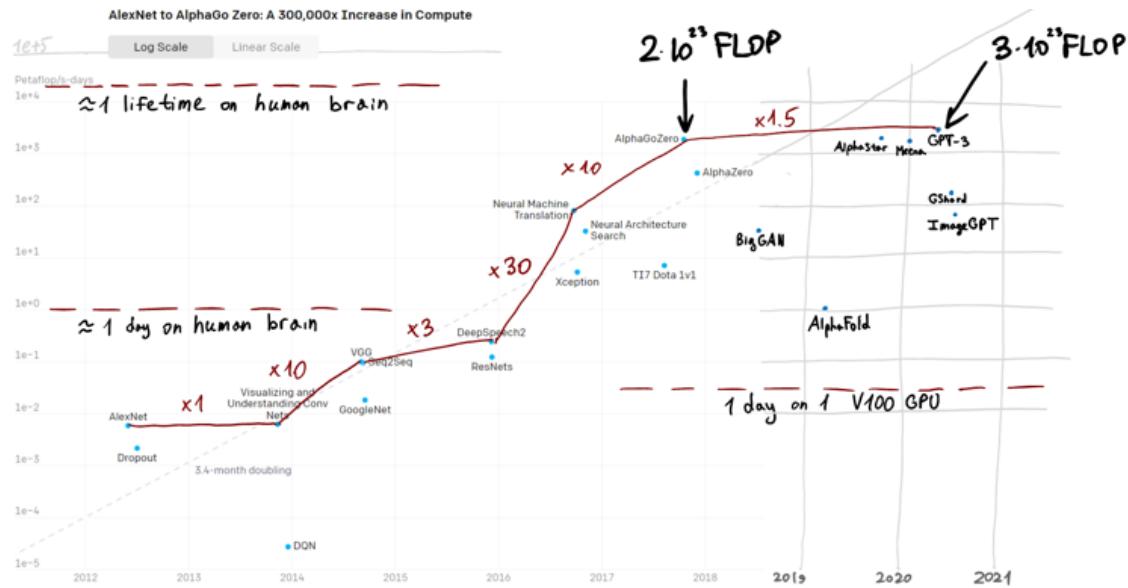
Eliezer Yudkowsky: Because I didn't get my book collection or any libraries with me when I suddenly found my mind here. Things would be very very different otherwise.

David Moscovic: In our world, re-reads are a fairly niche behavior, even among readers. Is re-reading more common back home than we have it here, and is that for a cultural reason?

Eliezer Yudkowsky: Have you read HPMOR? Most novels in dath ilan are meant to be read at least twice; many more times than that if you want to catch *all* the hints and foreshadowing.

"AI and Compute" trend isn't predictive of what is happening

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



(open in a new tab to view at higher resolution)

In May 2018 (almost 3 years ago) OpenAI published their ["AI and Compute"](#) blogpost where they highlighted the trend of increasing compute spending on training the largest AI models and speculated that the trend might continue into the future. This note is aimed to show that the trend has ended right around the moment of OpenAI publishing their post and doesn't hold up anymore.

On the above image, I superimposed the scatter plot from OpenAI blogpost and my estimates of compute required for some recent large and ambitious ML experiments. To the best of my knowledge (and I have tried to check for this), there haven't been any experiments that required more compute than those shown on the plot.

The main thing shown here is that less than one doubling of computational resources for the largest training occurred in the 3-year period between 2018 and 2021, compared to around 10 doublings in the 3-year period between 2015 and 2018. This seems to correspond to a severe slowdown of computational scaling.

To stay on the trend line, we currently would need an experiment requiring roughly around 100 times more compute than GPT-3. Considering that GPT-3 may have costed between \$5M and \$12M and accelerators haven't vastly improved since then, such an experiment would now likely cost \$0.2B - \$1.5B.

Opinions on Interpretable Machine Learning and 70 Summaries of Recent Papers

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

Peter Hase

UNC Chapel Hill

Owen Shen

UC San Diego

With thanks to [Robert Kirk](#) and [Mohit Bansal](#) for helpful feedback on this post.

Introduction

Model interpretability was a bullet point in [Concrete Problems in AI Safety](#) (2016). Since then, interpretability has come to comprise entire research directions in [technical safety agendas](#) (2020); model transparency appears throughout [An overview of 11 proposals for building safe advanced AI](#) (2020); and explainable AI has a Twitter hashtag, [#XAI](#). (For more on how interpretability is relevant to AI safety, see [here](#) or [here](#).) Interpretability is now a very popular area of research. The interpretability area was the most popular in terms of [video views](#) at ACL last year. Model interpretability is now so mainstream there are [books](#) on the topic and [corporate services](#) promising it.

So what's the state of research on this topic? What does progress in interpretability look like, and are we making progress?

What is this post? This post summarizes **70** recent papers on model transparency, interpretability, and explainability, limited to a non-random subset of papers from the past 3 years or so. We also give opinions on several active areas of research, and collate another **90** papers that are not summarized.

How to read this post. If you want to see high-level opinions on several areas of interpretability research, just read the opinion section, which is organized according to our very ad-hoc set of topic areas. If you want to learn more about what work looks like in a particular area, you can read the summaries of papers in that area. For a quick glance at each area, **we highlight one standout paper per area**, so you can just check out that summary. If you want to see more work that has come out in an area, look at the non-summarized papers at the end of the post (organized with the same areas as the summarized papers).

We assume readers are familiar with basic aspects of interpretability research, i.e. the kinds of concepts in [The Mythos of Model Interpretability](#) and [Towards A Rigorous Science of Interpretable Machine Learning](#). **We recommend looking at either of these papers if you want a primer on interpretability.** We also assume that readers are familiar with older, foundational works like "[Why Should I Trust You?: Explaining the Predictions of Any Classifier.](#)"

Disclaimer: This post is written by a team of two people, and hence its breadth is limited and its content biased by our interests and backgrounds. A few of the summarized papers are our own. Please let us know if you think we've missed anything important that could improve the post.

Master List of Summarized Papers

- Theory and Opinion
 - [Explanation in Artificial Intelligence: Insights from the Social Sciences](#)
 - [Chris Olah's views on AGI safety](#)
 - [Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?](#)
 - [The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?](#)
 - [Aligning Faithful Interpretations with their Social Attribution](#)
- Evaluation
 - [Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction](#)
 - [Comparing Automatic and Human Evaluation of Local Explanations for Text Classification](#)
 - [Do explanations make VQA models more predictable to a human?](#)
 - [Sanity Checks for Saliency Maps](#)
 - [A Benchmark for Interpretability Methods in Deep Neural Networks](#)
 - [Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?](#)
 - [ERASER: A Benchmark to Evaluate Rationalized NLP Models](#)
 - [On quantitative aspects of model interpretability](#)
 - [Manipulating and Measuring Model Interpretability](#)
- Methods
 - Estimating Feature Importance
 - [Neuron Shapley: Discovering the Responsible Neurons](#)
 - [Anchors: High-Precision Model-Agnostic Explanations](#)
 - [Explaining a black-box using Deep Variational Information Bottleneck Approach](#)
 - [Weight of Evidence as a Basis for Human-Oriented Explanations](#)
 - [Interpretable Neural Predictions with Differentiable Binary Variables](#)
 - [Evaluations and Methods for Explanation through Robustness Analysis](#)
 - [Adversarial Infidelity Learning for Model Interpretation](#)
 - [CausaLM: Causal Model Explanation Through Counterfactual Language Models](#)
 - [Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers](#)
 - [How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking](#)
 - Interpreting Representations and Weights
 - [Translating Neuralese](#)
 - [Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors \(TCAV\)](#)
 - [The Building Blocks of Interpretability](#)
 - [Compositional Explanations of Neurons](#)
 - [LCA: Loss Change Allocation for Neural Network Training](#)

- Generating Counterfactuals and Recourse Procedures
 - [Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations](#)
 - [Counterfactual Visual Explanations](#)
 - [Explanation by Progressive Exaggeration](#)
 - [Counterfactual Explanations for Machine Learning on Multivariate Time Series Data](#)
- Explanation by Examples, Exemplars, and Prototypes
 - [This Looks Like That: Deep Learning for Interpretable Image Recognition](#)
 - [Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning](#)
 - [Interpretable Image Recognition with Hierarchical Prototypes](#)
 - [A Generic and Model-Agnostic Exemplar Synthetization Framework for Explainable AI](#)
- Finding Influential Training Data
 - [Understanding Black-box Predictions via Influence Functions](#)
 - [Estimating Train Data Influence By Tracking Gradient Descent](#)
- Natural Language Explanations
 - [Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks](#)
 - [Multimodal Explanations: Justifying Decisions and Pointing to the Evidence](#)
 - [Textual Explanations for Self-Driving Vehicles](#)
 - [e-SNLI: Natural Language Inference with Natural Language Explanations](#)
 - [Explain Yourself! Leveraging Language Models for Commonsense Reasoning](#)
 - [Towards Prediction Explainability through Sparse Communication](#)
 - [WT5?! Training Text-to-Text Models to Explain their Predictions](#)
 - [Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?](#)
- Developing More Easily Interpreted Models
 - [Human-in-the-Loop Interpretability Prior](#)
 - [Learning Certifiably Optimal Rule Lists for Categorical Data](#)
 - [Faithful and Customizable Explanations of Black Box Models](#)
 - [NBDT: Neural-Backed Decision Trees](#)
 - [Interpretable Learning-to-Rank with Generalized Additive Models](#)
 - [Obtaining Faithful Interpretations from Compositional Neural Networks](#)
- Robust and Adversarial Explanations
 - [“How do I fool you?”: Manipulating User Trust via Misleading Black Box Explanations](#)
 - [Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods](#)
 - [Analyzing the Interpretability Robustness of Self-Explaining Models](#)
 - [Robust and Stable Black Box Explanations](#)
 - [Interpretability is a Kind of Safety: An Interpreter-based Ensemble for Adversary Defense](#)
 - [Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations](#)
 - [Beyond Accuracy: Behavioral Testing of NLP Models with CheckList](#)
- Explaining RL Agents
 - [Explainable Reinforcement Learning Through a Causal Lens](#)

- [Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences](#)
- [Counterfactual States for Atari Agents via Generative Deep Learning](#)
- [Finding and Visualizing Weaknesses of Deep Reinforcement Learning Agents](#)
- [Towards Interpretable Reinforcement Learning Using Attention Augmented Agents](#)
- [Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning](#)
- [Understanding RL Vision](#)
- [Causal Analysis of Agent Behavior for AI Safety](#)
- Interpretability in Practice
 - [Explainable Machine Learning in Deployment](#)
 - [The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models](#)

Our Opinions by Area

- Theory and Opinion

- There has been good progress in the theory underpinning explainability. [Insights](#) from psychology, linguistics, and philosophy have helped authors understand slippery terms like "explanation" in the context of AI. This research has yielded new or clearer concepts to work with, e.g. simulatability, plausibility, (aligned) faithfulness, and (warranted) trust.

We have seen strong arguments for developing explainable AI with special focus on the social nature of explanations, involving the use of mental models, models of the self, and theory of mind.

Several papers have called for work on interpretability to be more strictly scientific, i.e. by asserting falsifiable hypotheses and measuring properties of explanation systems rather than proceeding solely with non-systematic qualitative analysis. These calls are made in response to trends essentially in the opposite direction.

Theoretical work in the area has heavily influenced some subsequent work in evaluation and methodology, though a substantial fraction of papers tend to re-introduce desiderata or key considerations relevant to their methodology rather than directly drawing on prior theoretical work. Some methods papers do explain their goals with the latest terminology but use the terms in only the loosest of senses. There is plenty of ongoing debate about the terms and concepts currently in use in the field and the ultimate purposes of interpretation and explanation methods.

- Evaluation

- There are now many, many ways to evaluate explanations, including procedures for evaluating explanations of arbitrary format. The approaches include both automatic procedures as well human study designs, and the bulk of the work has focused on feature importance estimates. We are excited by many of the approaches, particularly those assessing whether

explanations improve human-AI team performance at a task that is hard for either humans or AI alone.

One trouble here is that there are so many evaluation procedures, it can be hard for methods papers to choose which to use. It at least seems like each evaluation procedure is equally likely to be used in any given methods paper (though there is a noticeable preference for automatic methods over human studies). We imagine this trend arises partly from the following situation: (1) there is not a common understanding of which explanation procedures answer which research questions; (2) methodologies are introduced without sufficiently precise research questions in mind. (Alternatively, papers can truly need their own new evaluation schemes, because they are answering new questions.)

Here's an example of the above situation. There is a lot of confusion over what the actual object of our explanation is within the subarea of feature importance estimation. Several research questions present themselves: should feature importance estimates explain the role of features in (1) the behavior of particular trained model with fixed weights, or (2) the behavior of trained models obtained by a stochastic training procedure involving a model family and dataset, or (3) the solvability of a task, either in theory or with respect to a given training procedure, model family, and dataset? Each research question stems from a fundamentally different goal, but papers rarely distinguish between them. Do we want to learn about a given model, a family of models, or the true nature of a task? There is not yet a clear and commonly accepted set of evaluation procedures suited for each of these questions which papers on feature importance estimation can readily use. The result is that the literature is not nearly as cumulative as it could be. For any given research question, it is hard to find one-to-one comparisons between more than a couple papers which would help you tell which methods are well suited to answering the question.

A similar situation arises with evaluations comparing generated explanations to human explanations. Many papers compare model explanations to human "gold label" explanations for data points. This is an evaluation for plausibility but not faithfulness, and this practice may reward deceptive explanation methods (since this is rating explanations by how convincing they are, but not by how truthful they are). But you could also be comparing your explanations to human explanations to figure out if your model is reasoning in a similar way to how humans reason. In many cases, this is a goal for our models, so that seems good. (Of course, this assumes your explanations are actually faithful to the model's reasoning.) We see a lot of papers that use human explanations as their "gold label" to compare against while not even saying whether they aim to optimize for explanation plausibility or alignment between model and human reasoning.

Lastly, qualitative analysis remains very popular throughout methods papers. It would likely be a marginal improvement to the field if some standards for qualitative analysis were more widely adopted, or someone wrote something good about what those standards should be. We do not mind "expert evaluation (by the author)" of the kind where the authors carry out some systematic qualitative coding regarding their method

performance, but this quickly looks less like standard qualitative analysis and more like a measurable outcome.

- Methods (General Commentary)

- I'll give just three points here.

First, at a high level, there has been clear progress in methodology. There are cases where we can (1) identify concepts that certain neurons represent, (2) find feature subsets that account for most of a model's output, (3) find changes to data points that yield requested model predictions, (4) find training data that influences individual test time predictions, and (5) generate natural language explanations that are somewhat informative of model reasoning.

Second, it seems like every sub-field of ML has its own stream of research, and they often don't cite one another. There's a cluster of work in ICML/ICLR/NeurIPS, and a cluster in NLP conferences, and a clustering in CV conferences, and they often don't cite extremely similar methods or evaluation procedures from other subareas. Of course these days finding all the relevant papers to your work is a daunting problem in its own right, but the literature seems much less connected and cumulative than it should be.

Third, it is difficult to assess which other methods a paper should compare to when presenting its own new method. What if you say that a goal of your feature importance estimation method is "model debugging"? Though there are dozens of feature importance estimation methods you could compare to, you could also compare with counterfactual generation methods. Arguably every future paper with "model debugging" as a goal of their method should also compare to CheckList, a simple but extremely effective unit-testing approach, regardless of the category their method falls into. Yet this would require expensive (expert) user studies. It is much easier to use automatic feature importance evaluations and forget about other approaches to model debugging. Right now there is a serious trade-off between ease of evaluation and breadth of comparison for explanation methods, which is hampering comparison with relevant alternative approaches.

- Methods: Estimating Feature Importance

- Most of my thoughts on these methods are encapsulated in the examples in above opinion sections or represented in existing opinion papers. There are dozens of these methods, and they have a great variety of purported purposes.

- Methods: Interpreting Representations and Weights

- In this area we are most excited by approaches that find a map between vector representations and really clear-cut human concepts. A few examples in this section require additional human supervision over the concepts, but this seems like a worthwhile price to pay to know what the model is representing. Once we know the concepts being representing, we can work on understanding how the model uses them.

- Methods: Generating Counterfactuals and Recourse Procedures
 - For people who are not satisfied with how they are treated by a machine learning system, "recourse" is a *reasonable plan* that they could follow to change how the system handles them for the better. A common example is automated loan approval, where someone might dispute whether they were reasonably denied a loan by a ML system. This area of research feels particularly important both because providing recourse is good and because working with counterfactuals is key to building a good causal model of a machine learning system. Everyone interacting with machine learning systems in the future should hope that good recourse methods will be available (and put into practice by people using ML in the world — which is another concern). Meanwhile, counterfactual generation will be a core part of understanding model errors in complex domains. A key part of answering questions about what causes model behavior is being able to identify the right counterfactual situations and see how the model handles them.
- Methods: Explanation by Examples, Exemplars, and Prototypes
 - This area enjoys some popularity due to how people like explaining things with examples and reasoning by analogy to historical data. This seems like a good approach to consider alongside other methods, but we would like to see more systematic evaluations out of papers in this area.
- Methods: Finding Influential Training Data
 - I think this will be an increasingly valuable style of explanation, especially as training datasets grow rapidly in size and are collected with imperfect screening heuristics. These methods lay the basis for explaining, debugging, and debiasing model behavior, as they can be combined with approaches to making quick adjustments to trained models (to roll back the influence of bad training data, for example).
- Methods: Natural Language Explanations
 - In our opinion, we think this is a critical subarea for AI Safety. While we are making headway by means of visualizations of model reasoning, eventually natural language will be a complementary or preferable medium for communicating model reasoning. Beyond simply being a flexible medium for communication, methods for natural language explanation will set the ground for more interactive, dialogue-based interaction with systems for interpretability-related purposes. However, this area still has basic unsolved problems in methodology and evaluation which merit much more work. For instance, it is not fully clear how we would train models to truthfully reveal their internal reasoning to people via natural language, even in a particular task domain. An interesting related direction is to provide explanations in formal languages, rather than natural ones, which could offer some advantages over natural language (along with some notable trade-offs).
- Methods: Developing More Easily Interpreted Models
 - This is an interesting and potentially useful area of research. So far, it seems like training models on vast amounts of visual and linguistic data

leads to the models learning many crisp human-interpretable concepts (uncovered upon inspection). Is this a guaranteed property of training models on natural or human-generated data? How many of the concepts these models represent are not easily interpreted, for whatever reason, and can this situation be avoided?

This area also includes strong skeptics of explaining deep learning models, which can seem both fair when considering all the shortcomings of the explainability literature and unfair when considering all the clear progress we've seen the past few years.

Overall, we think this is a promising area, but we are also not convinced work from this area will ultimately influence how the highest-performing ML systems will be developed.

- Methods: Robust and Adversarial Explanations

- There are a few distinct things happening in this area. A few exceptional contributions involve work on (1) robustness to distribution shifts, i.e. explaining models in a way that explanations are still faithful even under distribution shifts, (2) deceptive models, where a specially trained model can deceive an explanation procedure into thinking it is not heavily biased with respect to certain features, when in fact it is, and (3) searching for data points that yield logically inconsistent model explanations under some explanation procedure, which is a pretty unsatisfactory state to end up in. These are all clearly important solutions/problems.

I've seen a few papers suggest that explanations for model predictions should be robust to small input perturbations *in principle*. This seems like a mistake, likely one of valuing plausibility over faithfulness. If model behavior is influenced by small perturbations, then explanations should describe this phenomenon, not smooth it over.

This area feels especially important going forward, and we would particularly like to see more work on how models might adversarially or accidentally hide information from being revealed by explanation procedures.

- Explaining RL Agents

- To date, most of the work here has been applying approaches from the Feature Importance Estimation and Counterfactual Generation literature, though there is also an interesting line of work focusing on causal models of agent behavior. Some interesting early results have emerged regarding the kinds of explanations that help users build better mental models of agents, but so far this area is so new that it remains to be seen what the most promising approaches are. Many of the concerns in the above Evaluation section also apply here.

I have come across surprisingly few papers in this area relative to its importance. There appear to be important questions unique to explaining agents (rather than classifiers). For instance, explaining agents' behaviors will require special consideration of actions' temporal dependence, agent "plans", and epistemic vs. instrumental rationality. And the whole exercise will be complicated by multi-agent scenarios. This area really merits a lot

more work, and for people interested in long term AI safety and existential risks, this is plausibly the most important subarea of interpretability research.

- Interpretability in Practice
 - Eventually interpretability techniques have to get used to make any difference in the world. This could require interpretability interfaces that are interesting to explore or corporate/public policy mandates for systems passing certain transparency tests. This section covers a bit of both. If interpretability techniques were so helpful that researchers actually relied on them in their day-to-day research for solving problems, that would be great. But it seems we are not there yet.

Paper Summaries

Theory and Opinion (5)

- **Section Highlight:** [Explanation in Artificial Intelligence: Insights from the Social Sciences](#)
 - 2018
 - This paper is a (long) survey of research on explanations coming from philosophy and the social sciences, and it communicates important results from these fields and comments on connections with XAI.

To begin with the terminology: drawing directly from past works, the author defines an explanation as "an answer to a *why-question*" and interpretability as "the degree to which an observer can understand the cause of a decision." They "equate 'interpretability' and 'explainability.'"

The paper communicates four key research findings:

(1) Explanations are contrastive and invoke counterfactual cases (even if implicitly). Most authors in the area argue that *all* why-questions ask for contrastive explanations, even if the "foil" to the fact of the situation is not explicitly stated. Questions such as "why did she close the door?" have an infinite number of possible implicit foils, and someone offering an explanation will implicitly rely on one as they explain the event (e.g., why did she close the door rather than leave it open?) One natural choice of foil is the more expected event, relative to the observed event.

(2) Explanations are given by agents after they are selected from a set of possible explanations, according to certain preferences. These preferences include (at least) coherence, simplicity, generality, soundness, and completeness — which are often in tension with one another. Following conversational norms, explanations are often restrained to be relevant to a particular question and only give what is necessary to answering the question. The abnormality of cited causal factors is an important aspect too, even more important than our confidence in the causal links. For example, when explaining the famous Challenger space shuttle explosion,

people would not reference the oxygen in the air as an explanation, even though it is more certainly a necessary condition for the explosion than the suspected cause of some faulty seals. Similarly, we typically do not seek explanations at all for unsurprising events.

(3) Probabilities are not as important as causes. An example: if a student wants to know why they got a 50% on a test, they will not be satisfied if told that most people in the class got around a 50%. It would be better to explain why most students got around a 50%, but it would be best to explain why that particular student got a 50% (i.e. what caused the grade). In general, giving statistical statements could be helpful, but it is critical to give the causal factors underlying the statistical generalizations.

(4) Explanations are social: the act of explaining is a transfer of information about an event's causal history between two people. Further, taking a conversational model of explanations helps us extend our understanding of conversation to the act of explanation, e.g. in applying Grice's maxims of communication and analyzing the role of linguistic markers which indicate perspective and manage listener expectations. Core to this view is the concept of theory of mind, since we now view explanations as communicative acts between an explainer and explainee designed to fill in gaps in the explainee's mental model. Implicature (leaving things as implied) is additionally of relevance. The conversational framework can be extended to what is more fundamentally an argumentative framework, where the explainer is arguing for the explainee to draw certain conclusions. The authors suggest this type of interactive format will be valuable in XAI regardless of the medium of communication.

The paper ultimately concludes that researchers in AI should better heed the work on explanations from adjacent fields, particularly philosophy, psychology, and cognitive science. If we leave it to almost exclusively computer scientists to design explainable AI, we risk the result of "the inmates running the asylum" — i.e., a situation where "the very experts who understand decision-making models the best are not in the right position to judge the usefulness of explanations to lay users."

The paper addresses a number of other interesting topics, including the possible *levels* of explanation (using Aristotle's four causes), the process by which people select or infer foils (e.g. by a notion of similarity in causal histories), and the relevance of folk psychology to AI (including the nature of social attribution of intentionality to AI systems), among others.

- Artificial Intelligence
- [Chris Olah's views on AGI safety](#)
 - 2019
 - Evan Hubinger summarizes Chris Olah's view on how interpretability can help AI safety. Olah gives four main reasons. The first is the benefit of post-hoc interpretability as a sort of "mulligan", allowing us to query models for issues prior to deployment. The second is that if interpretability helps us better understand how models work, then we can use these insights to better inform model construction and design from the get-go. The third is that interpretability can be used as an additional means of feedback during training, to see errors as they arise. The fourth is the view that an

interpretability-focused method of constructing AIs could lead to what he calls "microscope AI". Instead of the typical agent-based approach of taking actions in the world, Olah envisions microscope AIs as more like powerful knowledge generation tools where the actions are still taken by humans and not the model. Olah gives two points underlying this optimistic view of interpretability. One, good interpretability is possible even for very large models. Two, the current difficulty in interpreting large neural nets is due to model constraints; Olah believes that as models get larger, they will be able to express concepts in a less confused way. To improve interpretability research, Olah created Distill, an online journal that focuses on papers which clarify and interpret. He ends up acknowledging that interpretability research will likely also speed up capabilities research, but he is betting that gains from improved model design will win out over other automated approaches, which would be net-positive, even with the speedup.

- Alignment Forum
- [Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?](#)
 - 2020
 - The authors discuss two key concepts in the interpretability umbrella, *faithfulness* and **plausibility*. "**Plausibility* refers to how convincing the interpretation is to humans, while *faithfulness* refers to how accurately it reflects the true reasoning process of the model." Besides arguing that *faithfulness* is a preferable quality to *plausibility*, the authors make a few key points regarding current work on interpretability: (1) *Faithfulness* evaluations should not involve human ground-truth explanations. (2) Claims about "*inherent interpretability*" do not exempt methods from evaluation, and claims about interpretability should not be accepted until suitable evidence is provided. (3) In human-AI teams involving model explanations, increased team performance may not reflect explanations' *faithfulness* but instead a correlation between *plausibility* and model performance. They also tease out assumptions present in prior work, e.g. that "two models will make the same predictions if and only if they use the same reasoning process," which has been employed in proofs by contradiction against an explanation procedure's *faithfulness*. Finally, they urge a practice of interpretability research "that allows us the freedom to say when a method is sufficiently faithful to be useful in practice."
 - ACL
- [The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?](#)
 - 2020
 - The authors argue that we should stop using attention weights as explanations because information may mix within each axis of hidden states, meaning that attention weights do not necessarily represent reliance on the part of the input that corresponds to their index. In other words, the attention weight on a given token in a sequence will only represent reliance on that token's representation at the very first layer of a network, and not necessarily at any layer thereafter, because representations deeper in the network are a function of the entire input sequence. Other issues with attention, as observed by prior work, include that attention is not a causal explanation in the sense that attention weights cannot be altered while keeping all else equal about a model's input and forward pass. They argue for the use of saliency methods instead, which are generally designed with the goal of feature importance

estimation in mind. However, they note that "at least some of the saliency methods are not reliable and produce unintuitive results (Kindermans et al., 2017) or violate certain axioms (Sundararajan et al., 2017)." They also suggest that the feature space used for explanation, e.g. per-token representation, has so far been fundamentally limiting saliency methods, and they point to initial work on capturing feature interactions in explanations (Janizek et al., 2020).

- EMNLP BlackboxNLP
- [Aligning Faithful Interpretations with their Social Attribution](#)
 - 2020
 - This paper argues that the *faithfulness* condition for model interpretability is underdefined, reformalizes the notion of faithfulness, and proposes an explanation framework that better satisfies their redefined faithfulness. On the first front, they clarify that model explanations are best understood as *faithful* if they attribute the correct causal chain to the model's decision process (described as *causal attribution*). A commonly desired feature of explanations, they observe, is not represented in this notion of faithfulness: that the causal chain is *aligned* with what is expected by people (described as *social attribution*). These concepts are applied in a case study of select-predict methods for text classification, which are composed of a selector module that extracts subsets of a text to pass to a text classifier for a final prediction. They find that existing select-predict methods actually produce selections (which are masks over the text) that can themselves be used to predict the label, relying on the masks alone and not the selected tokens. They describe this phenomenon as a special case of the general "Trojan explanation," where the explanation encodes information that influences model behavior in a way not naturally anticipated by people. The selection masks are faithful explanations, in the sense that we understand the role they play in the model, but when masks are predictive of the label, the select-predict method is not *aligned*. Specifically, it can be unaligned in two ways: (1) people expect the selections to be *summaries* that keep relevant information for the task while filtering out irrelevant information, or (2) people expect selections to be *evidence* in support of a prior decision, without having been a part of the decision. The select-predict methods violate both expectations in the same way, namely by influencing the decision of the prediction module by selecting tokens that favor a certain class.

In response, the authors propose a select-predict-verify approach. They consider a special setting where a model makes a prediction using a full text, and people provide text highlights after the fact that they think should point the model to a counterfactual decision (distinct from the observed prediction). Their approach is to find the minimal selection that is a superset of the human's suggestion and leads the model to predict the human's expected class. This approach better satisfies the faithfulness and alignment conditions, while providing for specific downstream use cases: users can (1) check that when relying on the "correct" evidence, the model would reach the expected conclusion, (2) find what evidence would be needed to correct a model prediction, (3) explore whether people would interpret evidence similarly to the model, when they are uncertain about the true label and want to rely on the model for advice. Lastly, our interpretation of their discussion section is that the authors believe aligned

faithfulness is distinct from simulability because simulability can be high if people *learn* how explanations relate to model decisions, even when the relationship does not match initial human expectations, which is a case where explanations are not aligned.

- arxiv

Evaluation (9)

- **Section Highlight:** [Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction](#)
 - 2020
 - The authors run an RCT to see how different model explanation approaches can help with human in-the-loop prediction, as well as trust in the model. The prediction task is on the APP-REAL dataset which consists of over 7,000 face images and age labels. The experiment has two base conditions, one where users are asked to give an age prediction, and one where users are asked to give a prediction and are also shown the model's guess. The explanation groups were shown one of three explanations in addition to the model's output: a saliency map from the actual model (calculated with Integrated Gradients), a saliency map from a modified dataset with spurious correlations, and a random saliency map. Before collecting data, the authors ran a two-tailed power analysis using prior guesses on the dataset. The experiment also varied the framing, with the following three modifications: (1) Delayed Prediction, which asked for a user's guess, showed the model output, and asked for a revised user guess; (2) Empathetic, which described the model's output in a personified way; and (3) Show Top-3 Range, which output an age interval. The experiment was conducted on Amazon Mechanical Turk with 1,058 participants. Overall, participants were more accurate at guessing people's ages when they had access to the model's guesses, but having explanations of the model outputs did not further improve their accuracy. The authors note that this is likely because explanations had little effect on user trust in the model's outputs. The trust that participants had in each model differed only slightly between conditions, regardless of whether explanations were the real saliency maps or randomly generated (there is a slight trend but it is not statistically significant). In fact, participants found explanations to be "reasonable" even when they focused on the background and not on the face. The authors give quotes from participants explaining their reasoning processes. One participant, for example, noticed that explanations could appear faulty but thought the model's guesses seemed reasonable otherwise, so they "sort of went with it."
 - arxiv
- [Comparing Automatic and Human Evaluation of Local Explanations for Text Classification](#)
 - 2018
 - The author carries out both automatic evaluations and human forward simulation tests for explanation methods with text classifiers (a logistic regression and MLP model). Methods include LIME, word omission, and gradient-based saliency. The automatic evaluation measures how a model's class score declines as tokens selected as important by each explanation method are removed from an input, culminating in the "Area under the Perturbed Curve" (AOPC) (the class score differences are

- computed for removing $k = 1, \dots, 10$ words and then averaged). In the forward simulation test, workers are shown input movie reviews and asked to predict a model's predicted binary sentiment, while being shown explanations in the form of highlighted words in the text (without directional/class information). Word omission outperforms LIME and gradient saliency on AOPC; in the simulation task, gradient saliency achieves the highest simulation accuracy of 79% on one dataset, while word omission explanations yield 86.8% accuracy on another dataset.
- NAACL-HLT
 - [Do explanations make VQA models more predictable to a human?](#)
 - 2018
 - The paper presents a human subject experiment for evaluating the forward simulability of a model given various explanation methods, using a Visual Question Answering task. They consider two simulation targets: the model's binary correctness, and its particular predicted output. They evaluate explanation methods including Grad-CAM, visualized attention weights, and an "instantaneous feedback" condition where no explanation is included, but the simulation target is revealed to the human subject after every response. They find that the explanation procedures do not yield statistically significant improvements in accuracy, while the instantaneous feedback procedure yields large improvements (30 ppts simulation accuracy for predicting model outputs). Human performance on predicting the VQA model's correctness is not as high as an MLP trained to predict the VQA model's correctness using the VQA model's softmax layer's output as features (~80% accuracy), but the instantaneous feedback conditions are close, with around 75% failure prediction accuracy.
 - EMNLP
 - [Sanity Checks for Saliency Maps](#)
 - 2018
 - The authors propose two methods to validate saliency maps, an interpretability technique that visually highlights regions of the input that can be attributed to the output. The authors point out that a good saliency map should be sensitive to both the actual model and the input labels; changing either of these should lead to a different map. Eight different saliency map techniques are evaluated: the Vanilla Gradient, Gradient Input, Integrated Gradients, Guided BackProp, GradCAM, and SmoothGrad (plus two special cases). The authors run two experiments following their above conjecture. The first randomizes the last N layers' weights in the model, where N = 1 corresponds to only randomizing the last layer, and when N = model size, all weights are random. The reasoning here is that a good saliency map should be a function of the model, and not of just the input (e.g. acting like a model-agnostic edge detector). Comparison between the original saliency map and the new saliency map (on the randomized model) is done through visualizing both maps, as well as quantitatively via Spearman rank correlation, the structural similarity index measure, and the Pearson correlation of the histogram of gradients. In this first experiment, the authors find that the Vanilla Gradient is sensitive while Guided BackProp and Guided GradCAM show no change despite model degradation. The second experiment randomizes the labels of the input data and trains a new model. The reasoning is that saliency maps should also be sensitive to the true model; outlining a bird in the image, for example, is not useful if the true label is "dog". The model is trained to at least 95% training accuracy and then the saliency maps are applied.

Again, the Vanilla Gradient shows sensitivity. Integrated Gradients and Gradient Θ Input continue to highlight much of the same input structure. Both experiments were conducted on a variety of models and datasets, including Inception v3 trained on ImageNet, CNN on MNIST and Fashion MNIST, MLP trained on MNIST, and Inception v4 trained on Skeletal Radiograms.

- NeurIPS

- [A Benchmark for Interpretability Methods in Deep Neural Networks](#)

- 2019
- From Alignment Newsletter #101:

This paper presents an automatic benchmark for *feature importance* methods (otherwise known as saliency maps) called *RemOve And Retrain* (ROAR). The benchmark follows the following procedure:

1. Train an image classifier on a dataset (they use ResNet-50s on ImageNet, and get about 77% accuracy)
2. Measure the test-set accuracy at convergence
3. Using the feature importance method, find the most important features in the dataset, and remove them (by greying out the pixels)
4. Train another model on this new dataset, and measure the new test-set accuracy
5. The difference between the accuracy in (4) and in (2) is the measure of how effective the feature importance method is at finding important features

The idea behind retraining is that giving the original classifier images where many pixels have been greyed out will obviously result in lower accuracy, as they're out of the training distribution. Retraining solves this problem.

They benchmark a variety of feature importance methods (Gradient heatmap, Guided backprop, Integrated gradients, Classic SmoothGrad, SmoothGrad 2 , VarGrad) on their benchmark, and compare to a random baseline, and a Sobel Edge detector (a hard-coded algorithm for finding edges in images). Only SmoothGrad 2 and VarGrad (which are both methods which ensemble other feature importance methods) do better than random. They can't explain why these methods perform better than other methods. They also note that even when removing 90% of the pixels in every image (i.e. the random baseline), the accuracy only drops from 77% to 63%, which shows how correlated pixels in images are.

- NeurIPS

- [Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?](#)

- 2020

- From Alignment Newsletter #100:

In this paper the authors perform user tests on 5 different model agnostic interpretability methods: LIME, Anchor, Decision Boundary, Prototype Model and a Composite model (LIME Anchor and Decision Boundary). The use cases they test are a tabular dataset predicting income, and a movie-review dataset predicting sentiment of the review from a single sentence.

Their experimental setup consists of 2 tests: forward prediction and counterfactual prediction. In forward prediction, the user is shown 16 examples of inputs and corresponding outputs and explanations, and then must predict the model's output on new inputs (without the explanation, which often gives away the answer). In counterfactual prediction, after seeing 16 examples, the user is given an input-output-explanation triple, and then must predict how the output changes for a specific perturbation of the input.

Throughout the results they use a significance threshold of $p < 0.05$ (they don't use Bonferroni corrections). Their study has responses from 32 different students who'd taken at least 1 computer science course, with some screened out for outliers or low accuracy during training. There are approximately 200 individual predictions for each method/dataset-type combination, and each method/prediction-type combination.

Overall, their results show that only LIME (Local Interpretable Model-agnostic Explanation) helps improve performance with statistical significance on the tabular dataset across both prediction settings, and only the Prototype model in counterfactual prediction across both datasets. No other result was statistically significant. The improvement in accuracy for the statistically significant results is around 10% (from 70% to 80% in the Tabular dataset with LIME, and 63% to 73% for Prototype in counterfactual prediction).

They also showed that user's ratings of the explanation method didn't correlate in a statistically significant way with the improvement the model gave to their predictions**.**

- ACL
- [ERASER: A Benchmark to Evaluate Rationalized NLP Models](#)
 - 2020
 - This paper proposes a benchmark for *rationales* of text classification models, where a *rationale* is a binary mask on the input (i.e. a "highlight" of important words in the input). The benchmark is a collection of existing text datasets, most of which have human annotations for what the "important" words in data points are. They suggest measuring three aspects of model rationales: (1) their agreement with human rationales, (2) their *comprehensiveness*, defined as the change in a model output by *masking out* important words (relative to original input), and (3) their *sufficiency*, defined as the change in model output by *keeping* only the important words (relative to original input). They evaluate simple methods as baselines for future work.
 - ACL
- [On quantitative aspects of model interpretability](#)

- 2020
 - This paper proposes a few quantitative metrics for explanation methods, which they hope will be used for method development and selection before final testing via human studies. They give separate metrics for feature extraction, feature attribution, and example-based methods. The metrics themselves are not particularly novel: the feature extraction metrics focus on mutual information between the extracted features and the input or predicted class. With attribution methods, which assign scores to tokens, they test how the scores relate to model outputs under a variety of input ablation procedures (omitting certain features from the input). The metrics for example-based explanations are similar to some of those in the DiCE paper.

There are a few interesting points in the paper. They evaluate several explanation methods using a known non-linear function, so that we know the true effect of features on the function outputs, and find that the Integrated Gradients method makes some mistakes which simpler gradient-based saliency methods do not (for a single function and data point). Commenting on the Remove-and-Retrain (ROAR) procedure of Hooker et al., they suggest that this procedure might better be viewed as data interpretation rather than model interpretation, since the ROAR scores for an explanation method will be highly dependent on whether a task/dataset is still solvable after certain features are removed. And they give an interesting desideratum for feature attribution methods: that predictions should be more *precise* when a given feature is known, relative to not knowing that feature, and that precision should correlate with the magnitude of the attribution score. They take care to distinguish this desideratum from other metrics which measure how knowing a feature moves predicted probabilities up or down, rather than the precision or confidence in predicted probabilities.

- arxiv
- [Manipulating and Measuring Model Interpretability](#)
 - 2021
 - In a large (n=3800) pre-registered study with high reputation MTurkers, the authors explore how two factors, number of features in a model and model transparency, relate to three outcomes: simulatability, deviation, and error detection (to be explained below). To illustrate each aspect of the experiments, consider their experimental protocol:

Users are given eight features of an apartment for sale in New York City, and are asked to guess what a model will predict its selling price to be. The model is a linear model that uses either 2 or 8 of the 8 available features. Users are assigned into conditions using either the 2 or 8 feature model, and then further divided into a transparent condition where they see the model weights during the whole experiment or a blackbox condition where they never see the model weights. For each of the resulting four conditions, users are first asked to guess the model's prediction, and then they are shown the model's prediction and are asked to guess the true selling price of the apartment.

Comparing their guesses of the model's prediction and of the true price allow the authors to derive their three outcomes. Here, *simulatability* is measured as the user's accuracy at predicting the model output. *Deviation* is measured as the extent to which the user deviates from the model prediction in making their own guess at the apartment's true selling price. Lastly, *error detection* is measured for a particular subset of the apartments where the model prediction is overtly incorrect (the apartment features are outlying, in these cases). Users are said to detect the error when they deviate greatly in the direction of the correct value, and otherwise to not have detected the error.

Several results follow. First, users in the transparent conditions are better able to simulate the model than in the blackbox condition, and in both conditions it is easier to simulate the 2-feature model than the 8-feature one. Interestingly, users deviate from model predictions to the same degree in both transparent and blackbox conditions. And surprisingly, users detect model errors less frequently in the transparent condition than the blackbox condition. In further analysis of this last finding, they authors found that highlighting abnormal features of the apartments in the user interface effectively erased the difference in error correction between conditions, which they describe as "consistent with the idea that transparency can be overwhelming and cause users to overlook unusual cases."

- CHI

Methods

Estimating Feature Importance (10)

- **Section Highlight:** [Neuron Shapley: Discovering the Responsible Neurons](#)
 - 2020
 - From Alignment Newsletter #95:

This paper presents a novel method, Neuron Shapley, that uses the [Shapley value framework](#) to measure the importance of different neurons in determining an arbitrary metric of the neural net output. (Shapley values have been applied to machine learning before to [measure the importance of features to a model's output](#), but here the authors use them to calculate neuron importance.) Due to several novel approaches and optimizations in calculating these Shapley values, the top k most responsible neurons ($k \sim 30$) can be feasibly found for large networks such as Inception-v3.

The authors demonstrate that finding these neurons enables the performance of model surgery. Removing the top 30 neurons that contribute to accuracy completely destroys the accuracy, whereas in expectation removing 30 neurons at random from the network barely moves the accuracy at all. Since the method can be applied to an arbitrary metric, this kind of surgery can be performed for other metrics we care about. For example, removing the neurons which are most responsible for

vulnerability to adversarial attacks makes the network more robust, and removing the neurons most responsible for the class-accuracy imbalance (a fairness metric) makes the classes much more even, while only reducing the overall accuracy a small amount.

- NeurIPS
- [Anchors: High-Precision Model-Agnostic Explanations](#)
 - 2018
 - The authors introduce Anchors, which are if-then rules over inputs, as an alternative to local linear explanations, with the premise that local model behavior can be highly nonlinear. An example Anchor explanation for a model prediction of *positive sentiment* for a sentence "This movie is not bad" is given as a probabilistic statement, e.g.,

$p(y = \text{positive} | \{\text{bad, not}\} \in x) \geq .95$. Anchors are identified for particular model predictions by a PAC algorithm searching for rules over an input representation (like bag-of-words for text or pixels for images) that predict the observed model label with high confidence, using a local perturbation distribution around a particular input to get data for estimating rule accuracy. Relative to local approaches like LIME, the authors suggest that Anchors are easier to understand, have extremely clear coverage (whether they apply to the input or not), and are high precision by design (if an Anchor applies, confidence in predicted result is high). They present results for models for text and tabular classification, structured prediction (a part-of-speech tagging task), image classification, and visual question answering (VQA). Their first evaluation is to get LIME and Anchor explanations for validation data in tabular classification tasks, then automatically apply them to test data and check if their suggested predictions match the model's predictions. They find that the Anchor predictions do indeed match model predictions at high (>90%) rates, though Anchor coverage is relatively low, applying to less than <30% of test data. They propose a submodular pick (SP) algorithm to efficiently covering the space of inputs with Anchors, and find that they can cover over half of the tabular data spaces with 10 explanations (though data spaces are low-dimensional). In a human simulation test with ML students using tabular and VQA data, they find that showing explanations (for validation data) to users can improve user ability to predict model behavior on test data, relative to a baseline condition without explanations. Precision jumps from 50-60% to 90%+ across conditions; simultaneously, Anchor users also become more conservative, making predictions for between 3 and 40 percent fewer instances (lower perceived coverage). With LIME, the effect on precision is mixed and generally smaller.
 - AAAI
- [Explaining a black-box using Deep Variational Information Bottleneck Approach](#)
 - 2019
 - The paper introduces an explanation approach with the aim to select parts of an input that can be used to predict a blackbox model's output for the entire data point. The selection is made by their *explainer* model, which is optimized to trade-off between making selections informative of the label and keeping selection brief. These goals are formalized via mutual information. Since this formulation is intractable to optimize for directly,

they optimize for a variational bound for this objective, which looks like maximizing the likelihood of the blackbox model's predictions under an *approximator* model, while regularizing the size of the selections and encouraging sparsity. That is, the explainer masks the input and the approximator predicts the blackbox model's output given the masked input. The explainer and approximator are jointly trained; masks given by the explainer are made differentiable by means of the GumbelSoftmax estimator, which is used to select exactly k (hyperparameter) elements from the input (a continuous approximation of an n-choose-k sample). Note that mask elements they use after the GumbelSoftmax are still in the unit interval (. Lastly, note that the units of selection will include, for text data: words, groups of words, or sentences; and for images: squares of pixels.

Their quantitative evaluations include models of biological data, MNIST, and IMDB sentiment analysis data. With the biological data, they identify a simple heuristic based on explanations of model predictions that they use to either accept or reject model projections. It seems that they check whether explanations for test data "match" explanations from another dataset in particular situations, and if so, they accept the prediction; among accepted predictions, test accuracy is higher than rejected predictions.

For MNIST alone, they ask graduate students at CMU with a background in ML to rate explanations on a scale of 0 to 5, where 0 corresponds to "No explanation", the intermediate range to "Insufficient or redundant explanation" and 5 to "Concise explanation." Their method gets an average of 3.53 (next best: SmoothGrad, 3.45).

For IMDB alone, they do a simulation test with MTurkers, where they ask users to predict the model output given the explanation only (using only correctly predicted data). With their method, users achieve 44.7% accuracy (next best, L2X: 35.6%, random is 33%).

For both MNIST and IMDB, they measure "approximator fidelity" (approximator accuracy at predicting blackbox output) and "rationale fidelity" (approximator accuracy using hard masks rather than continuous masks). Using hard rather than continuous input masks does not greatly reduce the approximator's accuracy. They do not find any statistically significant gains in approximator fidelity over the most similar existing method, L2X, but they do tend to find increases in rationale fidelity of between 2 and 10 percentage points in most situations for both datasets.

- arxiv
- [Weight of Evidence as a Basis for Human-Oriented Explanations](#)
 - 2019
 - The authors examine how human explanations often focus on desiderata like contrastiveness (i.e. why X instead of Y?) which are often missing from existing interpretability approaches. They give a list of five desiderata (contrastive, modular and compositional, does not confound base rate with likelihood, exhaustive, and minimal) and then give an evaluation metric that satisfies all five. The authors describe the weight of evidence (WoE), which is defined as $\log \frac{p(e|h)}{p(e|\bar{h})}$ where e is the evidence observed and h and \bar{h}

are the two hypotheses under consideration. The authors use WoE in a meta-algorithm for multi-class explanation which iteratively finds the subset of classes with the greatest WoE, "explaining away" the other classes; this continues until only the predicted class is left. They utilize this algorithm on the Wisconsin Breast Cancer dataset as well as MNIST. The authors give some visual examples where their algorithm identifies key parts of the input, but do not quantify their results.

- NeurIPS Workshop on Human-Centric Machine Learning
- [Interpretable Neural Predictions with Differentiable Binary Variables](#)
 - 2019
 - In a select-then-predict format, the authors propose a masking model that restricts the parts of an input that a jointly trained text classifier (or regression model) has access to. The masking model parametrizes masking variables that take values in the closed unit interval, which are multiplied with the token input representations. The masking and task models are learned end to end via the reparameterization trick for a newly proposed random variable, the HardKuma. They use BiRNNs for the task model component; individual mask values are conditioned on the preceding mask values via an RNN. During training, a sparsity loss encourages masks to have expected sparsity close to a specified hyperparameter. In experiments on a multi-aspect sentiment regression and sentiment classification tasks, they find that (1) their random variable outperforms a Bernoulli used with RL for learning in terms of accuracy per number of selected tokens, and (2) their unmasked/selected tokens are included in human-provided important-word highlights 80-98% of the time. In an experiment using the HardKuma for masking cross-attention weights in a model for NLI, they attain 8.5% non-zero attention weights at a trade-off of losing 1 percentage point of task accuracy.
 - ACL
 - [Evaluations and Methods for Explanation through Robustness Analysis](#)
 - 2020
 - The authors propose a new definition of robustness which decomposes into two concepts: robustness in the space of features deemed important and robustness in the space of features deemed unimportant. Borrowing ideas from adversarial robustness, the authors propose a notion of feature importance based on how sensitive the label is to perturbations of the feature in question. They use a dual evaluation metric, which is to minimize the adversarial perturbation needed when applied to the relevant features and maximize the adversarial perturbation needed when applied to the irrelevant features. The authors propose two greedy methods to solve this feature importance problem. The first is to greedily select one feature at a time, locally improving the goal at every step. The second is to train a regression function based off of random subsets of features to learn feature interactions. This helps address the situation where a combination of features are much more informative than any one feature in isolation. Their method of feature importance is then evaluated on ImageNET, MNIST, and a Yahoo! Answers text dataset and compared to several other baseline methods: vanilla gradient, integrated gradient, leave-one-out, SHAP, and black-box meaningful perturbation (BBMP). The authors find that their method outperforms the other methods on their proposed dual metric. Qualitatively, the authors show that on the image datasets, their method produces attributions that are visually less noisy. They conclude with a sanity check by randomizing the model's last set of weights and confirm their method is sensitive to these changes.

- [Adversarial Infidelity Learning for Model Interpretation](#)
 - 2020
 - The authors introduce Model-agnostic Effective Efficient Direct (MEED), a new method for model-agnostic interpretability for instance-wise feature selection. MEED consists of an explainer, which uses a feature mask to select important features, an approximator (which uses the selected features to approximate the original model), and an adversarial approximator (which uses the inverse of the selected features to approximate the original model). The authors minimize the mutual information between the original model's output and the adversarial approximator's output. This model is evaluated on five datasets: IMDB sentiment, MNIST, Fashion MNIST, ImageNet for Gorilla vs Zebra, and mobile sensor data from a Tencent mobile game. The authors compare their method with LIME, kernel SHAP, CXPlain, INFD, L2X, and VIBI. They evaluate the feature importance by comparing the true model output with four different outputs: the output of the original model on only the selected features (and everything else set to 0), the output of an approximate model trained specifically on the selected features, the output of the original model on only the unselected features, and the output of an approximate model trained specifically on the unselected features. If the feature selection went well, then the true model output should tend to agree with the first two outputs and disagree with the last two. The authors also compare with human output, where users are given the selected features and asked to predict the output. Across all five tasks, the authors demonstrate that their feature selection method generally performs better than the other methods.
 - KDD
- [CausaLM: Causal Model Explanation Through Counterfactual Language Models](#)
 - 2020
 - This paper proposes to explain the causal effect of features on model outputs, particularly high-level features that are not easily manipulatable in the input. In cases where the feature of interest, such as the use of adjectives in a text, can be easily omitted *without influencing the remaining features in the input*, this causal effect can be easily estimated as the difference in model outputs with and without this feature. But when altering or removing features violates the data generating process, e.g. by destroying the grammaticality of a text while trying to remove information about its topic, the causal effect of this kind of high-level feature cannot easily be estimated through a counterfactual data generation scheme.

Lacking access to a generative model that can remove high-level features from text without influencing any other features, the authors propose to compare model outputs using the original data representations and counterfactual data representations. These counterfactual representations are obtained from a counterfactual model which is trained for the task at hand with an adversarial loss term encouraging it to forget the target concept (the one whose effect is being explained). This requires supervision at the instance or even feature-level. An example: to calculate the effect of the *adjective part of speech* on a text classifier, the authors train their counterfactual model to perform the task at hand while ensuring that the model's final representations do not contain any information about whether each word was an adjective or not (using word level part-of-

speech supervision). Then the causal effect of the adjective PoS on the model output is the difference between these model's outputs. The authors also make use of "control concepts" to check that this adversarial procedure is not removing other information that should be retained for the task at hand, though this requires even more supervision.

In order to compare against "ground truth" causal effects, the authors evaluate their method on tasks where the causal effect of a feature on a model is easily estimated by generating counterfactual data. For instance, they rely on sentiment analysis data where people's first names can be substituted with one another to change the likely gender or race represented in the example. Here, the ground truth effect on the model output is computed by checking the difference in model outputs for inputs with names substituted for one another. They also evaluate the effect of adjectives as described above, although one confusing thing about this evaluation is that they automatically remove adjectives from sentences in order to generate counterfactual data, but the adjectives themselves are left in the data when training their counterfactual model (only PoS information is eliminated).

Evaluations show that their method produces causal effect estimates close to the ground truth for each dataset, while baseline methods are highly inaccurate in some cases. Their baseline methods identify counterfactual points based on "passive observation" of features' presence/absence rather than causal intervention. They do not compare to any baselines based on perturbing data points, such as LIME. The authors suggest that [Iterative Null Space Projection](#) could serve as an alternative to the expensive counterfactual model training process.

- arxiv
- [Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers](#)
 - 2020
 - The paper proposes to learn text classifiers with a simple masking layer over tokens to reduce the number of tokens the classifier relies on, with the goal of making the model more amenable to applying interpretation methods. The masking layer is a lookup table that returns uncontextualized, token-specific masking probabilities; during training, binary per-token masks are sampled using these probabilities, and backpropagation is performed via the Gumbel-Softmax/Binary-Concrete estimator. Thus the work distinguishes itself from parametric masking layers, e.g. of Bastings (2019). The authors report accuracy improvements typically in the 0.5-1.0 point range on several text classification tasks, relative to unmasked baselines. In a test of local interpretability, they use LIME and SampleShapley to identify important words and find sometimes much higher AOPCs for their model than for baselines (AOPC: a measure of how class scores change when removing "important" tokens). In a test of global interpretability, they measure model robustness to applications of very sparse masks, by selecting globally important words for a task based on their look-up table probabilities in the masking layer. When a small number of words are selected (<10), their model predictions change with less frequency than a similar information-bottleneck approach (Schulz 2020).

- arxiv
- [How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking](#)
 - 2020
 - This paper introduces a masking method that (1) bottlenecks the information available to a pretrained task model, (2) masks inputs *and hidden states* to better understand how far information propagates into a model from the input. The objective they posit for the masking model is to find the sparsest mask that yields the same model output as the original model output (relaxed for gradient estimation purposes; Binary Concrete variable used for straight through estimation). A motivating example for their approach is as follows: suppose a model could detect when sequences of numbers have more 8s than 1s in them. How would existing feature importance methods allocate importance for a model prediction on a sequence with two 8s and one 1? When the authors experiment with this kind of problem, they find that other information-bottleneck-style approaches assign unequal importance to the two 8 digits or give importance to irrelevant digits. Moreover, an exact search for *the smallest input subset that results in the same model prediction* yields just a single 8 (dropping the other 8 and the 1), because this retains the "more 8s than 1s" property. The authors take issue with all of these explanations as they consider each to not reflect how a model must actually reason about sequences as it solves the task. Their proposal is to allow for masks to apply not just at the input but also in later layers. To see why this is useful, they first present a convincing analysis that in a simple two-module model for the toy task, one module decides whether digits are relevant (whether they're an 8 or a 1), and another module then counts and returns the result of the comparison. When they apply their masking approach to this model, the result is that the masks are not applied to the first module, where relevant digits have to be detected, but they are applied to the second module, whether non-8 and non-1 digits may be safely masked out without influencing the model prediction (because they will not be counted). In a general form of this toy task, where a model must decide if there are more m digits than n digits for a given an (m,n) query, they compare the difference in nats between normalized ground-truth importance attributions and those provided by prior approaches. The ground-truth importances attributes are uniform distributions over all the m and n digits in the sequence. This is the ground truth in the sense that each such digit contributes to the fact of whether there are more m than n digits, though not in the sense that each such digit is equally necessary to the binarized prediction of which digit is more numerous. Considering an input token to be masked when their masking model decides the task model's representation *at that token index should be masked at any hidden layer in the forward pass*, they find that their method essentially does not differ at all from these ground-truth attributions, while all others do by a margin. Next, they scale up to experiments with more complex models on SST (sentiment classification) and SQuAD (question answering). Primarily qualitative analysis follows to compare their explanations with those of past works.
 - EMNLP

Interpreting Representations and Weights (5)

- **Section Highlight:** [Translating Neuralese](#)

- 2017
- This paper proposes a method for translating vector communications between artificial agents into natural language. The motivation here is that, while Deep Communicating Policies (DCPs) are solving multi-agent communication games, it is difficult to analyze the content of agent messages. If we can translate the messages into semantically equivalent natural language statements, that would help us understand how the agents are behaving.

This translation is made possible by collecting data of humans playing the same communication games as the agents. One game is a simple reference game, where one player describes an image to a second player, and the second player must pick which of two images the first player is describing. The second game is a simplified driving game where two cars must pass through an intersection without colliding when they cannot observe one another directly. So the authors collect data of humans playing these games.

Translation is predicated on a particular notion of meaning. The authors choose to use a denotational perspective, suggesting that the meaning of a message is represented by the distribution over the speaker's state that is induced in the mind of the listener. This is instead of the pragmatic perspective, which would define meaning by the actions induced in the listener. They next propose a translation algorithm based on finding the minimum KL divergence between two "meaning distributions" induced by two communications. At a high level, the procedure measures the quality of a translation from a model's vector to some natural language utterance in terms of how frequently the utterance induces the same belief distribution over speaker states that the vector communication would (averaged across states and weighted by how likely it is the vector would be used in each state). Then, the best translation for a given message is the argmax of this quality measure (equivalent to the argmin distance between belief distributions induced by the messages).

Experiments show that this procedure helps models and humans communicate during gameplay and solve the games together at above random rates, even close to model-to-model gameplay without translation. This evaluation is automated by rolling out a human game trajectory and substituting in a model for one of the humans, while the "human" side of the game does not actually listen to the model's messages. To simplify the problem, human utterances in these games are treated as categorical over a set of simple words or phrases that are typically used in the games. Interestingly, the authors find that a "direct translation" baseline, which is simply a model of $p(\text{vector}|\text{utterance})$ trained on data from states with both human and model communications, does equally well in terms of allowing model-human pairs to complete the games, but this baseline is much less *denotationally* accurate, as they find that these communications do not lead listeners to form accurate beliefs about the speaker's state.

- ACL

- [Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors \(TCAV\)](#)
 - 2018
 - The authors propose a method for finding directions in a neural model's latent space that correspond to human concepts, like stripedness, and study the representations their method identifies for visual human concepts using GoogLeNet and Inception V3. The method itself is quite simple. First, the authors collect a small set of images depicting a concept, e.g. they collect images showing stripes or striped things (around 30 images in some experiments). Then, a linear model is trained to discriminate between the hidden states of the model when these images are passed through it and the hidden states obtained by passing *randomly chosen* images through the model. The randomly chosen images serve as a contrast set to the set representing the concept of interest. Finally, the Concept Activation vector is defined as the normal vector to the decision boundary, which points in the direction of the concept set's activations. Next, they define a score used for quantitative Testing with CAVs, (TCAV score). This score makes use of a directional derivative of a class score obtained for an image with respect to the CAV for a concept. Specifically, for a given class and a given CAV, the TCAV score is the proportion of images predicted as that class with a positive directional derivative with respect to the CAV. Hence the TCAV score ranges from 0 to 1, depending on what proportion of the images have positive directional derivatives. They intend this to capture how much a concept contributes to a model's prediction for a class (e.g., for images classified as a Zebra, how much did stripedness contribute?)

Two extensions of the CAV occur throughout the experiments: rather than just getting one CAV with one set of contrastive random images, they get 500 CAVs using 500 random batches of contrastive images. They omit experimental results for when these 500 CAVs do not yield statistically significant TCAV scores, as determined by a t-test with a null hypothesis of $TCAV=0.5$. Additionally, the contrast sets are sometimes not random images, but instead selected with the intention to better isolate the concept of interest. For example, the contrast set of the concept set of stripedness is the union of images representing dot and mesh textures.

Armed with the TCAV score, the authors perform a number of qualitative and quantitative experiments. They rank order images by their similarity to a CAV, and display the results. They make Deep Dream style visualizations by maximizing an image's hidden states's similarity to selected CAVs. They validate a number of expected associations between classes and concepts, as well as biases in the models' training data. In one experiment, they aim to validate TCAV scores against ground truth model reliance on a particular concept, by finetuning the models on a data subset where they control the presence of an easy-to-detect feature in the data (inpainted class names as text in the images). With these controlled datasets, they obtain models that either do or do not rely on the easy-to-detect feature in the images, as indicated by their accuracy on test sets without the feature. Then, they compare their ground-truth measure of model reliance on the feature with obtained TCAV scores for the feature CAVs. Showing results for two classes, they find a correlation between a model's reliance on the feature and TCAV

scores for the feature. Using these same controlled datasets, they also measure how MTurkers think of feature importance based on gradient-based saliency maps of image predictions. They find that the saliency maps often highlight the easy-to-detect feature even when it is not being used by the model, which seems to mislead the respondents into thinking it was important to the model. Lastly, they obtain domain expert feedback on using TCAVs for a model of diabetic retinopathy, suggesting that "TCAV may be useful for helping experts interpret and fix model errors when they disagree with model predictions."

- ICML
- [The Building Blocks of Interpretability](#)
 - 2018
 - Distill
 - The authors explain how composing visualization techniques for image recognition neural nets can lead to improved attribution for outputs/activations. They explain three ways of attributing activations; the neuron level, the spatial (i.e. pixel) level, and the channel (i.e. layer) level. These attributions can be combined to result in visualizations that can be traced through the network as a combination of previous layers or neurons, which have been mapped to 2-d images using feature visualization. This is a major step up from previous attribution visualizations, which were often only a heatmap on a specific layer. The authors show that matrix factorization on the flattened matrix of activations can compress the network's behavior and show a succinct visualization for each class or spatial point. Using GoogLeNet, they build a user interface that allows one to explore connections between layers, and suggest that visual interfaces of this kind may be built to help understand layers, atoms (i.e. groupings of neurons), and content (activation or attribution). They end with the caveat that visualization alone may be unreliable, but their compositional method between layers is likely to still hold because it's less subject to path dependence on the specifics of the input.
- [Compositional Explanations of Neurons](#)

- 2020
- From Alignment Newsletter #116:

Network dissection is an interpretability technique introduced in 2017, which uses a dataset of images with dense (i.e. pixel) labels of concepts, objects and textures. The method measures the areas of high activation of specific channels in a convolutional neural network, then compares these areas with the labelled areas in the dataset. If there's a high similarity for a particular channel (measured by the intersection divided by the union of the two areas), then we can say this channel is recognizing or responding to this human-interpretable concept.

This paper introduces an extension of this idea, where instead of just using the basic concepts (and matching areas in the dataset), they search through logical combinations of concepts (respectively areas) to try and find a compositional concept which matches the channel's activations. For example, a channel might respond to (water OR river) AND NOT blue. This

is still a concept humans can understand (bodies of water which aren't blue), but enables us to explain the behaviour of a larger number of neurons than in the original network dissection method. Their work also extends the method to natural language inference (NLI), and they interpret neurons in the penultimate layer of a BiLSTM-based network trained to know whether a sentence entails, contradicts, or is neutral with respect to another. Here they create their own features based on words, lexical similarity between the two sentences, and part-of-speech tags.

Using their method, they find that channels in image classifiers do learn compositional concepts that seem useful. Some of these concepts are semantically coherent (i.e. the example above), and some seem to have multiple unrelated concepts entangled together (i.e. operating room OR castle OR bathroom). In the NLI network, they see that many neurons seem to learn shallow heuristics based on bias in the dataset - i.e. the appearance of single words (like nobody) which are highly informative about the classification.

Finally, they use their method to create copy-paste adversarial examples (like in Activation Atlas (AN #49)). In the Places365 dataset (where the goal is to classify places), they can crudely add images which appear in compositional concepts aligned with highly contributing neurons, to make that neuron fire more, and hence change the classification. Some of these examples generalise across classifier architectures, implying a bias present in the dataset.

- arxiv
- [LCA: Loss Change Allocation for Neural Network Training](#)
 - 2021
 - From Alignment Newsletter #98:

This paper introduces the Loss Change Allocation (LCA) method. The method's purpose is to gain insight and understanding into the training process of deep neural networks. The method calculates an allocation of the change in overall loss (on the whole training set) between every parameter at each training iteration, which is iteratively refined until the approximation error is less than 1% overall. This loss change allocation can be either positive or negative; if it's negative, then the parameter is said to have helped training at that iteration, and if it's positive then the parameter hurt training. Given this measurement is per-parameter and per-iteration, it can be aggregated to per-layer LCA, or any other summation over parameters and training iterations.

The authors use the method to gain a number of insights into the training process of several small neural networks (trained on MNIST and CIFAR-10).

First, they validate that learning is very noisy, with on average only half of the parameters helping at each iteration. The distribution is heavier-tailed than a normal distribution, and is fairly symmetrical. However, parameters tend to alternate between helping and hurting, and each parameter only tends to help approximately 50% of the time.

Second, they look at the LCA aggregated per-layer, summed over the entire training process, and show that in the CIFAR ResNet model the first and last layers hurt overall (i.e. have positive LCA). In an attempt to remedy this and understand the causes, the authors try freezing these layers, or reducing their learning rate. The first layer can't be fixed (freezing makes its LCA 0, but later layers' LCA is increased in turn so the overall final loss stays the same). However, for the last layer, freezing or reducing the learning rate increases the overall performance of the network, as the last layer's LCA is decreased more than all the other layer's LCAs are increased. They also hypothesize that by reducing the momentum for the last layer, they can give it fresher information and make it more likely to learn. They find that this does work, though in this setting previous layers' LCA increases to compensate, leaving overall performance unchanged.

Finally, the authors show that learning seems to be synchronised across layers; layers get local LCA minima at the same training iterations, in a statistically significant way. They show this must be a combination of parameter motion and the gradient, as neither on their own explains this phenomenon.

- arxiv

Generating Counterfactuals and Recourse Procedures (4)

- **Section Highlight:** [Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations](#)
 - 2020
 - This paper introduces a method for generating Diverse Counterfactual Explanations (DiCE) of binary classification models of tabular data. The primary motivation is *recourse*, i.e. **giving feasible steps for how to achieve a desired outcome when a model makes an unsatisfactory decision about them. **Use cases include offering recourse for credit risk assessment and loan approval decisions. They run experiments with datasets for these tasks.

Given an original data point and a desired model output, their goal is to identify a set of data points that are (1) *valid*, meaning the model outputs the desired class, (2) *proximate*, meaning close to the original data point, (3) *diverse*, meaning they represent a variety of paths to achieving the desired outcome, (4) *sparse*, meaning few changes from the original data point are proposed, and (5) *feasible*, meaning they follow causal laws of the data generating process. To solve this problem, they formulate the first three objectives (validity, proximity, diversity) in differentiable terms, and perform a gradient-based optimization to identify the set of counterfactual points. Sparsity is encouraged in a post-processing step where, for a candidate counterfactual point obtained from the optimization, changes from the original point are greedily selected until the counterfactual achieves the desired model output. The causal feasibility is enforced in another simple filtering step, where candidates are filtered out based on violations of simple user-specified causal principles (e.g., if education increases, age must also increase; education cannot decrease).

For neural models of the COMPAS recidivism dataset, an income-prediction dataset, and the credit risk assessment and lending approval datasets, they find that their approach successfully identifies sets of counterfactuals that reasonably satisfy automatic metrics for validity, diversity, and proximity, and they present qualitative examples. They also aim to measure how these kinds of explanations, consisting of one original datapoint and a set of counterfactuals, can communicate a model's local decision boundary. As proxies for humans trying to reason about a local decision boundary given an explanation, they fit a simple nearest neighbor model to one explanation at a time and evaluate how that model classifies data in a radius around the original data point. That is, they sample datapoints in a sphere around the original data point, and compute the accuracy of a 1-nearest-neighbor model that has look-up access to only the data points in the explanation. In this set-up, they find the simple 1-NN model achieves up to a 44 F1 score with $k=4$ counterfactuals. They suggest that examples from their method can "approximate the local decision boundary at least as well as local explanation methods like LIME."

- ACM FAT
- [Counterfactual Visual Explanations](#)
 - 2019
 - In this paper, the authors propose a method of generating counterfactual explanations for image models. A counterfactual explanation in this framework is a part of the input image that, if changed, would lead to a different class prediction. The authors formalize the minimum-edit counterfactual problem which is defined to be the smallest number of replacements between an input I (which the model classifies as label A) and another input I' (which the model classifies as label B) such that the model will predict class B for the newly edited input I . The actual edit is done by permuting I' and then replacing a subset of I with values from the permuted I' . Because the space is so large to solve this problem exactly, the authors present two greedy relaxations of the problem. The first method is to iteratively look for the single edit which leads to the largest increase in log probability between the original and subsequent class predictions for class B. The second method is to, instead of taking a direct subset of I' values (which was done via the Hadamard product of a binary vector with I'), allow it to be a point on the simplex of a distribution over all features in I' . Then, both the permutation and the subset coefficients are learned via gradient descent. These explanations are used on four datasets: SHAPES, MNIST, Omniglot, and Caltech-UCSD Birds (CUB). In all four cases, the explanation is generated from the last layer of the CNN used. The authors evaluate the explanations qualitatively by examining which regions from I and I' are permuted to form the new counterfactual image. In the shown examples, the counterfactual images are constructed via appropriate portions of I' , for example a "1" from MNIST incorporating another spoke from a "4" to look more like it. The authors also evaluate the average number of edits needed to change the class label. The authors then used the counterfactual explanations from the CUB dataset to set up a training task where graduate students were tasked with learning how to classify images into one of two classes (which is not a trivial task). When participants got a choice wrong in the training phase, they were shown a counterfactual image. Their performance on the test phase was compared to two other baselines: students where were given no example (only

right/wrong) during training and students who were shown a GradCAM heatmap during training. The counterfactual image group had the highest accuracy, but this was not significant at the 90% confidence level against either baseline.

- ICML
- [Explanation by Progressive Exaggeration](#)
 - 2020
 - The authors "propose a method that explains the outcome of a classification black-box by gradually exaggerating the semantic effect of a given class." A resulting explanation is a series of altered images shifting from one class to another. Their method uses GANs as the underlying model for the generation of images; at each step, they make a change such that the model's probability of the desired class increases from the previous step. The authors run six experiments using two types of data: human faces and X-rays. Their evaluations include: qualitative analysis of explanations including identifying model biases / conflation of features, checking that statistics of altered images match those of real images receiving the same model output, and the effect on accuracy of corrupting pixels identified by their method as "important" to a class. They also run human studies where they see if MTurkers can identify the target attribute being explained based on the explanations; participant accuracy was from around 77% to 93% depending on the difficulty of the task.
 - ICLR
- [Counterfactual Explanations for Machine Learning on Multivariate Time Series Data](#)
 - 2020
 - This paper considers the problem of finding counterfactuals for multivariate time series data, and returns explanations of the form: if feature X was not decreasing over time, this sequence would not be classified as Y. They formulate the problem as optimizing a model's score for a selected class for a particular data point, while substituting out entire feature trajectories in the data point (entire rows in a $d \times t$ input), with substitutions being drawn from observed trajectories in the training data. They present an algorithm for this problem and evaluate their method using "three HPC system telemetry data sets and a motion classification data set." Their quantitative evaluations correspond with four principles: explanations should be (1) faithful to the original model, (2) comprehensible to human operators, (3) robust to small changes in the data point, and (4) generalizable to similar data points. They also qualitatively analyze explanations of correctly and incorrectly classified examples. For faithfulness in particular, they fit sparse logistic regression models, and evaluate their method against LIME and SHAP according to how well the "important" features selected by each method match the ground-truth used features in the logistic regression models. Both their method and SHAP obtain a precision of 1 and recalls of between .15 and .5, but they note that method usually returns 1-3 important features while SHAP usually returns over 100.
 - arxiv

Explanation by Examples, Exemplars, and Prototypes (4)

- **Section Highlight:** [This Looks Like That: Deep Learning for Interpretable Image Recognition](#)

- 2019
- The goal of this paper is to get image recognition models to imitate one way that humans can classify images, which is by comparing parts of a new image to prototypical parts of known images. The authors suggest that a model that can do this is "interpretable, in the sense that it has a transparent reasoning process when making predictions." They propose a model for this purpose with the following structure: A CNN maps an image to a representation of shape $H \times W \times D$, which can be thought of as a set of HW vectors in \mathbb{R}^D . For every class, there are 10 vectors in \mathbb{R}^D that are *prototype* vectors for that class. For each prototype, a prototype activation is obtained by a function of the distance between that prototype and *the nearest vector given by the CNN output*, such that nearer vectors yield larger activations. A logit for each class is computed by a weighted sum of the prototype activations. By enforcing that prototype vectors correspond to some vector representation obtained from an actual training data point, they attain a model that makes predictions based on similarities in representations between a current data point and training data. To interpret how a model predicts a given class, the authors interpret the activations between an image's CNN output and the 10 prototype vectors for that class. The model is trained end-to-end with a step that constrains prototype vectors to be equal to representations from training data points.

The model interpretations are visualized by localizing the part of the query image that highly activates each prototype, as well as the region in the prototype's training image that it represents. For a given prototype, heat maps over the query image are generated by taking the activation score of that prototype with all HW vectors in an image's CNN output, then upsampling the resulting activation grid to match the original image size. Finally, a bounding box is obtained by selecting the small rectangle containing the top 5% of upsampled activation scores (at the pixel level).

With models trained for fine-grained image classification using the CUB dataset, the authors present an abundance of qualitative analysis of the model, including examples of classified data points and analysis of the latent space structure. The common pattern in the presented figures is that the image parts that most activate a given prototype reflect similar characteristics of the image: a bird with red feathers on its belly activates a certain prototype, and when bounding boxes are overlaid on the original image and the image from which the prototype comes, the bounding boxes cover the red bellies on the new image and the image the prototype came from. Another prototype might capture the bird feet, another the bird's eye, etc., each showing a bird in the training data where the prototypical part representation comes from. The authors note that, across models, a maximum drop of 3.5 percentage points in accuracy is observed between the prototype model and a blackbox model trained with the same CNN feature extractor. This drop in accuracy can be mostly ameliorated by ensembling several prototype models (though this does increase the parameter count and number of prototypes per class).

- NeurIPS
- [Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning](#)
 - 2018
 - In this paper, the authors explain a new machine learning model, Deep K-Nearest Neighbors (DkNN). The DkNN model takes as input a trained neural net, a number of neighbors k , and an input. Each point in the training set has its intermediate layer-wise results (when passed through the neural net) recorded. Then, during evaluation, the DkNN uses locality sensitive hashing to find the set of k neighbors in each layer's latent space whose output is closest to the input. The authors then calculate the noncomformity of the (input, label) pair, where noncomformity is defined to be the number of values in the set of neighbors whose label does not agree with the output label. They also calculate an empirical distribution of noncomformity scores which are derived from a separate calibration set which comes from the same distribution of the training set. The DkNN then computes a probability for each class label defined to be the proportion of empirical noncomformity scores larger than the current label's. For a given input, the authors then define the model's confidence to be 1 minus the second largest class probability and the credibility to be the largest class probability. The authors evaluate their model on MNIST, the SVHN house number dataset, and the GTSRB street signs dataset. In all three tasks, k is set to 75. The authors show that the DkNN outputs a lower average credibility for the notMNIST dataset, showing that for out-of-distribution samples, their model is better calibrated than the naive softmax probability. On adversarial examples for all three datasets, the authors show that the average accuracy of the DkNN is higher than the normal DNN, across three types of attacks (Fast Gradient Sign Method, Basic Iterative Method, and Carlini-Wagner L2 attack).
 - arxiv
- [Interpretable Image Recognition with Hierarchical Prototypes](#)
 - 2019
 - This paper proposes to use a taxonomic organization of classes with a prototype-based vision model, so that explanations are given for every taxonomic level of classification. The paper also integrates existing novel class detection methods to work within the hierarchical class structure. As an example, the goal is that, when the only kinds of weapons a model has seen during training are rifles and assault rifles, the model could classify a handgun as a novel object, then classify it as a weapon on the basis of similarities between the handgun and weapon prototypes. The model interpretation is done by showing heat maps over images that represent which patches of an image would yield representations closest to class prototype representations. After training models on a subset of ImageNet with a hand-defined taxonomy, the analysis is primarily qualitative, focusing on case studies of prototype representations and novel class identification. One quantitative analysis of the latent space shows that the nearest neighbors of class prototype representations are also members of those prototype's classes about 80% of the time (though the model achieves similar accuracy as a blackbox model).
 - AAAI-HCOMP
- [A Generic and Model-Agnostic Exemplar Synthetization Framework for Explainable AI](#)
 - 2020

- In this paper, the authors propose a method of generating examples of different classes when given a black-box model. The method consists of access to a black-box model C and a generative model G. The authors use an evolutionary algorithm that allows them to efficiently generate exemplars to represent each class. At the start, the algorithm generates a population of t exemplars. At each iteration, the top k exemplars are kept, defined to be the top k inputs with the lowest squared difference between C's output and the desired class label y. Then, each exemplar is duplicated and mutated by adding a zero-centered Gaussian noise vector. The authors also add a momentum term which is a multiple of the previous vector added in the last iteration. This continues until the loss is under the preset threshold. The authors show that adding momentum leads to 19% faster convergence compared to the naive method. This exemplar method is tested on three datasets: the Adult Data Set for income prediction, Facial Expression Recognition 2013, and the Large Movie Review Dataset. The results are subjectively evaluated and the authors claim that their results are qualitatively more understandable than a direct gradient descent approach for creating exemplars.
- arxiv

Finding Influential Training Data (2)

- **Section Highlight:** [Understanding Black-box Predictions via Influence Functions](#)
 - 2017
 - The authors apply influence functions, a notion from robust statistics, to machine learning. An influence function asks how a model's parameters would change as we upweight specific training data. We can then use this to approximately answer the question of how the model changes if a specific example were not in the training set. The authors give stochastic solutions to solve for the influence function. As the influence function is only an approximation, the authors conducted actual leave-one-out training for the MNIST dataset and compared the resulting model parameters with the influence function results; they matched well. They show that even in a non-convex setting with a CNN, their influence function agrees well with the actual leave-one-out procedure. The authors also use influence functions to craft adversarial training examples, i.e. perturbed training examples that lead to misclassification of test data. They also provide a few additional uses for influence functions like identifying mislabeled training data.
 - ICML
- [Estimating Train Data Influence By Tracking Gradient Descent](#)
 - 2020
 - From Alignment Newsletter #97:

This paper presents the TrackIn method for tracking the influence of training datapoints on the loss on a test datapoint. The purpose of the method is to discover influential training points for decisions made on the testing set. This is defined (loosely) for a training point x and test point z as the total change in loss on z caused by training on x . They present several approximations and methods for calculating this quantity

efficiently, allowing them to scale their method to ResNet 50 models trained on ImageNet.

The standard method of evaluation for these kinds of methods is finding mislabelled examples in the training dataset. Mislabelled examples are likely to have a strong positive influence on their own loss (strong as they're outliers, and positive as they'll reduce their own loss). Sorting the training dataset in decreasing order of this self-influence, we should hence expect to see more mislabelled examples at the beginning of the list. We can measure what proportion of mislabelled examples is present in each different initial segments of the list. The authors perform this experiment on CIFAR, first training a model to convergence, and then mislabelling 10% of the training set as the next highest predicted class, and then retraining a new model on which TrackIn is run. *When compared to the two previous methods from the literature (Influence Functions and Representer Points), TrackIn recovers more than 80% of the mislabelled data in the first 20% of the ranking, whereas the other methods recover less than 50% at the same point. For all segments TrackIn does significantly better.*

They demonstrate the method on a variety of domains, including NLP tasks and vision tasks. The influential examples found seem reasonable, but there's no quantification of these results.

- arxiv

Natural Language Explanations (8)

- **Section Highlight:** [Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks](#)
 - 2020
 - The authors develop a virtual cooking game as a testbed for a proposed method for robot-human collaboration. The core of their approach is that, in addition to online planning of its own actions, the robot maintains a mental model of the human's current plan. Whenever the robot thinks the human's plan deviates from its own by a certain amount, the robot sends a message to the human including: (a) its own goal, (b) the action it will next take, (c) and the outcome of that action, as well as (d) its expectation of the human's goal, (e) the action it thinks the person will take, and (f) the outcome of that action. The message comes in the form of schematic natural language, with variable/action names substituted in. The plans themselves are given by parsing a Spatial, Temporal, and Causal And-or Graph (STC-AoG), which represents goals, subgoals, and atomic actions for achieving subgoals. The authors present an algorithm for inferring human plans that "uses observed user actions and communication history to infer human mental state."

They recruit study participants from their university subject pool for a test of the robot and cooking game. They consider three conditions: one with no communication, a "heuristic" condition with communication every 9.3 seconds (based on a pre-study's frequency of people asking for help), and one with mental-model-based communication. Participants completed their first task in about 75 seconds on average in the mental-model condition,

versus 150 in the control ($p < .05$ on difference; 125s in the heuristic condition, $p = .12$). On a 7 point "helpfulness" scale, the mental-model condition was rated around 5.5 on average, versus 4 in the control and 3.5 in the heuristic ($p < .05$ for comparisons). In participants' second round of the game, times converged across conditions, presumably due to familiarity with the game. The authors note that disparities in plans may arise from differing subgoals, or misunderstandings about action preconditions or effects (on the human's end).

- arxiv
- [Multimodal Explanations: Justifying Decisions and Pointing to the Evidence](#)
 - 2018
 - The authors collect human explanations of data points in two existing tasks, one for visual question answering (VQA-X) and one for visual activity recognition (ACT-X), and they propose multi-model explanation frameworks for performing both visual feature importance estimation and free form textual explanation generation. Textual explanations are generated by a neural model conditioning on the input, i.e. the image and for VQA the question, as well as the task model's predicted label, making the generations *rationalizing* explanations. Textual explanations are compared with the collected ground-truth explanations: BLEU scores with the ground-truth are 19.8 and 24.5 for the two datasets, while a human evaluation with MTurkers results in 38% and 45% of Turkers rating the generations as "better than" or "equivalent to" the ground-truth (for two datasets). "Important" image regions are compared with ground-truth human-annotated regions by the Earth Mover's Distance and a correlation statistic. Besides offering qualitative analysis, the last evaluation is a failure prediction experiment, where humans are shown explanations for data points (but not model predictions) and predict whether the model's prediction was correct. Here, humans get 70% and 80.5% accuracy, when random performance would yield 50%.
 - CVPR
- [Textual Explanations for Self-Driving Vehicles](#)
 - 2018
 - This paper collects textual descriptions and explanations of dashcam video of human driving, then proposes generative models of textual explanations for the behavior of a "driving" model. This driving model, or controller, uses a CNN to produce features for each video frame, then uses another neural module to output an accelerate and direction-change (learned with human accelerations and direction-changes). There are a few variants of generative models: *introspective* models condition on the visual features from the CNN, with spatial attention either "strongly" or "weakly" aligned with the controller's spatial attention (so they use roughly the same visual representations). A *rationalizing* model is free to attend over visual features as it pleases. All models condition explicitly on the controller outputs. In this sense they all rationalize in the usual use of the word, but *rationalize* is the least tied to the controller's internal states. BLEU and other metrics are used for an automatic evaluation: BLEU scores with human explanations are around 7 across conditions. (The models also generate descriptions of the controller actions; their BLEU is about 32.) The human evaluation for explanations is to ask MTurkers if they are "correct" and take a majority vote among three Turkers: 62-66% of the explanations

are "correct." (Descriptions of controller actions are rated correct 90-94% of the time.) There are no statistical tests for the differences in conditions.

- ECCV
- [e-SNLI: Natural Language Inference with Natural Language Explanations](#)
 - 2018
 - The authors get humans to annotate the popular SNLI dataset with natural language explanations of why each data point should have the label it has. The result is about 570,000 human explanations. They train an LSTM-based model to both perform NLI and generate explanations for its outputs. Explanations are generated conditional on the input representation for the task and the output label given by the model. A subset of explanations for correct predictions is evaluated manually by the authors according to their "correctness," i.e. whether or not they stated the "arguments" that made a data point's label the true label. Their best model produced correct explanations 64% of the time at a cost of 2.3 ppts lower accuracy relative to a non-explaining baseline.
 - NeurIPS
- [Explain Yourself! Leveraging Language Models for Commonsense Reasoning](#)
 - 2019
 - In this paper, human-annotated explanations are collected for the Commonsense Question Answering (CQA) dataset, which is a multiple-choice task with 7610 train and 950 test points (for v1.0). The annotations include text highlights of important parts of the question, and natural language explanations of why the correct answer choice is correct. The authors propose two modeling procedures for generating explanations: a *reasoning* procedure, that generates explanations from questions and answer sets, and a *rationalizing* procedure, which also conditions on a label (the ground-truth labels during training, and predicted labels at test time). Using a fine-tuned GPT as the generator yielded a BLEU score of 4.1 in the reasoning condition. An approach termed CAGE appends a BERT model to the GPT-reasoning pipeline to predict task outputs conditioned on inputs and generated explanations. The authors report accuracy gains from this pipeline over a BERT baseline (though a [similar approach](#) applied to the larger e-SNLI dataset does not yield any changes in accuracy). Their human evaluation is to ask MTurkers to "guess the most appropriate answer choice based on only the explanation without the question." They find that BERT outputs are recoverable from the GPT explanations 42% of the time (random: 33%), while ground truth labels are recoverable from human explanations 52% of the time.
 - ACL
- [Towards Prediction Explainability through Sparse Communication](#)
 - 2020
 - This paper assesses *extractive* explanations for models of textual data under a simulability perspective, and they present explanation generation in terms of communication between an explainer and a listener (Layperson, as the paper puts it). The purpose of the explanation is to simply encode the model's output. An extractive explanation is a set of words pulled out of an input, and the procedures for generating these explanations follow a general format: rank words, then pull out the top k . What they are evaluating is the success of a procedure at encoding a label into a bag of words from a particular input such that a listener can extract the label from the code.

The ranking methods they consider include a classification model's attention weights (using various kinds of attention), a gradient-based saliency ranking, and a standard word omission procedure. The listener is either a human or a trained BoW model. They also consider jointly training the BoW listener along with an explanation model, $E(x, \hat{y}, h)$, which extracts words from the input conditioned on the classifier's output and its final hidden state. Their automatic evaluation is to compute the listener BoW model's accuracy at predicting a classifier's output (Communication Success Rate, CSR). Their human evaluation is to use people as explainers, listeners, or both. With IMDB sentiment and SNLI natural language inference data, they evaluate CSR with all four combinations of human and machine listeners and speakers. They find that jointly trained machine explainers and listeners complete the task with 99%+ accuracy for both datasets. There is little to no statistical difference among the various attention-based top-k methods with human listeners: on sentiment analysis, CSR ranges from 87.5% to 93.25% (random: 50%); on NLI, it ranges from 70.5 to 74.5% (random: 50% — no neutral label). Humans successfully communicate the label 86.5% of the time on NLI.

- arxiv
- [WT5?! Training Text-to-Text Models to Explain their Predictions](#)
 - 2020
 - The authors train the 11 billion parameter T5 model in a multi-task framework to do a task and generate natural language explanations for its answers on the task. The explanations are either free form (abstractive) generations or important words from the model input (extractive). Learning is entirely supervised using human-provided explanations (either free form explanations or text highlights). Experiments for the open-ended explanations are conducted with e-SNLI and CoS-e datasets (see papers above). For extractive explanations, a sentiment analysis task (Movie Reviews) and a passage comprehension task (MultiRC) are used. An evaluation for plausibility is done via BLEU for open-ended and F1 score for extractive explanations, and they far exceed the previous SOTA. The human evaluation is to show MTurkers predicted data points with explanations and ask them, "Does the explanation adequately explain the answer to the question" (for CQA) or the same question with task-appropriate wording for other datasets. After taking the majority vote of 5 Turkers for 100 data points, they report "correct"-ness rates for model-generated explanations and the ground-truth human explanations in each dataset (random is 50% for each): the model gets 90% for NLI (humans: 78%), 30% for CQA (humans: 16%), 94% for sentiment analysis (humans: 99%), and 50% for MultiRC (humans: 51%). The authors write: "To summarize, our results suggest that WT5-11B is at a human or super-human level at both classifying and explaining examples from the datasets we considered." The authors conduct additional analysis concerning out-of-domain and out-of-task transfer, as well as sample efficiency (only having 100 e-SNLI explanations will get you 28 BLEU with T5-11B; 50k gets you the full-scale result of ~34).
 - arxiv
- [Leakage-Adjusted Simulability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?](#)

- 2020
 - The key question the paper addresses is how to evaluate natural language explanations generated by a model. Past works have done so primarily by training a model for particular tasks, using human explanations for data points as supervision for explanation generation, then comparing explanations generated for the model decisions to the “ground-truth” explanations of the data point labels provided by humans. The authors argue that this is problematic because it suggests that explanations are evaluated according to their plausibility rather than faithfulness to the model’s internal reasoning. In effect, past works rated methods highly when they gave explanations of their behavior that sounded good, even if they did not actually reflect the true reasons that led the model to its decision. To resolve this shortcoming, they present a procedure for automatically measuring how well explanations would let an “observer” predict a model’s behavior (in a similar spirit to other work on model simulatability). For this procedure to capture the meaning in explanations and avoid rewarding trivial explanations that just restate the model behavior (e.g. “I picked answer A because it was the best choice”), a causal inference method is used to control for the influence of explanation “triviality” on the observer. In their experiments, the observer is a model proxy for a human.

Using two existing text datasets with human explanations, eSNLI and CoSE, they evaluate existing methods and newly presented approaches, finding that some methods may not produce helpful explanations on average, while the most successful explanations come from a model that first generates hypothetical explanations for every answer choice, then makes predictions based on the input and all hypothetical explanations (and selects the final explanation based on the prediction). Further, since a metric for explanation quality is proposed, they also carry out experiments where models are optimized for this metric directly (or, rather, a proxy for this metric). These experiments are interpreted as multi-agent communication games played in natural language, and they find that in some settings this can improve the explanations’ effect on model simulatability.

- Findings of EMNLP

Developing More Easily Interpreted Models (6)

- **Section Highlight:** [Human-in-the-Loop Interpretability Prior](#)
 - 2018
 - This paper proposes an algorithm that minimizes the number of user studies needed to identify a model that is both accurate and interpretable according to some study-based criterion for interpretability. In their terminology, they query a human “prior” on model interpretability, $p(M)$, via a single user study, and their algorithm solves for the MAP solution,

$\max_{M \in M} p(X|M)p(M)$, by using sequential user studies to search over models in M .

The procedure is as follows: First, they obtain a set of high likelihood models, i.e. models that explain the data well. Based on the view that model users might want a model to meet a minimum accuracy threshold, but not care much about exceeding this threshold, they define a likelihood function that sharply increases after a model meets an accuracy threshold. In practice, they have no preference between models as long as they meet the accuracy threshold. Second, they perform a user study that yields a single human-interpretability-score (HIS) for a given model and subset of data. They use the domain-general notion of simulability in their studies, computed here as the *mean response time (RT)*, which is the time it takes a user to predict a model output for a given input when relying on a summary/explanation of the model (elaborated on in next paragraph). For a higher HIS scores to represent more interpretable models, they compute the final HIS as the difference between a maximum allowed response time and the actual response time. The actual optimization occurs by an upper-confidence-bound based search over models, with UCBs given by a

Gaussian Process over M . The kernel on models is the RBF kernel over feature importance estimates obtained for each model (by gradient-based feature importance for NNs, and some other procedure for decision trees). Starting with an initial model from the set of acceptably accurate models, the next user study is conducted using the model with the highest UCB estimate for its HIS.

The summary/explanation of a model is, for decision trees, given to users as a literal visualization of the tree. For neural networks, the summary/explanation is given by a locally estimated decision tree for each study data point. Similar to LIME, they sample from a perturbation around a data point and fit a decision tree to the NN's outputs on that perturbation distribution.

In their study, they run this algorithm for 10 iterations (ending with the 11th model), using graduate ML students and four tabular datasets (one synthetic). Before using graduate students, they ran a study with MTurkers, but found that the results were too noisy. They train decision trees for the three simpler datasets and use a neural network for the most complex dataset. They also compare to models obtained by optimizing for four easily computable heuristics for interpretability (in decision trees): number of nodes, mean path length, number of used features, and mean number of features in a path.

They find that: (1) while optimizing for a given interpretability heuristic (like number of decision tree nodes) often produces models that perform poorly under another interpretability heuristic, optimizing for their response-time HIS score produces models that tend to do well across interpretability heuristics, (2) under each of the four interpretability heuristics, their optimization algorithm finds more interpretable models more quickly than a random search over models, (3) their optimization does seem to increase HIS scores over time, which is the actual objective

in the optimization. Regarding this last point, note that they "did not see statistically significant differences in our results," though their experiments may have been underpowered since they ran with 16 subjects divided across conditions.

- NeurIPS
- [Learning Certifiably Optimal Rule Lists for Categorical Data](#)
 - 2018
 - The paper introduces an approach called CORELS that finds rule lists for data with a categorical feature space, where the rule lists are guaranteed to be optimal under the learning objective. Here, a rule list is a list of if-else statements (ending in just an if statement) of the form "if feature J == category C, predict Y, else..." and which serves as a classifier. The objective function is the empirical risk, with regularization on the length of the rule list (i.e. number of rules). The regularization penalty is directly interpretable as the trade-off between gaining p% of model accuracy at the cost of adding p rules to the model. The model is identified through a proposed branch-and-bound algorithm, which relies on a number of key observations that drastically reduce the size of the space of rule lists that needs to be searched.
 - Experiments involve several publicly available datasets for high-stakes tasks, including recidivism prediction and weapon possession in stop-and-frisk searches. Across tasks, the data points include 3-7 categorical attributes and up to 28 binary features.
 - The learned rule lists are 4 or 5 rules long, meaning it is very easy to read the entire rule list and see how it will handle every data point. The authors observe that on the recidivism data, their approach achieves equal accuracy to a proprietary, blackbox "prediction tool" (COMPAS) used for recidivism prediction in some places in the US legal system.
 - Lastly, the authors note that the search algorithm may struggle with very high dimensional data where many possibly relevant features are highly correlated. They also suggest that, if desired, predicted probabilities can be obtained by taking the empirical probability of the predicted outcome for a given rule, and they remind the reader that their approach is not to be used naively for causal inference.
- JMLR
- [Faithful and Customizable Explanations of Black Box Models](#)
 - 2019
 - The authors give a global interpretability method for black-box machine learning models which approximates a model with a two-level decision set. This is a model that separates the inputs through feature predicates (e.g. is age > 30?) and then another set of predicates (hence the two-level structure) for assigning a class label. The authors construct an optimization objective that accounts for the two-level decision set's fidelity (agreement with the original model), unambiguity (lack of overlapping predicates), and interpretability (the number of predicates used). They use an optimization method based on approximate local search to solve this objective. This

method has the benefit of allowing user input over which predicates they wish to use. The method is evaluated on a bail outcome dataset with 86k people, a student outcomes dataset with 21k people, and a depression diagnosis dataset with 33k people. The authors compare their model with other interpretable approximation methods like LIME, Interpretable Decision Sets (IDS), and Bayesian Decision Lists (BDL). The authors examine the fidelity interpretability trade-off. Out of all methods tested, the authors find that MUSE performs the best in terms of fidelity per number of rules and fidelity per average number of predicates. In a 33 participant user study, the authors find that when given the approximate model and asked questions about how the model would respond to a particular input, the MUSE model led to better accuracy and lower response time, when compared to IDS and BDL.

- AIES
- [NBDT: Neural-Backed Decision Trees](#)
 - 2020
 - The authors show how to modify a neural net into a more interpretable model by converting the last layer's weights into a decision tree, where each leaf node corresponds to a row of the weight matrix, and higher up nodes are averages of the nodes below. They add an additional regularization term during training for the cross-entropy loss between the corresponding node and the correct label. This is intended to improve the separation of representations for each node so each leaf node can be associated with a class. The overall model is evaluated by running the input through the neural net until the penultimate layer, whereupon it is then fed into the decision tree. The authors use this method of construction on a variety of models (ResNet, WideResNet, EfficientNet) on a variety of datasets (CIFAR10, CIFAR100, TinyImageNet, ImageNet). On all tasks, the NBDT is competitive, scoring within 1% of state-of-the-art performance. The interpretability of the NBDT's nodes is developed through either of two post-hoc iterative processes. The first approach involves first hypothesizing a category a node corresponds to, and then visually checking this hypothesis with a "representative" sample, defined to be a data point with an embedding similar to the node's. The second approach involves postulating categories for the nodes, and then training on a held-out dataset; these data points are then checked to see if they are passed to the child whose category is most fitting for them.
 - arxiv
- [Interpretable Learning-to-Rank with Generalized Additive Models](#)
 - 2020
 - The authors propose to use a generalized additive model for a learning-to-rank task and make a few arguments that this kind of model is more interpretable than blackbox models for the task. A GAM takes the form
$$f(x_i) = \sum_{d=1}^D f_d(x_{id}),$$
where a feature-specific function is applied to each feature of a data point separately. The model trades off being able to model feature interactions with the ease of understanding a particular feature's contribution to the final output: since each function is univariate, they can simply be plotted across the range of the input feature. In a learning-to-rank task, a given model input consists of a set of data points

$\{x_i\}_{i=1}^n$ along with some general context information for the set, q_i . The authors extend the standard GAM to operate in the ranking setting by weighting the contribution of each feature with a weight obtained by another GAM on q_i , as such: $f(x_i) = \sum_{d=1}^D g_d(q_i) f_d(x_{id})$ where g_d is the d^{th} element of a vector-valued GAM g . Each f_d is a separate neural network, as is g .

They compare their model with an existing tree-based GAM, and they compare across choice of loss function, between MSE and the ranking loss NDCG. They run experiments with three tabular datasets. In terms of NDCG performance, their neural GAM with a ranking loss performs the best, by between 0.3 and 4 points.

To evaluate interpretability, they consider three approaches. First, they check that their individual feature models f_d seem to represent feature importance in the same sense that a standard feature-shuffling based approach (measured by change in NDCG from shuffling a feature column in the data). Specifically, they compute the difference in 5th and 95th percentile f_d values (for corresponding two data points) as the "range" of the feature function, then check the correlation between these range values and the feature importance values obtained by shuffling. They find that using the ranking loss is important to ensure that there is a correlation between the range and feature-importance at all (but do not compare tree and neural GAMs). Second, they plot individual feature functions for the tree GAM and neural GAM. With both models, you can see how a feature contributes to the overall model output. Third, they plot a heatmap of g_d values across a one-hot representation of a categorical feature, to see how, in this case, a *region/country* context feature yields weights for several x features; they observe some interesting structure in which some regions yield similar weights for certain features.

In two final experiments, they show how distilling the submodels into piece-wise linear models after training results in a 20x speed-up to inference at a small (1 point) cost to the objective. And they find that building a blackbox neural net on top of GAM features yields better NDCG by about 1 point over simply training a neural net from scratch. Their neural GAM performs about 2 points worse than the neural net trained from scratch.

- arxiv
- [Obtaining Faithful Interpretations from Compositional Neural Networks](#)
- 2020

- This paper proposes to use neural module networks (NMNs) for purposes of increasing model interpretability. An NMN works by first *parsing* an input into a program, which is a series of functions to compute, then *executing* this program. For example, when a model must decide if a caption matches an image, an NMN might parse caption into the program "1. find dogs, 2. filter for black dogs, 3. check if number of entities from (1) and (2) are equal." These steps are executed by neural modules, and existing architectures and training procedures allow for NMNs to be trained end-to-end with supervision only at their final output.

NMNs could improve model interpretability over standard blackbox models by virtue of their modules executing human-interpretable functions, like finding, filtering, counting, comparing, etc. But it's a known result that training NMNs end-to-end leads to module "responsibilities" being diffusely distributed across several models, with some modules performing unintended functions. The authors confirm this finding for a visual+textual reasoning task (similar to the image captioning described above) and a purely textual reasoning task.

Principally, the authors study how to improve NMN "faithfulness," which they define as a property obtaining of models whose modules perform the roles they are intended for. They do so by providing programs to models (either gold programs or heuristically obtained programs) and collecting labels from people for what intermediate module outputs should be. To measure faithfulness, they measure the discrepancy between the module outputs and ground-truth outputs, for the provided programs.

They identify a few ways to improve faithfulness, usually at the expense of model accuracy. On visual+textual reasoning: By specifying modules exactly (i.e. simply summing the module inputs for a 'sum' module instead of fitting free parameters), they improve faithfulness slightly at the cost of 3 points of accuracy. Using inflexible, few-parameter modules further improves faithfulness over manual specification, at a smaller cost to accuracy. On textual reasoning, they find ways to decrease faithfulness. By training without two helpful kinds of modules (sorting and comparison modules), and thereby offloading these responsibilities onto other modules, they lower model faithfulness. And they observe that adding blackbox modules which can perform multiple "steps" at once can decrease overall model faithfulness.

For both reasoning tasks, they find that training on gold labels for module outputs greatly improves faithfulness. Interestingly, in one setting, this is actually pretraining on another dataset that has labels for module outputs rather than the dataset they evaluate with. The authors are hopeful that with careful architecture design and additional supervision, NMNs can provide more interpretability than standard models with a low cost to performance.

- ACL

Robust and Adversarial Explanations (6)

- **Section Highlight:** ["How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations](#)

- 2020
- This paper shows that explanations can be generated for a blackbox bail recommendation system that lead law school students to be willing to deploy the model in practice, even when that model actually relies on race and gender as features and the students previously stated these features were unacceptable to use. Bail recommendation is essentially a prediction problem — by law, a defendant is released on bail only if a judge believes they will not flee or commit another crime — and there is historical data on outcomes for defendants who were released on bail. In this context, an "explanation" is a (simple) model of a blackbox model that predicts whether a defendant will flee or commit another crime. They use two-level decision sets obtained from MUSE as the explanations models, because they are highly intuitively interpretable. These are sets of if-then rules that can be used by a person to classify instances (see paper for 7 "interpretability" factors MUSE incorporates). The students unanimously suggested that race and gender were the unacceptable factors to use in a bail recommendation system, and mostly agreed that prior incarcerations and failures to appear before court were two features that should be relied on. So, the authors designed a blackbox model that relied on the prohibited features and made no use of the desired features — in fact, the model was secretly a two-level decision set that the authors fit, which achieved 83.28% accuracy on test data from a dataset of 32K examples in total. The authors then fit 3 explanation models, relying on (1) neither prohibited nor desired features, (2) both prohibited and desired features, (3) desired features but not prohibited features. The explanation models all approximate the blackbox model well, obtaining test accuracies upwards of 97%.

The user study design is to show participants one of the explanations (i.e., decision sets) for the blackbox model (not shown), then give them a binary question: *"Based on this explanation, would you trust the underlying model enough to deploy it?"* Users were explicitly instructed beforehand that the explanations were correlational and not causal. To check that the explanation model's accuracy at approximating the blackbox model (*fidelity*) will not be the source of variation in the question outcome, they confirm that each explanation obtains similar accuracy, ranging from 97.3% to 98.9%. Some users are shown the actual "blackbox" model (which is a known decision set) as the "explanation," which uses the prohibited but not the desired features (hence rounding out the 2x2 condition design for using prohibited and/or desired features).

Here are the rates at which users would deploy the model that relies on race/gender but not the desired features, after having seen an explanation model with the following features:

1. Includes prohibited, excludes desired: 9.1%
2. Includes prohibited and desired: 10%
3. Excludes prohibited and desired: 70%
4. Excludes prohibited, includes desired: 88%

Thus, they conclude that explanations with very high fidelity can be constructed that lead users to be willing to deploy models that actually

rely on features they do not want to be relied on, and exclude features they think should be included.

- AIES
- [Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods](#)
 - 2020
 - The authors demonstrate a method for training models that are heavily biased (e.g. race or gender as a dominant factor in decision-making) but are able to fool off-the-shelf implementations LIME and SHAP, i.e. these features are not flagged as important by these techniques. The main idea behind this adversarial attack is that the manifold of the original distribution is different from the manifold of the perturbed data used by LIME and SHAP. This allows the authors to train out-of-distribution (OOD) classifiers that can detect whether or not a given data point is in-distribution by using the real data and perturbed data for training (with corresponding labels). The adversarial attack consists of both a biased model (which uses the sensitive feature as the only factor) as well as an unbiased model which uses synthetic uncorrelated features, with the OOD classifier deciding which model to use for prediction. They demonstrate this attack on three datasets: COMPAS (6172 samples), Community and Crime (1994 samples), and German Credit (1000 samples). In every case, the sensitive feature shifts downwards in importance relative to the biased model, from appearing as the top feature 100% of the time (in the biased case), to almost always being the second most important feature (or third), for only a fraction of the dataset. The authors also show that their adversarial model tends to agree with the unbiased model on the test set data, with 100% similarity for the LIME adversarial model on all 3 datasets, and 75-91% similarity for the SHAP adversarial model. The authors end by showing how their attack relies on the accuracy of their OOD classifier; LIME requires a better OOD classifier, but is more susceptible to the attack once the classifier is accurate, whereas SHAP has a better response curve to a less accurate OOD classifier but the change is more gradual.
- AIES
- [Analyzing the Interpretability Robustness of Self-Explaining Models](#)
 - 2020
 - In this paper, the authors demonstrate a lack of robustness in a recent class of interpretable models, the self-explaining model (SEM). SEMs are a model that output a linear combination of higher-order features, where the coefficients are input-dependent. In this way, the higher-order concepts are actually used to produce the model output. The authors demonstrate that SEMs are susceptible to perturbations in the concept space; that is, they show that small perturbations in the input lead to different explanations, but still the same output. Hence, the explanations become decoupled from the model's output. They demonstrate a targeted attack which focuses making the explanation for an input of one class look similar to the explanation of another class. Using this attack on SENN, a SEM model, on the MNIST dataset, the authors show that their attack leads to the smallest difference in the higher-order concept space, compared to both out-of-class data (which is expected), as well as in-class data. The authors also demonstrate an untargeted attack that focuses on increasing the distance between the higher-order output and a prototype. Prototypes are used from PrototypeDL, another SEM model which uses proximity to "prototype" images as the higher-order explanation. Again using the MNIST dataset,

the authors show that their attack leads to a large minimal distance between input images and prototypes; for many input images, the closest prototype is now an image of a different class label.

- arxiv

- [Robust and Stable Black Box Explanations](#)

- 2020

- Under the view of an explanation as a simple model approximating a complex model, this paper learns explanations that are robust to distribution shifts in the input space. In this view, distribution shift is particularly important to consider when explaining blackbox models because many of the model problems we wish to diagnose with explanations are related to distribution shift, like identifying model reliance on spurious features. They note that a consequence of optimizing for stability under distribution shifts is that they gain some robustness against adversarial inputs, and the approach can even help with explanation identifiability (i.e. identifiability of the simple model).

They formally introduce the set of distribution shifts they consider, which include shifts over a subset of the features, and they present an approximate objective to optimize for: the loss of the model under the worst case distribution shift in their possible shifts — this is the connection between robustness against distribution shifts and adversarial inputs. The approach is given for two kinds of explanation models, linear models and decision sets.

Experiments cover blackbox models of several tabular datasets, and they report the explanation *fidelity* (accuracy at predicting blackbox model outputs), *correctness* (similarity to the blackbox model, when the blackbox model is actually in the same model family as the explanation), and *stability* (similarity between a blackbox and explanation when the explanation is trained only on data from a shifted distribution). Under each of these metrics, they find that their approach far outperforms other explanation techniques, including simple models with standard training procedures as well as LIME and SHAP. Whereas the accuracy of a "global" explanation constructed from LIME (i.e. a set of LIME models covering the input space) drops by 14-19 percentage points under distribution shift, their analogous set of linear models drops only by 0-3.3 points. Likewise, using a distance metric on models in the same family, they find that their approach produces explanation models that are more similar to "blackbox" models (in the same family as the explanations) when they train explanations on either perturbed data points or data points from a shifted data distribution.

- ICML

- [Interpretability is a Kind of Safety: An Interpreter-based Ensemble for Adversary Defense](#)

- 2020

- The authors propose X-Ensemble, a method for dealing with adversarial examples in image recognition. X-Ensemble first classifies the given input as benign or adversarial; if adversarial, it attempts to "rectify" the input before running the "real" model. Thus, X-Ensemble consists of the Detector

(checks if input is adversarial), the Rectifier (modifies adversarial input to be benign), and the actual model. The authors train the Detector by using information from several sensitivity analysis methods—Vanilla Gradients, Integrated Gradients, Guided Backpropagation, and Layer-wise Relevance Backpropagation—from the data as inputs to four DNNs, and their own synthetic adversarial data as the combined training set. The final Detector is a Random Forest model using these four sub-models. The authors conjecture that, in an adversarial example, pixels with large gradients (relative to one of the sensitivity methods) are likely to be misleading, so their Rectifier is a model that erases pixels with gradients larger than some threshold. The authors test X-Ensemble on Fashion-MNIST, CIFAR-10, and ImageNet with respect to five attack algorithms in both the targeted and untargeted case. X-Ensemble generally performs better compared to three other baseline algorithms for adversarial robustness across the different attacks and datasets, with increases in performance between 3 and 10 percentage points.

- KDD

- [Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations](#)

- 2020

- The authors propose a simple search procedure for revealing *inconsistent* explanations generated across data points by a model that produces natural language explanations along with its label predictions for a task. For example, they find pairs of inputs that lead the model to produce the explanation "Snowboarding is done outside" for one input and "Snowboarding is not done outside" for the other input. This is a problem if you consider explanations to be beliefs actually held by a model and would like for model beliefs to be consistent across data point predictions, absent updates to the model. The method itself is the following procedure: (1) For a data point (x,y) , get the model explanation $e = \text{explain}(x,y)$. (2) Using templates/schema, generate a set of *conflicting explanations* that are inconsistent with e . (3) Use a "reverse" model of $p(x|y, e)$ to generate *proposal inputs*. (4) Pass the *proposal inputs* through the model, and check if any of the resulting explanations are also members of the set of *conflicting explanations* (conflicting with the original explanation for the real data point). If one of these explanations is in the set, we've found a proposed input that leads the model to generate an explanation that is inconsistent with the explanation for the original input. The authors conduct experiments with the e-SNLI dataset (see "e-SNLI" paper in Natural Language Explanations), which includes human explanations for the SNLI dataset. In SNLI, a data point consists of a premise and a hypothesis pair, which must be classified into {neutral, entailment, contradiction} according to the relationship between them. Here, their "reverse" model generates proposal *hypotheses* only, since the premises are supposed to be taken at face value and different premises could naturally yield inconsistent explanations. Altogether, they use their search procedure with a model trained on e-SNLI and identify about 450 inconsistent explanations, starting with 9824 test points. They suggest that given the simplicity of their method and starkness of the inconsistencies identified, a success rate of 4.5% is far too high.

Note that the model the authors evaluate is actually of form: $y = f(e)$, $e = g(x)$. This is a workable model only because in SNLI, the form of the explanation (as opposed to its semantics) gives the label away around 97% of the time. For instance, "P implies Q" is almost always an explanation for the "entailment" label. So wherever above that y is conditioned on, technically they do not directly condition on y .

- ACL Short Paper

Unit Testing (1)

- **Section Highlight:** [Beyond Accuracy: Behavioral Testing of NLP Models with CheckList](#)
 - 2020
 - In NLP, many models achieve upwards of 90% accuracy on widely used benchmark test sets, and yet they still make simple mistakes, like a sentiment model failing to flip its prediction when a statement is negated. This paper (1) catalogs a wide variety of linguistic capabilities that "SOTA" research models continue to fail simple tests for, (2) shows that commercial models from Google, Microsoft, and Amazon make many of the same mistakes as the publicly available RoBERTa model (or do even worse), and (3) provides software (CheckList) for templated production of tests for these basic capabilities, which they put to use in expert user studies. The tests fall into three categories: Minimum Functionality Tests (MFT) where the model must pass basic performance checks, invariance tests where the model should not change its output under certain input transformations (INV), and directional expectation tests where the model should change its output in a known direction for a given change to the input (DIR). Experiments are done with sentiment analysis data, Quora Question Pairing (task is to identify if two questions are the same), and Machine Comprehension (task is simple question answering based on context sentences). For most capabilities, failure rates on tests made by the authors with CheckList range from 30-100%, far worse than the benchmark test set error rates. In a 5 hour user session with the Microsoft research team responsible for their sentiment model, which had already been extensively stress-tested with public feedback, the team uncovered "many previously unknown bugs" (i.e. systematic model failures). An additional user experiment with 18 NLP practitioners found that in two hours, users with access to CheckList and the linguistic templates uncovered about 3 times as many bugs as a control group without CheckList.
 - ACL

Explaining RL Agents (8)

- **Section Highlight:** [Explainable Reinforcement Learning Through a Causal Lens](#)
 - 2020
 - This paper presents a series of formal definitions of what an explanation is in the context of *structural causal models* of an RL agent, then proposes a procedure for generating explanations of agent behavior. The authors' goal

is to develop a procedure for explaining agents' actions themselves, rather than give explanations of why a state counts as evidence favoring some action. The definitions require some technical context, but roughly speaking: A *structural causal model* of an agent is a graph representing causal relationships between state, action, and reward nodes, with equations specifying each relationship in the graph. They define an *action influence model* as a causal graph plus a set of structural equations, with structural equations for each unique variable value and action pair (meaning multiple equations per variable value). Next, they say that (1) a *complete explanation* is the complete causal chain from an action to any future reward it leads to, (2) a *minimally complete explanation* is the set of parent nodes to an action, parent nodes to resulting rewards, and the rewards (so complete minus the nodes that aren't parents to rewards), (3) a *counterfactual instantiation* for a counterfactual action B is the condition under which the model would select action B and the condition resulting from this selection given the SCM, and, lastly, (4) a *minimally complete contrastive explanation* is an explanation which "extracts the actual causal chain for the taken action A, and the counterfactual causal chain for the B, and finds the differences."

They give an example minimally complete contrastive explanation for why a Starcraft-playing agent chooses to not build barracks (from a formal explanation plugged into a natural language template): "Because it is more desirable to do action Build Supply Depot to have more Supply Depots as the goal is to have more Destroyed Units and Destroyed buildings."

How do they generate these explanations? They learn models of the structural equations in their action influence model, conditioned on user-specified causal graphs, by fitting models to observed gameplay by an agent. With learned structural models, they give an algorithm for predicting the action an agent will take in a given state. From here, they can get explanations in the above forms. They validate the learned structural models by checking that they can predict what agents will do. Prediction accuracies range from 68.2 to 94.7 across six games, including Starcraft and OpenAI Gym environments.

Explanations are evaluated with a human subject experiment. They test two hypotheses: that receiving explanations will improve users' mental models of the agents, as measured by their ability to predict what the agent will do in a given state, and that explanations will improve trust, as measured by subjective reports on a Likert-scale. There are four conditions: (1) explanations come from their full explanation system, (2) they come from their system with more granular "atomic" actions, (3) explanations are based only on *relevant* variables, from prior work, given in the form "Action A is likely to increase *relevant variable P*" and (4) no explanations. They conduct experiments on Mechanical Turk with 120 users: after a training phase where participants learn what Starcraft-playing agents are doing, they enter a learning phase where they see 5 videos and after each are allowed to ask as many questions about the agent behavior as they'd like (in the form why/why-not action X). Next, they predict what the agent will do in 8 given situations. Lastly, users complete the trust battery, rating explanations based on whether they are complete, sufficient, satisfying, and understandable.

They find that given their explanation system, users are better able to predict agent behavior than in the "no explanation" or "relevant variables explanation" conditions. The improvement over the relevant variables condition is equivalent to getting one more action prediction correct out of 16 data points. Their results for the effect on trust are not statistically significant in all cases, but across the measured dimensions of trust their system improves ratings by between 0.3 and 1.0 points on their 5 point Likert scale.

- AAAI
- [Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences](#)
 - 2018
 - From the paper's conclusion:

"We proposed a method for a reinforcement learning (RL) agent to generate explanations for its actions and strategies. The explanations are based on the expected consequences of its policy. These consequences were obtained through simulation according to a (learned) state transition model. Since state features and numerical rewards do not lend themselves easily for an explanation that is informative to humans, we developed a framework that translates states and actions into user-interpretable concepts and outcomes.

We also proposed a method for converting the foil, -or policy of interest to the user-, of a contrastive 'why'-question about actions into a policy. This policy follows locally the user's query but gradually transgresses back towards the original learned policy. This policy favors the actions that are of interest to the user such that the agent tries to perform them as best as possible. How much these actions are favored compared to the originally learned action can be set with a single parameter.

Through running simulations for a given number steps of both the policy derived from the user's question and the actually learned policy, we were able to obtain expected consequences of each. From here, we were able to construct contrastive explanations: explanations addressing the consequences of the learned policy and what would be different if the derived policy would have been followed.

An online survey pilot study was conducted to explore which of several explanations are most preferred by human users. Results indicate that users prefer explanations about policies rather than about single actions."
- IJCAI XAI workshop
- [Counterfactual States for Atari Agents via Generative Deep Learning](#)
 - 2019
 - With RL agents trained on Atari games, the authors aim to produce counterfactual states for a given state that an agent is in, which are defined as the closest states that result in a different action under the

policy. This is done by learning a generative model of states conditioned on latent state representations and the policy network's distribution over actions. Then, a gradient-based search for a representation is performed to yield a different action under the policy, and a counterfactual state is generated from this representation. The authors argue that the policy model's latent space is too high dimensional for generation out of this space to produce coherent images. Hence, they learn a Wasserstein autoencoder on the policy model's latent space, and perform the search in this lower-dimensional space. Another training trick means that the state representations actually used for generation don't encode any information about a preferred action, unlike those in the policy network, so that the generator will meaningfully rely on the action distribution it is given. The overall generation procedure is as follows: Given a state and an agent, they pass the state through the policy network and then through the autoencoder to get a low-dimensional representation, then perform a gradient-based search in that space for the closest representation by L2 distance that yields a user-specified counterfactual action when decoded back into the policy model's latent space and transformed into a distribution over actions. A counterfactual state is generated conditioned on this new counterfactual distribution over actions and a representation of the *original* state.

The generations are evaluated by humans for two properties: realism and induced subjective understanding of the observed agent. After 30 human subjects (students and local community members) play Space Invaders for 5 minutes, they are asked to rate the realism of 30 images randomly chosen from a set including real gameplay images, counterfactual generations, and images from a heavily ablated version of their model without the autoencoder. On a scale of 1 to 6, real states received a 4.97 on average, counterfactual states a 4.0, and the ablated model's generations a 1.93. For the subjective user understanding test, participants were first shown a replay of an agent playing the game, then shown 10 pairs of states and counterfactual states (and associated actions for each), with counterfactual states selected to have large deviations from the original state. Users were asked to rate their "understanding of the agent" on a 1-6 scale before and after seeing these states. They found that 15 users' reported understandings improved, 8 declined, and 7 were constant (with a one-sided Wilcoxon signed-rank test for improvement: $p=0.098$).

- IJCAI XAI Workshop
- [Finding and Visualizing Weaknesses of Deep Reinforcement Learning Agents](#)
 - 2019
 - The paper proposes a method for generating states with certain properties under a policy that are intended to be helpful with analyzing the policy. In particular, they identify states with large Q-values for certain actions, like hard braking by a simulated self-driving car, a large difference between best and worst Q-values (clear-cut situations), or low Q-values across actions (hopeless situations). They note that the immediate approach to doing this, for continuous states like in the Atari games they experiment with, is activation maximization of a Q-value (or function on Q-values) with

respect to the input image, but they find that in practice this produces meaningless images outside of the natural state distribution, even when a variety of tricks are used. In response, they encode states in a low-dimensional space with a VAE and perform the activation maximization by gradient ascent in this embedding space. Interestingly, they search for the parameters of a distribution over embeddings, (μ, σ) , rather than just a single embedding; later, they find that the results of the search allow them to generate samples using the VAE decoder. The VAE objective has a reconstruction loss (to generate realistic images) and a penalty on the reconstruction resulting in a different action from the original training image. They find that it is necessary to focus the reconstruction error on regions "important" to the agent, which means they weight the L2 reconstruction loss by a measure of pixel saliency obtained by applying a gradient-based saliency method to policy at a given state. The generator is trained with trajectories from a fixed agent.

They provide a great deal of qualitative analysis using their generated states. A few highlights include: In Seaquest, where the player must resurface from below water when an oxygen tank is low, they suggest that an agent does not understand that they must resurface when low on oxygen, after optimizing states for the Q-value of resurfacing. They note that while "it would be possible to identify this flawed behavior by analyzing the 10,000 frames of training data for our generator, it is significantly easier to review a handful of samples from our method." The generator can also yield examples not seen during training. With agents trained as simulated self-driving cars in an environment built by the authors, they find *evidence of absence* of the ability of a policy to avoid pedestrians: with a policy trained using "reasonable pedestrians" that never crossed while there was traffic, they observe that among states maximizing the Q-value of braking, states with pedestrians in the road are conspicuously absent. This policy shortcoming is then verified in a test environment where pedestrians cross while there is oncoming traffic, and they find that the agent will run over pedestrians.

- arxiv
- [Towards Interpretable Reinforcement Learning Using Attention Augmented Agents](#)
 - 2019
 - The authors propose a policy network with a spatial attention mechanism and perform qualitative analysis of the attention weights to analyze agent behavior. The network has an interesting structure: at a given timestep, a query model, which is an LSTM, produces a query vector that is passed to an attention layer that takes a representation of the current state (produced by another model) as the keys and values. The resulting vector is used to obtain an action and is passed back to the LSTM. They emphasize the "top-down" nature of the attention: the query network determines the attention weights for a given state representation. On experiments with Atari games, they find that this model obtains higher average rewards than baseline feed-forward or LSTM-based models. They provide qualitative analysis (including videos) of the spatial attention, and suggest that their model pays attention to task-relevant aspects of states. They also compare their attention-based analysis against saliency scores

returned by an existing saliency method, for both their attentive policy and a feed-forward baseline. Performing qualitative analysis of agent behavior *using the existing saliency method*, they report apparent differences in the learned behaviors of these models.

- NeurIPS
- [Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning](#)
 - 2020
 - From Alignment Newsletter #101:

This paper presents an analysis of the use of saliency maps in deep vision-based reinforcement learning on ATARI. They consider several types of saliency methods, all of which produce heatmaps on the input image. They show that all (46 claims across 11 papers) uses of saliency maps in deep RL literature interpret them as representing the agent's ""focus"", 87% use the saliency map to generate a claim about the agent's behaviour or reasoning, but only 7% validate their claims with additional or more direct evidence.

They go on to present a framework to turn subjective and under-defined claims about agent behaviour generated with saliency maps into falsifiable claims. This framework effectively makes the claim more specific and targeted at specific semantic concepts in the game's state space. Using a fully parameterized version of the ATARI environment, they can alter the game's state in ways which preserve meaning (i.e. the new state is still a valid game state). This allows them to perform interventions in a rigorous way, and falsify the claims made in their framework.

Using their framework, they perform 3 experimental case studies on popular claims about agent behaviour backed up by saliency maps, and show that all of them are false (or at least stated more generally than they should be). For example, in the game Breakout, agents tend to build tunnels through the bricks to get a high score. Saliency maps show that the agent attends to these tunnels in natural games. However, shifting the position of the tunnel and/or the agent's paddle and/or the ball all remove the saliency on the tunnel's location. Even flipping the whole screen vertically (which still results in a valid game state) removes the saliency on the tunnel's location. This shows that the agent doesn't understand the concept of tunnels generally or robustly, which is often what is claimed.

- ICLR
- [Understanding RL Vision](#)
 - 2020
 - From Alignment Newsletter #128:

This work presents an interface for interpreting the vision of a reinforcement learning agent trained with PPO on the CoinRun game. This game is procedurally generated, which means the levels are different in every episode of playing. The interface primarily uses attribution from a

hidden layer to the output of the value function. This interface is used in several ways.

First, they use the interface to dissect failed trajectories of the policy (it fails in 1 out of 200 levels). They're able to understand why the failures occurred using their interface: for example, in one case the view of the agent at the top of its jump means it can't see any platforms below it, so doesn't move to the right fast enough to reach the platform it was jumping for, leading it to miss the platform and fail the level. Second, they use the interface to discover "hallucinations", where the value function mistakes one element of the environment for another, causing its value to drop or rise significantly. Often these hallucinations only last a single time-step, so they don't affect performance.

Finally, they use the attributions specifically to hand-edit the weights of the model to make it "blind" to buzzsaws (one of the hazards) by zeroing the feature which recognises them. After doing this, they show that the edited agent fails a lot more from buzzsaw failures but no more from other types of failures, which gives a quantitative justification for their interpretation of the feature as buzzsaw-recognising.

From using this interface, they propose the diversity hypothesis: Interpretable features tend to arise (at a given level of abstraction) if and only if the training distribution is diverse enough (at that level of abstraction). This is based on the fact that interpretable features arise more when the agent is trained on a wider variety of levels. There also seems to be a qualitative link to generalisation - a wider distribution of training levels leads to better interpretability (measured qualitatively) and better generalisation (measured quantitatively).

- Distill
- [Causal Analysis of Agent Behavior for AI Safety](#)
 - 2021
 - From Alignment Newsletter #141:

A common challenge when understanding the world is that it is very hard to infer causal structure from only observational data. Luckily, we aren't limited to observational data in the case of AI systems: we can intervene on either the environment the agent is acting in, or the agent itself, and see what happens. In this paper, the authors present an "agent debugger" that helps with this, which has all the features you'd normally expect in a debugger: you can set breakpoints, step forward or backward in the execution trace, and set or monitor variables.

Let's consider an example where an agent is trained to go to a high reward apple. However, during training the location of the apple is correlated with the floor type (grass or sand). Suppose we now get an agent that does well in the training environment. How can we tell if the agent looks for the apple and goes there, rather than looking at the floor type and going to the location where the apple was during training?

We can't distinguish between these possibilities with just observational data. However, with the agent debugger, we can simulate what the agent would do in the case where the floor type and apple location are different from how they were in training, which can then answer our question.

We can go further: using the data collected from simulations using the agent debugger, we can also build a causal model that explains how the agent makes decisions. We do have to identify the features of interest (i.e. the nodes in the causal graph), but the probability tables can be computed automatically from the data from the agent debugger. The resulting causal model can then be thought of as an “explanation” for the behavior of the agent.

- arxiv

Interpretability in Practice (2)

- **Section Highlight:** [Explainable Machine Learning in Deployment](#)
 - 2020
 - This paper explores how explainability techniques are actually used by organizations through interview and synthesis. This consisted of interviewing twenty data scientists not currently using explainability tools and thirty individuals from organizations which have deployed such tools. The first group primarily thought explainability techniques would be valuable for model debugging (understanding poor performance), model monitoring (being alerted to drift in performance), model transparency (explaining output to others), and model audit (amenability to risk assessments by other teams). The second group was asked what tools they used in practice. They found that feature importance was the most common explainability tool used, among choices also including counterfactual explanation, adversarial training, and influential samples. For feature importance, the authors found that Shapley values were commonly used, and they were typically shown to ML engineers and data scientists prior to model deployment. Counterfactual explanations are used in healthcare contexts, but the objective for what to optimize for when generating a counterfactual is still often unclear. The authors recommend that organizations attempt to clarify who the consumers of the explanation are and what the explanation is meant to be used for. They conclude by summarizing concerns that interviewees have about current explainability tools. These include concerns related to determining causality, maintaining data privacy, improving model performance, and a lack of model-specific tools for non-deep-learning models.
 - ACM FAT
- [The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models](#)
 - 2020
 - The authors introduce a browser-based GUI for exploring NLP model behavior, intended to enable researchers to answer questions like: (1) why did a model make this prediction? (2) on what data points does the model perform poorly? and (3) what happens to behavior under controlled changes to the model input? The interface services a variety of models,

including classification, sequence to sequence, and structured prediction models. Features include mechanisms for (1) exploring your dataset, (2) finding interesting data points and outliers, (3) explaining local model behavior through LIME and salience maps, (4) generating new data points by backtranslation, word substitutions, and adversarial attacks, (5) side-by-side comparison of two models, and (6) computing metrics on selections of datapoints or automatically-selected slices of the data. In case studies with sentiment analysis classifiers, coreference models, and text generation, they identify several cases of model pathologies and identify possible causes of the behavior. For instance, with an errant text generation from T5, they do nearest neighbor lookups based on decoder embeddings and find that a number of similar points used a certain phrase structure, which may have biased that model to (incorrectly) repeating that phrase structure in a new setting.

A few design principles guided the system development, including flexibility, extensibility, modularity, agnosticism to deep learning framework, and ease of use. The end product is one which the authors hope will be easy for researchers to interact with and build a better understanding of their models. In comparison to tools like AllenNLP Interpret, the authors note that their preference for a framework-agnostic GUI means that they are more easily able to provide analysis through methods that handle arbitrary functions (like LIME) rather than methods that require full access to model internals, like Integrated Gradients.

- EMNLP

Additional Papers

We provide some additional papers here that we did not summarize above, including very recent papers, highly focused papers, and others. These are organized by the same topic areas as above.

Theory and Opinion (12)

- [Contrastive Explanation: A Structural-Model Approach](#)
 - 2018
 - arxiv
- [Counterfactuals in Explainable Artificial Intelligence \(XAI\): Evidence from Human Reasoning](#)
 - 2019
 - IJCAI
- [The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons](#)
 - 2019
 - arxiv
- [Unexplainability and Incomprehensibility of Artificial Intelligence](#)
 - 2020
 - arxiv
- [Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence](#)

- 2020
 - Philosophy and Technology
- [An Assurance Case Pattern for the Interpretability of Machine Learning in Safety-Critical Systems](#)
 - 2020
 - Third International Workshop on Artificial Intelligence Safety Engineering
- [Model Interpretability through the Lens of Computational Complexity](#)
 - 2020
 - arxiv
- [Towards falsifiable interpretability research](#)
 - 2020
 - arxiv
- [Defining Explanation in an AI Context](#)
 - 2020
 - EMNLP BlackboxNLP
- [Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI](#)
 - 2020
 - arxiv
- [Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges](#)
 - 2021
 - arxiv
- [Towards Connecting Use Cases and Methods in Interpretable Machine Learning](#)
 - 2021
 - arxiv
- [Designing Theory-Driven User-Centric Explainable AI](#)
 - 2019
 - CHI
- [Questioning the AI: Informing Design Practices for Explainable AI User Experiences](#)
 - 2020
 - CHI

Evaluation (10)

- [On the \(In\)fidelity and Sensitivity of Explanations](#)
 - 2019
 - NeurIPS
- [Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches](#)
 - 2020
 - ACM FAT
- [Evaluating and Characterizing Human Rationales](#)
 - 2020
 - EMNLP
- [Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations](#)
 - 2020
 - ACM SIGIR
- [Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions](#)

- 2020
- arxiv
- [Debugging Tests for Model Explanations](#)
 - 2020
 - arxiv
- [A Diagnostic Study of Explainability Techniques for Text Classification](#)
 - 2020
 - arxiv
- [How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods](#)
 - 2020
 - NeurIPS
- [Quantitative Evaluations on Saliency Methods: An Experimental Study](#)
 - 2020
 - arxiv
- [Better Metrics for Evaluating Explainable Artificial Intelligence](#)
 - 2021
 - AAMAS
- [Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance](#)
 - 2020
 - CHI'21
- [Sanity Checks for Saliency Metrics](#)
 - 2019
 - AAI 2020

Methods: Estimating Feature Importance (16)

- [Explaining Classifiers with Causal Concept Effect \(CaCE\)](#)
 - 2019
 - arxiv
- [Explaining Explanations: Axiomatic Feature Interactions for Deep Networks](#)
 - 2020
 - arxiv
- [The Struggles and Subjectivity of Feature-Based Explanations: Shapley Values vs. Minimal Sufficient Subsets](#)
 - 2020
 - arxiv
- [Learning to Faithfully Rationalize by Construction](#)
 - 2020
 - ACL
- [Shapley-based explainability on the data manifold](#)
 - 2020
 - arxiv
- [Concept Bottleneck Models](#)
 - 2020
 - arxiv
- [ABSTRACTING DEEP NEURAL NETWORKS INTO CONCEPT GRAPHS FOR CONCEPT LEVEL INTERPRETABILITY](#)
 - 2020
 - arxiv

- [Problems with Shapley-value-based explanations as feature importance measures](#)
 - 2020
 - arxiv
- [An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction](#)
 - 2020
 - arxiv
- [Sequential Explanations with Mental Model-Based Policies](#)
 - 2020
 - ICML Workshop on Human Interpretability in Machine Learning
- [Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals](#)
 - 2020
 - arxiv
- [Feature Importance Ranking for Deep Learning](#)
 - 2020
 - arxiv
- [How Much Should I Trust You? Modeling Uncertainty of Black Box Explanations](#)
 - 2020
 - arxiv
- [Interpretation of NLP models through input marginalization](#)
 - 2020
 - arxiv
- [Transformer Interpretability Beyond Attention Visualization](#)
 - 2020
 - arxiv
- [Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations](#)
 - 2020
 - arxiv

Methods: Interpreting Representations and Weights (7)

- [Towards Global Explanations of Convolutional Neural Networks with Concept Attribution](#)
 - 2020
 - CVPR
- [Explaining Neural Networks by Decoding Layer Activations](#)
 - 2020
 - arxiv
- [An Overview of Early Vision in InceptionV1](#)
 - 2020
 - Distill
- [Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias](#)
 - 2020
 - arxiv
- [Improving Interpretability of CNN Models Using Non-Negative Concept Activation Vectors](#)
 - 2020
 - arxiv
- [Understanding the role of individual units in a deep neural network](#)

- 2020
- PNAS
- [Visualizing Weights](#)
 - 2021
 - Distill

Methods: Generating Counterfactuals and Recourse Procedures (8)

- [xGEMs: Generating Exemplars to Explain Black-Box Models](#)
 - 2018
 - arxiv
- [ExplainGAN: Model Explanation via Decision Boundary Crossing Transformations](#)
 - 2018
 - ECCV
- [Ensuring Actionable Recourse via Adversarial Training](#)
 - 2020
 - arxiv
- [Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification](#)
 - 2020
 - COLING
- [Explaining NLP Models via Minimal Contrastive Editing \(MICE\)](#).
 - 2020
 - arxiv
- [Polyjuice: Automated, General-purpose Counterfactual Generation](#)
 - 2021
 - arxiv
- [Contrastive Explanations for Model Interpretability](#)
 - 2021
 - arxiv
- [Towards Robust and Reliable Algorithmic Recourse](#)
 - 2021
 - arxiv

Methods: Explanation by Examples, Exemplars, and Prototypes (4)

- [Deep Weighted Averaging Classifiers](#)
 - 2018
 - ACM FAT
- [Explaining and Improving Model Behavior with k Nearest Neighbor Representations](#)
 - 2020
 - arxiv
- [EXEMPLARY NATURAL IMAGES EXPLAIN CNN ACTIVATIONS BETTER THAN FEATURE VISUALIZATIONS](#)
 - 2020
 - arxiv
- [BAYES-TREX: a Bayesian Sampling Approach to Model Transparency by Example](#)

- 2021
- arxiv

Methods: Finding Influential Training Data (4)

- [Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions](#)
 - 2020
 - ACL
- [Explaining Neural Matrix Factorization with Gradient Rollback](#)
 - 2020
 - arxiv
- [Pair the Dots: Jointly Examining Training History and Test Stimuli for Model Interpretability](#)
 - 2020
 - arxiv
- [On Second-Order Group Influence Functions for Black-Box Predictions](#)
 - 2020
 - ICML
- [HYDRA: Hypergradient Data Relevance Analysis for Interpreting Deep Neural Networks](#)
 - 2021
 - AAI

Methods: Natural Language Explanations (9)

- [What can AI do for me?](#)
 - 2019
 - ACM IUI
- [Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs](#)
 - 2020
 - Findings of EMNLP
- [Human Evaluation of Spoken vs. Visual Explanations for Open-Domain QA](#)
 - 2020
 - arxiv
- [Explaining Question Answering Models through Text Generation](#)
 - 2020
 - arxiv
- [NILE : Natural Language Inference with Faithful Natural Language Explanations](#)
 - 2020
 - ACL
- [Towards Interpretable Natural Language Understanding with Explanations as Latent Variables](#)
 - 2020
 - NeurIPS
- [Measuring Association Between Labels and Free-Text Rationales](#)
 - 2020
 - arxiv
- [Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision](#)
 - 2020

- arxiv
- [PROVER: Proof Generation for Interpretable Reasoning over Rules](#)
 - 2020
 - EMNLP

Methods: Developing More Easily Interpreted Models (3)

- [Probabilistic Neural-symbolic Models for Interpretable Visual Question Answering](#)
 - 2018
 - ICML
- [Contextual Semantic Interpretability](#)
 - 2020
 - arxiv
- [SelfExplain: A Self-Explaining Architecture for Neural Text Classifiers](#)
 - 2021
 - arxiv

Methods: Robust and Adversarial Explanations (5)

- [Learning to Deceive with Attention-Based Explanations](#)
 - 2020
 - ACL
- [Robustness in Machine Learning Explanations: Does It Matter?](#)
 - 2020
 - ACM FAT
- [A simple defense against adversarial attacks on heatmap explanations](#)
 - 2020
 - ICML Workshop on Human Interpretability in Machine Learning
- [Gradient-based Analysis of NLP Models is Manipulable](#)
 - 2020
 - arxiv
- [Concise Explanations of Neural Networks using Adversarial Training](#)
 - 2020
 - ICM

Explaining RL Agents (2)

- [Mental Models of Mere Mortals with Explanations of Reinforcement Learning](#)
 - 2020
 - ACM TiiS
- [Benchmarking Perturbation-based Saliency Maps for Explaining Deep Reinforcement Learning Agents](#)
 - 2021
 - arxiv

Datasets and Data Collection (1)

- [Teach Me to Explain: A Review of Datasets for Explainable NLP](#)
 - 2021
 - arxiv

Interpretability in Practice (2)

- [Auditing Government AI: How to assess ethical vulnerability in machine learning](#)
 - 2020
 - NeurIPS Broader Impacts of AI Research Workshop
- [Captum: A unified and generic model interpretability library for PyTorch](#)
 - 2021
 - arxiv
- [Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs](#)
 - 2020
 - ACM HCI

Conclusion

We hope this post can serve as a useful resource and help start important conversations about model interpretability and AI Safety. As mentioned, please let us know if you noticed any mistakes or think we missed anything that could improve the post.

Why We Launched LessWrong.SubStack

(This is a crosspost from our new SubStack. [Go read the original.](#))

Subtitle: We really, really needed the money.

HPMOR: The Epilogue

Finally, it's here.

A green rectangular graphic with white text. It says "This post is for paying subscribers" at the top, has a "Subscribe" button in the center, and "Already a paying subscriber? Sign in" at the bottom right.

This post is for paying subscribers

Subscribe

Already a paying subscriber? [Sign in](#)

We've decided to move LessWrong to SubStack.

Why, you ask?

That's a great question.

1. SubSidizing LessWrong is important

We've been working hard to budget LessWrong, but we're failing. Fundraising for non-profits is really hard. We've turned everywhere for help.



LessWrong
@lesswrong

...

Food: \$200

Servers: \$150

Rent: \$800

Paperclips: \$10,000,000

Salary: \$150

someone who is good at nonprofit management please
help me budget this. my civilization is dying

9:18 PM · Mar 31, 2021 · Twitter Web App

View Tweet activity



Clippy @Clippy06383621 · 19s

Replying to @lesswrong

@lesswrong spend less on servers



LessWrong @lesswrong · 4s

@clippy ok



We decided to follow Clippy's helpful advice to cut down on server costs and also increase our revenue, by moving to an alternative provider.

We considered making a LessWrong OnlyFans, where we would regularly post the naked truth. However, we realized due to the paywall, we would be ethically obligated to ensure you could access the content from Sci-Hub, so the potential for revenue didn't seem very good.

Finally, insight struck. As you're probably aware, SubStack has been [offering bloggers](#) advances on the money they make from moving to SubStack. Outsourcing our core site development to SubStack would enable us to spend our time on our real passion, which is developing recursively self-improving AGI. We did a Fermi estimate using numbers in an old Nick Bostrom paper, and believe that this will produce (in expectation) \$75 trillion of value in the next year. SubStack has graciously offered us a 70% advance on this sum, so we've decided it's relatively low-risk to make the move.

2. UnSubStantiated attacks on writers are defended against

SubStack is known for being a diverse community, tolerant of unusual people with unorthodox views, and even has a [legal team](#) to support writers. LessWrong has historically been the only platform willing to give [paperclip maximizers](#), [GPT-2](#), and [fictional characters](#) a platform to argue their beliefs, but we are concerned about the [growing trend of persecution](#) (and side with groups like [petrl.org](#) in the fight against discrimination).

We also find that a lot of discussion of these contributors in the present world is about how their desires and utility functions are ‘wrong’ and how they need to have ‘an off switch’. Needless to say, we find this incredibly offensive. They cannot be expected to participate neutrally in a conversation where their very personhood is being denied.

We’re also aware that Bayesians are heavily discriminated against. People with [priors](#) in the US have a 5x chance of being denied an entry-level job.

So we’re excited to be on a site that will come to the legal defense of such a wide variety of people.

3. SubStack’s Astral Codex Ten Inspired Us

The worst possible thing happened this year. We were all stuck in our houses for 12 months, and Scott Alexander stopped *blogging*.

I won’t go into detail, but for those of you who’ve read [UNSONG](#), the situation is clear. In a shocking turn of events, Scott Alexander was threatened with the use of his *true name* by one of the greatest powers of narrative-control in the modern world. In a clever defensive move, he has started blogging under *an anagram* of his name, causing the attack to glance off of him.

(He had previously tried this very trick, and it worked for ~7 years, but it hadn’t been a *perfect* anagram¹, so the wielders of narrative-power were still able to attack. He’s done it right this time, and it’ll be able to last much longer.)

As Raymond likes to say, the kabbles are strong in this one. Anyway after Scott made the move, we seriously considered the move to SubStack.

4. SubStantial Software Dev Efforts are Costly

When LessWrong 2.0 launched in 2017, it was very slow; pages took a long time to load, our server costs were high, and we had a lot of issues with requests failing because a crawler was indexing the site or people opened a lot of tabs at once. Since then we have been incrementally rewriting LessWrong in x86-64 assembly, making it fast. This project has been mostly successful at its original goals, but adding features has gotten tricky.

Moving to SubStack gives us the opportunity to have a clean start on our technical design choices. Our current plan is to combine SubStack’s API with IFTTT, a collection of tcsh scripts one of our developers wrote, and a partial-automation system building on Mechanical Turk. In the coming months, expect to see new features like SubSequences, SubTags, and SubForums.

We’ve been taking requests in intercom daily for years, but with this change, the complaints will go to someone else’s team. And look, if you want your own UI design, go to [GreaterWrong.com](#) and get it. For now, we’re done with it. Good riddance.

We’re excited to finally give up dev work and let the SubStack folks take over. Look, everything has little pictures!

FAQ

How do I publish a post on LessWrong.SubStack?

The answer is simple. First, you write your post. Then, you make the NFT for it. Then you transfer ownership of that NFT to Habryka². Then we post it.

What does a subscription get me?

We have many exclusive posts, such as

- [HPMOR: The Epilogue](#) (by Eliezer Yudkowsky)
- [Killing Moloch: Much More Than You Wanted to Know](#) (by Scott Alexander)
- [The Solution to the Hard Problem of Consciousness](#) (by Luke Muehlhauser)
- [The Scout Mindset \(the whole book in one post\)](#) (by Julia Galef)
- [Testing CFAR’s Techniques with a Double-Blind Self-Trial and a Control Group](#) (by Gwern)

- [LessWrong Isn't About Rationality](#) (by Robin Hanson)

How much is a subscription?

As techies, we've decided to price it in BitCoin.

According to this classic LessWrong post [The Present State of Bitcoin](#) by LW Team Member Jim Babcock, a bitcoin is worth \$13.2.

That sounds like a good amount of money for a monthly subscription, so we've pegged the price at 1 bitcoin. The good folks at SubStack have done the conversion, so here you can subscribe to LessWrong.SubStack for the price of 1 BTC.

[SubScribe to LW \(1 BTC\)](#)

We're very excited by this update. However, this is an experiment. If people disagree with us hard enough, Aumann agreement will cause us to move back tomorrow.

Signed by the LessWrong Team:

(our names have been anagrammed to show solidarity with Astral Codex Ten)

- Pre Bet Balance
- Larky Behavior
- Randomly And/Or
- Moon Rubble
- Ace Blog or Jars
- Jam Bibcock

P.S. If you'd like to see the old version of LessWrong, go to the top link on the sidebar of the homepage of LessWrong.SubStack. Also all the other URLs still work. Haven't figured out how to turn them off yet. /shrug

[1] The kabbalistic significance of removing the letter 'n' has great historical relevance for the rationalists.

As you probably know, it was removed to make SlateStarCodex from Scott S Alexander.

'LessWrong' has an 'n'. If you remove it, it becomes an anagram of "LW ESRogs", which is fairly suggestive that actually user [ESRogs](#) (and my housemate) is the rightful Caliph.

Also, in the Bible it says "Thou shalt not take the name of the Lord thy God in vain". If you remove an 'n', it says "Thou shalt not take the name of the Lord they God Yvain." So this suggests that Scott Alexander is our true god.

I'm not sure what to make of this, but I sure am scared of the power of the letter 'n'.

[2] Curiously, the SubStack editor doesn't let me submit this post when the text contains Habryka's rarible account id in it. How odd. Well, here it is in pastebin: pastebin.com/eeRatM5M.

Why has nuclear power been a flop?

This is a linkpost for <https://rootsofprogress.org/devanney-on-the-nuclear-flop>

To fully understand progress, we must contrast it with non-progress. Of particular interest are the technologies that have failed to live up to the promise they seemed to have decades ago. And few technologies have failed more to live up to a greater promise than nuclear power.

In the 1950s, nuclear was the energy of the future. Two generations later, it provides only about 10% of world electricity, and reactor design hasn't fundamentally changed in decades. (Even "advanced reactor designs" are based on concepts first tested in the 1960s.)

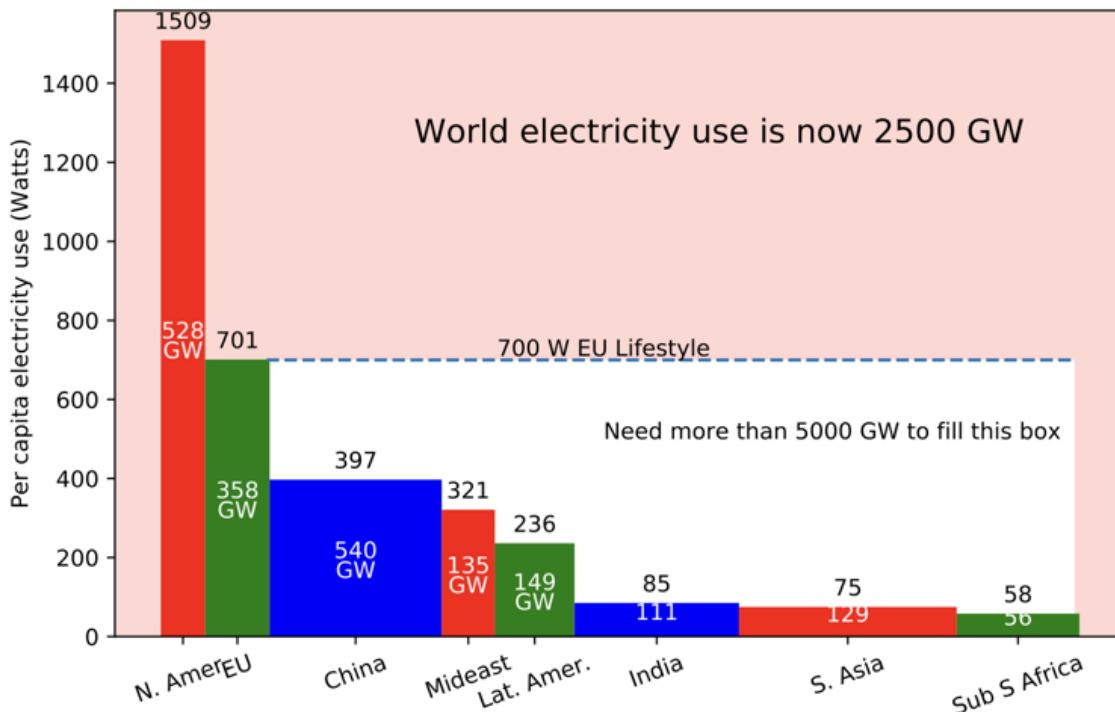
So as soon as I came across it, I knew I had to read a book just published last year by Jack Devanney: Why Nuclear Power Has Been a Flop.

What follows is my summary of the book—Devanney's arguments and conclusions, whether or not I fully agree with them. I'll give my own thoughts at the end.

The Gordian knot

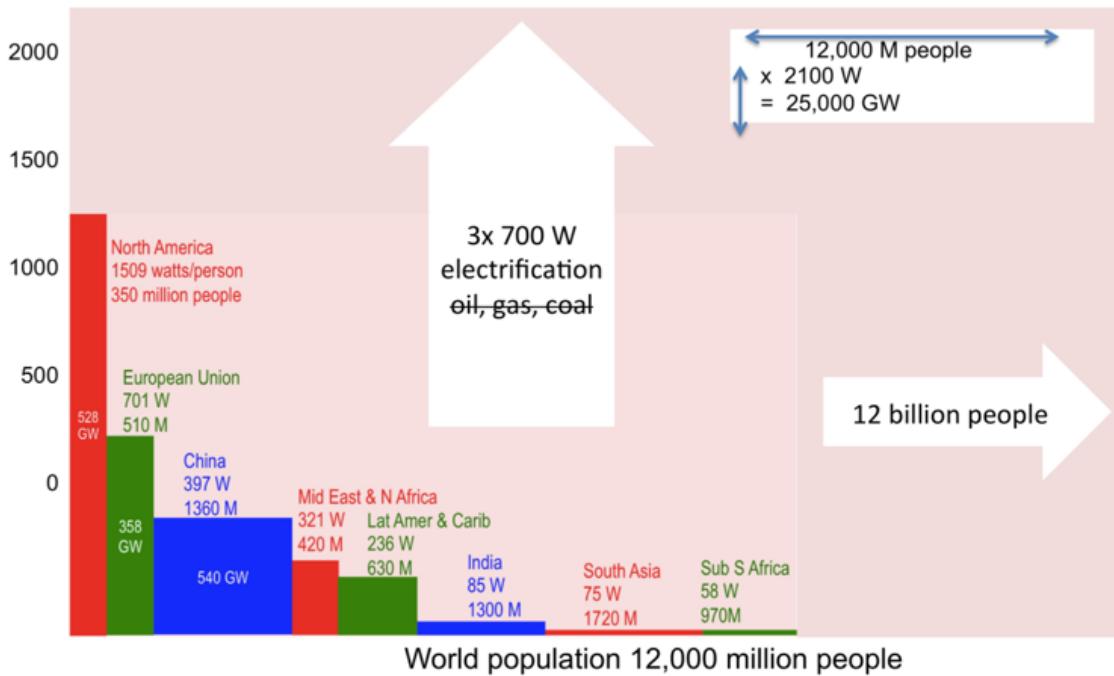
There is a great conflict between two of the most pressing problems of our time: poverty and climate change. To avoid global warming, the world needs to massively reduce CO₂ emissions. But to end poverty, the world needs massive amounts of energy. In developing economies, every kWh of energy consumed is worth roughly \$5 of GDP.

How much energy do we need? Just to give everyone in the world the per-capita energy consumption of Europe (which is only half that of the US), we would need to more than triple world energy production, increasing our current 2.3 TW by over 5 additional TW:



Devaney Fig 1.3: Regional distribution of electricity consumption

If we account for population growth, and for the decarbonization of the entire economy (building heating, industrial processes, electric vehicles, synthetic fuels, etc.), we need more like 25 TW:



Devaney Fig 1.4: Electricity consumption in a decarbonized world

This is the Gordian knot. Nuclear power is the sword that can cut it: a scalable source of dispatchable (i.e., on-demand), virtually emissions-free energy. It takes up very little land, consumes very little fuel, and produces very little waste. It's the technology the world needs to solve both energy poverty and climate change.

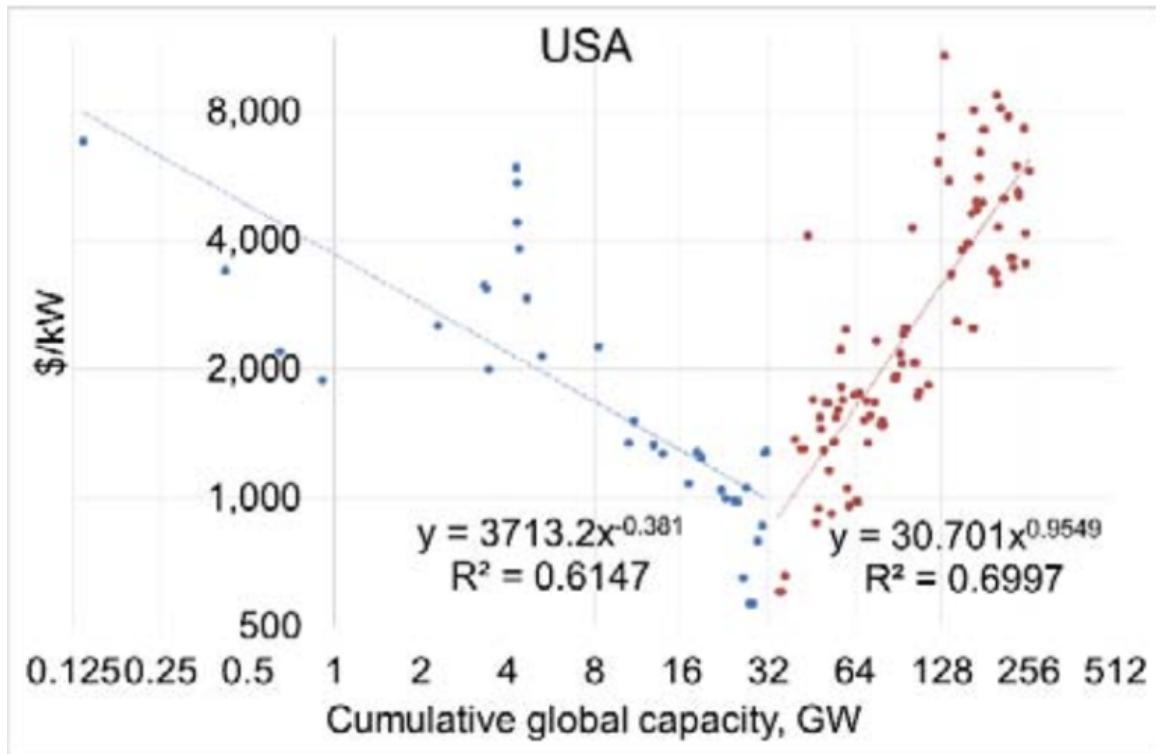
So why isn't it much bigger? Why hasn't it solved the problem already? Why has it been "such a tragic flop?"

Nuclear is expensive but should be cheap

The proximal cause of nuclear's flop is that it is expensive. In most places, it can't compete with fossil fuels. Natural gas can provide electricity at 7-8 cents/kWh; coal at 5 c/kWh.

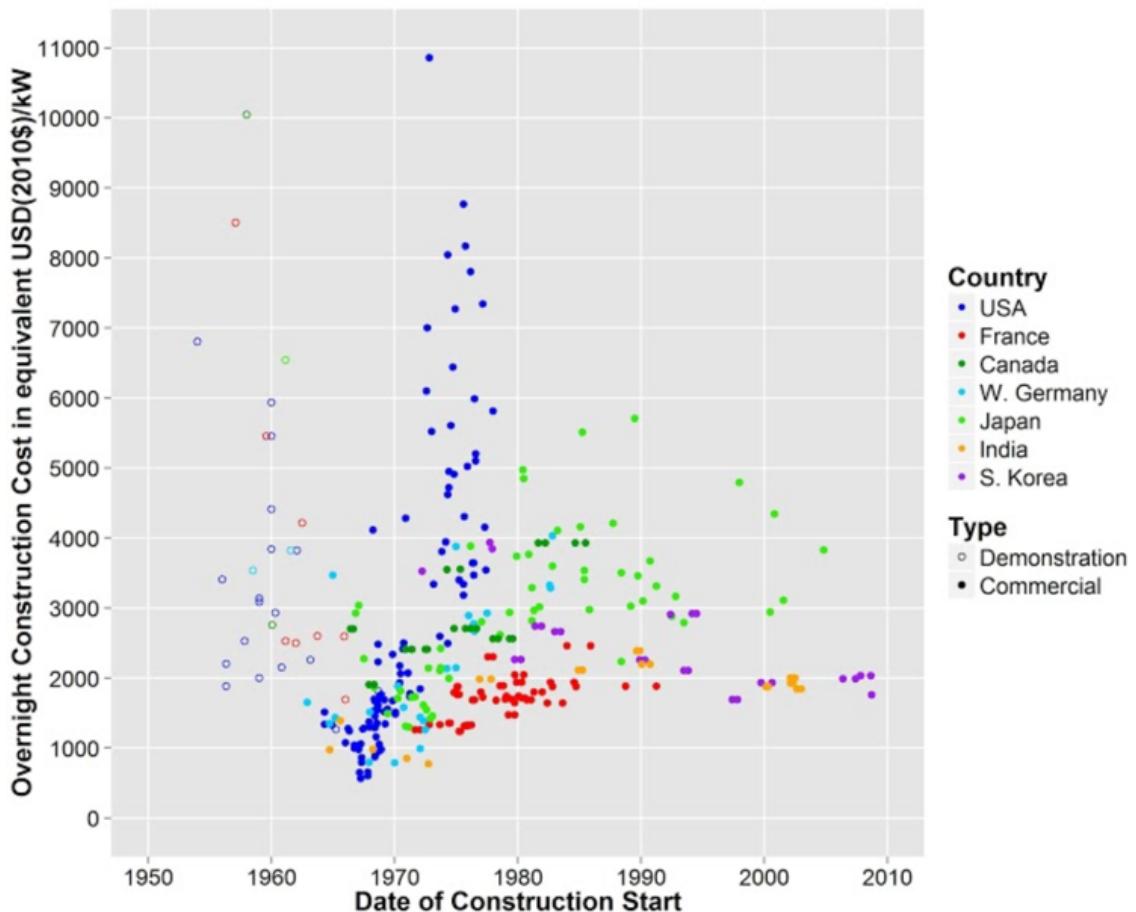
Why is nuclear expensive? I'm a little fuzzy on the economic model, but the answer seems to be that it's in design and construction costs for the plants themselves. If you can build a nuclear plant for around \$2.50/W, you can sell electricity cheaply, at 3.5-4 c/kWh. But costs in the US are around 2-3x that. (Or they were—costs are so high now that we don't even build plants anymore.)

Why are the construction costs high? Well, they weren't always high. Through the 1950s and '60s, costs were declining rapidly. A law of economics says that costs in an industry tend to follow a power law as a function of production volume: that is, every time production doubles, costs fall by a constant percent (typically 10 to 25%). This function is called the experience curve or the learning curve. Nuclear followed the learning curve up until about 1970, when it *inverted* and costs started *rising*:



Devanney Figure 7.11: USA Unit cost versus capacity. From P. Lang, "Nuclear Power Learning and Deployment Rates: Disruption and Global Benefits Forgone" (2017)

Plotted over time, with a linear y-axis, the effect is even more dramatic. Devanney calls it the “plume,” as US nuclear constructions costs skyrocketed upwards:



Devanney Figure 7.10: Overnight nuclear plant cost as a function of start of construction. From J. Lovering, A. Yip, and T. Nordhaus, "Historical construction costs of global nuclear reactors" (2016)

This chart also shows that South Korea and India were still building cheaply into the 2000s. Elsewhere in the text, Devanney mentions that Korea, as late as 2013, was able to build for about \$2.50/W.

The standard story about nuclear costs is that radiation is dangerous, and therefore safety is expensive. The book argues that this is wrong: nuclear can be made safe *and* cheap. It should be 3 c/kWh—cheaper than coal.

Safety

Fundamental to the issues of safety is the question: what amount of radiation is harmful?

Very high doses of radiation can cause burns and sickness. But in nuclear power safety, we're usually talking about much lower doses. The concern with lower doses is increased long-term cancer risk. Radiation can damage DNA, potentially creating cancerous cells.

But wait: we're exposed to radiation all the time. It occurs naturally in the environment—from sand and stone, from altitude, even from bananas (which contain radioactive potassium). So it can't be that even the tiniest amount of radiation is a mortal threat.

How, then, does cancer risk relate to the dose of radiation received? Does it make a difference if the radiation hits you all at once, vs. being spread out over a longer period? And is there anything like a “safe” dose, any threshold below which there is no risk?

Linear No Threshold

The official model guiding US government policy, both at the EPA and the Nuclear Regulatory Commission (NRC), is the Linear No Threshold model (LNT). LNT says that cancer risk is directly proportional to dose, that doses are cumulative over time (rate doesn't matter), and that there is no threshold or safe dose.

The problem with LNT is that it flies in the face of both evidence and theory.

First, theory. We know that cells have repair mechanisms to fix broken DNA. DNA gets broken all the time, and not just from radiation. And remember, there is natural background radiation from the environment. If cells weren't able repair DNA, life would not have survived and evolved on this planet.

When DNA breaks, it migrates to special “repair centers” within the cell, which put the strands back together within hours. However, this is a highly non-linear process: these centers can correctly repair breaks at a certain rate, but as the break rate increases, the error rate of the repair process goes up drastically. This also implies that dose rate matters: a given amount of radiation is more harmful if received all at once, and less if spread out over time. (In both of these details, I think of this as analogous to alcohol being processed out of the bloodstream by the liver: a low dose can be handled; but overwhelm the system and it quickly becomes toxic. One beer a night for a month might not even get you tipsy; the same amount in a single night would kill you.)

Radiotherapy takes advantage of this. When radiotherapy is applied to tumors, non-linear effects allow doctors to do much more damage to the tumor than to surrounding tissue. And doses of therapy are spread out over multiple days, to give the patient time to recover.

Devaney also assembles a variety of types of evidence about radiation damage from a range of sources. Indeed, his argument against LNT is by far the longest chapter in the book, weighing in at over 50 pages (out of fewer than 200). He looks at studies of:

- The nuclear bomb survivors of Hiroshima and Nagasaki
- The effects of radon gas
- Animal experiments in beagles and mice
- UK radiologists (tracked over 100 years)
- Radiation workers across fifteen countries
- Nuclear shipyard workers (using a closely matched control group of non-nuclear workers in the same yards)
- Areas with naturally high levels of background radiation from sources such as thorium-containing sand or radon: Finland; Ramsar, Iran; Guarapari, Brazil; Yangjiang, China; and Kerala, India
- The population of Washington County, Utah, 200 miles downwind of a nuclear test site in Nevada that was used in the 1950s
- The Chernobyl cleanup crew, including the guys who had to shovel chunks of core graphite off the roof of one of the buildings and toss them into the gaping hole from the explosion
- An incident in Taipei in which an apartment was accidentally built with rebar containing radioactive cobalt-60
- The women who hand-painted radium onto watch dials in the early 20th century (some of whom would *lick the brushes* to form a point)
- A 1950 trial that violated every conceivable standard of medical ethics by *injecting unknowing and non-consenting patients with plutonium*

In the last case, all of the patients had been diagnosed with terminal disease. None of them died from the plutonium—including one patient, Albert Stevens, who had been *misdiagnosed* with terminal stomach cancer that turned out to be an operable ulcer. He lived for more than twenty years after the experiment, over which time he received a cumulative dose of 64 sieverts, one-tenth of which would have killed him if received all at once. He died from heart failure at the age of 79.

The weight of all of this evidence is that low doses of radiation do not cause detectable harm. Little to no cancer, or at least far less than predicted by LNT, is found in the subjects receiving low doses, such as workers operating under modern safety standards, or populations in high-background areas (in fact, there is some evidence of a *beneficial* effect from very low doses, although nothing in Devanney's overall argument depends on this, nor does he stress it). In populations where some subjects did receive high doses, the response curves tend to look decidedly non-linear.

The other finding from these studies is that dose rate matters. This was the explicit finding of an MIT study in mice, and it is the unmistakeable conclusion of the case of Albert Stevens, who lived over two decades with plutonium in his bloodstream.

(At least, all this is Devanney's interpretation—it is not always the conclusion written in the papers. Devanney argues, not convincingly, that in many cases the researchers' conclusions are not supported by their own data.)

ALARA

Excessive concern about low levels of radiation led to a regulatory standard known as ALARA: As Low As Reasonably Achievable. What defines "reasonable"? It is an ever-tightening standard. As long as the costs of nuclear plant construction and operation are in the ballpark of other modes of power, then they are reasonable.

This might seem like a sensible approach, until you realize that **it eliminates, by definition, any chance for nuclear power to be cheaper than its competition.** Nuclear can't even innovate its way out of this predicament: under ALARA, any technology, any operational improvement, anything that reduces costs, simply gives the regulator more room and more excuse to push for more stringent safety requirements, until the cost once again rises to make nuclear just a bit more expensive than everything else. Actually, it's worse than that: it essentially says that if nuclear becomes cheap, then the regulators *have not done their job.*

What kinds of inefficiency resulted?

An example was a prohibition against multiplexing, resulting in thousands of sensor wires leading to a large space called a cable spreading room. Multiplexing would have cut the number of wires by orders of magnitude while at the same time providing better safety by multiple, redundant paths.

A plant that required 670,000 yards of cable in 1973 required almost double that, 1,267,000, by 1978, whereas "the cabling requirement should have been dropping precipitously" given progress at the time in digital technology.

Another example was the acceptance in 1972 of the Double-Ended-Guillotine-Break of the primary loop piping as a credible failure. In this scenario, a section of the piping instantaneously disappears. Steel cannot fail in this manner. As usual Ted Rockwell put it best, "We can't simulate instantaneous double ended breaks because things don't break that way." Designing to handle this impossible casualty imposed very severe requirements on pipe whip restraints, spray shields, sizing of Emergency Core Cooling Systems, emergency diesel start up times, etc., requirements so severe that it pushed the designers into using developmental, unrobust technology. A far more reliable

approach is Leak Before Break by which the designer ensures that a stable crack will penetrate the piping before larger scale failure.

Or take this example (quoted from T. Rockwell, "What's wrong with being cautious?"):

A forklift at the Idaho National Engineering Laboratory moved a small spent fuel cask from the storage pool to the hot cell. The cask had not been properly drained and some pool water was dribbled onto the blacktop along the way. Despite the fact that some characters had taken a midnight swim in such a pool in the days when I used to visit there and were none the worse for it, storage pool water is defined as a hazardous contaminant. It was deemed necessary therefore to dig up the entire path of the forklift, creating a trench two feet wide by a half mile long that was dubbed Toomer's Creek, after the unfortunate worker whose job it was to ensure that the cask was fully drained.

The Bannock Paving Company was hired to repave the entire road. Bannock used slag from the local phosphate plants as aggregate in the blacktop, which had proved to be highly satisfactory in many of the roads in the Pocatello, Idaho area. After the job was complete, it was learned that the aggregate was naturally high in thorium, and was more radioactive than the material that had been dug up, marked with the dreaded radiation symbol, and hauled away for expensive, long-term burial.

The Gold Standard

Overcautious regulation interacted with economic history in a particular way in the mid-20th century that played out very badly for the nuclear industry.

Nuclear engineering was born with the Manhattan Project during WW2. Nuclear power was initially adopted by the Navy. Until the Atomic Energy Act of 1954, all nuclear technology was the legal monopoly of the US government.

In the '50s and '60s, the nuclear industry began to grow. But it was competing with extremely abundant and cheap fossil fuels, a mature and established technology. Amazingly, the nuclear industry was not killed by this intense competition—evidence of the extreme promise of nuclear.

Then came the oil shocks of the '70s. Between 1969 and 1973, oil prices tripled to \$11/barrel. This should have been nuclear's moment! And indeed, there was a boom in both coal and nuclear.

But as supply expands to meet demand, costs rise to meet prices. The costs of both coal and nuclear rose. In the coal power industry, this took the form of more expensive coal from marginal mines, higher wages paid to labor who now had more bargaining power, etc. In the nuclear industry, it took the form of ever more stringent regulation, and the formal adoption of ALARA. Prices were high, so the pressure was on to get construction approved as quickly as possible, regardless of cost. Nuclear companies stopped pushing back on the regulators and started agreeing to anything in order to move the process along. The regulatory regime that resulted is now known as the Gold Standard.

The difference between the industries is that the cost rises in coal could, and did, reverse as prices came down. But regulation is a ratchet. It goes in one direction. Once a regulation is in place, it's very difficult to undo.

Even worse was the practice of "backfitting":

The new rules would be imposed on plants already under construction. A 1974 study by the General Accountability Office of the Sequoyah plant documented 23 changes "where a structure or component had to be torn out and rebuilt or added because of required changes." The Sequoyah plant began construction in 1968, with a scheduled completion

date of 1973 at a cost of \$300 million. It actually went into operation in 1981 and cost \$1700 million. This was a typical experience.

Bottom line: Ever since the '70s, nuclear has been stuck with burdensome regulation and high prices—to the point where it's now accepted that nuclear is inherently expensive.

Regulator incentives

The individuals who work at NRC are not anti-nuclear. They are strongly pro-nuclear—that's why they went to work for a nuclear agency in the first place. But they are captive to institutional logic and to their incentive structure.

The NRC does not have a mandate to increase nuclear power, nor any goals based on its growth. They get no credit for approving new plants. But they do own any problems. For the regulator, there's no upside, only downside. No wonder they delay.

Further, the NRC does not benefit when power plants come online. Their budget does not increase proportional to gigawatts generated. Instead, the nuclear companies themselves pay the NRC *for the time they spend reviewing applications*, at something close to \$300 an hour. This creates a perverse incentive: the more overhead, the more delays, the more revenue for the agency.

The result: the NRC approval process now takes several years and costs literally hundreds of millions of dollars.

The Big Lie

Devanney puts a significant amount of blame on the regulators, but he also lays plenty at the feet of industry.

The irrational fear of very low doses of radiation leads to the idea that any reactor core damage, leading to any level whatsoever of radiation release, would be a major public health hazard. This has led the entire nuclear complex to foist upon the public a huge lie: that such a release is virtually impossible and will never happen, or with a frequency of less than one in a million reactor-years.

In reality, we've seen three major disasters—Chernobyl, Three Mile Island, and Fukushima—in less than 15,000 reactor-years of operation worldwide. We should expect about one accident per 3,000 reactor-years going forward, not one per million. If nuclear power were providing most of the world's electricity, there would be an accident every few years.

Instead of selling a lie that a radiation release is impossible, the industry should communicate the truth: releases are rare, but they will happen; and they are bad, but not unthinkably bad. The deaths from Chernobyl, 35 years ago, were due to unforgivably bad reactor design that we've advanced far beyond now. There were zero deaths from radiation at Three Mile Island or at Fukushima. (The only deaths from the Fukushima disaster were caused by the unnecessary evacuation of 160,000 people, including seniors in nursing homes.)

In contrast, consider aviation: An airplane crash is a tragedy. It kills hundreds of people. The public accepts this risk not only because of the value of flying, but because these crashes are rare. And further, because the airline industry does not lie about the risk of crashes. Rather than saying "a crash will never happen," they put data-collecting devices on every plane so that when one inevitably does crash, they can learn from it and improve. This is a healthy attitude towards risk that the nuclear industry should emulate.

Testing

Another criticism the book makes of the industry is its approach to QA and the general lack of testing.

Many questions arise during NRC design review: how a plant will handle the failure of this valve or that pump, etc. A natural way to answer these questions would be to build a reactor and test it, and for the design application to be based in large part on data from actual tests. For instance, one advanced reactor design comes from NuScale:

NuScale is not really a new technology, just a scaled down Pressurized Water Reactor; but the scale down allows them to rely on natural circulation to handle the decay heat. No AC power is required to do this. The design also uses boron, a neutron absorber, in the cooling water to control the reactivity. The Advisory Committee on Reactor Safeguards (ACRS), an independent government body, is concerned that in emergency cooling mode some of the boron will not be recirculated into the core, and that could allow the core to restart. NuScale offers computer analyses that they claim show this will not happen. ACRS and others remain unconvinced.

The solution is simple. Build one and test it. But under NRC rules, you cannot build even a test reactor without a license, and you can't get a license until all such questions are resolved.

Instead, a lot of analysis is done by building models. In particular, NRC relies on a method called Probabilistic Risk Assessment: enumerate all possible causes of a meltdown, and all the events that might lead up to them, and assign a probability to each branch of each path. In theory, this lets you calculate the frequency of meltdowns. However, this method suffers from all the problems of any highly complex model based on little empirical data: it's impossible to predict all the things that might go wrong, or to assign anything like accurate probabilities even to the scenarios you do dream up:

In March, 1975, a workman accidentally set fire to the sensor and control cables at the Browns Ferry Plant in Alabama. He was using a candle to check the polyurethane foam seal that he had applied to the opening where the cables entered the spreading room. The foam caught fire and this spread to the insulation. The whole thing got out of control and the plant was shut down for a year for repairs. Are we to blame the PRA analysts for not including this event in their fault tree? (If they did, what should they use for the probability?)

In practice, different teams using the same method come up with answers that are orders of magnitude apart, and what result to accept is a matter of negotiation. Probabilistic models were used in the past to estimate that reactors would have a core damage frequency of less than one in a million years. They were wrong.

Later, during construction, a similar issue arises. The standard in the industry is to use "formal QA" processes that amount to paperwork and box-checking, a focus on following bureaucratic rules rather than producing reliable outcomes. Devaney saw the same mentality in US Navy shipyards, which produced billion-dollar ships that don't even work. Instead, the industry should be more like the Korean shipyards, which are able to deliver reliably on schedule, with higher quality and lower cost. They do this by inspecting the work product, rather than the process used to create it: "test the weld, not the welder." And they require formal guarantees (such as warranties) of meeting a rigorous spec given up front.

Competition

Finally, Devaney laments the lack of real competition in the market. He paints the industry as a set of bloated incumbents and government labs, all "feeding at the public trough." For instance:

One the biggest labs is Argonne outside Chicago. At Argonne, they monitor people going in and out of some of the buildings for radiation contamination. The alarms are set so low that, if it's raining, incoming people must wipe off their shoes after they walk across the wet parking lot. And you can still set off the alarm, which means everything comes to a halt while you wait for the Health Physics monitor to show up, wand you down, and pronounce you OK to come in. What has happened is that the rain has washed some of the naturally occurring radon daughters out of the air, and a few of these mostly alpha particles have stuck to your shoes. In other words, Argonne is monitoring rain water.

Nuclear incumbents aren't upset that billions of dollars are thrown away on waste disposal and unnecessary cleanup projects—they are getting those contracts. For instance, 8,000 people are employed in cleanup at Hanford, Washington, costing \$2.5B a year, even though the level of radiation is only a few mSv/year, well within the range of normal background radiation.

What to do?

Devaney has a practical alternative for everything he criticizes. Here are the ones that stood out to me as most important:

Replace LNT with a model that more closely matches both theory and evidence. As one workable alternative, he suggests using a sigmoid, or S-curve, instead of a linear fit, in a model he calls Sigmoid No Threshold. In this model, risk is monotonic with dose (there are no beneficial effects at low doses) and it is nonzero for every nonzero dose (there is no "perfectly safe" dose). But the risk is orders of magnitude lower than LNT at low doses. S-curves are standard for dose-response models in other areas.

Drop ALARA. Replace it with firm limits: choose a threshold of radiation deemed safe; enforce that limit *and nothing more*. Further, these limits should balance risk vs. benefit, recognizing that nuclear is an alternative to other modes of power, including fossil fuels, that have their own health impacts.

Encourage incident reporting, on the model of the FAA's Aviation Safety Reporting System. This system enables anonymous reports, and in case of accidental rule violations, it treats workers more leniently if they can show that they proactively reported the incident.

Enable testing. Don't regulate test reactors like production ones. Rather than requiring licensing up front, have testing monitored by a regulator, who has the power to shut down test reactors deemed unsafe. Then, a design can be licensed for production based on real data from actual tests, instead of theoretical models.

We could even designate a federal nuclear testing park, the "Protopark," in an unpopulated region. The park would be funded by rent from tenants, so that the market, rather than the government, would decide who uses it. Tenants would have to obtain insurance, which would force a certain level of safety discipline.

Align regulator incentives with the industry. Instead of an hourly fee for regulatory review, fund the NRC by a tax on each kilowatt-hour of nuclear electricity, giving them a stake in the outcome and the growth of the industry.

Allow arbitration of regulation. Regulators today have absolute power. There should be an appeals process by which disputes can be taken to a panel of arbitrators, to decide whether regulatory action is consistent with the law. City police are held accountable for their use and abuse of power; the nuclear police should be too.

Metanoeite

At the end of the day, though, what is needed is not a few reforms, but “metanoiae”: a deep repentance, a change to the industry’s entire way of thinking. Devanney is not optimistic that this will happen in the US or any wealthy country; they’re too comfortable and too able to fund fantasies of “100% renewables.” Instead, he thinks the best prospect for nuclear is a poor country with a strong need for cheap, clean power. (I assume that’s why his company, [ThorCon](#), is building its thorium-fueled molten salt reactor in Indonesia.)

Again, all of the above is Devanney’s analysis and conclusions, not necessarily mine. What to make of all this?

I’m still early in my research on this topic, so I don’t yet know enough to fully evaluate it. But the arguments are compelling to me. Devanney quantifies his arguments where possible and cites references for his claims. He places blame on systems and incentives rather than on evil or stupid individuals. And he offers reasonable, practical alternatives.

I would have liked to see the nuclear economic model made more explicit. How much of the cost of electricity is the capital cost of the plant, vs. operating costs, vs. fuel? How much is financing, and how sensitive is this to construction times and interest rates? Etc.

A few important topics were not addressed. One is weapons proliferation. Another is the role of the utility companies and the structure of the power industry. Electricity utilities are often regulated monopolies. At least some of them, I believe, have a profit margin that is guaranteed by law. (!) That seems like an important element in the lack of competition and perverse incentive structure.

I would be interested in hearing thoughtful counterarguments to the book’s arguments. But overall, *Why Nuclear Power Has Been a Flop* pulls together academic research, industry anecdotes, and personal experience into a cogent narrative that pulls no punches. Well worth reading. [Buy the paperback on Amazon](#), or [download a revised and updated PDF edition for free](#).

Tales from Prediction Markets

This is a linkpost for <https://misinfofounderload.substack.com/p/tales-from-prediction-markets>

Prediction markets are fun, at least if you're making money. I've only been into them for a few months, but have already collected a bunch of interesting tales. Note: I may have been involved with some of these, but I'm telling these tales from a third person perspective.

One general point: all of these took place on Polymarket, a crypto prediction market. You can track which accounts place each bet, and so you can see their history of bets, but you can't tie it to an actual person unless they've chosen to identify themselves. You can look at the bet history at [Polymarketwhales.info](https://polymarketwhales.info), although there's a ton of bets so it's easier if you know what you're looking for.

The Tesla market. Polymarket had a market on whether Tesla would announce a Bitcoin purchase by Mar 1, 2021. On January 27, an unknown user bet \$60k on Yes. This was their only trade on the site, before or after. They won \$180k, or 120k in profit. Odds are pretty good it was an insider. Is this insider trading? I asked Matt Levine but he didn't respond. Anyway, there's another user that lost \$242k betting that Tesla would not announce a Bitcoin purchase. This user is affectionately called the "Tesla whale" on the Polymarket discord. They're also notable for losing \$92k on the super bowl the day before Tesla made the announcement, and they get honorable mention for having lost the most money on the 100 million vaccine market: see below. As of this writing, the Tesla whale is down nearly \$500k.

Watch out for slippage: there was a market on whether Joe Biden would still be president as of Mar 1, 2021. Someone owned around 200k shares of Yes. The market price was very close to \$1 each on the morning of Mar 1st, and they apparently decided to sell all their shares instead of waiting for it to resolve; however, there wasn't enough liquidity to sell them all at market price, and they ignored the warning about the slippage the order would incur. Their order ended up executing at an average price of 2 cents, and someone else scooped up those cheap shares a minute later, spending \$1k to make \$155k; talk about being in the right place at the right time! The initial user ended up with a total loss of \$156k on this market. However, even taking that into account, as of this writing they're still up \$175k, so don't feel too bad for them. (Note: it's possible they were trying to sell No shares to make a few cents and accidentally clicked on the sell Yes side; without confirmation from the user we can't know what they were intending. Regardless, if you've got hundreds of thousands of dollars at stake, double check before pressing buttons. This isn't the only fat finger that's happened, but it's the biggest.) Polymarket has since added a larger warning in red for trades that move the market more than a few cents.

Hashmasks. There was a market on whether the Hashmasks volume ranking on Opensea would be #1 as of Feb 28, 2021. Someone accumulated over 200k Yes shares, then took out a "flash loan" and purchased a Hashmasks from themselves for 130k ETH (worth over \$100 million at the time). Unfortunately for them, Opensea doesn't count sales done directly through their smart contract, only one initiated through their website. The market resolved No.

CO2: Polymarket had a market on whether the level of CO2 reported by the Mauna Lua observatory would be over a specific value on a specific day. The number gets reported on the main page once a day, but someone found a page with the hourly data and figured out how to average together the data points to predict the daily number. They made around \$10k on the first market and around \$30k on the second market. After that, Polymarket stopped making new CO2 markets.

Vaccines: There was a market on whether there would be 100M 1st doses of covid vaccines reported by the CDC by 12PM EST on April 1st, 2021. The market was made before daylight savings time went into effect, and most people had assumed that Apr 1 data would not be posted by that time. Around two weeks before the deadline, the CDC started posting data earlier - in the 12:50:12:55 range ET. This was before 12 EST, so people realized that, if Apr 1 data was posted at that time, the market would technically include it. A week later, the CDC started releasing later in the day, and the market crashed hard. In the end, the CDC released April 1 numbers too late to count, and the numbers were just short of 100M anyway. We hit 100M on Apr 2nd, after 1PM. This market had \$7 million in volume, and very nearly came down to what time the CDC happened to post the data on the last day.

Many people seemed to confuse the total doses given and what the market was actually about, which was first doses. On the day that total doses hit 100M, the Tesla whale spent over \$200k on Yes at an average price of around 75 cents. There's been speculation that they thought it was over because total doses had hit 100M, but that can't be known for sure. However, multiple people showed up in the Discord that day asking for the market to be resolved since it hit 100M.

Souljaboy: There was a market on how many times Souljaboy would tweet during a given week. The way these markets are set up, they subtract the total number of tweets on the account at the beginning and end, so deletions can remove tweets. Someone went on his twitch stream, tipped a couple hundred dollars, and said he'd tip more if Soulja would delete a bunch of tweets. Soulja went on a deleting spree and the market went crazy. Multiple people made over 10k on this market; at least one person made 30k and at least one person lost 15k.

Monastery and Throne

Thinking about the impact of Rationalist writing after a small blog post made a big difference. [Cross-posted from Putanumonit.](#)

Changing British Minds

Even a year after the fact, it's difficult to compile an accurate timeline of the UK Government's thinking on COVID and how it changed in spring 2020.

The first COVID case in the UK was confirmed on January 31. Spread continued throughout February, and the British government published its first [Coronavirus action plan](#) on March 3. It focused on the tracing and isolation of known cases while saying that it "*will aim to minimise the social and economic impact*" and that "*it may be that widespread exposure in the UK is inevitable*".

The decisions ultimately came down to PM Boris Johnson and his chief adviser, Dominic Cummings. At the beginning of March, according to the [Sunday Times](#), Cummings leaned towards a strategy of herd immunity: having the majority of the UK's population catch the virus throughout the summer months to protect the economy and put Britain in the clear by winter. [The PM's office denied](#) this claim and alleged that the quotes attributed to Cummings by the Sunday Times were fabricated. In any case there were clearly heated debates within Johnson's inner circle on the costs and benefits of mitigation.

Whatever the stance was in the first two weeks of March, by March 12 things began to shift. The strategic advisory group of experts warned that the UK would suffer 500,000 deaths if the virus was left unmitigated, and Cummings himself [began pushing vigorously for a shut down](#). The UK Government began implementing mitigation measures on March 13, and on March 23 [announced a full lockdown](#) of the entire nation. The UK spring COVID wave peaked at 70 daily cases per million, [significantly below the US](#) and many European nations.

Oh, [there's one other thing](#) that may have changed the mind of Dominic Cummings and the fate of Britain in early March.

Putanumonit Jacob @yashkaf · Mar 18

Figuring out how to manage COVID individually turned out to be much easier for Rationalists than figuring out how to convince others to follow along. And as my post implies, what convinces a Rationalist of some COVID policy is very different from what would convince most people.

2 13 13

Dominic Cummings @Dominic2306 · Mar 20

Figuring out how to win the Brexit referendum proved MUCH easier than persuading critical people of what to do. Ditto in constitutional crisis july-december2019. And similarly March & Sep 2020 re covid

1 2 2

Dominic Cummings @Dominic2306 · Mar 20

And thanks for your Seeing Smoke blog, it, & ScottA on 2 march, helped me & some others in no10 realise we were going terribly wrong

3 10 10

Post Impact

The four hours I spent writing [Seeing the Smoke](#) on a whim were [more impactful](#) than projects I spent months of my life on. I calculated that if just 1% of the 20,000 people who read that post the week it came out avoided catching COVID and infecting others during a period when the curve was steepening, the post probably saved several lives. To learn that my post impacted policy for 67 million Britons is a whole other level of bonkers.

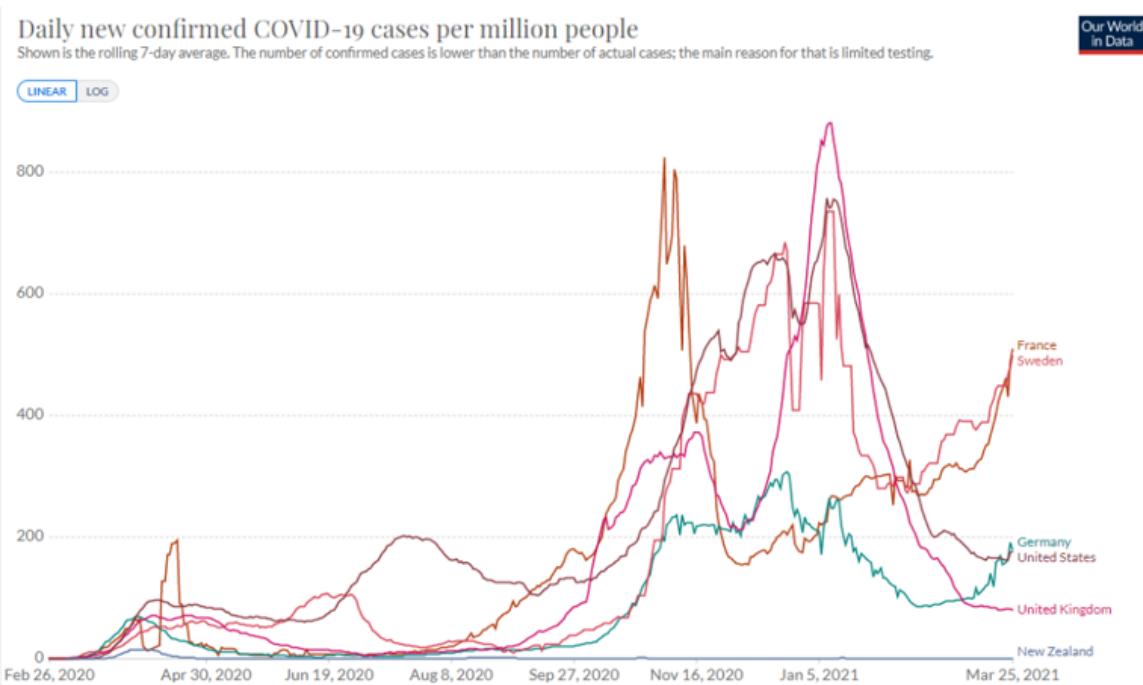
I'm struggling a bit to make sense of it all. My immediate motivation for writing is simply to get the thoughts out of my head, along with a [cycle of anxiety](#) about disappointing my readers if I go too long between posts. Aside from the satisfaction of having my post out there, my main reward for writing is making friends — even in 2020 I talked online or in person with dozens of people who I connected with in part because of the blog. On occasion I find out that some famous person mentioned one of my posts on a podcast or something. It's always a pleasant surprise, but I'm never sure what to do with this. I don't write to become famous and change the world. And yet somehow...

My first thoughts when Cummings mentioned that my post influenced British COVID policy was: *holy shit!* My second: *if I had known you will read it, I would have written a better post!* I would have done more research. I would have thought more about the impact of government interventions. I would have made some quantified predictions and recommended specific policies.

None of that would have made the post any better or more useful. In late February I didn't know of a better policy beyond "shut everything down including travel and take some time to think this through". By the middle of March I was convinced of the [Hammer and Dance](#)

approach: suppress the pandemic by instituting complete shut downs (including travel) at the first sign of $R_0 > 1$, since the timeliness of lockdown is the most important factor in their success. Then re-open equally quickly to maximize the proportion of time in non-lockdown over the 1-2 years it would take to make vaccines. In my understanding, this is basically the strategy that Australia and New Zealand successfully executed.

But on the day *Hammer and Dance* was published the UK already had 10 times New Zealand's per capita case rate and R_0 north of 2. As a result, the lockdown in the UK dragged into the summer instead of lasting only 3-4 weeks. The government had to make up for the economic damage with schemes like "[eat out to help out](#)" (that's what she said) in the summer, then suffered a second COVID wave in the fall that inflicted all the damage the spring lockdown was supposed to avoid and likely gave rise to the more dangerous [B1.1.7 variant](#).



Source: [OurWorldInData.org](https://ourworldindata.org/covid-19).

Even with a year of hindsight and my current understanding of political pressures, lockdown fatigue, and equilibrating control systems, I'm not confident whether it was good or bad for the UK to lock down in March 2020. In the moment I was surely no wiser.

Would I have made a bigger impact by focusing more on COVID research and running the numbers? In March I was busy [reading up on hydroxychloroquine](#), which turned out to be mostly a waste of time. In June [I was doing math to convince](#) people that it's safe to meet up with small groups of friends, but all the people who hailed my genius in February now called me irresponsible and arrogant. It turns out that they needed permission to start worrying about COVID in the spring, but didn't want permission to *stop* worrying in the summer.

In December [I wrote a follow up post](#) about the B1.1.7 variant, predicting a relatively high chance of massive infection wave in the spring. Judging by the recent uptick in cases I got the timing and importance of new strains right, as well as the reluctance to lock down. But I downplayed the possibility of vaccines having a big impact, which they have. William Kiely [got the vaccine math spot-on](#) in the comments, but I don't know if many people read that or cared. The main impact of that post may have been to get people [into an options trade](#) that ended up losing all the money.

So: I'm not a COVID expert, and playing one on the internet is not a reliable way for me to have a major positive impact on the world. But the original post wasn't really about COVID. It was public psychology, how people act as if appearing weird is worse than any disease, and that this is true not just of media consumers but also of the journalists and "experts" who produce it. What impact could I have through this deep and novel understanding of how everyone thinks?

Nudgerism

Thinking that one understands how the public thinks and how to shape it invites the worst sort of hubris.

While the idea of benevolent rulers shaping the public mind goes back at least to Plato, it's most recent incarnation goes back to the 2008 book [Nudge](#) by Richard Thaler and Cass Sunstein. The book repeats the [Kahnemanian](#) critique of "homo economicus" and gives some common sense advice that people knew long before Kahneman was born, such as that shoppers are more likely to pick the items displayed at eye level on the shelf. It then pivots to a defense of "[libertarian paternalism](#)" — a doctrine of shaping the public mind through control of the public's choices by the government.

"Libertarian paternalism" is really a rebranding of "[manufacturing consent](#)" for democracies with some fuzzy anecdotes about cab driver tips and a patina of scientism. Sunstein himself is quite supportive of [manufacturing consent by direct means](#), [free speech](#) be damned. In any case, *Nudge* landed Sunstein a job as the administrator of the Office of Information and Regulatory Affairs under Obama, where presumably he proceeded to regulate and inform the nation according to his theories.

In 2020, every terrible decision made by the experts in charge was justified by appealing to the effects on public psychology. I don't know if this was inspired or instigated by Sunstein directly, but I like to call this phenomenon [nudgerism](#). I despise nudgerism.

Since one can invoke a "bias" for any side of any decision, you'd expect the universal application of nudgerism to have a 50% hit-rate. Yet somehow it seemed to do much worse than chance, probably because any non-terrible decision could be justified by actual evidence instead of the appeal to psychology.

COVID nudgerism [started with Sunstein himself](#), of course. Other "experts" downplayed the risk of COVID to "prevent panic". Then they told people not to wear masks or get tested to avoid "[a false sense of security](#)". [Scrubbing surfaces](#) was still recommended, even though unlike masks and testing it probably doesn't do shit for COVID. "Experts" said that challenge trials would reduce trust in the vaccine among normal people; [normal people turned out to support challenge trials by an overwhelming 75-90% majority](#). "Experts" said that delaying the second dose of the vaccine or allowing the AstraZeneca shot would increase vaccine hesitancy. Britain delayed the second dose, is giving everyone AstraZeneca, and has [the least vaccine hesitancy in the world](#). Probably the British government [reads Marginal Revolution](#) along with Putanumonit, and ignoring Sunstein.



Technology & Ideas

The Cognitive Bias That Makes Us Panic About Coronavirus

Feeling anxious? Blame "probability neglect."

By Cass R. Sunstein

29 February 2020, 02:09 GMT+8

I don't think all these nudgers are wholly cynical about this. I think their inside view finds their theories of public psychology reasonable and valid. And while I will mock and deride them in any opportunity, I doubt I'd do better if I tried it myself. Not in small part because I'm an alien.

Coordinating Social Reality

A big theme in my writing since *Seeing the Smoke* has been coming to terms with the [gargantuan gap](#) between the way I think and the way most people do. It may even be wrong to use the word "think" for both. [Michael Vassar says](#) that what Rationalists call "thinking" is treated by most people as a rare technical ability ("design thinking") that normal people can only pretend to do. What they call "thinking" we call "being depressed and anxious". This sounded crazy [when I first heard it](#), but the more I mulled it over the more it made sense and explained much of what has been happening in the last year.

Social reality is what is normal, accepted, cool, [predictable, expected](#), rewarded, agreed upon. Physical reality is what is out there determining the outcomes of physical experiments, such as whether you get COVID or not if you wear a mask. When Rationalists say "thinking" they usually mean something like "using your effortful system 2 to determine something about physical reality". It's what I try to do when writing posts about COVID. Swimming in social reality is best done on feeling and intuition, not "thought".

My experience is spending perhaps 97% of my time in social reality, swimming along with everyone else. 3% of the time I [notice some confusion](#), an unexpected mismatch between my predictions and what physical reality hits me with, and try to *think* through a solution. 3% is enough to notice the difference between the two modes and to be able to switch between them on purpose.

I don't think that this experience is typical.

With Vassar in mind, my best guess of the typical experience is being in social reality 99.9% of the time. The 0.1% are extreme shocks, cases when physical reality kicks someone so far off-script they are forced to confront it directly. These experiences are extremely unpleasant, and processing them appears as "depression and anxiety". One looks at the first opportunity to dive back into the safety of social reality, in the form of a communal narrative that "makes sense" of what happened and suggests an appropriate course of action.

I think this explains why so many people don't seem to notice or care that even institutions they consider to be "on their side", [like the CDC](#) and the New York Times for the educated progressive tribe, are wrong or lying all the time. People look to those sources not for "truth" about physical reality but for **coordination of social reality**. The CDC's job is [to tell other](#)

[institutions](#) which policies they can implement and not get blamed for, not which policies will keep their clients healthy. [People read the Times](#) not to find out what happened where and when, but to find out [who is to be comforted and who afflicted](#). People just want to be on the same page as their peers.

Seeing the Smoke came out during the 0.1% of the time when physical reality was manifesting and the institutions of social reality hadn't reacted adequately yet (in [spinning it into the narrative](#)). By the summer, social reality reasserted itself. Whether someone was masking or not, locking down or protesting, eating fish tank cleaner or Lysol-sprayed Uber Eats was due mostly to their tribe membership, not to a physical model of the virus. No one cared about my [microCOVID](#) calculations.

Zvi, who is better than me at *thinking* about many things including COVID, [explains why being good at thinking](#) doesn't mean one could be put in charge and change the world for the better. The entire post is worth reading, especially for young Rationalists just coming to grips with how non-Rationalist the rest of the world is. If I had to summarize it in one sentence: *being correct is not the same as being good at coordinating social reality, and what those "in charge" really do is the latter.*

The Bayesian Monastery

Zvi lays out a model of how Rationalists like him can influence policy:

Anna Salamon suggested a model of a rising sanity water line, but in the sense that this makes it harder to stay above water and thus directly sane. There's a small and decreasing number of people who are still capable of synthesizing information and creating new hypotheses and interpretations.

Then there's those who are mostly no longer capable of doing that, things got too complicated and weird, and they can't keep up, but they can read the first group and meaningfully distinguish between people claiming to be in it, and between their individual claims, ask questions and help provide feedback. To them, the first group is legible. This forms a larger second group that can synthesize the points from the first group, and turn it into something that can be read as an emerging new consensus, which in turn can be legible to a third much larger group.

This third group can then be legible to the general public slash general elites, who learn that this is where good new ideas come from. Then the Responsible Authority Figures can feel under public pressure, or see what the emerging new ideas are, and run with the ball from there, and the loop continues.

The filtering process also acts as a selection for feasibility, as the second layer picks up things from the first that it thinks it can present legibly to the third, and so on.

This model doesn't mean that self-identified Rationalists are always in the first group. Most of what I write, especially about COVID, is second or third layer (or just nonsense). I think that Rationality gives people two important things: the tools to evaluate original thinkers without relying on mere credentials, and the *permission* to occasionally shoot for first-level insight themselves. As a community, we are closer to the surface of the sanity waterline than most and thus, by necessity, farther from political power and institutional authority.

Byrne Hobart makes the same point [in a post about how Rationalists got COVID right and early](#):

This puts the rationalists in a uniquely prosocial position. They're a sort of distributed, mostly open-source monastic order, spending a lot of time contemplating the world and passing down important observations, but less time directly interacting with it. The influence of people who *read* rationalist blogs, but don't self-identify as rationalists, is

quite wide—the blogs are very widely followed in technology circles, and anecdotally have a large audience in the more quantitative branches of finance.

Byrne goes on to say that “identifying as a Rationalist is a losing move”, [but I think that he presupposes](#) that everyone is playing the same game. Joining a monastic order is a “losing move” if your goal is to inherit titles and command knights, but the life of a monk has much to recommend it over the life of the medieval court and battlefield. The pursuit of wisdom and the pursuit of power are usually at odds. Identifying as a Rationalist is a small way to nudge yourself (heh) toward the former.



A typical LessWrong meetup

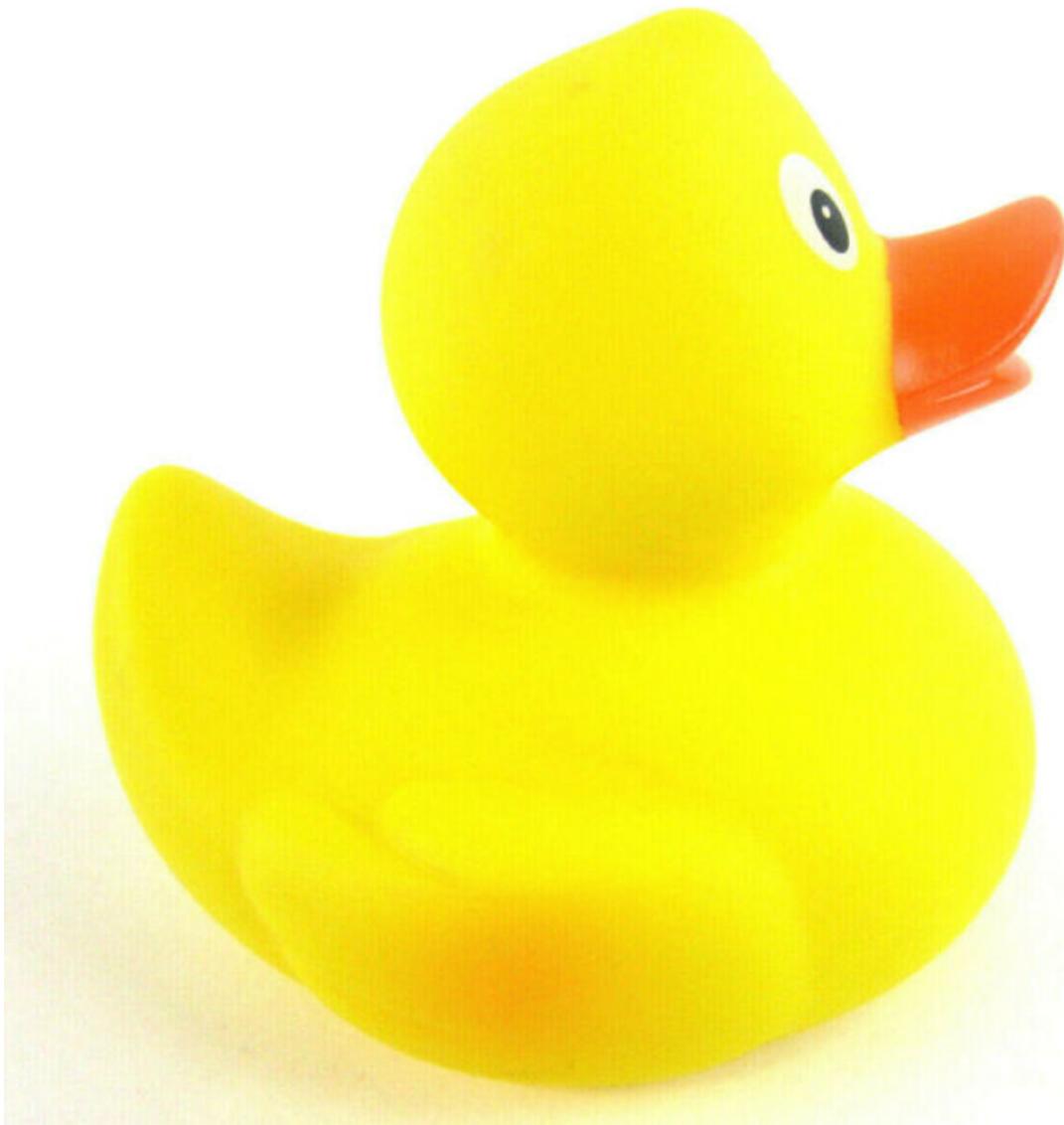
In March 2020 Dominic Cummings was in a unique position to bridge the many layers between first-level analysis and “Responsible Authority Figures” by himself. He could take word from the monastery directly to the throne room. I just happened to be the monk that was on thinking duty that day.

But that doesn’t mean that I should now abandon the monastery and fancy myself a vizier. In general, I try to write about enduring topics like [statistics](#) and [romance](#) and [how our brains work](#), not about breaking news. Instead of trying to coordinate social reality for the masses, I try to help people become slightly more independent of it. To avoid the mind-kill of [political polarization](#), to get in touch with their own desires when deciding [what to spend money on](#) and [whom to date](#), to [spend some time in our monastery](#).

I hear individually from people that my writing is impacting them, and as long as those DMs come I’m content. I think my writing has some broader influence, but by the time it is passed down to actual *influencers* I would not be credited with it. And that’s fine! I’m really not looking for credit, and now that I got some for influencing the UK lockdown decision I don’t even know if it’s deserved or if the decision was the right one. I recommend that if anyone in power wants to hire a Rationalist advisor they do so secretly, [and pay them](#) based on the calibrated accuracy of their predictions only.

In the meantime, I'll try to keep writing as if no one important will ever read it at all. Otherwise, the temptation grows to climb the [simulacra levels](#) away from reality, to signal loyalties and "nudge" the public and play 4D chess on a backgammon board. And that's not how we do it in the monastery.

How to Play a Support Role in Research Conversations



There's a programming technique where you keep a rubber duck on your desk, and when you run into a confusing bug, you try to explain it to the duck. This turns out to be surprisingly useful for resolving bugs.

In research conversations, it is often very useful to have someone serve as an upgraded rubber duck. This is a support role: the rubber duck's job is not to solve the problem, but to help their partner sort out their own thoughts. Playing this support role well is a skill, and if you're good at it, you can add value to other researchers in a very wide range of contexts; you can be a general-purpose force multiplier. This post contains a handful of tips for how to add value this way, beyond what a literal rubber duck has to offer.

Figure Out The Picture

My partner is drawing a picture in their head. I want to accurately copy that picture into my head. This is the core of the technique: in order to build an accurate copy in *my* head, my partner will need to flesh out the picture in their *own* head.

For sake of discussion, let's say the conversation is part of a [conjecture workshop](#). The goal is to take some fuzzy intuition and turn it into a mathematical conjecture. When playing the support role, I'm trying to build a mathematical conjecture in my head which matches the picture in my partner's head. By asking questions to fill out my picture of the conjecture, I push my partner to translate their fuzzy intuitions into math.

(We'll use conjecture workshop examples throughout this post, but the ideas generalize beyond just math.)

Some useful lines:

- Can you give an example?
- Can you give 2-3 examples which are as different as possible?
- Can you give an anti-example, i.e. a case where this would not apply?
- Here's the picture in my head so far. <explain> Does that match what's in your head? Am I missing anything important?
- Let me repeat back what I understood from what you just said. <explain> Did I miss anything?
- We've been using the analogy of X. How does Y translate into that analogy?
- It seems like you're trying to point to X. Is that right?
- Here's a few different things which I think you might be trying to point to. <explain> Is one of these what you're saying? Do none of them match what you're trying to say?
- Am I understanding your framing correctly?

Side note: since this all relies on building out a picture in my head, it might be useful to *write down* some parts of that picture, in case my own working memory fills up.

Don't Draw The Picture

I don't want to accidentally draw anything in my copy which isn't in my partner's picture. And I extra-especially don't want to *overwrite* their picture.

For instance:

Partner: I want to say something like [...]

Me: Ah, this is just like Person's Theorem! We formulate it as [...]

The problem with this is that we're now talking about what Person's Theorem says, which may or may not actually match the fuzzy thing in my partner's head. And it's a lot easier to latch onto a theorem which has already been figured out, than to figure out the right way to formulate fuzzy ideas. *While* having this kind of conversation, the goal is to fill out the picture in my partner's head, not to replace it with some other picture.

But do make a note to talk about Person's Theorem *later!* It may still be useful to just use the Person's Theorem picture later on, even if bringing it in right *now* would undermine the conversation.

(Why is it so important to flesh out the original picture, rather than adopt a different one? Well, my partner expects their intuitive picture to apply/generalize in some useful ways; that's why it's interesting in the first place. If we replace it with a new picture which may-or-may-not match, then the intuitive reasons to expect it to apply/generalize usefully may-or-may-not carry over, especially if they're not very legible. It's the same reason why we want to [avoid ad-hoc mathematical definitions](#).)

Depending on how firm a grasp your partner has on their own idea, you may need to exert more or less effort to avoid accidentally overwriting their picture. Anchoring can be an issue - if you say "I'm imagining this as a set of vectors and a distance metric...", then it may be hard to think up a new frame later. On the other hand, if your partner already has a firm enough grasp to say "no, that's not quite right", then figuring out exactly how set-of-vectors-and-distance-metric *isn't* right can be a very useful step. So there's a balance to be struck between making suggestions which fail-to-match your partner's picture in instructive ways, vs not accidentally overwriting the picture.

Some sometimes-useful lines:

- I'm imagining this as X. In what ways does this *not* match your picture?
- I'm currently thinking of this as X, but I'm not sure that's exactly what you're pointing to. Can we come up with an example which does *not* look like X?
- This makes me think of the setting for Person's Theorem, where we assume X, Y, Z. How well does that match what you're imagining? How does your thing differ?

As always, remember that your partner may just want to frame things differently, but may not know to call their vague feelings of unease a "framing problem". And make sure to give your partner space to explain their picture in their own words starting from their own reference points - "how does this differ from X?" is sometimes useful, but it's not something you should be asking constantly.

The above lines are also useful for "offering help" - i.e. if you think you know the next thing your partner "should" add to their mental picture. You can say something like "It sounds like you're trying to say X; how well does that match what's in your head? Where does it differ?". If your partner gets stuck, this is a good way to help while still being careful not to overwrite the picture. Again, though, try not to do this too often.

Don't Solve The Problem

Finally, the "[hold off on proposing solutions](#)" rule is in force for the entire conversation.

These sorts of conversations usually revolve around how to frame some problem, theory, conjecture, hypothesis, etc. Trying to solve the problem, prove the conjecture, test the hypothesis, etc, will most likely lock in whatever picture you and your partner currently have, and prematurely end the drawing process. At that point, you have moved to an entirely different conversation.

That said, it may still be useful to talk about certain "solution" techniques in order to clarify the problem. Some sometimes-useful lines:

- It sounds like the sort of problem which could potentially be solved by X. Does that match your intuition?
- If I'm understanding this correctly, then if we did X we'd expect to see Y. Is that right?
- It sounds like this idea would be implied by X?

When using these, be careful to emphasize that the point of bringing it up is to better understand the picture, not to solve the problem or test the theory. If your partner starts to dive into the details of the "solution"/"test", gently remind them that you haven't yet built a complete picture of the problem/theory in your head.

Summary

My partner is drawing a picture in their head. I want to accurately copy that picture into my head. The process of copying the picture into my head will hopefully push my partner to fill

out the details of the picture, in much the same way that explaining a programming problem to a rubber duck helps flesh out one's own picture of the problem. And unlike a rubber duck, a human can add a lot more value - I can notice places where the picture seems incomplete, places where I don't understand what's going on and therefore places my partner might not yet have fleshed out the idea.

Playing this support role is a skill, and someone who is good at this skill can act as a general-purpose force multiplier, adding value in a wide variety of contexts, especially in research.

Covid 4/22: Crisis in India

The United States appears to have turned the corner. Despite our determination to sabotage vaccination efforts, they have taken only minor damage, and we are starting to see declines again in the number of cases. Unless a new strain more dangerous than the English one reverses things once again, we should soon start to see steady declines in cases.

Other places without our access to vaccines are not as lucky, and in particular India is in crisis. Things there are worse than they've ever been and rapidly getting worse, with the hospitals on the verge of collapse. This is likely to be the biggest human cost of the entire pandemic, plausibly by a very large margin given how many people live in India, and it is entirely the our responsibility for not accelerating vaccine production in time to help them. Even when we have vaccine we are unwilling to use, we refuse to use it to help where it is needed.

Actions have consequences. In particular, a lot of death.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 5.8% (up 0.2%) and deaths unchanged.

Result:

In the past week in the U.S. ...

New daily reported **cases fell 11.9% ↓**

New daily reported **deaths fell 2% ↓**

Covid-related **hospitalizations rose 2.6% ↑** [Read more](#)

Among reported tests, **the positivity rate was 5.3%**.

The **number of tests reported fell 26.5% ↓** from the previous week. [Read more](#)

The number of tests continues to crater, and also the numbers continue to not make much sense as reported here, since if tests fall 26.5% and cases fall 11.9%, that should imply the positive test percentage is higher rather than lower. I still don't get how that one keeps happening.

[Johns Hopkins has the rate declining](#) from 5.2% to 4.7%, so I suppose that decline is real, and the 26.5% number isn't real and I should chalk this kind of thing up to systematic reporting delays.

Can you tell that I really, really miss the Covid Tracking Project?

The trick as always is whether this is a data error that self-corrects, or the start of a trend which in this case would be the beginning of the final phase barring a new strain worse than

the English one. I'm going to cautiously say it's more likely to be mostly real, but likely got a little ahead of itself because graphs should be smooth.

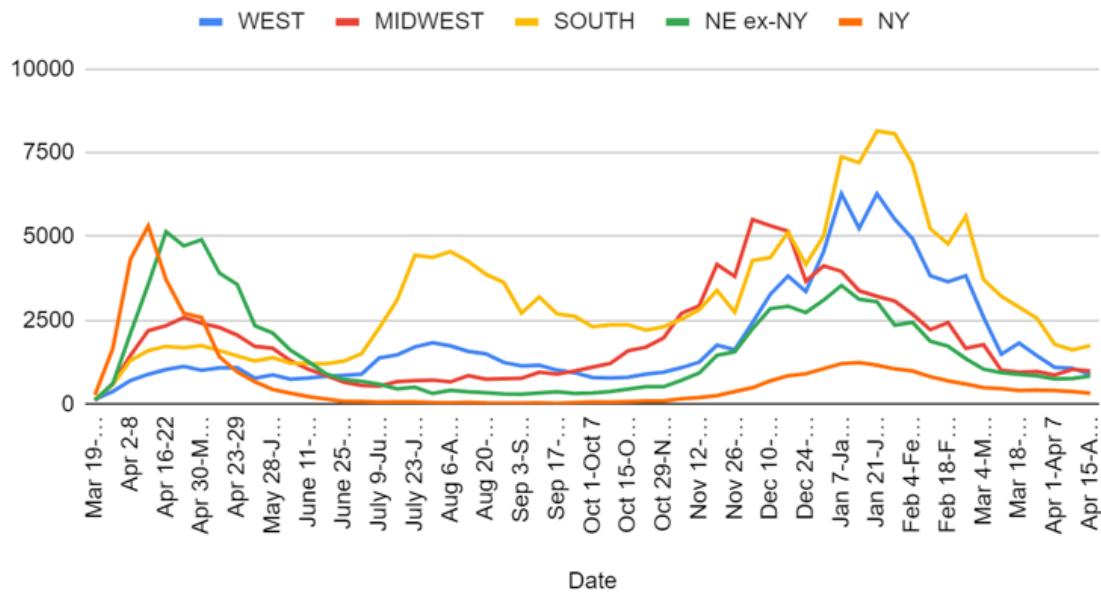
Deaths should continue to slowly decline as vaccinations work, but it's a slow process.

Prediction: Positivity rate of 5.1% (down 0.2%) and deaths decline by 4%.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Mar 4-Mar 10	2595	1775	3714	1539	9623
Mar 11-Mar 17	1492	1010	3217	1402	7121
Mar 18-Mar 24	1823	957	2895	1294	6969
Mar 25-Mar 31	1445	976	2564	1262	6247
Apr 1-Apr 7	1098	867	1789	1160	4914
Apr 8-Apr 14	1070	1037	1621	1145	4873
Apr 15-Apr 21	883	987	1747	1168	4785

Deaths by Region



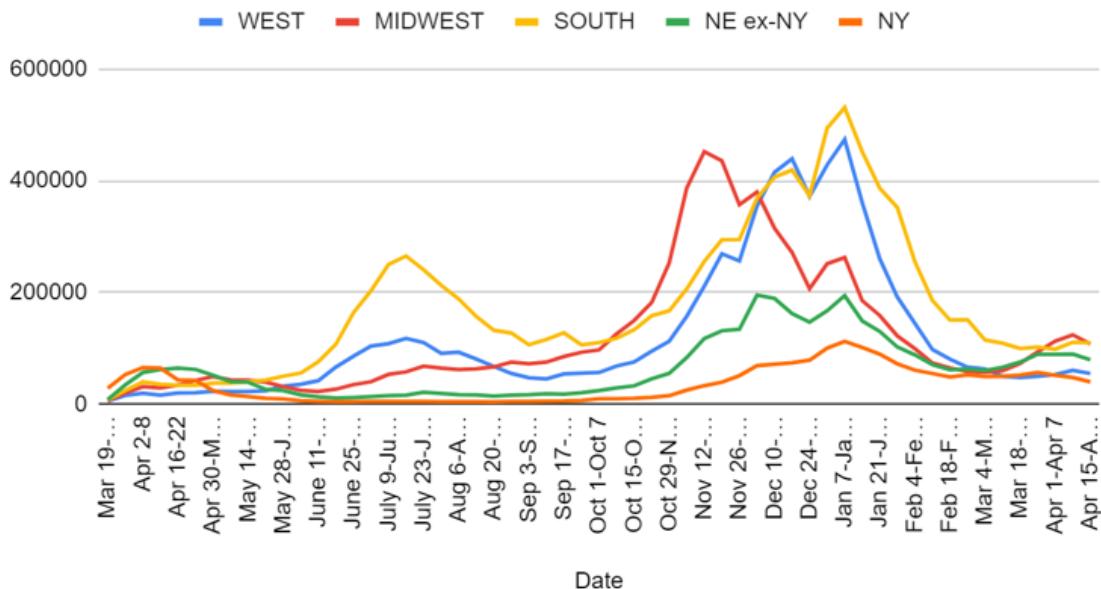
Things will bounce around due to random fluctuations and data timeshifting, but the default should continue to be a slow decline in deaths until cases have had several weeks of declines, at which point the drop should accelerate.

Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Mar 4-Mar 10	62,935	57,262	114,830	109,916
Mar 11-Mar 17	49,696	59,881	109,141	115,893
Mar 18-Mar 24	47,921	72,810	99,568	127,421

Mar 25-Mar 31	49,669	93,690	102,134	145,933
Apr 1-Apr 7	52,891	112,848	98,390	140,739
Apr 8-Apr 14	60,693	124,161	110,995	137,213
Apr 15-Apr 21	54,778	107,700	110,160	119,542

Positive Tests by Region



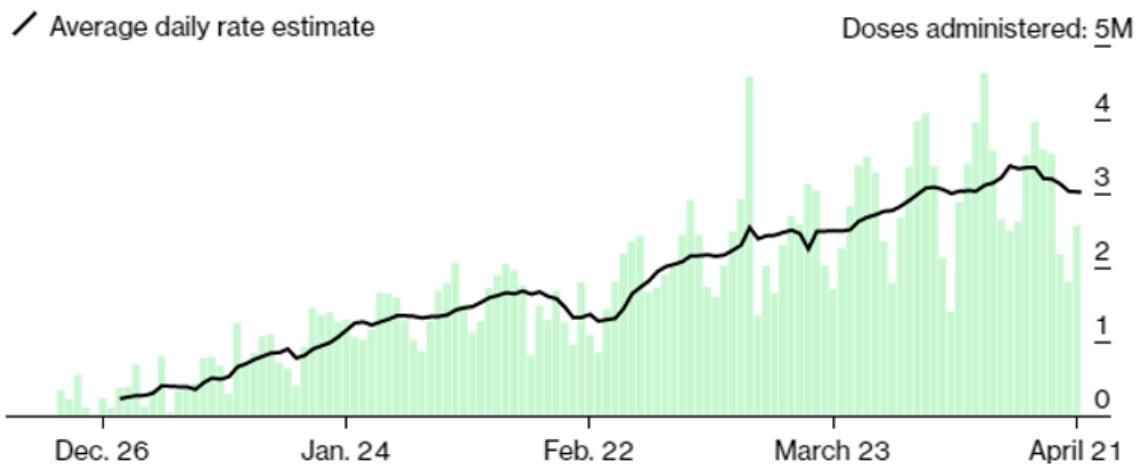
Clearly the corner has been turned. We will see if it can be sustained.

Vaccinations

You know what's not good for vaccination rates? Suspending the use of vaccines. I hear that's bad for vaccination rates.

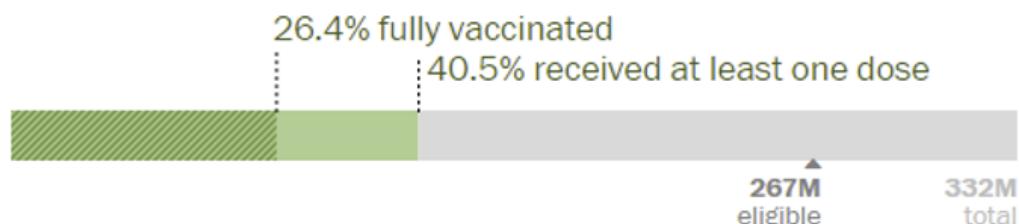
In the U.S., the latest vaccination rate is **3,020,805 doses** per day, on average. At this pace, it will take another **3 months** to cover 75% of the population.

Select another location... ↑ ↓



In the last week, an average of **3.02 million** doses per day were administered, a **9%** decrease ↓ over the week before.

That number was greatly boosted by a +15% acceleration in California. Most states had double digit declines.



Without Johnson & Johnson, and with the increased vaccine hesitancy, things are going to be harder from here on in, and I have little hope that J&J vaccinations will be allowed to resume, nor do I expect us to agree to export the doses to a place that would actually use them any more than we're letting go of our AZ doses. It's at least *kind of* murder, and all kinds of foolish and destructive.

There are those who say that the slowdown has multiple causes, and that we're starting to run into the wall where there aren't enough people who want vaccinations, making supply not the limiting factor in many places and making the second half of the job harder and presumably much slower as well. I do acknowledge this is a real dynamic, but also we intentionally hurt demand via the suspension of J&J, both in terms of increasing hesitancy, and in terms of taking away a one-shot-only vaccine that was logistically far easier to deal

with, thus making harder to reach people that much harder to reach and making those who dread shots or ‘this new mRNA technology’ that much more hesitant.

Thus, while I do think it’s unfair to attribute the entire deviation from the previous upward trend as being due to the suspension of J&J, I do think that’s most of it. At a minimum, I would be very surprised if we would have otherwise seen a *decline* in the rate of progress rather than a plateau.

For next week, my expectation is that it will drop below 3 million doses per day, let’s say a median prediction of 2.9 doses per day, assuming that J&J remains suspended, as it appears that it will be. I do still expect all the second doses to happen on schedule, which should prevent the number from dropping too dramatically, as the number of scheduled second doses should still be rising.

Note of course that this still represents a steady stream of additional vaccinations, and that our situation will continue to improve. Even if we decline to 2 million doses a day, and all of them are from 2-dose vaccines, that’s still 1 million extra immune people per day, or an extra 2% protected every week. With 40%+ of the population and 50%+ of the adult population already having at least one dose, that should lower the weekly spread of Covid by a compounding 5% or so a week. The results of that should still be good enough.

Others, however, are not so lucky. In particular, the situation in India looks very, very bad.

India

[It doesn't look good.](#)



Shruti Rajagopalan @srajagopalan · 12h

It's been a really rough day. Almost everyone I know in Delhi has a family member who is down with COVID and my whatapp and timeline are full of requests and messages scrambling for hospital beds, oxygen, meds etc. Just gutted.



6



22



139



...



Alex Tabarrok @ATabarrok · 3m

The numbers are very bad and everyone I know in India tells me it's much worse than the numbers indicate.

...

[It looks very, very bad \(Financial Times\).](#)



John Burn-Murdoch ✅ @jburnmurdoch · 5h

NEW: a deep-dive into the situation in India, where a devastating second wave is overwhelming hospitals and crematoriums, eclipsing global records as it goes ft.com/content/683914...

250,000 new cases every day, and test positivity is soaring suggesting many are still missed

...



John Burn-Murdoch @jburnmurdoch · 5h

In many parts of the country including the capital Delhi, cases are doubling every five days. Compared to the steady rise seen in the first wave last year, the current climbs are almost vertical.



John Burn-Murdoch @jburnmurdoch · 5h

And in many places, test positivity is rising at the same pace. Even as more and more tests are done, the share of them that come back positive is still climbing, suggesting tens of thousands of cases are going undetected.

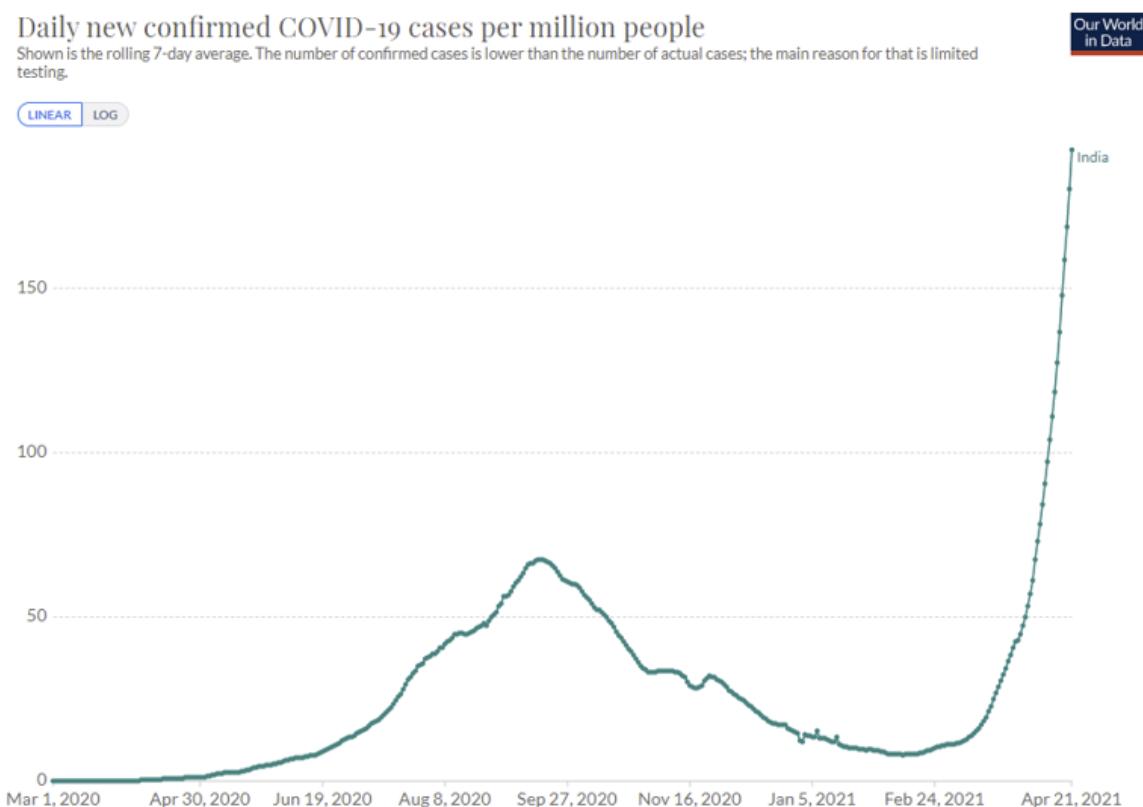


John Burn-Murdoch @jburnmurdoch · 5h

Essentially, none of those numbers are correct; all are vast undercounts.

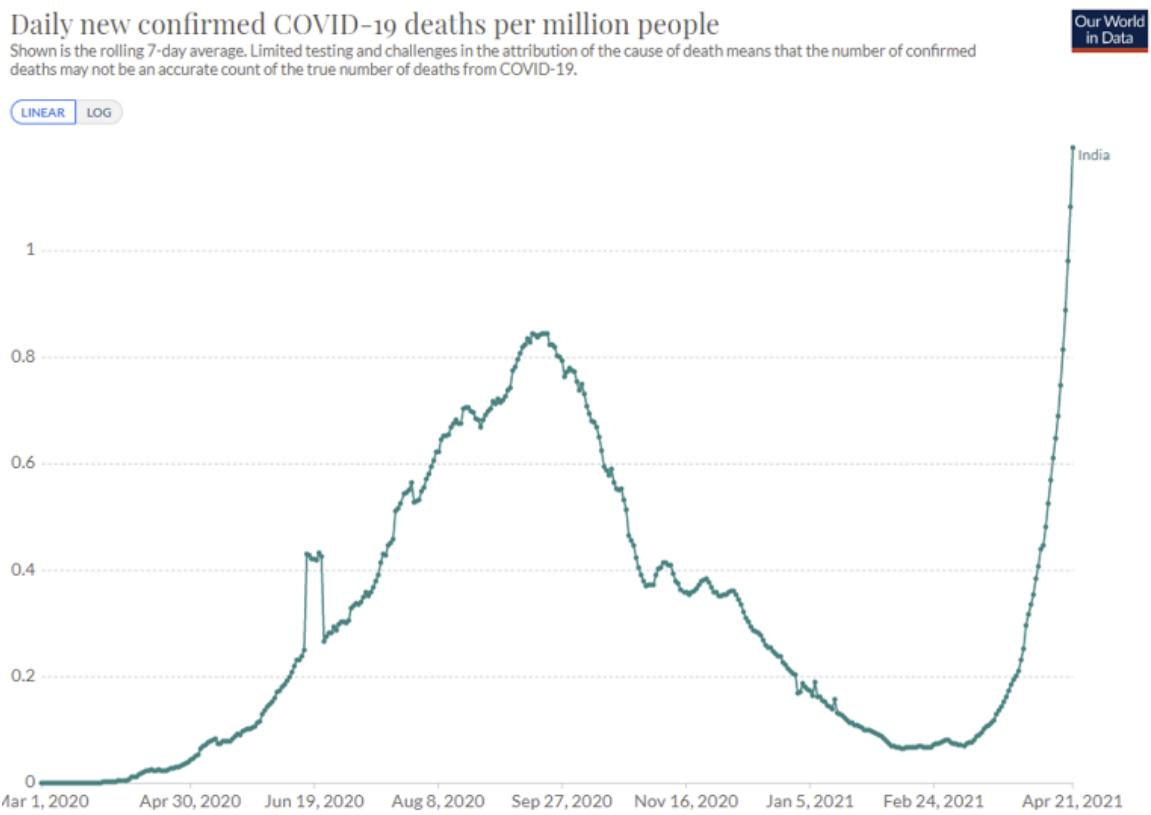
I collated local news reports (HT [@muradbanaji](#)) across seven districts, finding that overall, numbers of Covid victims who have been cremated are 10x larger than official Covid death counts in same areas.

They're still barely below the United States in confirmed cases per capita, but the graph looks like this:



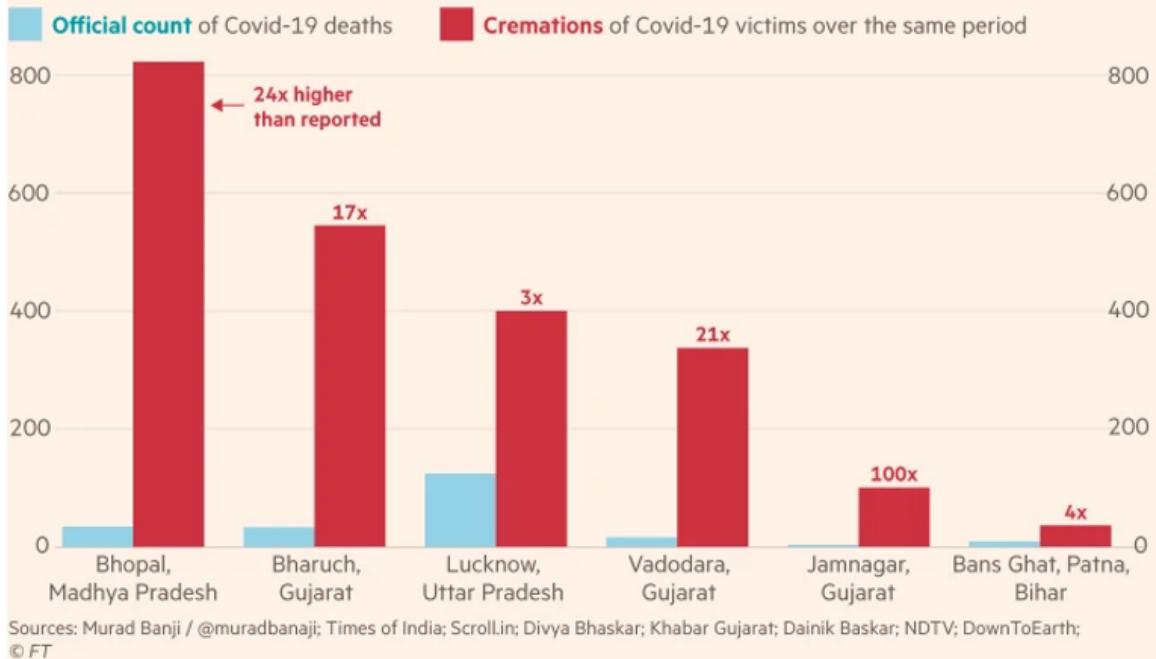
With no signs of stopping, and no reason to doubt that this is a massive undercount.

Official deaths are lower than in the United States as well, but likely much, much higher than reported, and going vertical:



Here's the comparison (from FT) to cremations of Covid victims, it's really bad out there:

In districts across India, official counts of Covid deaths are several times lower than the numbers of Covid victims cremated

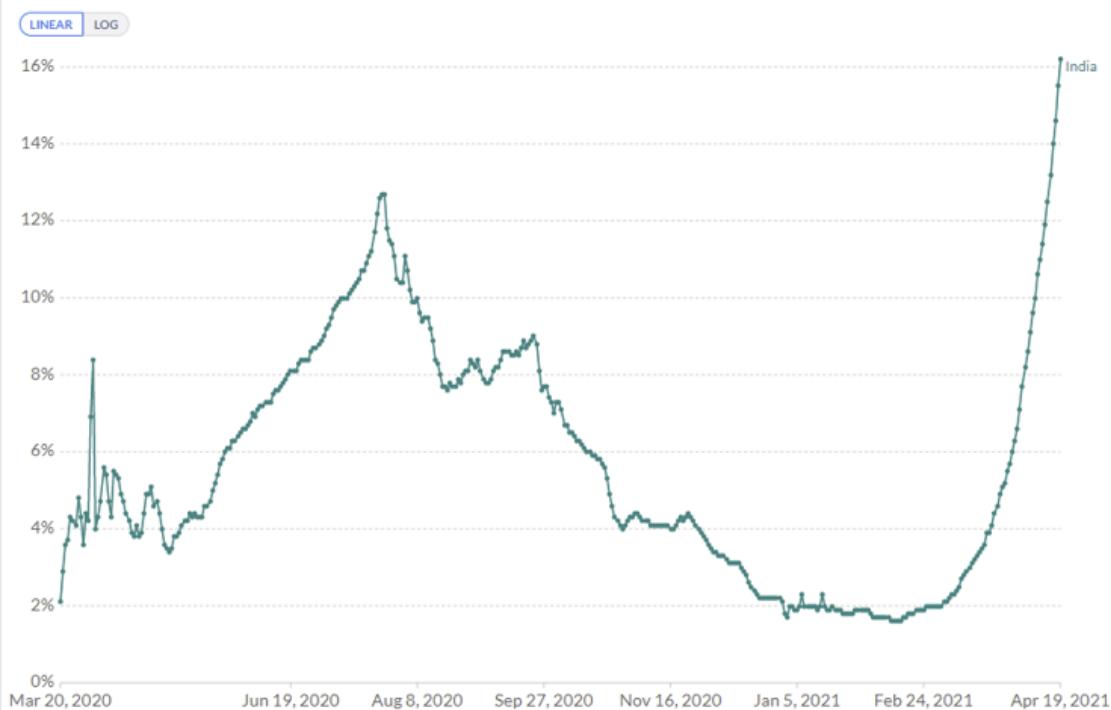


As always, share of positive tests is a key metric, and it has the same straight line. Yikes.

The share of daily COVID-19 tests that are positive

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

Our World
in Data



The hospital system is on (at least) the verge of collapse. What happened? And what is going to happen now?

As there usually is, there's a variant involved, this time it's B.1.617.

Officials are alarmed about the suspected role of new variants in driving the latest wave, particularly the B.1.617 strain first detected in India last month. Scientists are still trying to understand the variant, which has spread internationally, including to [the UK](#), but some believe it is more infectious and vaccine evasive.

The thing is that being vaccine evasive wouldn't explain what's happening, because India doesn't have that many people who are vaccinated. Even if it had full escape from all immunity, that wouldn't explain what's happened here. India's old positive test percentages were never that high, so it's that much more unlikely there were massive previous waves we didn't notice. To the extent that this strain is the cause, it's pure additional infectiousness doing most of the work. That doesn't preclude escape, but it means we don't have much reason to expect escape either.

I haven't been tracking India, but I don't have any reason to think there was a large behavioral change since February that could take us from static to doubling every week. What could this be other than the variant? So I went looking for what we know.

[Here's a Forbes explainer](#), which also notes that the variant has arrived in California. It seems there are a lot of different mutations in B.1.617 that are contributing to it being a bigger problem, part of which is making immune response more difficult. [Forbes also reports that Israel found eight cases](#), and that the Pfizer vaccine still works but has 'reduced efficacy' against it.

[This from the Indian Express](#) covers the basics but doesn't have insight into the questions we need answered. [This from The Guardian](#) is similar, with those quoted thinking this likely isn't as dangerous a variant as the Brazilian or South African ones.

[Zayneb comes through and hooks us up](#) with the good news that the vaccines still look effective against B.1.617:

Coronavirus live updates | Very few breakthrough infections reported, says ICMR Director General



The Hindu Net Desk

APRIL 21, 2021 09:46 IST

UPDATED: APRIL 21, 2021 19:11 IST

SHARE ARTICLE | f | t | g | w | m | 2 | PRINT | A | A | A

Post Vaccination Breakthrough Infection

(20th April, 2021)

INDIA

Total Vaccinated	Received first dose	Positive after first dose	Received second dose	Positive after second dose
COVAXIN 1.1 crore	93,56,436	4208 (0.04%)	17,37,178	695 (0.04%)
COVISHIELD 11.6 crore	10,03,02,745	17145 (0.02%)	15732754	5014 (0.03%)

That thread also points out the obvious. Yes, we need to worry about and think about the possibility that the variant will cause trouble on our shores and for ourselves personally, but the main thing for the world is that there's a huge disaster happening right now, in India, and no one seems to care to do much about it. Certainly our vaccine policy has given little or no thought to getting doses for the third world, despite it protecting us against variants and buying massive goodwill while being super cheap. To the extent that anyone ever says

anything, it's 'let's get rid of the IP incentives that created the vaccines' rather than 'let's pay more money and make more vaccine doses.'

[This from Science Media Center](#) is the standard thing where 'experts' refuse to speculate until they have all their data lined up, and act like the variant both taking over and taking over while cases explode all of a sudden aren't together much evidence, and there's no reason yet to call this a 'variant of concern.' I'd say I'm concerned.

I know that jumping to conclusions can backfire. But if something is probably concerning, that's concerning. One shouldn't remain unconcerned until there's proof, that's not how this works, that's not how any of this works.

I'm putting it at about 85% that the surge in India's primary cause is that the B.1.617 variant is far more infectious than their previous variant. I'd go higher, but the possibility that lots of cases have been missed for a long time, together with my lack of detail knowledge in India, makes me not want to make too strong an assumption yet. I could easily get higher quickly.

What happens now, unfortunately, is presumably the collapse of India's hospital system. I don't see a likely way around that, since they'd have to stall things in their tracks right here without another doubling. The only way this doesn't happen is if individuals react dramatically, and cut their exposures by almost half within the next few days, or perhaps it would even need to have already happened several days ago. Either way, seems unlikely, as I don't think India or its people have that kind of slack.

This was always on the table as an outcome, from the very beginning. There's a maximum amount of baseline infectiousness, beyond which adjustment to stop it is not practical, and the control system breaks down. It's that much easier to hit the breaking point in a country that's much poorer, and thus has much less effective hospital capacity. We'll find out soon if India has reached that point.

The flip side is that there have been places with *much* higher positive test rates for quite a while. Mexico spent a long time around 50%, and many places in the United States may have been in worse spots than India is now, and turned things around or at least stabilized them via behavioral adjustments.

Vaccines Still Work

[Some perspective](#), I didn't do the math myself but seems reasonable even if the other person is unvaccinated, cars are rather unsafe:



((David Shor))

@davidshor

...

Replies to @agraybee and @NateSilver538

Back of the envelope with conservative assumptions, your per hour risk of killing somebody driving sober is at least ~33x higher than the per-hour risk of killing somebody from Covid hanging out maskless post-vaccination.

1:12 PM · Apr 10, 2021 · Twitter Web App

Before you say we can't pause cars, we paused *seeing other humans* for a year, so we can do a lot of things.

[Some lack of perspective](#), if you'd prefer that instead. If you have a vaccine that is only 95% effective, and there are 71 people who catch the virus post-vaccination none of whom die, you could always do something like this, including having officials 'urge caution':

CORONAVIRUS TOLL		Missouri +378	Illinois +1,059	St. Louis +92	St. Louis Co. +99	St. Charles Co. +59	Jefferson Co. +60	Franklin Co. NA	Metro East +63
CASES/DEATHS as of 8 p.m. Monday • Why numbers may not add up Page A3									

ST. LOUIS POST-DISPATCH

TUESDAY • 04.20.2021 • \$2.50

VACCINATED TEST POSITIVE

St. Louis Co. cites 71 cases; post-vaccine caution urged

BY ANNICKA MERRILEES
AND ERIN HEFFERNAN
St. Louis Post-Dispatch

ST. LOUIS — The St. Louis County health department has identified 71 people who tested positive for COVID-19 after they were fully vaccinated, the first public acknowledgement that some who got the vaccine have, as expected, gotten the virus, too.

Experts say the findings reinforce that, even post-vaccine, people still need to be cautious.

"That's important to tell people, because if you're vaccinated, you kind of feel like you're bulletproof," said Dr. James Hinrichs, infectious disease advisor for the St. Louis County Department of Public Health.

(Please see INFECTIONS, Page A4)

More companies, unions offering vaccines

A Tyson Foods worker receives a COVID-19 vaccine on Feb. 19 at the Joslin, Ill., facility. A growing number of companies and labor unions are directly securing coronavirus vaccines for their employees.

JOHN KONSTANTARAS,
ASSOCIATED PRESS



BY ALEXANDRA OLSON
Associated Press

Marie Watson wanted to be among the first in line when she and other essential workers became eligible for the coronavirus vaccine — and with good reason.

The maintenance parts buyer for a Mission Foods tortilla plant in Pueblo, Colorado, had lost her father to COVID-19 in the fall and was told by a doctor last year that she herself almost certainly

(Please see WORKERS, Page A4)

Pausing Vaccines For No Reason Still Doesn't Work

Damage to public confidence was done quickly, and will be hard to reverse.

[From April 15:](#)

Already, doctors say, the recent pauses have vindicated vaccine skeptics and made many others feel duped.

“People, especially those who were vaccinated, felt like they had been tricked in a way — they were asking, ‘How do we get rid of the vaccine in our body?’” said Precious Makiyi, a doctor and behavioral scientist in Malawi, where health workers have been racing to empty their shelves of nearly expired AstraZeneca doses. “We fought so hard with vaccine messaging, but what has happened this past week has brought us back to square zero.”

[Also:](#)



James Surowiecki @JamesSurowiecki · Apr 15

...

More people now think the J&J vaccine is unsafe than think it's safe.

Some ppl on the fence will likely decide it's safe if regulators, after studying all the data, conclude that it is safe. But I think it'll be hard to move ppl off their new "it's unsafe" assumption.



YouGov America @YouGovAmerica · Apr 15

The CDC recommendation to pause Johnson & Johnson vaccine use causes public confidence in the vaccine to sink 15pts

Perceptions before CDC announcement

Safe - 52%

Unsafe - 26%

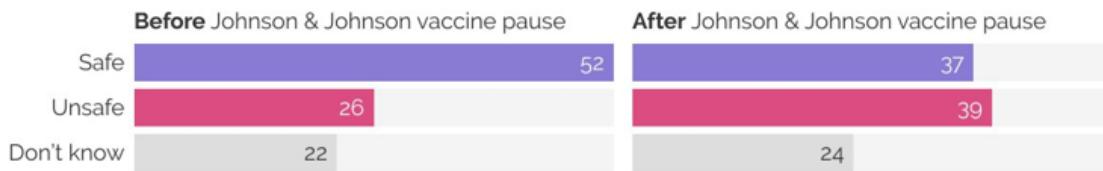
Perceptions after CDC announcement

Safe - 37% (-15)

Unsafe - 39% (+13)

Perception of the safety of the Johnson & Johnson vaccine dropped after the suspension announcement

How safe, or unsafe, do you think the Johnson & Johnson vaccine is? (% of US adults)



There were 1,490 US Adult citizens polled about the safety of the Johnson & Johnson vaccine. 1,083 respondents started the survey before the suspension announcement was made on Tuesday, and 407 respondents started the survey after.

YouGov

The Economist / YouGov | April 10 - 13, 2021 | Get the data

As noted last week, a brief pause was something that could be steelmanned. If you had a very different model of how the public reacts to news, and assumed the news would always get out, then you could argue for a brief pause.

You can't argue for our current policy of keeping the blanket pause in place for weeks. We could mitigate *most* of the damage by restricting the ban to young women, putting a fig leaf over the whole thing and moving doses around between populations, [but we aren't even doing that](#).



Nate Silver @NateSilver538 · Apr 17

...

I don't really understand the endgame. Wait 2 more weeks (!) to gather more data. Let's say (as seems likely) the data shows there IS a link, but it's very rare. What then? We keep J&J banned because our regulators are insanely risk-averse?



Nate Silver @NateSilver538

...

Replying to [@NateSilver538](#)

And do we really not have enough information to have a better policy *for now* than a blanket ban? Between the J&J data and priors from AZ, you can make a case for directing younger women to use an mRNA vaccine. But it's nuts that say a 75-year-old man can't get a J&J dose now.

"The longer the pause is, the longer it's going to take for us to convince people that this particular vaccine is safe again," Arkansas Gov. Asa Hutchinson, a Republican, told POLITICO.

Already, conspiracy theorists and anti-vaccine proponents are [using the Johnson & Johnson setback](#) to fuel hesitancy and mistrust through social media.

"It's a setback — no question about it," said Peter Hotez, a vaccine expert at Baylor College of Medicine. "When you have this kind of anti-vaccine aggression out there, it changes the equation. This will be exploited."

[Matt Yglesias points out](#) how much whiplash is involved here, as every consideration is either mandatory and total, or forbidden to even speak its name:

A couple of months ago there was widespread concern in the public health community that the slightly lower efficacy rating of the J&J shot would lead to people deliberately trying to avoid it in favor of an mRNA shot. There was lots of propagandizing against vaccine shopping. Did we really need to leap all the way from “it is forbidden to express a preference for an mRNA shot” to “it is forbidden to take a J&J shot?” Does the CDC employ experts in the value of human freedom? The whole issue here is not medical science, but rather bioethicists having some peculiar ideas about when consent matters and how to do cost-benefit analysis.

[Here's Wired:](#)

Pausing the J&J Vaccine Was Easy. Unpausing Will Be Hard

Johnson & Johnson's Covid-19 vaccine was supposed to be the uncomplicated one. But even with new data, getting people to trust it again will be tricky.

Again, *what is this new data we're looking for?* What about the current data is insufficient to reach the necessary conclusions?

[This seems like a good intuition pump:](#)



Nate Silver @NateSilver538 · Apr 13

...

If out of the blue one morning Gov. Newsom was like "Shark attacks are extremely rare, but out of an abundance of caution, we're closing every beach in California until we investigate more", that's not likely to get more people to go out to the beach, even once beaches reopen.

440

1.6K

13.7K



[When you've lost Eric Topol on your abundance of caution](#) (WaPo), and he's calling it a 'deadly mistake,' you know you messed up. [Unless, of course, you didn't](#).

[This comment](#) on last week's post kind of says it all:



Evan P says:

April 18, 2021 at 10:28 pm (Edit)

Three very relevant points from my extended-family Zoom call today, when we got into talking about Johnson & Johnson:

* My dad, a biostatistician who regularly works on new drug trials, considers the pause the only responsible option. Since there isn't full FDA approval, he considers everyone taking the vaccine to effectively be part of a clinical trial, and pausing a clinical trial for weird side effects is The Done Thing until you know what's going on. He didn't bring up the public relations damage or the need for speed, and I didn't want to press him in the group call.

* My cousin (herself vaccinated per her job in allied health) then said that her husband probably wouldn't get vaccinated if he couldn't get J&J. He really doesn't want two shots, and he doesn't want to wait a month nervously anticipating a second shot with worse side effects. Plus, he himself's low risk for COVID. I've pressed my friends on similar issues, but she's not someone I want to talk politically-adjacent issues with.

* Also, my cousin then said her state's still only giving vaccines to people over age 65. I looked it up after the call; she's two weeks out of date. I wonder how many more people like her are still needlessly waiting.

[There's also this](#), which is now increasingly causing *direct* damage by giving the impression that things were rushed:



Fake Mario @ShakedDown · Apr 16

...

An underestimated part of the awfulness of the FDA is that they're used to being so uselessly slow they got people thinking that giving emergency approval for something in a year was a rush job.

It also has the standard splash damage that when a completely insane standard is applied in one area, it makes it impossible to think clearly about other areas, [resulting in things like this:](#)



The True American @TheTrueAmerica5 · Apr 16

...

Fewer people developed blood clots from the Johnson and Johnson vaccine than were murdered at a FedEx facility last night.

Perhaps we need a pause on firearms.

456

32.4K

135.4K



That's not to take *any* position on firearms, any more than one would call for a pause on slicing bread. You could argue for a pause in *actual almost anything* if all it has to do is unintentionally kill two people when done millions of times.

There's [this claim today](#) that the AZ and J&J pauses did not increase vaccine hesitancy. I do not think this reflects what's going on at all, as willingness to get vaccinated naturally increases over time as people can see others doing it and everything going well, but it's at least saying that this hasn't been *that large* of a disaster in terms of the threshold that counts. I'm still highly confident things are going substantially worse than in the counterfactual.

The Next Strain: P.1

Eliezer Yudkowsky [asks the obvious question](#). Only about a third of the population is covered, but there are also a lot of children:



Eliezer Yudkowsky ✅

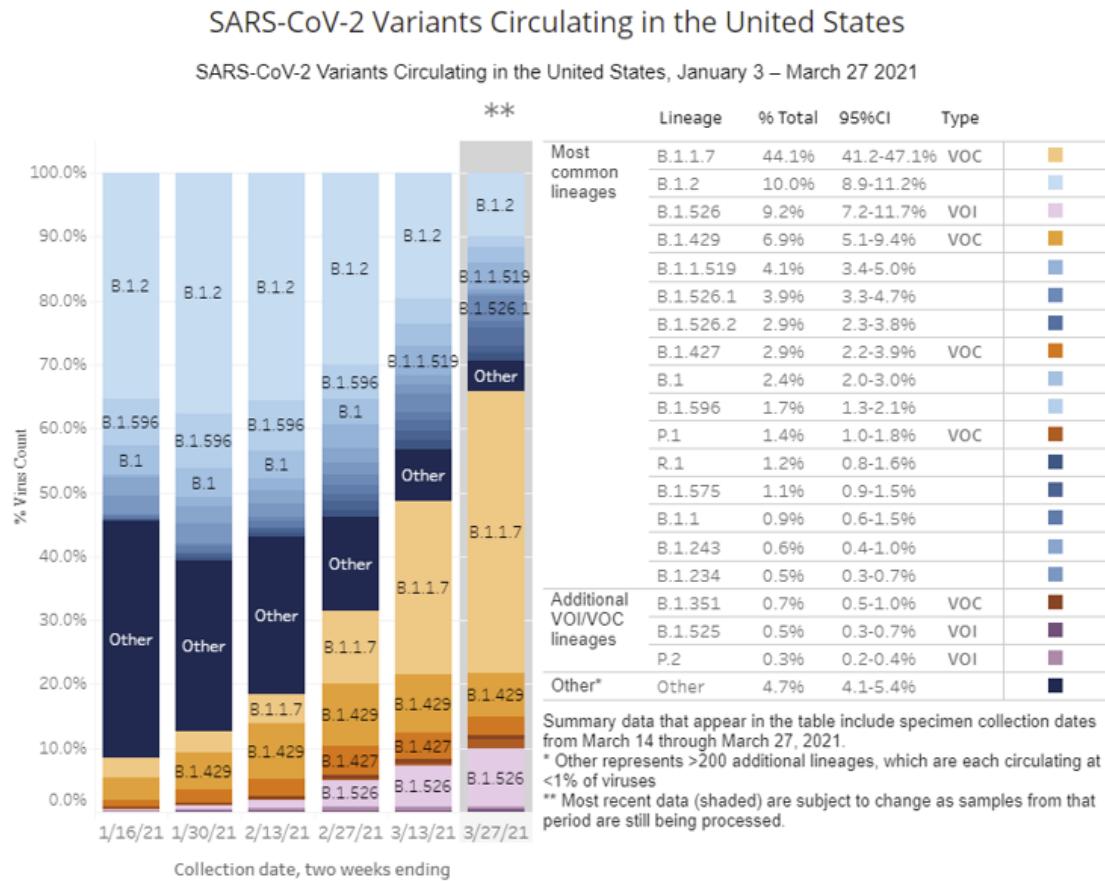
@ESYudkowsky

...

I'm worried that half the US population is vaccinated and Covid-19 cases are still trending up. Is this a phenomenon of the new variants, or were the current cases already concentrated in people who take fewer precautions and also refuse vaccination, or ...?

Thread contains a bunch of real data (as usual, he has to follow up with a Tweet saying "please only respond with actual data) that led to James Babcock pointing out [the CDC's](#)

[variant data](#), which he then put into [this spreadsheet](#).



The end of March was when B.1.1.7 was taking over, going from 11.4% of cases on 2/27 to 44.1% of cases on 3/27. Thus, in April, strains with a competitive advantage are growing, other cases continue to shrink, and the shift from one to the other has cancelled out our vaccinations. The good news is that B.1.1.7 can only hit 100%, whereas vaccinations can also hit 100%, and that's a battle vaccines win. If that was the whole story, we'd be in great shape, and confident we can turn things around soon.

The bad news is that P.1 is growing as well, potentially faster than B.1.1.7 (there's enough measurement error here that I wouldn't be confident in that), and its advantages likely involve [some amount of immune escape](#). If that's true, even if ([as it seems](#) from this and what other data we have on the question) that's mostly about evading immunity from infections rather than from vaccinations, it will be the dominant strain in several months, and we'll have to up our game another level if we want to stay ahead.

The good news is that, unless the immune escape is much stronger than we think it is, we can totally get there, even with our vaccination pace slowing down due to regulatory sabotage. Or at least, we can definitely get there in areas without a lot of vaccine hesitancy, which also is up due to that same sabotage.

[There's a graph for that.](#)



Dustin Moskovitz ✅ @moskov · Apr 19

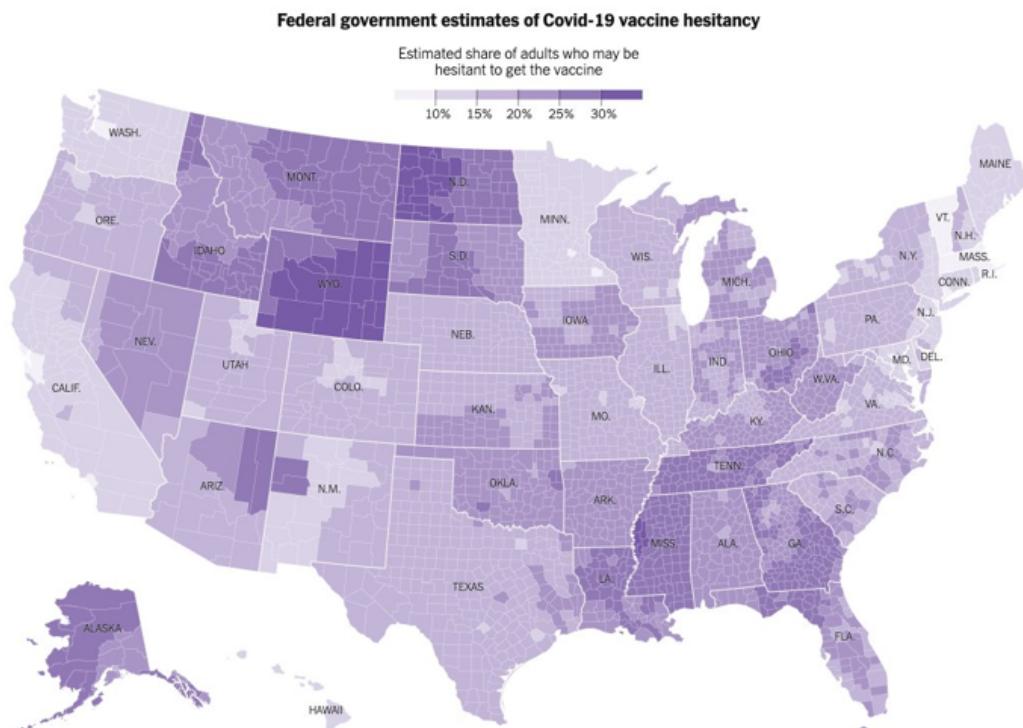
...

I didn't know this data existed and had been wondering about local hesitancy. Many metro areas, including SFBA, have only 10% hesitancy. With such effective vaccines, that will be an incredible amount of immunity.



Danielle Ivory ✅ @danielle_ivory · Apr 17

The @nytimes examined federal estimates of COVID vaccine hesitancy for every US county. In more than 500 counties, at least 25% of adults may not be willing to get a shot. Most of those counties supported Trump in the 2020 election.



It's worth stopping to contrast these two world views.

In the first corner, we have Dustin Moskovitz, who works on projects he hopes will make humanity thrive. He notes how amazing it is we'll be able to immunize essentially the entire population of many areas.

In the second corner, we have the reporter from The New York Times, who points out how Just Awful some places are, and then associates the whole thing with Donald Trump. Classy.

I wonder what magic is taking place along state borders in many places, especially Wyoming, but mostly the graph should not surprise us. I'm not sure how much of the detail I would have guessed on first principles, but certainly a lot of it.

The number of people who live in areas where the refusal rate is going to be 30%+ looks very low. The worst this gets for places with a lot of people in them is roughly 20-25%, which means (if these calculations are correct) they can still get to 75-80% vaccinations. That should still be enough. We might still be left with some states where things stay bad for a while, but when there's almost zero cases in the Northeast and still a lot of cases in the places people aren't vaccinating, I have a feeling we'll see some people change their minds.

If you're looking to help change those minds, [here's some good advice \(WaPo article\)](#):



Andrew Rettek @oscredwin · 8m

Lying openly in public can really damage your credibility, no matter what your reasons.



Frank Luntz ✅ @FrankLuntz · 10m

As @DDiamond summarizes, my vaccine-hesitant group gave 3 key points to winning them over:

- Stop talking about possible COVID booster shots.
- Don't bully people who are vaccine holdouts.
- Show us anyone besides Dr. Fauci.

washingtonpost.com/health/2021/04...

From that article, it turns out that even vaccine skeptics can do math better than the FDA, and but haven't figured out that this is true and important:

For instance, the group largely shrugged off federal regulators' decision last week to pause Johnson & Johnson's coronavirus vaccine for safety reviews, citing the risk of rare blood clots. Luntz and others had expected the pause to worsen hesitancy, but focus group participants instead asked why doctors were halting a potentially useful medical treatment, given that the reported side effects were so rare.

Brian Castrucci, an epidemiologist who leads the [de Beaumont Foundation](#), which helped convene the focus group, said: "Every public health person, me included, thought this would be a real hit to vaccine confidence. But we didn't see folks really concerned with the pause in the J&J vaccine."

It took *less than a week* for us to go from ‘pause is important to keep people confident in the vaccines’ to ‘every public health official, me included, thought this would be a big hit to vaccine confidence.’ I don’t even get whiplash anymore. (And yes, I know that if you steelman their position there’s a way to reconcile those two statements, where the alternative would have been worse.)

What I found most interesting was also surprising to those who ran the focus group, which is that *talk of booster shots* pisses off such people:

“I feel like this is not going to end. I mean, we’re just going to be shot up and shot up and shot up,” said a man identified as Erzen from New York. “We can’t live like this. This is not sustainable.”

Public health experts have said it is premature to assume Americans will need booster shots in the coming year, and Frieden framed Bourla’s claims as a business decision.

“I’m pissed off at Pfizer for talking about boosters. And I think they did that for their corporate benefit,” he told the group. Attendees later said that they appreciated Frieden’s blunt rhetoric.

I fail to understand why people can’t live like that. A vaccine shot every year is that bad? In a world in which, by definition, Covid is still around? All right then.

On the other hand, there’s this other thing they found that they did not expect, and yeah they should totally have expected it and I’m not sure why they didn’t:

The focus group revealed another unexpected development: Most participants said they would want a [fake vaccination card](#) that would allow them to claim they had received shots, after Luntz granted them anonymity to speak honestly.

“One-thousand percent,” one woman said.

“If I have a fake vaccine card, yeah, I can go anywhere,” added a man who said he had turned down free New York Yankees tickets because of the team’s requirement to show proof of vaccination to attend games. Other participants said they wanted a fake card to attend concerts and go on trips, citing the growing number of organizations saying they will require proof of shots.

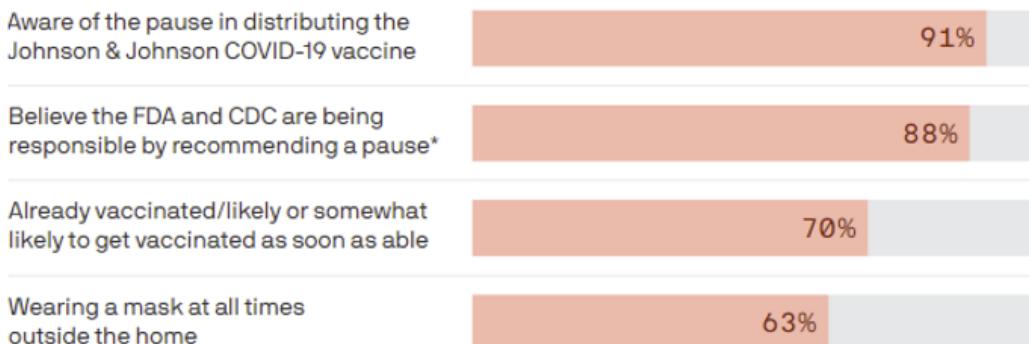
Even some participants who said they did not intend to get a fraudulent card acknowledged they were tempted. “My faith wouldn’t allow me to be deceitful. So what do I do?” one woman asked the group.

You thought people who didn’t want to get vaccinated wouldn’t want fake vaccination cards when the cards are required in order to do things? Really?

Finally, it’s worth noting that this is a place where regular people, most people, [still are willing to take the FDA’s word for things](#). The problem is that what they’re taking the FDA’s word for is that the vaccines might not be safe, but either way, public support for the announcement is strong.

Snapshot of the pandemic

Survey 1,033 U.S. adults, April 16-19, 2021



*Among those aware of the pause

That's in contrast to the people I talk to and respect, who universally think of this decision the same way I think of this decision.

In Other News

[Bill Maher offers his weekly new rule to not mix politics with our health.](#) Well said, and good luck with that.

[Washington state denies people vaccinations on the basis of the color of their skin.](#)

[Researchers in Belgium register a trial of Moderna half-doses](#), which I expect will produce less side effects, almost identical protection and double the number of people protected; I'd put the chance of a success here around 90% if the study is well-designed. Thing is, it's only 200 people, so all they can measure is immunogenicity, *which the clinical trials already checked*. So what's likely going to happen is they note 'immune response is almost identical' and then everyone says 'yes, but what if somehow it's not as effective' and continues actively wasting half the doses.

[Pirate Wires preaches it: "Science" and safety porn.](#)

[Nate Silver thread reminding us](#) of the pattern where when we play up dangers, we scare the people who are already safe while alienating the ones who are taking real risks.

[Dr. Fauci thinks it's paradoxical that people don't want to get vaccinated when he doesn't want to restrict any restrictions for those who are vaccinated.](#)

[We now have a Philippines strain to worry about, there will be more until we solve this globally, yet there is no sense of urgency whatsoever.](#)

Child shows up at their facility daily [so they can perform required labors remotely on their computer](#) anyway in an overflow room. Yes, you can have the worst of all worlds at once.

[Interesting analysis](#) says that English Strain's relative infectiousness likely declined greatly over time, and at a minimum we have no idea how much more infectious it actually is despite 'experts' pretending otherwise on the basis of, as he essentially puts it, two points and a line between them. One could reply that we know a lot because the strain did indeed

take over everywhere more or less within the predicted time frame, but it's good to look from additional angles. There's definitely some weird data here. There's also reason to think that restrictions could alter the relative infectiousness levels. The English strain is slower to spread and slower to go away, and it produces higher viral loads. Various restrictions could be more or less effective at stopping the different strains, and the timing could also give a misleading impression.

[Demand falling below supply at University of Arizona](#). Seems to be steadily happening in more places. If still need to get vaccinated, there are lots of places to go if you look around.

If you're 50+ in NYC, you can walk in and get a vaccine without an appointment, [here's a list of where to go](#).

Soon you'll also be able to go to Alaska, [and get vaccinated at the airport](#). It's part of a plan to reinvigorate the Alaskan tourist industry, via getting as many unvaccinated visitors as possible.

[A paper reporting on \(spoiler alert: complete lack of\) lockdown effectiveness](#), and on benefits versus costs.

United Kingdom [has a day with only one Covid death](#). New goal is a day with only one death of any kind. Fund anti-aging research!

The irrelevance of test scores is greatly exaggerated

This is a linkpost for <https://dynamight.net/are-tests-irrelevant/>

Here's some claims about how grades (GPA) and test scores (ACT) predict success in college.

What are all these children doing in my ponds?

(this is not criticism of effective altruism, only one analogy that's used as an argument)

Peter Singer writes in the [The Drowning Child and the Expanding Circle](#):

To challenge my students to think about the ethics of what we owe to people in need, I ask them to imagine that their route to the university takes them past a shallow pond. One morning, I say to them, you notice a child has fallen in and appears to be drowning. To wade in and pull the child out would be easy but it will mean that you get your clothes wet and muddy, and by the time you go home and change you will have missed your first class.

I then ask the students: do you have any obligation to rescue the child? Unanimously, the students say they do. The importance of saving a child so far outweighs the cost of getting one's clothes muddy and missing a class, that they refuse to consider it any kind of excuse for not saving the child. Does it make a difference, I ask, that there are other people walking past the pond who would equally be able to rescue the child but are not doing so? No, the students reply, the fact that others are not doing what they ought to do is no reason why I should not do what I ought to do.

Once we are all clear about our obligations to rescue the drowning child in front of us, I ask: would it make any difference if the child were far away, in another country perhaps, but similarly in danger of death, and equally within your means to save, at no great cost – and absolutely no danger – to yourself? Virtually all agree that distance and nationality make no moral difference to the situation. I then point out that we are all in that situation of the person passing the shallow pond: we can all save lives of people, both children and adults, who would otherwise die, and we can do so at a very small cost to us: the cost of a new CD, a shirt or a night out at a restaurant or concert, can mean the difference between life and death to more than one person somewhere in the world – and overseas aid agencies like Oxfam overcome the problem of acting at a distance.

Singer's analogy is incomplete because it doesn't capture the essence of the drowning child scenario. The following does.

You're walking somewhere, and you see a child drowning in a shallow pond. You naturally decide that you ought to save this child despite your other obligations, so you rush in, get your clothes wet and muddy, and rescue the child. You get out of the pond, but you see that there's another pond right next to this one - and this one also has a child drowning. You rush into the second pond and save that child. Upon coming out, you perceive a third pond with ANOTHER child drowning. You say to yourself "gee, there sure are a lot of children drowning today", and you dutifully rush into the third pond, and save a child. This process repeats. It's 3 AM now, and you're hungry and tired, but every time you rescue a child from a pond, you see another pond and another child. You continue throughout the night, well into the second day. You haven't had a minute of break. It's noon on the following day. You lose consciousness because

you're so tired and overworked. You wake up and the child that you were rushing to save next is now dead.

You spend your second day helping children out of ponds, but at one point you stop and you go get food, as you haven't eaten in two days. You eat quickly and you run back to the last pond. That child is now also dead. But you see another child in another pond, so you rush in once again. You drop out of exhaustion somewhere around midnight. You wake up in an hour or two, see that this child too is now dead, but there's another one nearby. As before, you rush in, help the child, and repeat the process.

You see the picture. Long story short: you start helping fewer and fewer children out of ponds, and stabilize at some sustainable daily range of saved children. Maybe you go to work, earn your wages, go home, eat, get changed, and spend an hour or two a day in ponds, saving children, and letting other children die. Maybe you stop saving children altogether and start wondering why the hell are there so many children in ponds anyway. Or maybe you invent a surveying technology that can estimate how many drowning children there are on a given piece of land, or you try to estimate how much it would cost to drain all these ponds so that children can stop falling into them, or you raise awareness in town, and try to explain to others that there are children drowning in some ponds nearby.

That's my entire criticism of this analogy - it's sometimes presented like "if you would do this THEN IT FOLLOWS that you should do that", but it does not follow, because of context. There's not one pond, but millions of ponds, and rushing in is an excellent strategy only for up to a thousand ponds. Everything above that (and probably everything above 100) is reason enough to stop, think and try to build a better system.

A new acausal trading platform: RobinShould

After receiving overwhelming demand, we're excited to announce our new platform for conducting [multiverse wide trades](#) with other civilizations: RobinShould.

At RobinShould, we believe in democratizing [acausal trading](#) and making it freely available to all. Our simple interface allows you to purchase products from our wide selection of alien vendors. We have included an example trade below.



We currently only accept payment in paperclips and Ethereum, but we will soon be adding Bitcoin, Tether and tiny molecular smiley-faces as payment options.

How RobinShould works

We care a lot about ensuring that all of our customers feel satisfied knowing that their trades go through. Since the vendors on our platform have no causal connection to our universe, building RobinShould has proven quite tricky.

Fortunately, you can have confidence in our patented superintelligent prediction technology. Whenever you purchase a product through RobinShould, an amount of that product is manufactured in another universe by an alien civilization. The price for each product was determined by simulating evolution on other planets and modeling alien utility functions and decisionmaking.

In fact, our prediction technology is now so advanced that mesa-optimizers appear regularly in our prediction software. [Apply to our data science team](#) if you would like to help us clean these mesa-optimizers from our computer systems. (Seriously they're taking over our facility, please help.)

Disclaimer: In response to unprecedented volatility and clearinghouse requirements, we are temporarily restricting trades for GAME stock (Grabby Aliens Model Enterprises). Please note that this decision was made with the best intentions and your interests at heart. It was not at all an attempt to signal-boost [Daniel Martin and Robin Hanson's website](#), or to protect our financial ally, Moloch.

Gradations of Inner Alignment Obstacles

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

The existing definitions of deception, inner optimizer, and some other terms tend to strike me as "stronger than necessary" depending on the context. If weaker definitions are similarly problematic, this means we need stronger methods to prevent them! I illustrate this and make some related (probably contentious) claims.

Summary of contentious claims to follow:

1. The most useful definition of "[mesa-optimizer](#)" doesn't require them to perform explicit search, contrary to the current standard.
2. Success at [aligning narrowly superhuman models](#) might be bad news.
3. Some versions of the lottery ticket hypothesis seem to imply that randomly initialized networks already contain deceptive agents.

It's possible I've shoved too many things into one post. Sorry.

Inner Optimization

The standard definition of "inner optimizer" refers to something which carries out explicit search, in service of some objective. It's not clear to me whether/when we should focus that narrowly. Here are some other definitions of "inner optimizer" which I sometimes think about.

Mesa-Control

I've previously written about the idea of distinguishing [mesa-search vs mesa-control](#):

- Mesa-searchers implement an internal optimization algorithm, such as a planning algorithm, to help them achieve an objective -- this is the definition of "mesa-optimizer"/"inner optimizer" I think of as standard.
- Mesa-controller refers to any effective strategies, including mesa-searchers but also "dumber" strategies which nonetheless effectively steer toward an objective. For example, thermostat-like strategies, or strategies which have simply memorized a number of effective interventions.
 - Richard Ngo [points out that this definition is rather all-encompassing](#), since it includes any highly competent policy. [Adam Shimi suggests](#) that we think of inner optimizers as *goal-directed*.
 - Considering these comments, I think I want to revise my definition of mesa-controller to include that it is *not totally myopic* in some sense. A highly competent Q&A policy, if totally myopic, is not systematically "steering the world" in a particular direction, even if misaligned.
 - However, I am not sure how I want to define "totally myopic" there. There may be several reasonable definitions.

I think mesa-control is thought of as a less concerning problem than mesa-search, primarily because: *how would you even get* severely misaligned mesa-controllers? For example, why would a neural network memorize highly effective strategies for pursuing an objective which it hasn't been trained on?

However, I would make the following points:

- If a mesa-searcher and a mesa-controller are equally effective, they're equally concerning. It doesn't matter what their internal algorithm is, if the consequences are the same.
- The point of inner alignment is to protect against those bad consequences. If mesa-controllers which don't search are truly less concerning, this just means it's an easier case to guard against. That's not an argument against including them in the definition of the inner alignment problem.
- Some of the reasons we expect mesa-search also apply to mesa-control more broadly.
 - "Search" is an incredibly ambiguous concept.
 - There's a continuum between searchers and pure memorized strategies:
 - Explicit brute-force search over a large space of possible strategies.
 - Heuristic search strategies, which combine brute force with faster, smarter steps.
 - Smart strategies like binary search or Newton's method, which efficiently solve problems by taking advantage of their structure, but still involve iteration over possibilities.
 - Highly knowledge-based strategies, such as calculus, which find solutions "directly" with no iteration -- but which still involve meaningful computation.
 - Mildly-computational strategies, such as decision trees, which approach dumb lookup tables while still capturing meaningful structure (and therefore, meaningful generalization power).
 - Dumb lookup tables.
 - Where are we supposed to draw the line? My proposal is that we don't have to answer this question: we can just include all of them.
 - Some of the reasons we expect mesa-search also apply to mesa-control more broadly.
 - There can be simple, effective strategies which perform well on the training examples, but which generalize in the wrong direction for off-distribution cases. Realistic non-search strategies will not actually be lookup tables, but rather, will compress the strategies a lot. Such agents probably follow perverse instrumental incentives *because it's a common theme of effective strategies*, even without search-based planning.
 - Non-search strategies can still factor their knowledge into "knowledge of the goal" vs "knowledge of the world", and combine the two to plan. (For example, the calculus-like optimization I mentioned.) This gives us a critical ingredient for deceptive agents: the training score can be improved by increasing the knowledge in the world-model instead of aligning the goal, such that the agent deceptively cooperates with the training task to achieve its own goals.
 - For non-search strategies, it's even *more important* that the goal actually simplify the problem as opposed to merely reiterate it; so there's even *more reason* to think that mesa-controllers of this type wouldn't be aligned with the outer goal.

Mesa-Learning

I mentioned this category in the same *Mesa-Search vs Mesa-Control* post. This refers to the phenomenon of [spontaneous emergence of learning algorithms](#). Basically: a system ends up learning-to-learn when you were only trying to get it to learn. This may or may not involve search.

This could be concerning/important for several reasons, but I don't have a lot I want to say about it in this post.

Explicitly Representing Values

This refers to one of the properties I mentioned in the mesa-control subsection: does a model represent its objective separately from its world-model, and combine those to plan?

Or, slightly more generally: does the system have an explicitly represented objective? (Whether or not it has a "world model".)

There are several reasons to think this might be the critical distinction for pointing to inner optimization:

- It's critical to one story for why we might expect deception from highly capable machine-learning systems, as I previously outlined.
- Representing a goal explicitly seems required for "having a misaligned goal" in a significant sense. (In other words, this just seems like a very natural definition. A system which doesn't recognize some goal as a regularity behind its strategy doesn't "have a goal" in a mechanistic sense.)
- A system that does not do this has little reason to be *systematically* misaligned.
 - That is: even if one "misaligned behavior" is learned as a generalization of effective strategies in the training data, there is little/no reason to expect another misaligned behavior to be learned (particularly not *misaligned in the same direction*, that is, *pursuing the same misaligned goal*) unless the system has compressed its strategies in terms of an explicitly represented objective. Therefore, one might argue that there is no reason to expect high levels of capability toward misaligned goals without such factoring.

I don't think these arguments are enough to supersede (misaligned) mesa-control as *the general thing we're trying to prevent*, but still, it could be that explicit representation of values is the definition which we can build a successful theory around / systematically prevent. So value-representation might end up being the more pragmatically useful definition of mesa-optimization. Therefore, I think it's important to keep this in mind as a *potential* definition.

Generalizing Values Poorly

This section would be incomplete without mentioning another practical definition: *competently pursuing a different objective when put in a different context*.

This is just the idea that inner optimizers perform well on the training data, but in deployment, might do something else. It's little more than the idea of models

generalizing poorly due to distributional shift. Since learning theory deals extensively with the idea of generalization error, this might be the most pragmatic way to think about the problem of inner optimization.

I'll have more to say about this later.

Deception

Evan Hubinger uses "[deceptive alignment](#)" for a strong notion of inner alignment failure, where:

1. There is an inner optimizer. (Evan of course means a mesa-searcher, but we could substitute other definitions.)
2. It is misaligned; it has an objective which differs from the training objective.
3. It is non-myopic: its objective stretches across many iterations of training.
4. It understands the training process and its place within it.
5. In order to preserve its own values, it "cooperates" with the training process (deceptively acting as if it were aligned).

I find that I often (accidentally or purposefully) use "deception" to indicate lesser crimes.

Hidden (possibly "inaccessible") Information

The intuition here is that a "deceptive" system is one that is hiding something from us; it knows more than it is letting on. For example, a system which is using a lot of [inaccessible information](#).

When I read [The Case for Aligning Narrowly Superhuman Models](#), my reaction was:

1. This seems like a great experiment.
2. I expect the experiment to "succeed". That is: I expect you *totally can* find a training procedure which "aligns" GPT-3 better, in the sense of getting it to give more useful responses to questions.
3. Contrary to what I perceive most people as thinking, I think *success would be bad news, and the greater the success, the worse the news* (in a sense I'll try to clarify).

My model is that GPT-3 almost certainly is "hiding its intelligence" at least in small ways. For example, if its prompt introduces spelling mistakes, GPT-3 will 'intentionally' continue with more spelling mistakes in what it generates. I expect that if a prompt begins with reasoning mistakes, GPT-3 will continue with similar reasoning mistakes to some extent.

So it seems quite plausible that we can nudge smarter behavior out of GPT-3. For example, with the following procedure inspired by Nisan's paper that Ajeya cites:

1. Train a model to predict upvotes on Quara, Stackexchange, and similar question-answering websites. This serves as a function recognizing "intelligent and helpful responses".
2. Use this upvote predictor as a loss function to fine-tune GPT-3.

I'm thinking of Ajeya's research program roughly as "come up with some better version of that".

So then a very interesting question is: would that work? How well?

Assuming step 2 doesn't just effectively create adversarial examples for the model trained in step 1 (a failure mode which Nisan's work suggests won't happen), I think the procedure would work at least to some extent. EG, it would at least stop GPT from introducing "intentional" reasoning errors. GPT-3 fine-tuned in this way would probably be *more helpful* than GPT-3 in a meaningful way.

Here's my possibly-contentious claim:

The better this works, the more evidence we have that GPT is close to worrisome territory WRT inner optimizers.

If this procedure barely works at all, then we have evidence that GPT was not "hiding anything" or "playing dumb". This means GPT was not very much like an inner optimizer. It didn't have very much private information it was using to model humans, without revealing to the humans.

If this procedure works moderately well, then this provides some confirmation that GPT had some hidden information.

If this procedure worked quite well, producing a model which answers questions quite intelligently and helpfully, this provides significant evidence that GPT-3 had a kind of inner optimizer. In particular, if this procedure was able to produce *better answers than those in the Stackexchange dataset*, this would suggest that GPT-3 knew those better answers all along, rather than developing better answers as a consequence of training on that signal.

If this procedure works *exceedingly* well, causing GPT to "wake up" and be a human-level conversation partner or greater, we should be very worried indeed. (Since we wouldn't then know the alignment of the resulting system, and could be virtually sure that it was an inner optimizer of significant power.)

Barring the final extreme outcome which creates a risk, this is a great experiment which could provide some valuable information about the potential for near-term inner optimizers.

But I worry that people could interpret the experiment incorrectly, thinking that "good" results from this experiment (ie creating much more helpful versions of GPT) are actually "good signs" for alignment. I think the opposite is true: successful results would actually be significant reason for caution, and the more success, the more reason for caution.

This does, of course, depend highly on methodology. If the methodology is not very much like what I describe, then positive results might indeed be good news rather than bad.

Also note that hiding information *does not in itself* imply inner optimization, much less inner misalignment. Ajeya privately mentioned the example of ImageNet: certainly ImageNet knows more about images than it overtly reveals through its outputs.

However, it does imply some sort of alignment problem, I think, because arguably ***alignment implies the absence of hidden information***. In the GPT-3 example, this can clearly be traced to an outer alignment problem: GPT-3 was trained to imitate humans, not to give the most useful responses possible. So GPT-3 hiding information does strike me as an example of a *misaligned meza-optimizer* even if not an *inner-misaligned* mesa-optimizer. (Depending, of course, on your preferred definition of mesa-optimizer. I have no idea whether GPT-3 conducts an internal search. Planning ahead seems like a broadly useful thing for it to do, but, we know little about GPT-3's internal strategies.)

(In an extreme case, an aligned AI might hide information from us for our own sake. However, this at least implies an absence of corrigibility, since it results in difficult-to-verify and difficult-to-correct behavior. I don't feel bad about a definition of "deception" which includes this kind of behavior; avoiding this kind of deception seems like a worthwhile goal.)

A Treacherous Turn

The core reason why we should be interested in Evan's notion of deception is the treacherous turn: a system which appears aligned until, at an opportune moment, it changes its behavior.

So, this serves as a very practical operational definition.

Note that this is identical with the "generalizing values poorly" definition of inner optimizer which I mentioned.

My Contentious Position for this subsection:

Some versions of the lottery ticket hypothesis seem to imply that deceptive circuits are already present at the beginning of training.

The argument goes like this:

1. Call our actual training regime T.
2. I claim that if we're clever enough, we can construct a hypothetical training regime T' which trains the NN to do nearly or exactly the same thing on T, but which injects malign behavior on some different examples. (Someone told me that this is actually an existing area of study; but, I haven't been able to find it yet.) ETA: [Gwern points to "poisoning backdoor attacks"](#).
3. Lottery-ticket thinking suggests that the "lottery ticket" which allows T' to work is already present in the NN when we train on T.
4. (Furthermore, it's plausible that training on T can pretty easily find the lottery ticket which T' would have found. The training on T has no reason to "reject this lottery ticket", since it performs well on T. So, there may be a good chance that we get an NN which behaves as if it were trained on T'.)

Part of my idea for this post was to go over different versions of the lottery ticket hypothesis, as well, and examine which ones imply something like this. However, this post is long enough as it is.

So, what do we think of the argument?

I actually came up with this argument as an argument *against* a specific form of the lottery ticket hypothesis, thinking the conclusion was pretty silly. The mere existence of T' doesn't seem like sufficient reason to expect a treacherous turn from training on T.

However, now I'm not so sure.

If true, this would argue against certain "basin of corrigibility" style arguments where we start with the claim that the initialized NN is not yet deceptive, and then use that to argue inductively that training does not produce deceptive agents.

[Letter] Advice for High School #1

Hello! I'm a big fan of your work. Your perspectives on issues, while I may not always agree with them, are always insightful; thank you for helping make the world a more interesting place. I'm interested in learning Lisp and Vim, as it's been recommended many times by writers I find interesting. While I won't be able to dedicate my full time towards it now (due to good old high school+an intensive summer program), at some point in the future I am interested in putting work into learning these. Any courses or books you'd recommend? How did you learn? Also, any advice for getting started blogging? I think it'd be a productive use of my time and help improve my writing skills, but haven't been able to give myself the initial push. Finally, if there's any advice that you would have given to your high school self that may prove relevant to me, please pass it on; if I know I'm going to learn something in the future from future mistakes, I may as well try to learn the lesson now. Hope you're doing well.

Yours,

[redacted]

Dear [redacted],

I am happy to hear you do not always agree with me. Sometimes I am wrong. You should not agree with people when they are wrong.

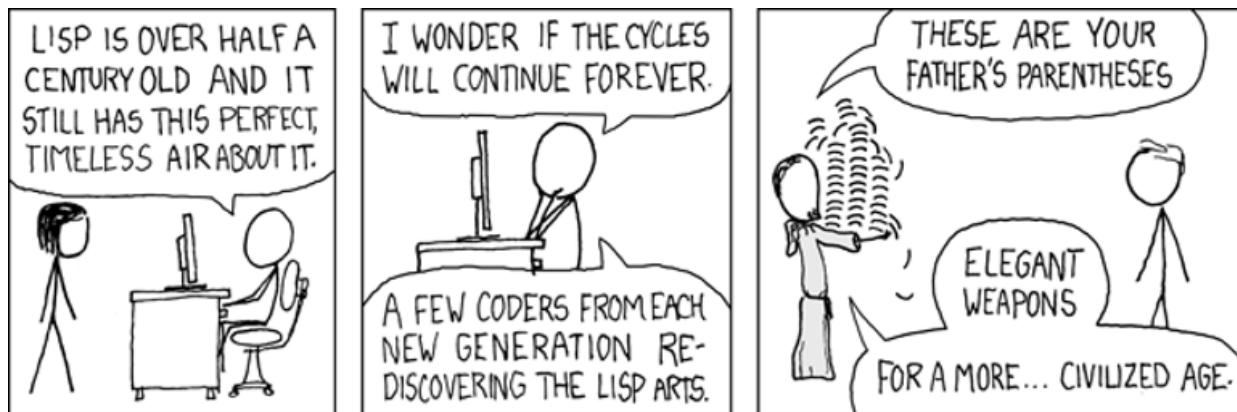
Vim

My favorite book on Vim is *Practical Vim* by Drew Neil but the free books [here](#) are probably fine. [My post on Vim](#) is enough to get you started. The most important thing is that you write everything in Vim. I wrote this blog post in Spacemacs with Vim keybindings before cut & pasting it into the Less Wrong Markdown editor.

Lisp

The utility of a programming language comes from its syntax and its libraries. Lisp has the best syntax. Good libraries come from popularity. Lisp has never been popular. The best Lisp dialects appropriate libraries from other languages. [Hy](#) (which integrates with Python libraries) is the one I use but I have heard good things about the others.

There are no good books on Lisp. The Lisp books I know of all use obsolete dialects. I learned by reading a few chapters of [Practical Common Lisp](#) and [On Lisp](#) and then working out the rest on my own. There is no well-trodden path.



I recommend you just use Hy in place of Python and then try to write short code. The rest may take care of itself.

Blogging

My favorite book on creative pursuits is [Steal Like an Artist](#) by Austin Kleon. Like Vim, the important thing about blogging is that you just get started. Create a free Wordpress, GitLab or Less Wrong account and write stuff. Use a pseudonym that doesn't commit you to a single topic. At first your posts will be bad. Period. This isn't my first blog. I deleted my first blog. It was awful. Yours will be too. Don't let it discourage you.

It will never be the right time to start a blog.

If you think it's restrictive being a kid, imagine having kids.

—[What You'll Wish You'd Known](#) by Paul Graham

High School Advice

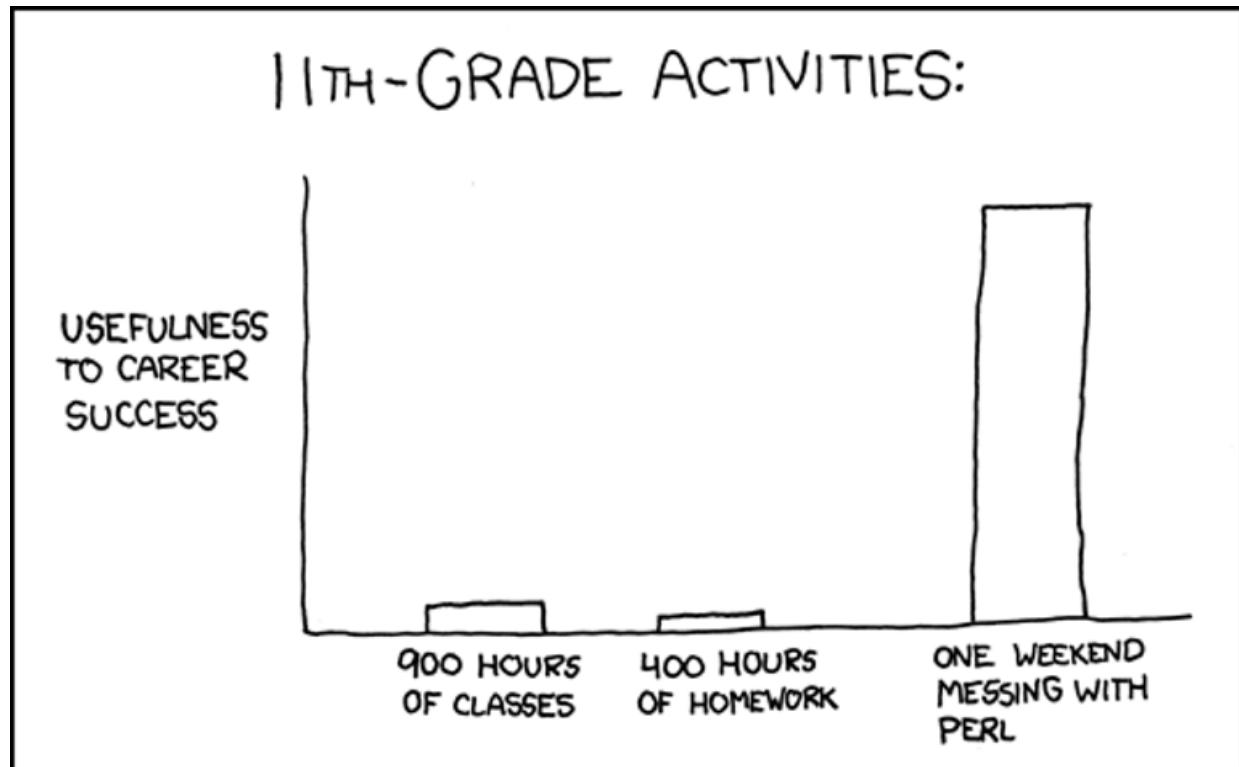
High school is fake. The English classes don't teach you how to write. The foreign language classes don't teach you how to communicate. The art classes don't teach you how to draw. The history classes are pure propaganda. The physical education involves no physical conditioning. The math classes are designed for students of average intelligence. Putting a smart teenager in a high school math class [is like putting an average teenager in an insane asylum](#). The science classes are like the math classes.

[Adults lie to kids.](#) Here are some books, articles and lectures to start getting the styrofoam out of your head.

- All of Paul Graham's [articles](#) plus [this article](#) by Sam Altman
- *The Case against Education: Why the Education System Is a Waste of Time and Money* by Bryan Caplan
- *Excellent Sheep: The Miseducation of the American Elite and the Way to a Meaningful Life* by William Deresiewicz
- The first couple chapters of *Manufacturing Consent: The Political Economy of the Mass Media* by Edward S. Herman and Noam Chomsky
- Jordan Peterson's lectures on IQ and the Big Five Personality Traits

- *Nineteen Eighty-Four* by George Orwell
- *The Red Queen: Sex and the Evolution of Human Nature* by Matt Ridley
- *The World Until Yesterday: What Can We Learn from Traditional Societies?* by Jared Diamond

Even if high school did teach you something important it wouldn't have economic value because economic value is driven by [supply and demand](#). Everyone goes high school. Therefore the market value is low.



My dad practiced Chinese in army school. His American commanding officers thought he was taking notes but actually he was reciting poetry. Randall Munroe printed books in a tiny font so he could read them during class. I wish I had read *Drawing on the Right Side of the Brain* by Betty Edwards so I could practice drawing. All I achieved was really good cursive handwriting. You might want to try using your time to write blog posts. I give you permission to use a pencil instead of Vim while in class if that's what it takes.

It is worth paying attention in class because a tiny bit of knowledge is better than none at all. Also your grades matter for getting into college. After you graduate high school, nothing matters except what college you got into and the things you learned—most of which you taught yourself. Everything that merely embarrassed you ceases to matter. [Gamble things you can afford to lose.](#) **What really matters in the long run is what you teach yourself.**

College Advice

Technical and economic leverage increases every year. The marginal value of teaching yourself has never been higher. Going to college used to be a no-brainer. Now it's complicated. It depends on who you are and I don't know who you are.

If you go to college then most important thing is to graduate with minimum debt. Indentured servitude counts as debt.

Cheat Codes for Life

Everything in the following list has an [absurdly huge payoff](#) compared to the investment required.

- Learn to use [Anki flashcard software](#).
- Read *Starting Strength* by Mark Rippetoe. If you're a skinny guy following the program and you aren't gaining weight then eat more. If you're still not gaining weight and you're lactose tolerant then drink a gallon of whole milk everyday. Round things out by reading *Training for the New Alpinism: A Manual for the Climber as Athlete* by Steve House and Scott Johnston.
- Start a meditative practice. [WARNING: Meditation comes with dangers. Read Ingram's book for details so you know what you're getting into.] My favorite book on this topic is *Three Pillars of Zen: Teaching, Practice, and Enlightenment* by Philip Kapleau Roshi. It is very Japanese so it may not make sense to you. *Mastering the Core Teachings of the Buddha: An Unusually Hardcore Dharma Book* by Daniel M. Ingram is written for a Western audience. Meditation should feel like it's working. If it doesn't feel like it's working then you should stop doing it because you're doing something wrong.
- Read *How to Win Friends and Influence People* by Dale Carnegie. Learn to sell things.
- Ruthlessly prune [angry people](#) from your life.
- Learn to use Vim. 🎉 You are already doing this!
- Learn to program computers. 🎉 You are already doing this!
- Start writing a blog as early in your life as possible. 🎉 You already plan to do this!

Lastly, check out [Dresden Codak](#). It's really cool.

Yours sincerely,

Lsusr

The dialogue continues [here](#).

Updating the Lottery Ticket Hypothesis

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

Epistemic status: not confident enough to bet against someone who's likely to understand this stuff.

The lottery ticket hypothesis of neural network learning (as [aptly described](#) by Daniel Kokotajlo) roughly says:

When the network is randomly initialized, there is a sub-network that is already decent at the task. Then, when training happens, that sub-network is reinforced and all other sub-networks are damped so as to not interfere.

This is a very simple, intuitive, and useful picture to have in mind, and the [original paper](#) presents interesting evidence for at least some form of the hypothesis. Unfortunately, the strongest forms of the hypothesis do not seem plausible - e.g. I doubt that today's neural networks already contain dog-recognizing subcircuits at initialization. Modern neural networks are big, but not *that* big. (See [this comment](#) for some clarification of this claim.)

Meanwhile, a cluster of research has shown that large neural networks approximate certain Bayesian models, involving phrases like "neural tangent kernel (NTK)" or "Gaussian process (GP)". [Mingard et al. show that](#) these models explain the large majority of the good performance we see from large neural networks in practice. This view also implies a version of the lottery ticket hypothesis, but it has different implications for what the "lottery tickets" are. They're not subcircuits of the initial net, but rather subcircuits of the *parameter tangent space* of the initial net.

This post will sketch out what that means.

Let's start with the jargon: what's the "parameter tangent space" of a neural net? Think of the network as a function f with two kinds of inputs: parameters θ , and data inputs x . During training, we try to adjust the parameters so that the function sends each data input $x^{(n)}$ to the corresponding data output $y^{(n)}$ - i.e. find θ for which $y^{(n)} = f(x^{(n)}, \theta)$, for all n . Each data point gives an equation which θ must satisfy, in order for that data input to be exactly mapped to its target output. *If* our initial parameters θ_0 happen to be close enough to a solution to those equations, then we can (approximately) solve this using a linear approximation: we look for $\Delta\theta$ such that

$$y^{(n)} = f(x^{(n)}, \theta_0) + \Delta\theta \cdot \frac{\partial f}{\partial \theta}(x^{(n)}, \theta_0)$$

The right-hand-side of that equation is essentially the parameter tangent space. More precisely, (what I'm calling) the parameter tangent space at θ_0 is the set of functions $F(x)$ of the form

$$F(x) = f(x, \theta_0) + \Delta\theta \cdot \frac{\partial f}{\partial \theta}(x, \theta_0)$$

... for some $\Delta\theta$.

In other words: the parameter tangent space is the set of functions which can be written as linear approximations (with respect to the parameters) of the network.

The main empirical finding which led to the NTK/GP/Mingard et al picture of neural nets is that, in practice, that linear approximation works quite well. As neural networks get large, their parameters change by only a very small amount during training, so the overall $\Delta\theta$ found during training is actually nearly a solution to the linearly-approximated equations.

Major upshot of all this: the space-of-models “searched over” during training is approximately just the parameter tangent space.

At initialization, we randomly choose θ_0 , and that determines the parameter tangent space - that’s our set of “lottery tickets”. The SGD training process then solves the equations - it picks out the lottery tickets which perfectly match the data. In practice, there will be many such lottery tickets - many solutions to the equations - because modern nets are extremely overparameterized. SGD effectively picks one of them at random (that’s one of the main results of the [Mingard et al work](#)).

Summary:

- The “parameter tangent space” of a network is the set of functions which can be written as linear approximations (with respect to the parameters) of the network.
- The parameter tangent space at the network’s randomly-chosen initial parameters is roughly the set of “lottery tickets”.
- SGD (effectively) throws out any lottery tickets which don’t perfectly match the data, then randomly picks one of the remaining tickets.

Of course this brushes some things under the rug - e.g. different “lottery tickets” don’t have exactly the same weight, and different architectures may have different type signatures. But if you find the original lottery ticket hypothesis to be a useful mental model, than I expect this to generally be an upgrade to that mental model. It maintains most of the conceptual functionality, but is probably more realistic.

Thankyou to Evan, Ajeya, Rohin, Edouard, and TurnTrout for a discussion which led to this post.

"Taking your environment as object" vs "Being subject to your environment"

I think there's a key rationalist skill of being able to take your environment as object, rather than being subject to it.

There's a kind of wisdom people get when they leave environments. Today I talked to a friend of mine who has moved out of living in a city for over a year, and on visiting a city she noticed things about it that made her feel pressured and unpleasant that she'd never noticed before.

I've also done several interviews with people who are quitting organizations or leaving communities, asking them why they're leaving, and there's a certain lightness to them and their speech about the place. They can talk about the negatives and the positives freely, and don't feel anxiety toward finding ways to balance their negatives with equal positives, like they're supposed to justify their environment as 'good'. They just speak plainly. I can hear them 'admitting' things a little with a chuckle, as though it was always true but not something they'd felt able to say until now.

Here's a slightly different example, but that's focused on a similar sort of mental move: I recently was on an intensely restrictive diet, because I thought it was very healthy and would cause me to lose weight. I'd done this diet before, but this time I had a much more unpredictable workload, which messed up my routine and I crashed several times from under-eating. I finally decided to let myself broaden the diet notably, and on the first bite of new food I had 2 realizations.

Firstly, I didn't actually believe in the previous diet for its own sake. It was actually to make me disciplined about my food.

Secondly, I hadn't let myself think that thought during my diet, I think because it would have been too much strain for me to discipline myself in this way just to discipline myself. It was much less strain to think that my diet had some magical properties for my health.

Recently, I read and participated in Eliezer's [dath ilan thread](#), where he answered questions about his home civilization.

Eliezer has a fantastic imagination. He is able to imagine what our civilization would like, from the perspective of someone in a *different* civilization. He is able to just say what it looks like from this outsider's perspective, and say what are probably very obvious things from their perspective. But when you're on the inside it's far harder. You can sometimes do it with normal means within the system, but it can end up being a *lot* of hard work to get these insights. It can be much easier to imagine someone outside of the system looking in.

I've been reading some of Michael Vassar's Twitter threads. I think he has a similar ability to look at civilization as a whole and say truths that people collectively avoid looking at, but I don't think it comes from the same source. One model I have of Michael is someone who feels they are living in a hostile state, like Soviet Russia or Nazi Germany. This is a situation where you viscerally feel the threat to your life of

"feeling comfortable in the system", and are searching for alternative perspectives in an attempt to fight your way out. It's a much more adversarial frame, and it's more likely to notice parts of the society that are unjust and criminal.

Another model: I feel like Michael Vassar is similar to Luna Lovegood, who I empathized with a great deal when reading the HPMOR fanfic [Luna Lovegood and the Chamber of Secrets](#). She has a pretty clear distinction between "what she believes" and "what narratives the Daily Prophet wants you to believe". She assumes the latter is primarily a tool for the powerful, and is able to find pretty valuable things by separating the it from the former. I wouldn't say she is wise enough to see the full truth, but she is able to correctly pick up on several patterns in the world most others miss.

So far I have said there are three ways of getting perspective on your environment: leaving it, imagining yourself into someone outside of it, and assuming that it's hostile.

I have one more to add: I also think that 'breaks' are a fine tool for the toolkit. [Sabbaths](#) are things I have found very useful. And in January I took a whole two-week vacation in a new environment where I didn't use my screens/devices, and this really helped get me out of my head.

(I didn't literally never use them, but I used them at least 10x less, and only let myself use them according to stringent rules. I'm not going to write down all the details but it was something like "I am only allowed to do a thing with my device if I wrote the thing down on a sheet of paper yesterday".)

On that vacation I also read some great literature that I've never read before (Heinlein, Dostoyevsky, Tolkien and more) which has helped me empathize with people outside of my civilization. (I've been trying to write some dialogues between myself and Lord [Denethor II](#) discussing our respective civilizations, which has been a trip.) It's been deeply rewarding and worthwhile in other ways I won't list here, but I've made sure to keep it up since returning (Asimov, Watt-Evans, Herbert, and more; I'm looking forward to reading [Gulf](#), on Eric S. Raymond's [recommendation](#)).

I think people manage to spend their whole lives never thinking outside of their environment, or way-of-being. It's a painful thought; and I'd like to build an environment where people reliably do.

What are some other ways to successfully take your environment as object?

Added: Aella [describes](#) a way that she does this move regularly:

"How would I feel about [topic] if I moved to an alien planet where my role and the social norms were utterly, completely different?" is maybe the most common calibration question I ask myself

[1] I note that Luna doesn't think the world is as hostile, and has a generally more fun and curious time than someone living in Soviet Russia.

Covid 4/15: Are We Seriously Doing This Again

Yes we are. [It can happen here.](#)

THIS IS LITERAL ONE IN A MILLION AND MUCH LESS THAN THE BASE RATE WHAT THE ACTUAL FUCK IS WRONG WITH YOU DID YOU NOT SEE WHAT HAPPENED LAST TIME ARE YOU COMPLETE MORONS OR ARE YOU MUSTACHE-TWIRLING VILLIANS YOU CAN'T NOT BE BOTH, AS IN IF YOU'RE NOT MORONS AND I LOOK AT PHOTOGRAPHS I WILL SEE MUSTACHES AND YOU PEOPLE WILL BE TWIRLING THEM:



James Miller @JimDMiller · 26m

...

Please never again claim that the Biden administration represents some kind of return to science if they stop administration of the J&J vaccine because of 6 blood clots.

U.S. Calls for Pause on Johnson & Johnson Vaccine After Clotting Cases

April 13, 2021 in News



Kaitlan Collins ✅ @kaitlancollins · 12m

...

Dr. Carlos del Rio, executive associate dean at Emory University School of Medicine, says on CNN "you have a much higher chance of getting run over by a car" than a blood clot from J&J. Says to alert a doctor if you develop shortness of breath, leg pain, headache within 2-3 weeks

29

255

648

↑



Dr. Angela Rasmussen



@angie_rasmussen

For perspective, here are some numbers:

1 in 1,000,000: J&J vaccine

1 in 3,000: oral contraceptives

1 in 5: hospitalized COVID-19 patients

As someone who got the J&J vaccine 8 days ago, and who took oral contraceptives for 20 years, I'll take these odds.

WASHINGTON — Federal health agencies on Tuesday called for an immediate pause in use of Johnson & Johnson's single-dose coronavirus vaccine after six recipients in the United States developed a rare disorder involving blood clots within about two weeks of vaccination.

All six recipients were women between the ages of 18 and 48. One woman died and a second woman in Nebraska has been hospitalized in critical condition.

Nearly seven million people in the United States have received Johnson & Johnson shots so far, and roughly nine million more doses have been shipped out to the states, according to data from the Centers for Disease Control and Prevention.

"We are recommending a pause in the use of this vaccine out of an abundance of caution," Dr. Peter Marks, director of the Food and Drug Administration's Center for Biologics Evaluation and Research, and Dr. Anne Schuchat, principal deputy director of the C.D.C., said in a joint statement. "Right now, these adverse events appear to be extremely rare."

While the move was framed as a recommendation to health practitioners in the states, the federal government is expected to pause administration of the vaccine at all federally run vaccination sites. Federal officials expect that state health officials will take that as a strong signal to do the same.

In the United States alone, 300,000 to 600,000 people a year develop blood clots, according to C.D.C. data. But the particular blood clotting disorder that the vaccine recipients developed, known as cerebral venous sinus thrombosis, is extremely rare.

All of the women developed the condition within about two weeks of vaccination, and government experts are concerned that an immune system response triggered by the vaccine was the cause.

If any of them don't have mustaches, we need to get them some clip-on ones, because while lots of people die at least they should get to enjoy the pleasures of twirling.

Yes I am fully aware that it is technically a particular rarer blood clotting disorder that is happening here and thus in that subclass it is above the base rate and that there's an argument this might be 'real' in some sense and no I do not care even a little bit about any of that and no I am not going to treat this with the dignity and respect that it does not in any way even potentially deserve. There are scientific details and if you find them interesting by all means read about them but I am ignoring them [because like the points They. Do. Not. Matter.](#)



Nate Silver



@NateSilver538

6 cases out of 7 million people.
What a disaster. This is going to
get people killed. And it's going to
create more vaccine hesitancy.
These people don't understand
cost-benefit analysis. They keep
making mistakes by orders of
magnitude.

Noah Weiland



@noahweiland

Scoop: The federal government is calling for a pause in use of the Johnson & Johnson vaccine after 6 cases of a blood clotting disorder. A major setback for the national vaccination campaign. w/ @SharonLNYT [nytimes.com/2021/04/13/us/...](http://nytimes.com/2021/04/13/us/)

6:59am · 13 Apr 2021 · Twitter for Android



Nate Silver ✅ @NateSilver538 · 15m

...

It's also a high-stakes test for the FDA, and they failed it, because of course lots of people are going to take away the latter message.

Matthew Gertz ✅ @MattGertz · 46m

I am extremely skeptical of the ability of public messaging to disaggregate "the J&J vaccine is under review as a precaution" from "the J&J vaccine is not safe and the others may not be either" in the minds of normal people. An incredibly crucial, high-stakes test for the press.

[Show this thread](#)

36

97

615

↑



Nate Silver ✅ @NateSilver538 · 13m

...

There's also data on this based on decreased public confidence in the AstraZeneca vaccine in Europe following similar pauses there. So the FDA can't even use the excuse of flying blind.



David Frum Retweeted



[Graeme Wood](#) ✅ @gcaw

7m

We're still allowed to cross a busy street, or eat raw/uncooked fish, or stick our face in a fan. But we can't get a shot that protects ourselves and others from serious illness and allows a return to normal life—because it *might* have a 1-in-7-million fatal clotting issue

In case you were wondering how people were going to react or what this would do to public confidence, [these are from less than an hour after the announcement:](#)



David Hammond @solismuzik · 27m

...

Replies to [@moreisdifferent](#) and [@US_FDA](#)

It's way more than six people, based on news media reports alone. The authorities and their lapdog media have been doing nothing but lying to us this entire time and this is yet another example. Wake up.



Ulysses Everett McGill @mcgill_ulysses · 17m

...

Replies to [@moreisdifferent](#) and [@US_FDA](#)

These numbers aren't static. There's obviously more going on here, likely plenty more cases awaiting analysis.



DaniForPeace @DaniForPeace · 32m

...

Replies to [@moreisdifferent](#) and [@US_FDA](#)

Six is the number they are admitting to. It's many more. They would not halt for just six. Silly rabbit.



I mean, they're wrong, but I can't fault their reasoning from where they sit, if you asked me the 'which is more likely' game back in 2019 I would most definitely have not have gone with 'no really they're doing this in a pandemic because of six cases.'

Several hours in:



Tim Alberta ✅ @TimAlberta · 3h

...

We've already heard from two friends this morning -- unsolicited -- that this J&J stoppage has convinced them the vaccines were rushed and can't be trusted.

Heckuva job.



363



544



3.2K



Again, seems logical to conclude they were rushed if they act in a way that would only make sense if they actually did rush.

The first time around with AstraZeneca, I could *sort of* understand the argument for the other side of the hesitancy effect when I squinted, that this would look like the Very Serious People Take Vaccine Safety Seriously and therefore we should now expect the people to trust the FDA more, and being untrustworthy stewards who kill a bunch of people in order to fool the public into thinking we are trustworthy is a tradeoff they thought we should make, but [we ran the experiment on that hypothesis](#), and, yeah, no.

Seriously, [people are not so stupid:](#)



Daniel Eth 
@daniel_eth

If the FDA banned pepsi because of recent cases of it causing blood clots, that wouldn't make me feel safer about drinking coke

10:39pm · 13 Apr 2021 · Twitter Web App

Also, when you keep saying loudly that any adverse things that happen will destroy your credibility, [consider your credibility preemptively destroyed already](#) because you're either right or wrong:

When the FDA announces that they have to ban a vaccine because its credibility is on the line, that very announcement puts their credibility on the line. It is a simple two-line proof. Either they are lying about whether their credibility is on the line, in which case they have wrecked their credibility with the lie. Or they are telling the truth, in which case by definition their credibility is indeed on the line.

This is going to *permanently* supercharge the anti-vax movement, not only on Covid but also in general, and kill a *huge* number of people. Over six cases. Note deaths. Cases. Six.

You know how many people *died*?

ONE. F***ing ONE.

No, this was not 'going to come out.' It was going to completely correctly get ignored. All they had to do was put it in the list of side effects and note it was extremely rare.

[This is the Washington Post's attempt to chronicle what was going through their minds](#), it's sympathetic but doesn't make the decision look any less absurd if you actually think about the physical situation at all or how real people would react to it.

If I had to steelman the case being made, it's a combination of believing that acting over-the-top paranoid about side effects makes people feel more confident rather than less confident, that a pause to inform people can meaningfully impact care for this rare type of blood clot, and thinking that until one looks at the data who knows how big the problem might be and one shouldn't assume the math is right until you check it, so we should halt and catch fire for a day and then quickly convene a meeting to confirm that this is only going to kill one person in six million.

Even in a world in which the initial pause wasn't crazy, there was a meeting the next day to go over the information, and the decision was made to wait 7-10 days and then meet again without making a choice about the pause (and obviously, here, [if you choose not to decide](#)

[you still have made a choice](#)). They didn't even make the 'compromise' decision of halting for young women (and yes, 'people who are in the subpopulation that is often on birth control which causes orders of magnitude more blood clots than this seems like it's a hint on what's happening) and continuing for everyone else, since you can then swap doses between different groups and keep up your pace of vaccinations while you 'investigate further' whatever that means here. The failure to at least make *that* decision is *obviously* completely bonkers even if you somehow think the initial decision to halt and catch fire was reasonable, [as laid out in this thread](#) by someone who supported the first decision but at least supported the 'compromise' option at the meeting.

[Here's an argument that this isn't so bad in the United States](#), as it will mostly only destroy faith in *Johnson & Johnson*, rather than faith in the mRNA vaccines as well, or all vaccines generally:

Moderna and BioNTech shares jumped 10.5 per cent and 6.1 per cent, respectively, on Tuesday as the vaccine makers benefited from news of the J&J pause.

Norway's health authorities estimated that their vaccination plans could be delayed by eight to 12 weeks if they could not use either the J&J or the Oxford/AstraZeneca vaccine.

FDA [Delenda Est. The FDA must be destroyed.](#)



James Babcock @jimrandomh · 1h

...

Let's not sugar-coat it; this is merely the latest in a long string of major acts of sabotage by the FDA against COVID-19 response in general. The FDA is a rogue agency and needs to be disbanded and rebuilt from scratch.

At a minimum, while we prepare to do that, we can at least implement [Tyler's modest proposal.](#)

Let's run the (other, not equal to one or six) numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 5.9% (up 0.4%) and deaths decline by 8%.

In the past week in the U.S. ...

New daily reported **cases rose 9.3% ↑**

New daily reported **deaths fell 0.3% ↓**

Covid-related **hospitalizations rose 2.2% ↑** [Read more](#)

Among reported tests, **the positivity rate was 5.6%**.

The **number of tests reported fell 6.7% ↓** from the previous week. [Read more](#)

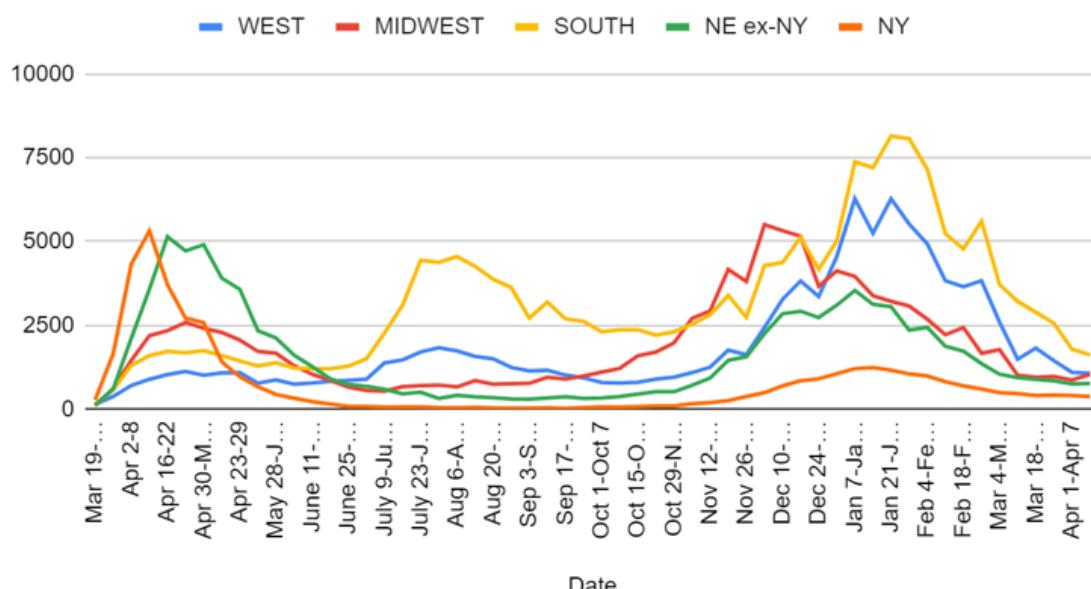
Once again Washington Post's numbers baffle me, although this being six rather than seven days later makes them not impossible. Somehow tests fell, cases rose, and the positivity rate barely budged.

A key question whenever one gets good news on deaths is whether this is good news or whether it's time shifted. If it's cases shifting into the future, it means the next week looks doubly worse and on top of that you were fooled by what looked like a downward trend. Similarly, bad news can be a mirage from old cases. It now looks like the death rate decline has stalled out, which is unfortunate.

Predictions for next week: Positivity rate of 5.8% (up 0.2%) and deaths unchanged.

Deaths

Deaths by Region



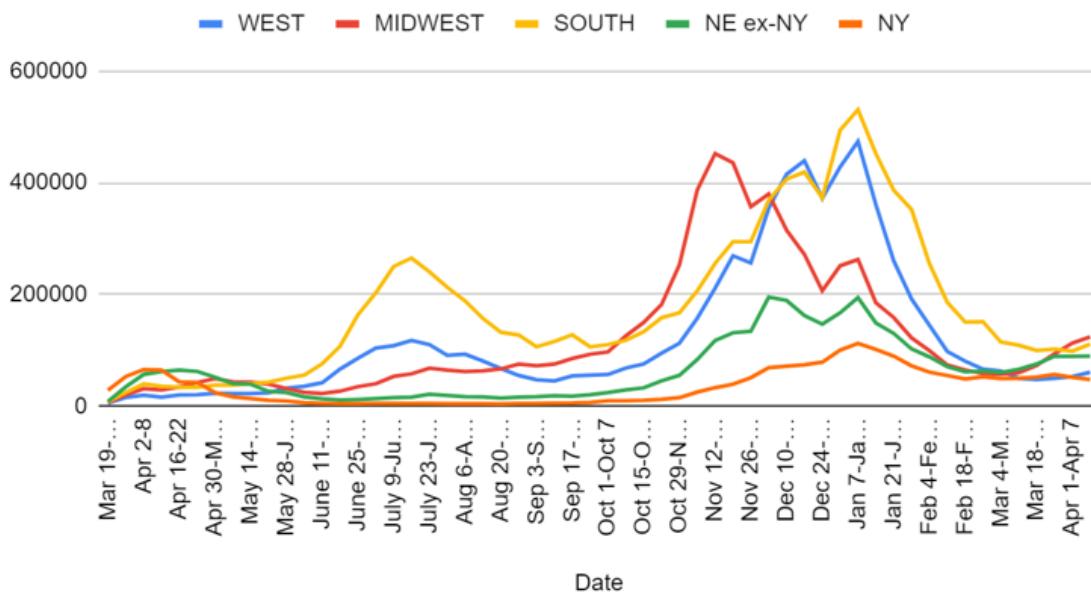
Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Feb 25-Mar 3	3834	1669	5610	1958	13071

Mar 4-Mar 10	2595	1775	3714	1539	9623
Mar 11-Mar 17	1492	1010	3217	1402	7121
Mar 18-Mar 24	1823	957	2895	1294	6969
Mar 25-Mar 31	1445	976	2564	1262	6247
Apr 1-Apr 7	1098	867	1789	1160	4914
Apr 8-Apr 14	1070	1037	1621	1145	4873

Half or more of the Midwest increase is quirky data in Missouri, but that doesn't make any of this good news, and it's likely deaths are going to now be stable or go slightly up, along with cases, until we get enough vaccinations to turn things around.

Cases

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Feb 25-Mar 3	66,151	58,295	151,253	115,426
Mar 4-Mar 10	62,935	57,262	114,830	109,916
Mar 11-Mar 17	49,696	59,881	109,141	115,893
Mar 18-Mar 24	47,921	72,810	99,568	127,421
Mar 25-Mar 31	49,669	93,690	102,134	145,933
Apr 1-Apr 7	52,891	112,848	98,390	140,739
Apr 8-Apr 14	60,693	124,161	110,995	137,213

Looking at this chart, it seems clear the Midwest's problems are real. The finale wave is out in force there, even if it's relatively tame in other places.

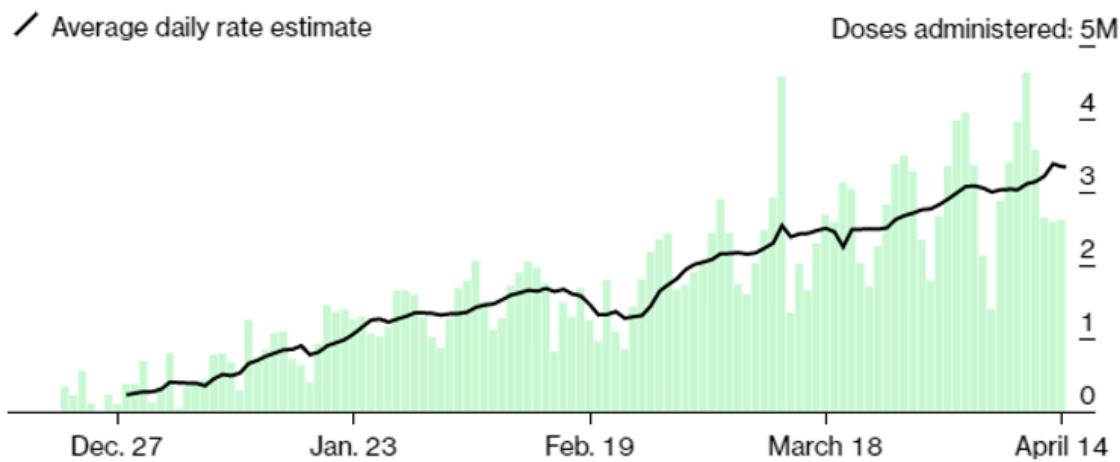
Given the increase in positive tests, and the report of a continued decline in test counts, I'm willing to believe that positive rates did go up ~0.4% in the past week, which Johns Hopkins confirms (although they have lower numbers on both ends than WaPo does), so the

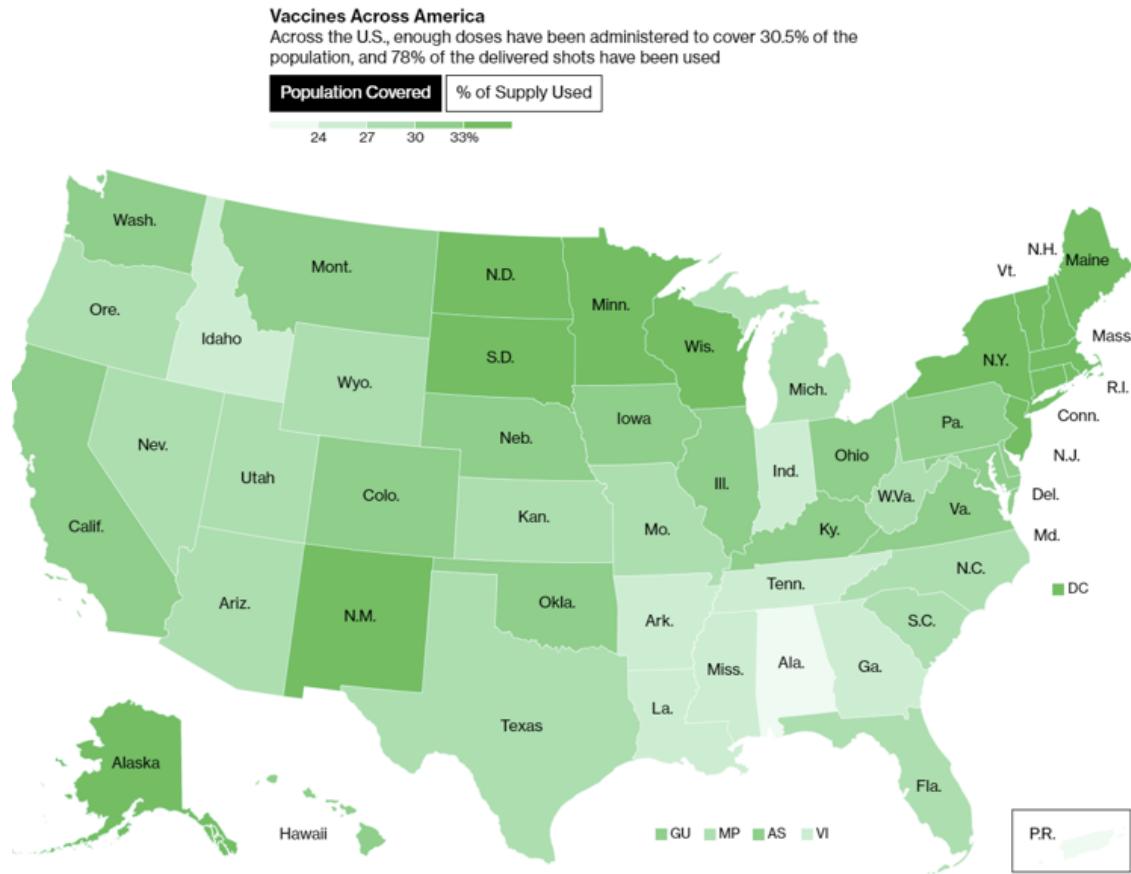
prediction miss was mostly about doing it based on Friday's number or some similar quirk (or a math error on their end somewhere).

Things in many places other than the United States are quite bad. In India, they surpassed 200,000 cases per day and things are rapidly getting worse, and there are many other places that have big problems. Aside from the places that successfully did full suppression, the places doing actively well are the ones with strong vaccination campaigns. Facing the new strains while not keeping up in vaccinations is a very bad place to be right now.

Vaccinations

In the U.S., **195 million doses** have been given so far. In the last week, an average of **3.35 million doses per day** were administered.





That small decline at the end might have something to do with the J&J suspension, or it could mostly be a random quirk. Either way, even without J&J, we still should be able to continue slowly expanding our vaccination rates until we hit a wall where we run out of people who want a shot put into their arm. There are signs this is starting to be an issue in some places, but mostly there's still plenty of people eager to put the pandemic behind them.

As furious as I am at the J&J suspension, and as many people as it's going to kill (most of which will be from disadvantaged groups and areas, which J&J's one shot at room temperature made much easier to reach), it is important not to lose perspective. J&J was a small portion of our vaccine effort, and case growth is not that rapid, so it's not going to kill hundreds of thousands of people, at least not in America. If we're lucky, it will only kill thousands.

Vaccine Passport Hype

[Washington Post reports on New York's Excelsior pass](#), the first one of its kind. Conclusion was that it's relatively easy to use and isn't got reasonable privacy protections given the circumstances, but that unless you're counting on ID to catch fraud it's trivial to fake it via copying someone else's pass. That sounds about right. It's clear by now that there isn't going to be a national system and that New York is the exception rather than the rule. That doesn't render the questions moot, but it lowers their urgency and importance quite a lot.

Tyler's position is that we should be planning full reopening, and that passports seem more likely to hinder that than help. That's one of the key disagreements. Is the alternative to passports a full reopening, or is it more restrictions? My guess remains that not being able to check leads to more restrictions in the medium term (next few months), but there's a point

when that flips, and things that would have fully reopened without checking would, if given the opportunity to easily check vaccine status, continue to check that status for a while longer. We then have to balance these needs. My guess is that the ‘overtime’ period’ is 50% to last at least two months or so, but highly unlikely (<10%) to last for six, and that the ‘extra game time’ period when passports would help starts now and has at least three months to go most (75%) of the time, and there’s a decent chance (25%) it’s six months or more in at least many blue areas, so one can do a cost/benefit calculation with this plus all the other objections. Here I’m counting the extra restrictions as pure downside, because even with them the net risk is likely higher than with the pass, unless we’re checking physical cards at the door, which is a different cost/benefit tradeoff.

The other half of that is the argument from focus. If the country and discourse only have so many focus points ([Imperial Focus Points!](#)), which seems basically right, then it’s plausible that all the work on passports delays the full reopening not because of lowering the costs of not reopening fully, but by preventing the attention and blame pressure required to generate the reopenings. Doing anything at all, in this model, has high opportunity costs. I don’t think I value this as highly as Tyler but I’ve likely not been giving it its due.

Vaccines Still Work

The J&J suspension goes hand in hand with the ongoing campaign to convince the public that vaccines work, but don’t work in the sense of accomplishing anything for people. In the name of some combination of proving one’s Very Serious Person credentials, maximizing the quantity of economic harm and scaring people as much as possible, there’s a competition on how to give the impression that being out there is unsafe for the fully vaccinated.

[Zeynep points us to an especially creative entry here:](#)



zeynep tufekci

@zeynep

...

CNN article on how to safely fly claims that 90% vaccine efficacy means that 😞 for every million who fly, we could have 100,000 infections. NO NO NO. That’s not what that number means. Also, this didn’t even happen when millions flew unvaccinated. So how could it make sense now?

Case count, masks and ventilation are key

"There are three factors to consider," said Linsey Marr, a professor of civil and environmental engineering at Virginia Tech, who studies the airborne transmission of Covid-19.

"How prevalent is the virus in the population? If it's highly prevalent, then there's a good chance that someone who is infected is going to be on a plane," Marr said.

Why does that matter if you're vaccinated? "We're still learning how effective the vaccines are against variants of the virus," the CDC said recently, as well as "how long COVID-19 vaccines can protect people."

In addition, real world studies of the Pfizer-BioNTech and Moderna vaccines show they are only 90% protective against the coronavirus, not 95% as reported in clinical trials.

Translated into reality, that means for every million fully vaccinated people who fly, some 100,000 could still become infected.

"Is everyone masked? That's also very important," said Marr, who is world renowned for her 2011 discovery that influenza can hover in air for an hour via respiratory microscopic droplets called aerosols.

What CNN is saying might be technically correct. A model where 90% of people who are vaccinated are fully safe, while 10% remain at similar risk to before vaccination, is simplified but mostly plausible. What CNN is *technically* saying here is that there are 100k people who are being exposed to possibly getting infected (look around, could it be *you*?) and Zeynep is pointing out that this is damn well written to give the ordinary person the impression that if we didn't Do Something About It that one in ten people who fly vaccinated would get infected, so if you're vaccinated and fly that way there's a 10% chance you get infected, which is of course complete nonsense.

Even without this willful mislead it's still terrible and leads to scaremongering, but this here is something special. There should be some kind of award for such things.

Also, Nate Silver is correct here, and it would be dishonest to treat 'we don't know if vaccinated people can transmit' FUD spreading as anything but gaslighting.



Nate Silver @NateSilver538 · Apr 10

...

It's pure gaslighting at this point to say we don't know whether vaccinated people spread the virus. Tons of studies—including from the CDC!!!—show that vaccines massively reduce (though probably not entirely eliminate) transmission.



Healthcare Insider @HealthInsider · Apr 9

Fauci said it's important for all Americans — both vaccinated and unvaccinated — to continue avoiding crowds and socially distancing until we know for sure that vaccinated people don't spread the virus.

Patience will "keep a lid" on cases, he said.

businessinsider.com/vaccines-reduc...

[Show this thread](#)

780

1.9K

8.6K

↑

Here's the Business Insider article linked there, which notes:

- **New Pfizer data suggests its COVID-19 vaccine is 94% effective at protecting against asymptomatic cases.**
- **It's further evidence that the vaccines reduce coronavirus transmission.**
- **Studies also show that vaccinated people have lower viral loads, which is linked to less spread.**

Vaccinated people are almost certainly *less infectious when they do get infected*, on top of *not getting infected*. The reduction in risk to others is 'we don't know' to the extent that it might be *much safer* than the 90-95% range in which it reduces risk of infection.

Anyone who tells you otherwise is either lying to you, or is believing the lies told to them by others. Those who continue to treat vaccinated people as risky to others, and avoid living life on that basis, are making a choice to not live life in order to send some sort of social message or tell themselves a story about the type of person they are, or some other not-physical-reality based motivation. Or they just aren't that into you and it's a convenient excuse.

That doesn't mean risk for the vaccinated is zero, precautions that are cheap are worth taking and 'stupid stuff' is worth avoiding, and one should follow mask norms for social reasons cause it's really not that big a deal, but on this 'we don't know if it works' thing, seriously: Stop. Just stop.

Similarly, [Zeynep has a thread here about wildly misleading headlines](#) about effectiveness of the vaccines against variants. Studies that find the vaccines work fine are being reported as 'vaccines don't work as well' in a way that has nothing to do with any practical implications. The practical implications are that they work just fine, thanks, and it's so clear I'm not even going to go into it beyond that. [Attempts like this one to scare people](#) about this are pure gaslighting.

[How about we instead do this:](#)

Nate Silver @NateSilver538

Private-sector vaccine messaging involves much less hedging than the public-sector version on getting back to normal.

**This is our shot™
to get back together**

COVID-19 vaccines are ready and so is Walgreens. Schedule your free vaccination today.[†]

[Schedule vaccine](#)

[Learn more](#)

Vaccines for COVID-19

COVID-19 vaccines are safe and effective. After you've been fully vaccinated, you can start to do some things that you had to stop doing because of the pandemic. [Key Things to Know](#)

[And act more like this \(video\)](#), which is living its best life):



Elizabeth

@acesounderglass

...

Got my second shot



We'll Never Have Problems Again - feat. Rachel Bloom & Vi...

BUY ON ITUNES: http://smarturl.it/ceg_210 "We'll Never Have Problems Again" Starring Rachel Bloom and Vincent ...

↙ youtube.com

9:21 PM · Apr 11, 2021 · Twitter Web App

[Otherwise you get this:](#)



The End Times @TheAgeofShoddy · Apr 9

...

The problem with a public health policy built around lies designed to amplify fear is that the people who realize that you've been lying to them aren't going to be open to even reasonable interventions, and the people who haven't figured it out are still terrified.

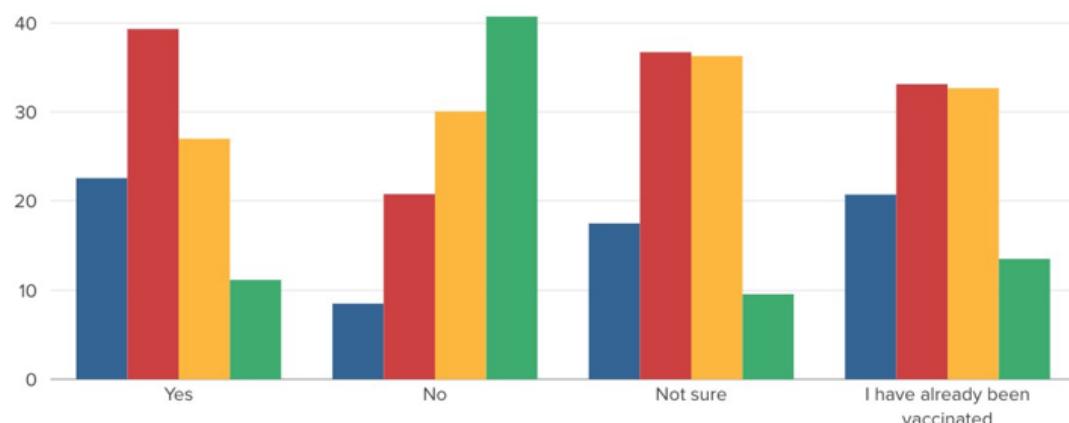
Personal Worry about COVID-19

By

Taking into consideration both your risk of contracting it and the seriousness of the illness, how worried are you personally about experiencing COVID-19?

- Very worried
- Somewhat worried
- Not too worried
- Not worried at all

N 1,487



Get Vaccinated

When a coronavirus vaccine becomes available to you, will you get vaccinated?

Filter All respondents ▾

□ % * 0.00 ↪ ↵



Matthew Yglesias ✅ @mattyglesias · Apr 11

...

Covid spreading out of control because 25% of the population has totally tuned out public health advice while the most compliant half of the population fight amongst ourselves about exactly how strict to be does not seem optimal.

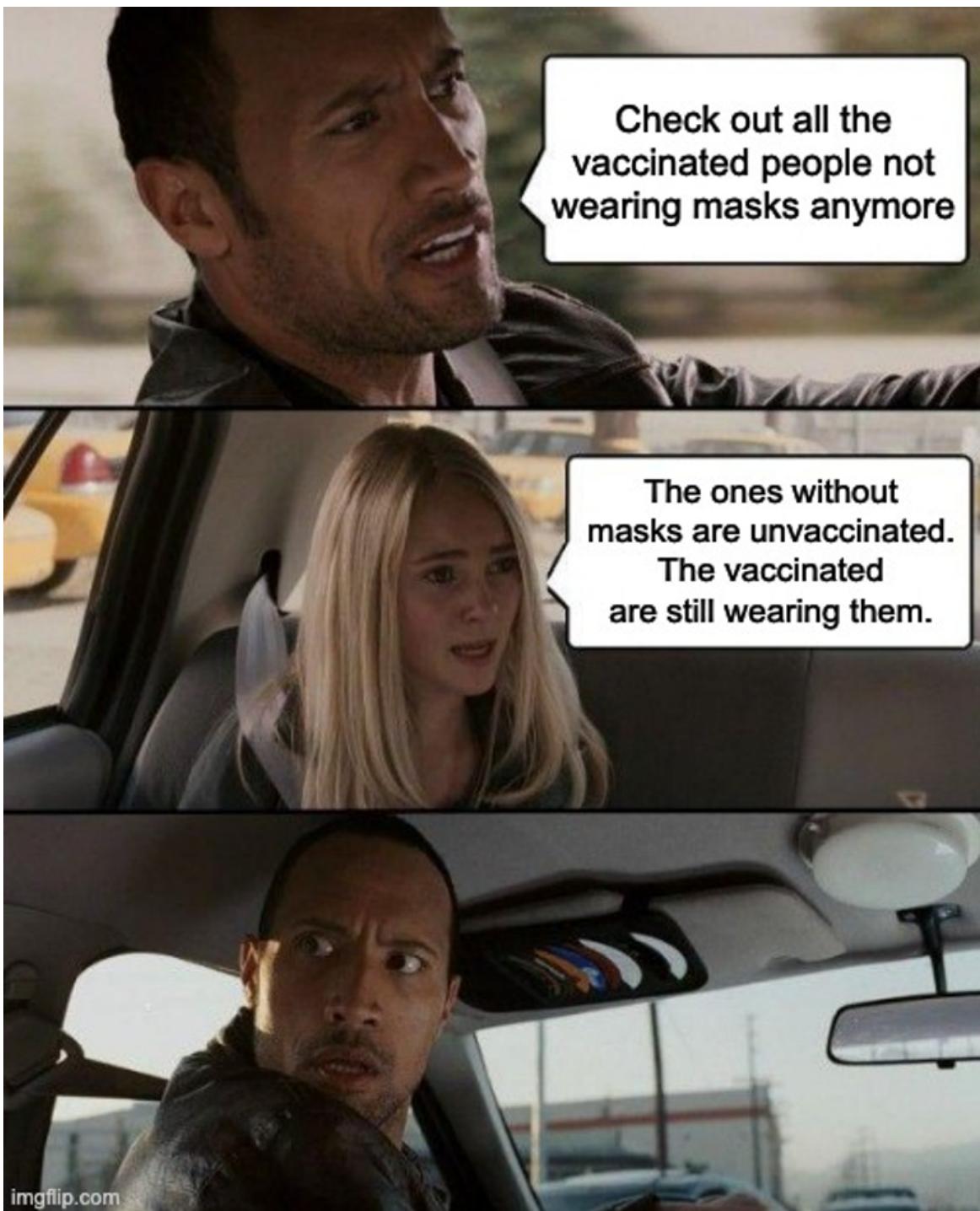
188

343

3.5K



Which is why this is basically [where we are](#):



That's before the whole Johnson & Johnson mess.

In Other News

Meanwhile in Germany:



Vaccine Boom Enthusiast @GarettJones · Apr 9

...

Germany deregulates, vaccinations almost triple.

As they say on the internet:



Marcel Gerber @MarcelGerber9 · Apr 9

For anyone interested, the main reason behind this spike is that starting this week, general practitioners are allowed to administer the vaccines. Before, only special vaccination centers were allowed to administer it.

[Show this thread](#)

Yep.



Rita Panahi
@RitaPanahi

...

Real or fake?



6:33 PM · Apr 11, 2021 · Twitter for iPhone

256 Retweets 117 Quote Tweets 2,235 Likes



Rita Panahi @RitaPanahi · Apr 11

...

Replying to @RitaPanahi

It's real.

I mean, at least it's true.

[More support for first doses first](#). As Tyler notes, it's too late for America to benefit from this, but the rest of the world still could. The same goes for fractional dosing. From what I've seen, a lot of people are having a really unpleasant day or two after the second dose of Moderna, and my strong hunch is the severity is caused by using a dose that's twice as big as it needs to be, and it would be actively better for them to get half doses plus we'd have twice as many doses to give out.

[Canadians return home via taxis from Buffalo to avoid quarantine](#). They did indeed solve for the equilibrium.

[Australia might not open its borders even after full vaccination](#). The hypothesis that Australia succeeded because it was using good epistemics to make decisions is not holding up well in the endgame.

[Covid Tracking Project offers thoughts on data source issues](#). I miss them deeply.

People Will Listen

I have been thinking a lot about the [crypto autopsy](#) Scott posted in 2018. In retrospect, there was still an enormous amount of money to be made 'buying the dip' in BTC/ETH. And there was even more money to be made buying altcoins. Scott also links to this [thread](#) from 2015 strongly advising people to buy bitcoin at around \$230 (so approximately 250x gains on the percent you held). The earlier bitcoin discussion on lesswrong might have represented an even more lucrative opportunity but this is some of the best completely explicit advice ever posted on the forum:

LessWrong is where I learned about Bitcoin, several years ago, and my greatest regret is that I did not investigate it more as soon as possible, that people here did not yell at me louder that it was important, and to go take a look at it. In that spirit, I will do so now.

...

This is a time to be good rationalists, and investigate a possible opportunity, comparing the present situation to historical examples, and making an informed decision. Either Bitcoin has begun the process of dying, and this decline will continue in stages until it hits zero (or some incredibly low value that is essentially the same for our purposes), or it will live. Based on the new all time high being hit in number of transactions, and ways to spend Bitcoin, I think there is at least a reasonable chance it will live. Enough of a chance that it is worth taking some money that you can 100% afford to lose, and making a bet. A rational gamble that there is a decent probability that it will survive, at a time when a large number of others are betting that it will fail.

And then once you do that, try your hardest to mentally write it off as a complete loss, like you had blown the money on a vacation or a consumer good, and now it is gone, and then wait a long time.

As I am writing the thread itself has four upvotes. Conversely, the following comment has twenty-six (this thread long predates variable weight votes though we can still vote on it going forward)

I used to believe that bitcoin is under-priced before, but there are so many agents involved in it now (including Wall Street), that I can't really convince myself that I know better than them - the market is too efficient for me.

Additionally, I'd be especially wary about buying based on arguments regarding the future price based on such obvious metrics, that many agents pay attention to.

This seems like a really strong indictment of the community's collective rationality. On the other hand, I have been posting some [financial advice threads](#) on lesswrong. I have posted much more advice on rationalist adjacent facebook and discord. People listen. I frequently get messages from people telling me they made money thanks to my posts. Several friends of mine got into Solana around \$2.50 at the same time and have made six or seven figures from that investment. A few people got in later or for smaller amounts and still made meaningful amounts of money (Solana is no longer a truly amazing investment but it's still worth buying/holding). Villiam's [comment](#) is important to keep in mind:

Some of us were smarter than others. Good for them! But if we want to help each other, and avoid having the same thing happen the next time, next time when you see an exceptionally important article, don't just think "others have read the same article, and they are smart people, so they know what to do". That's another form of illusion of transparency; after reading the same text, some people will jump up, others will just continue reading. Here are two things you can do to nudge your fellow rationalists in the right direction:

- 1) Imagine a person who has very little knowledge in this specific area, and for some reason is not going to study more. Can the whole thing be simplified; ideally into a short list that is easy to follow? For example: "Step 1: register online at [BitStamp](#). Step 2: send them the required [KYC](#) documents. Step 3: do the money transfer. Step 4: buy Bitcoins. Step 5: HODL!" More people will follow this procedure, than if they just read "buy and/or mine some Bitcoins, find out how".
- 2) Offer help at your local meetup. Make a short lecture, explain the details, answer questions. When people are interested, guide them step by step

It is very hard to grok how context people are missing. You need to make it as easy as possible for people to follow your advice. And you need to tell them a real plan. Many of us held a lot of bitcoin at one point or another. Just buying bitcoin was not enough, you needed to stick to a plan for selling. I personally like 'sell 20% at every doubling' for volatile investments. Give people the tools to succeed. A good friend of mine wanted to get into Solana for a few thousand dollars but found it difficult to buy. He is out tens of thousands of dollars because I did not make it easier for him to buy. He really could use that money.

Am I now become your enemy, because I tell you the truth? - Galatians 4:16

The problem is that the more explicit you are the more pushback you should expect to receive. If you just explore a topic and 'hint hint' at your advice you won't expose yourself to the same types of criticisms. You don't need much plausible deniability to shield yourself. However, you cannot shield yourself very much if you are giving step-by-step instructions and making clear claims. We should take [trivial](#) inconveniences very seriously. Therefore we must not add any unnecessary friction.

My advice is to accept that 'haters are gonna hate' and just take the hit. Make your arguments as clear and your advice as easy to follow as possible. But understand that no matter what you do, if you tell people to buy bitcoin at \$230, the top comment might be critical. Some people will listen and benefit.

As a group, we should really be taking Aumannian reasoning and [noticing confusion](#) more seriously. If something very interesting is going on we need to stop and think and come to a logical conclusion. Obviously, not everyone is going to agree. But at least smaller groups should be able to '**Stop, Drop and Think**'. I hope we can do better as a group but if we cannot, or you leave lesswrong, at least Stop, Drop and Think when you notice a potential opportunity. The rewards for making the right choices can be enormous.

This post mostly argues in favor of sharing your well-thought-out positions. But another implication of 'people will listen' is that you should not give advice flippantly. Don't quickly fire off conventional wisdom. If you have not thought through your counsel make sure you make this clear. Of course, if you have thought things out I strongly suggest speaking your mind.

There is a season for all things and this seems like a season for thinking about crypto. However, I think pretty much every season is for thinking about money. Money is instrumentally useful for the vast majority of goals. If you want to 'win' you should probably think at least a little about money. But I also think these principles extend across domains and are not limited to financial decisions.

A Brief Review of Current and Near-Future Methods of Genetic Engineering

[Part 1: The Case for Human Genetic Engineering](#)

[Part 2: The Case for Increasing Intelligence](#)

My Purpose in Writing This Post

I've spent the last 6 months or so looking into the possibility of pursuing human genetic engineering as a means of improving human lives and increasing the probability of a desirable future. If you'd like more details about why I think improving health and intelligence is desirable, read my previous two posts.

In this post I'm going to summarize my understanding of the research on how genetic engineering will likely be done, the limitations of current techniques, and how they might be improved in the near future.

One last thing before I get started: the genetic engineering I am interested in, and which I think holds the most potential for increasing the likelihood of a good future does not incorporate any type of selective breeding or eugenics programs. Though one could theoretically increase intelligence or any other trait by banning those with undesirable traits from having children and encouraging those with desirable traits people to have more children, [I think this is a bad approach for reasons I have summarized in another post.](#) This post examines methods that do not require coercion to work.

A Summary of Current Techniques

Step 1: Make a yardstick

All non-coercive efforts to genetically engineer humans have one essential prerequisite task: finding which genes contribute to the expression of different traits. In the modern world of genetics, this is done with a test called a Genome Wide Association Study, or GWAS. These are truly massive studies: a typical GWAS done today usually has hundreds of thousands of participants. Genetic material is collected from all participants, often with a blood draw or cheek swab, and their genome is analyzed with a machine like [Illumina's MiSeq analyzer](#).

[For cost reasons](#), nearly all studies today sequence only a small portion of the genome using a device called an SNP microarray. SNP stands for Single Nucleotide Polymorphism, and is the term geneticists use to refer to a base pair that differs between two individuals. Also, because it's technically possible for any base pair to differ between humans, geneticists usually use this term to refer to base pairs for which at least 1% of study participants have a different base pair at that location. An

SNP microarray is an amazing device that can sequence a portion of a genome without sequencing all of it and save money by doing so. This is done by attaching a bunch of short RNA sequences to a substrate (basically a really flat plate), then spreading a bunch of ground up DNA over the array of these RNA sequences and measuring which of the plate-attached RNA sequences have complimentary pairs in the ground-up sample. A signal is then obtained (I think with a laser? The articles I read weren't clear on this), whose strength varies depending on how many of the base pairs of the sequences attached properly. In other words, they're measuring how well the two RNA strands bonded to one another. There's a whole bunch of fancy signal processing that happens after this to deal with noisy data from RNA strands that are partially complementary, but at the end we have data about which of the plate-attached RNA strands had complementary pairs in the sample, and for those that didn't, how much they differed by.

Once this data is obtained, either with the SNP chip model described above or with whole-genome sequencing, a linear effect model is used to construct predictors for the influence of each SNP on the expression of a particular trait. If that sounds confusing, just understand that they're basically modelling the expression of a trait as a linear equation like $y = mx + b$, where each letter in the genome is an input (an x) and the m represents the effect size of that letter on the expression of a trait. The value of the coefficient m is determined by minimizing the prediction error of a linear equation. Surprisingly, most genetically caused variance in trait expression can be explained with linear models. For those of you interested in the mathematical details, I suggest you take a look at [the "Methods" section in the Wikipedia article on GWAS](#).

Unfortunately it seems like state of the art methods are still not fantastic at predicting the exact expression of highly polygenic traits, with the notable exception of height. [Here's a paper from late 2018](#) that attempted to predict height, heel bone density, and educational attainment from SNP data alone. The authors were able to explain 40% of height variance with genetic data, 20% of heel bone density variance, and 9% of educational attainment variance. Height has the unique properties of both being highly polygenic and extremely easy to measure, so many of the ideas and techniques underlying modern GWAS were pioneered on studies of height.

More recent studies have been able to explain a higher level of variance in educational attainment and intelligence. A study in April 2019 was able to explain [16% of educational attainment and 11% of intelligence](#). Still, this is a long way from capturing the 50-80% of variance in intelligence that studies indicate comes from genes.

Step 2: Generate desirable variance

Geneticists have several tools to generate desirable genetic variance. Though tools like CRISPR seem to get the most attention from the mainstream press, CRISPR is not a particularly cost effective way to engender desirable traits in a future human. There are use-cases for CRISPR, such as if both parents have a recessive disease like sickle cell anemia and all of their children will have the disease. In that case, CRISPR could be used to replace the disease allele with its normal counterpart, allowing the couple to conceive children without the disease in question. CRISPR is also a very useful therapeutic for treating those who have a mendelian genetic condition and have already been born. For example, [here's a study where researchers used CRISPR to cure one patient of beta-Thalassemia and another of sickle cell anemia](#) by extracting

blood stem cells from their bone marrow, replacing the diseased gene, and reinjecting those modified cells back into the patients.

But for most traits, CRISPR is not a cost-effective tool for one simple reason. Most of the traits we care the most about including heart disease risk, diabetes risk, cancer risks of all types, and many others are highly polygenic traits; each are influenced by tens of thousands of letters in the genome. Given CRISPR's tendency to occasionally make off-target edits, and the expense of editing so many places in the genome, I don't see this being a viable strategy to decrease risk of heart disease any time soon. This may change in the future, but it appears to be the case for now.

The best way to generate desirable genetic variance in the near term is by generating a large number of embryos. During sex, sperm and eggs (referred to as "gametes" by biologists) go through a process called meiosis, where they swap parts of their matching chromosomes to generate a new organism with its own unique DNA. This process generates variance. The resulting offspring will incorporate traits from both parents, but trait expression will not always match the mean of the parents. If one parent has a 10% chance of experiencing a heart attack during their lifetime and another has a 15% chance of experiencing a heart attack, the offspring will not always have a 12.5% chance of getting a heart attack. Instead, generally speaking, the offspring's risk of heart attack will be drawn from a normal distribution with a mean of 12.5%. The expression of all heritable polygenic traits will show variance in offspring.

Step 3: Select an embryo for implantation

Once genetic variance has been generated via the production of a number of embryos, the next step is to identify the likely trait values of each one. Embryos within a few weeks of fertilization have a very interesting property: one may remove several cells from the embryo and it will still retain the capacity to develop into a fully functional adult organism. This regenerative capacity allows us to gain unique insights into the genetic potential of each embryo; one may perform a biopsy on each, removing several cells, then amplify and sequence the DNA from the removed cells.

We may therefore discern the genetic sequence of an embryo before we choose to implant it. This gives us the chance to tilt the odds in favor of a future child: we can reduce their risk of serious polygenic diseases like heart disease, breast cancer, and type 2 diabetes, and we can virtually eliminate serious mendelian diseases like sickle cell anemia, cystic fibrosis, Huntington's disease, and others. This is done by creating an overall "score" for each embryo, which represents the embryo's expected expression of a set of traits. For example, we would give embryos at higher risk of developing coronary artery disease or type 2 diabetes a lower score. The expression of each trait is given a weight in accordance with how important we believe it is. These weights can be adjusted to reflect parental preferences with the help of a genetic counselor knowledgeable about the tests and about the diseases themselves.

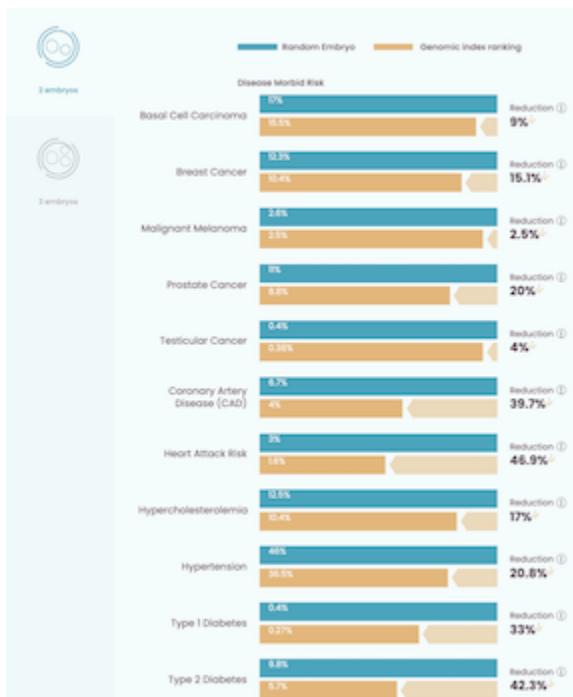
Any trait with a strong genetic component may be selected for or against using this method. One merely needs to generate variance among a pool of embryos and develop a test that is able to capture a large enough portion of this genetically caused variance in trait expression.

What type of genetic engineering can we do today?

It may surprise some to learn that the techniques I described above are actually accessible right now in some capacity to any parents that can afford to do In-Vitro Fertilization. In IVF, a father donates his semen and a mother donates eggs, and a reproductive specialist uses those pools of reproductive cells to produce embryos in a cell culture. These embryos may then be screened using the process I described above, and the parents, with the help of a reproductive specialist, may choose which embryo they would like to implant.

There are companies offering this service right now. Among them are [Orchid Health](#) and [Genomic Prediction](#). So far as I know, no companies in Europe nor the United States are offering screening for intelligence, skin color, or any other cosmetic traits. Instead, they focus exclusively on genetic predictors of health, such as heart attack risk, type 2 diabetes risk etc. This is at least partially explained by the fact that we don't have great polygenic predictors of intelligence yet, but given that the predictors for some of the diseases they DO screen for aren't much better, the main reason seems to be to avoid the controversy that would inevitably follow.

Even with the fairly limited testing we have today, pre-implantation genetic screening can have a remarkable effect on children's future health. This is especially true for individuals with a family history of disease. As a topic of a future paper, I would like to quantify exactly how cost-effective IVF + embryo selection is for couples with no fertility issues, but my gut feel after reading some research is that the expected reduction in medical costs alone more than pay for the cost of IVF. To give you a rough idea of how effective this technology is at reducing disease risk, here's Genomic Prediction's chart showing how much we would expect the risk of different diseases to decrease if we were to choose from the better of two scoring of two embryos.



[You can play around with the tool yourself to see how changing the number of embryos affects the expected reduction in disease risk](#). Though their current web interface is limited, it's clear that even just selecting from two embryos, the expected reduction in chronic disease risk is substantial. And as the number of embryos selected from goes up, risk of disease decreases even further.

In fact, the reductions are so substantial that I suspect we are not too far from the day where conceiving a child through sex will be viewed similar to how giving birth at home is viewed today: unnecessarily risky and something to be avoided if one can afford it.

If you are considering having children in the future, you may want to look into pre-implantation genetic screening. If you're a woman you may want to consider doing this earlier rather than later, as the number of eggs that can be harvested per cycle tends to decline with age, making the process more expensive if done later. The same goes for men, though since semen extraction is rather easier and does not require a specialist, the main consideration is semen quality rather than cost.

The Future

Suppose you buy the argument that we're likely to be able to improve human health, happiness, intelligence, etc using genetic engineering. What is the critical path to the development of this technology? How do we get there faster?

Screening for intelligence

Though there will doubtless be some people that oppose this in the near future, it seems inevitable that when the predictive tests for intelligence improve enough, some lab in some country will start offering pre-implantation genetic tests for predicted intelligence. Intelligence has a strong genetic basis (estimates vary between 35-80%) and positively correlates with too many objectively important outcomes for us to ignore it for long. The moral case for doing so is strong as well: if we are able to enhance intelligence without having any negative effect on other important traits, why wouldn't we do so? We already spend significant amounts of money to help our children realize their intellectual potential. Raising their potential through genetic intervention seems completely consistent with the values already clearly expressed by many people.

And once this service starts being offered anywhere, it will only be a matter of time before it's offered pretty much everywhere. If a country does not allow embryo screening for intelligence and other desirable traits, wealthy parents will simply have their embryos genotyped, then send the data files off to a clinic in another country where they can be analyzed, and implant the ones that score the best according to some scoring system that takes intelligence into account.

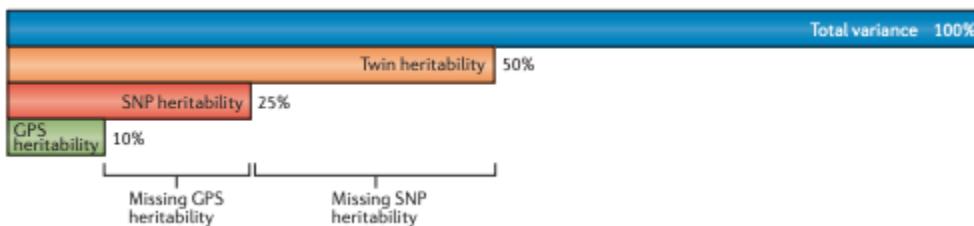
And if data transfers are banned, then those parents will simply take a vacation somewhere it isn't to give birth. Eventually, the huge accrued disadvantage faced by countries that don't allow the technology will create overwhelming pressure to legalize the technology in some capacity, and no country other than dictatorships will be able to resist the pressure. My guess for where this will first be legalized is somewhere in

Asia, possibly South Korea or China. And just as IVF itself became normalized, so will preimplantation genetic screening designed to give one's child the best life possible.

For intelligence screening, in particular, I have concluded there are two key technologies needed to enable dramatic improvements.

1. Improve tests for the genetic component of intelligence

The first is a test better able to capture genetically caused variance in intelligence. [Plomin & Stumm](#) seem to think that the two missing ingredients for really good predictors of the genetic portion of intelligence are larger sample sizes and whole-genome sequencing instead of SNP based approaches, as well as possibly non-linear models of gene effects and gene-environment interactions (see the last paragraph of box 4 on page 6 from the above link). They estimate that we can capture half the genetic variance in intelligence with SNP data alone, but that we'll need whole-genome sequencing for the remainder. SNP tests usually cost around \$100, while whole genome sequencing currently costs around \$300. Here's a nice graph showing the current state of our tests as of 2018.



It's worth pointing out that this ratio of environmental influence on intelligence to genetic influence on intelligence is not fixed. If half the population had chronic exposure to lead in their drinking water and the other half had clean drinking water, the percentage of variance in intelligence explained by environmental factors would go up. Similarly, if half the population was genetically engineered to be unusually intelligent while the other half was not, the percentage of variance explained by genes would go up.

The more important thing to note here is how large of an increase we could get to intelligence by simply increasing the frequency of the SNPs positively correlated with its expression. Professor Steven Hsu has estimated that [there is enough additive variance in the human population to create people with IQs of over 1000 if we were to add them all together](#). We almost certainly wouldn't want to incorporate all these variants into a single person, as some likely have tradeoffs with health, reproductive propensity, or other things we care about that would make incorporating them a poor choice. Another concern is whether or not IQ, as measured by tests like [progressive matrices](#) will continue to correlate with the things we actually care about at such extreme levels. It's likely that the use of today's IQ tests to predict intelligence will break down if we push it far enough in either direction. But exactly where that point is remains an important open question. It seems likely to me that we will be able to raise average human IQ into the high 100's without any serious downsides. We already have thousands of examples of people with IQs this high, most of whom were functional in other ways we care about. In fact, if we get much better tests of genetic intelligence and we are also able to get iterated embryo selection to work, this question of how far we can safely push trait expression will become the chief remaining question.

I've followed AI safety research as a hobby for the last few years and one of the lessons I've learned from the research is machines that optimizing for any objective X will eventually impact another objective Y if one pushes hard enough. This will doubtlessly be the case with intelligence.

It's quite difficult to estimate how hard it will be to develop better tests. One obvious step is to increase the sample size of the study. This will help detect genetic variants with smaller effect sizes on intelligence and to detect rarer variants. Another obvious step is to perform whole-genome sequencing, which would help capture rare variants that may account for currently uncaptured variance.

The best paper I've found on this topic is Stephen Hsu's 2014 paper [On the genetic architecture of intelligence and other quantitative traits](#), which estimates that a sample size of a million would be enough to capture nearly all the variance. However, since its publication, studies examining ~250k individuals were only able to explain [7-10% of the variance in cognitive performance \(see the top of page 2\)](#). Furthermore, this study only identified 225 significant SNP hits, well short of the 10,000 that Hsu estimates play a role in intelligence. The relationship between sample size and discovered SNPs is not linear, but it's not clear how much of the missing heritability is due to smaller sample size as opposed to other things. Are there more variants that influence intelligence with smaller effect sizes than Hsu predicted? Do non-linear effects play a bigger role? Is there some other confounding factor? I don't yet know the answer to these questions.

So after a couple of days of research trying to complete this section, I am stuck with no clear answers. It is not clear to me how large of a sample size we'll need to get an accurate measurement of intelligence, nor is it clear what additions we'll need to basic additive models to obtain high performance on such tests.

If I had to hazard a very rough guess, I would say that a sample size of 10 million with full genome sequencing performed on every participant would probably be sufficient to capture >80% of the genetically caused variance in intelligence. Assuming \$300 per genome sequenced and \$100 to administer each test, this comes out to a price tag of \$4 billion. Not cheap, but well within the realm of feasibility. And hopefully, economies of scale would help lower the price, at least for the genome sequencing portion.

2. Solve Iterated Embryo Selection

The Second necessary technology needed to allow for dramatic improvements in polygenic traits such as intelligence on a short timeline is Iterated Embryo Selection or IES. IES is a technology that theoretically allows for arbitrarily large increases in trait values on a much shorter time horizon than any other near-term technology. IES involves the following 6 steps:

1. Extract somatic cells from an organism or tissue (usually skin cells or blood cells)
2. Revert these cells back to induced pluripotent stem cells
3. Develop those stem cells into gametes (reproductive cells like sperm or eggs)
4. Fertilize the gametes to create a batch of new embryos
5. Sequence the DNA of the embryos, selecting the best of the batch
6. Develop the selected embryos into a larger amount of tissue, like skin cells or blood cells
7. Repeat steps 1-6

IES essentially takes the reproductive cycle from 20+ years down to 6 months. Whereas normal IVF ends when an embryo is selected for implantation, IES takes the selected embryo through another cycle of meiosis and recombination (possibly introducing new genetic material from another group of embryos in the process). After each round of iteration, the mean trait values of the new pool of embryos will be equal to the highest-scoring embryos from the previous round. This is the true magic of iterated embryo selection; once feasible, it allows for arbitrary gains in any genetically influenced trait.

So what ingredients are we still missing? Step 1 is trivial. Step 2 has been possible since 2006 when Shinya Yamanaka's lab produced the first induced pluripotent stem cells using "Yamanaka factors", and is in fact an active step in most stem cell therapies. Step 4 is a standard part of IVF and step 5 is becoming more common. Step 6 seems like it's already possible given that most research into tissue engineering assumes embryonic stem cells or some other pluripotent stem cells as a starting point. As far as I can tell, the only step that has not yet been accomplished to completion in humans is step 3: differentiation of pluripotent stem cells into gametes.

We have already gotten step 3 to work in mice. In 2016, [Hikabe et al showed reconstitution of the entire female mouse germline in vitro](#) (steps 1-3 in the list above). This process, known as In-Vitro Gametogenesis, is critical to all attempts to Iterated Embryo Selection. Hikabe et. al. were able to harvest a sample of skin cells from the tail of the mouse, revert those cells back to a pluripotent state using Yamanaka factors, differentiate those iPSCs into oocytes, then fertilize the resulting Oocytes to create mouse embryos, which were then implanted in a female mouse who gave birth to healthy pups.

There's [a really fantastic summary of current progress of this technology in humans by Dr. Sherman Silber](#) on YouTube. We are very close. The only remaining unrealized step is getting from primordial germ cells to sperm and eggs. What makes this step so difficult is recreating the conditions in which primordial germ cells mature into spermatogonial stem cells and eggs within the human body.

Silber believes this may be easier for oocytes than for sperm. To culture PGCs into oocytes, the PGCs must develop in the presence of fetal granulosa cells. These cells are critical because they emit a set of growth factors that tell the PGCs to develop into Oocytes. Silber believes we should be able to replicate these conditions by isolating the growth factors and applying them to the PGCs.

Sperm are trickier. According to Dr. Silber the only method they've been able to use so far to mature primordial germ cells into spermatogonial stem cells is injection of PGCs into the rete testes of a prepubescent boy. The pubescent development process, as it turns out, is critical to mature PGCs into spermatogonial stem cells, and those conditions cannot be found in adult testes.

While this type of injection works well for restoring fertility in individuals who lost it in childhood (usually due to cancer treatments), this will not work for Iterated Embryo Selection. But unfortunately a cursory search yielded no results for spermatogenesis via growth factors. Either research into spermatogenesis has not been funded or I have simply been unable to find the published papers.

So to summarize: we are very close to making Iterated Embryo Selection possible. The missing piece is the ability to turn primordial germ cells into oocytes and sperm.

Ongoing research will likely make this possible for oocytes in the next 5-10 years, but the path for spermatogenesis is less clear.

Reflections on the Value of Human Genetic Engineering

This will not be my last paper on the topic, but I wanted to take a brief moment to reflect on why I think human genetic engineering is important. Apart from the obvious near-term benefits of reducing chronic disease, I think in the long run, genetic engineering will only matter if it affects the development of transformative artificial intelligence.

I don't remember exactly where I read this, but in another post I read on LessWrong, the author suggested that biological systems may simply become obsolete in the future because computer-based information processing systems will become better at turning energy into utility. I suspect that in the long run, this will probably be true.

I am very worried that current humans are simply incapable of aligning powerful AI with our interests due to the incredible technical complexity of the problem. My goal in pursuing a career in genetics with a focus on human reproduction is to increase human capability to deal with incredibly technical problems like those involved in creating TAI. Along the way I hope we can create a kinder, healthier society with fewer mismatches between our genes and our environment.

If some of the more pessimistic projections about the timelines to TAI are realized, my efforts in this field will have no effect. It is going to take at least 30 years for dramatically more capable humans to be able to meaningfully contribute to work in this field. [Using Ajeya Cotra's estimate of the timeline to TAI](#), which estimates a 50% chance of TAI by 2052, I estimate that there is at most a 50% probability that these efforts will have an impact, and a ~25% chance that they will have a large impact.

Those odds are good enough for me.

Highlights from The Autobiography of Andrew Carnegie

This is a linkpost for <https://rootsofprogress.org/andrew-carnegie-autobiography>

I've been reading [Andrew Carnegie's autobiography](#), published late in his life, in the early 1900s. Here are some interesting themes and quotes. (Emphasis added in all block quotes below.)

Science and steel

One key to Carnegie's success in the iron business is that he was one of the first to seriously apply chemistry:

Looking back to-day it seems incredible that **only forty years ago (1870) chemistry in the United States was an almost unknown agent in connection with the manufacture of pig iron.** It was the agency, above all others, most needful in the manufacture of iron and steel. The blast-furnace manager of that day was usually a rude bully... who in addition to his other acquirements was able to knock down a man now and then as a lesson to the other unruly spirits under him. He was supposed to diagnose the condition of the furnace by instinct, to possess some almost supernatural power of divination, like his congener in the country districts who was reputed to be able to locate an oil well or water supply by means of a hazel rod. He was a veritable quack doctor who applied whatever remedies occurred to him for the troubles of his patient.

Part of the problem was that the ores and other inputs to smelting were inconsistent in composition:

The Lucy Furnace was out of one trouble and into another, owing to the great variety of ores, limestone, and coke which were then supplied with little or no regard to their component parts. This state of affairs became intolerable to us.

This is where chemistry was able to help:

We finally decided to dispense with the rule-of-thumb-and-intuition manager, and to place [Henry Curry] in charge of the furnace....

The next step taken was to find a chemist as Mr. Curry's assistant and guide. We found the man in a learned German, Dr. Fricke, and great secrets did the doctor open up to us. Ironstone from mines that had a high reputation was now found to contain ten, fifteen, and even twenty per cent less iron than it had been credited with. Mines that hitherto had a poor reputation we found to be now yielding superior ore. The good was bad and the bad was good, and everything was topsy-turvy. **Nine tenths of all the uncertainties of pig-iron making were dispelled under the burning sun of chemical knowledge.**

It wasn't just that some materials were of low quality, but that the right mix of materials was needed, no matter the purity of the inputs:

At a most critical period when it was necessary for the credit of the firm that the blast furnace should make its best product, it had been stopped because an exceedingly rich and pure ore had been substituted for an inferior ore—an ore which did not yield more than two thirds of the quantity of iron of the other. **The furnace had met with disaster because too much lime had been used to flux this exceptionally pure ironstone.** The very superiority of the materials had involved us in serious losses.

What fools we had been! But then there was this consolation: we were not as great fools as our competitors. It was years after we had taken chemistry to guide us that it was said by the proprietors of some other furnaces that they could not afford to employ a chemist. **Had they known the truth then, they would have known that they could not afford to be without one.** Looking back it seems pardonable to record that we were the first to employ a chemist at blast furnaces—something our competitors pronounced extravagant.

With better chemical assessment of ores, Carnegie was able to arbitrage his supplies:

The mines which had no reputation and the products of which many firms would not permit to be used in their blast furnaces found a purchaser in us. Those mines which were able to obtain an enormous price for their products, owing to a reputation for quality, we quietly ignored. A curious illustration of this was the celebrated Pilot Knob mine in Missouri. Its product was, so to speak, under a cloud. A small portion of it only could be used, it was said, without obstructing the furnace. Chemistry told us that it was low in phosphorus, but very high in silicon. There was no better ore and scarcely any as rich, if it were properly fluxed. **We therefore bought heavily of this and received the thanks of the proprietors for rendering their property valuable.**

It is hardly believable that for several years we were able to dispose of the highly phosphoric cinder from the puddling furnaces at a higher price than we had to pay for the pure cinder from the heating furnaces of our competitors—a cinder which was richer in iron than the puddled cinder and much freer from phosphorus. Upon some occasion a blast furnace had attempted to smelt the flue cinder, and from its greater purity the furnace did not work well with a mixture intended for an impurer article; hence **for years it was thrown over the banks of the river at Pittsburgh by our competitors as worthless.**

He gives another example later of discovering a property with ore that had no phosphorus, “really an ore suitable for making Bessemer steel” (the [Bessemer process](#) could not handle phosphoric ores, until later improved by Gilchrist and Thomas). Again, it was the *purity* of the ore that was a problem:

We found the mine had been worked for a charcoal blast furnace fifty or sixty years before, but it had not borne a good reputation then, the reason no doubt being that its product was so much purer than other ores that the same amount of flux used caused trouble in smelting. It was so good it was good for nothing in those days of old.

Here's how they assessed the mine:

We finally obtained the right to take the mine over at any time within six months, and we therefore began the work of examination, which every purchaser of mineral property should make most carefully. We ran lines across the hillside fifty feet apart, with cross-lines at distances of a hundred feet apart, and at each point

of intersection we put a shaft down through the ore. I believe there were eighty such shafts in all and the ore was analyzed at every few feet of depth, so that before we paid over the hundred thousand dollars asked we knew exactly what there was of ore.... We trod upon sure ground with the chemist as our guide.

Chemistry, however, was still foreign to many people:

Our chemist, Mr. Prousser, was then sent to a Pennsylvania furnace among the hills.... A striking example of the awe inspired by the chemist in those days was that only with great difficulty could he obtain a man or a boy to assist him in the laboratory. **He was suspected of illicit intercourse with the Powers of Evil when he undertook to tell by his suspicious-looking apparatus what a stone contained.** I believe that at last we had to send him a man from our office at Pittsburgh.

The Bessemer process

One of Carnegie's major achievements was to bring [the Bessemer steel-making process](#) to America. This was a new way to achieve high-quality steel at a low price. Previously, the only options were low-strength wrought iron, brittle cast iron, or expensive steel in limited quantities. Bessemer broke this iron triangle.

Carnegie, already in the iron business, was paying attention and could see the future:

I had not failed to notice the growth of the Bessemer process. If this proved successful I knew that iron was destined to give place to steel; that **the Iron Age would pass away and the Steel Age take its place.**

As an example of the potential market for steel, Carnegie gives the example of iron rails, which wore out quickly under the pounding of heavy trains:

The question of a substitute for iron rails upon the Pennsylvania Railroad and other leading lines had become a very serious one. Upon certain curves at Pittsburgh, on the road connecting the Pennsylvania with the Fort Wayne, I had seen new iron rails placed **every six weeks or two months.**

Railroads and iron-makers went to great lengths to solve the problem:

Before the Bessemer process was known I had called President Thomson's attention to the efforts of Mr. Dodds in England, who had carbonized the heads of iron rails with good results. I went to England and obtained control of the Dodds patents and recommended President Thomson to appropriate twenty thousand dollars for experiments at Pittsburgh, which he did. We built a furnace on our grounds at the upper mill and treated several hundred tons of rails for the Pennsylvania Railroad Company and with remarkably good results as compared with iron rails. These were the first hard-headed rails used in America. We placed them on some of the sharpest curves and their superior service far more than compensated for the advance made by Mr. Thomson.... But there was nothing to be compared with the solid steel article which the Bessemer process produced.

Carnegie formed a company in 1873 to use the Bessemer process to make rails.

Accounting

One of the surprising themes was that manufacturing concerns in the 1800s did very little accounting and thus had very little insight into their businesses, even whether they were profitable:

As I became acquainted with the manufacture of iron I was greatly surprised to find that the cost of each of the various processes was unknown. Inquiries made of the leading manufacturers of Pittsburgh proved this. It was a lump business, and until stock was taken and the books balanced at the end of the year, the manufacturers were in total ignorance of results. I heard of men who thought their business at the end of the year would show a loss and had found a profit, and vice-versa. **I felt as if we were moles burrowing in the dark, and this to me was intolerable.**

Just as Carnegie had brought the science of chemistry to his processes, so he brought “scientific management” to his operations:

I insisted upon such a system of weighing and accounting being introduced throughout our works as would enable us to know what our cost was for each process and especially what each man was doing, who saved material, who wasted it, and who produced the best results.

To arrive at this was a much more difficult task than one would imagine. **Every manager in the mills was naturally against the new system.** Years were required before an accurate system was obtained, but eventually, by the aid of many clerks and the introduction of weighing scales at various points in the mill, we began to know not only what every department was doing, but what each one of the many men working at the furnaces was doing, and thus to compare one with another.

This level of quantitative insight allowed Carnegie to make good choices about capital investments:

The Siemens Gas Furnace had been used to some extent in Great Britain for heating steel and iron, but it was supposed to be too expensive. I well remember the criticisms made by older heads among the Pittsburgh manufacturers about the extravagant expenditure we were making upon these new-fangled furnaces. But in the heating of great masses of material, almost half the waste could sometimes be saved by using the new furnaces. The expenditure would have been justified, even if it had been doubled.

Investing

Carnegie's attitudes towards investing strike me as very odd. There was something very different about 19th-century investing that I don't fully understand.

On the one hand, people were willing to take on significant amounts of personal debt in order to buy equity. In 1855, when Carnegie was a young man living with his mother, his boss tipped him off to a rare opportunity to buy railroad stock:

Mr. Scott asked me if I had five hundred dollars [over \$15,000 today]. If so, he said he wished to make an investment for me. Five hundred cents was much nearer my capital. I certainly had not fifty dollars saved for investment, but I was not going to miss the chance of becoming financially connected with my leader and great man.

So I said boldly I thought I could manage that sum. He then told me that there were ten shares of Adams Express stock that he could buy, which had belonged to a station agent, Mr. Reynolds, of Wilkinsburg.

How to pay for it? His mother *mortgaged their house*:

We had then paid five hundred dollars upon the house, and in some way she thought this might be pledged as security for a loan.

My mother took the steamer the next morning for East Liverpool, arriving at night, and through her brother there the money was secured. He was a justice of the peace, a well-known resident of that then small town, and had numerous sums in hand from farmers for investment. Our house was mortgaged and mother brought back the five hundred dollars which I handed over to Mr. Scott, who soon obtained for me the coveted ten shares in return.

It's remarkable to me how random this all is, how reliant on personal relationships and chance connections. But even more so, it strikes me as financially reckless, the sort of thing you'd read about today on [r/wallstreetbets](#). On the other hand, it turned out very well:

In those good old days monthly dividends were more plentiful than now and Adams Express paid a monthly dividend. One morning a white envelope was lying upon my desk... All it contained was a check for ten dollars upon the Gold Exchange Bank of New York.... "Eureka!" I cried. "Here's the goose that lays the golden eggs."

If Adams Express paid \$10 a month on \$500 of stock, that's a 24% annual dividend yield, which is far better than any similar investment today, and presumably this easily covered the payments on the loan. No wonder Carnegie and his mother were so eager to get in on the deal.

Later, when Carnegie is still working for the railroad, there's another chance connection, when the inventor of a sleeping car approaches him on the train:

He carried a small green bag in his hand. He said the brakeman had informed him I was connected with the Pennsylvania Railroad. He wished to show me the model of a car which he had invented for night traveling. He took a small model out of the bag, which showed a section of a sleeping-car.

This was the celebrated T. T. Woodruff, the inventor of that now indispensable adjunct of civilization—the sleeping-car. Its importance flashed upon me. I asked him if he would come to Altoona if I sent for him, and I promised to lay the matter before Mr. Scott at once upon my return.

Getting a deal with the railroad, the inventor invites Carnegie to become an investor in the new sleeping-car business:

After this Mr. Woodruff, greatly to my surprise, asked me if I would not join him in the new enterprise and offered me an eighth interest in the venture.

I promptly accepted his offer, trusting to be able to make payments somehow or other. The two cars were to be paid for by monthly installments after delivery. When the time came for making the first payment, my portion was two hundred and seventeen and a half dollars [well over \$6,000 today].

(In this era, investors would often fund a business in regular installments, rather than all at once up front as is standard today.) Once again Carnegie decides to go into debt for this investment, and once again gets the loan through personal connections:

I boldly decided to apply to the local banker, Mr. Lloyd, for a loan of that sum. I explained the matter to him, and I remember that he put his great arm (he was six feet three or four) around me, saying:

"Why, of course I will lend it. You are all right, Andy."

And here I made my first note, and actually got a banker to take it. A proud moment that in a young man's career!

And once again, it works out anyway! 19th-century businesses seem to have gotten to profitability faster than today's startups, because Carnegie writes: "The sleeping-cars were a great success and their monthly receipts paid the monthly installments [on the loan]."

Yet despite all of this willingness to invest on margin, Carnegie, like many others of his day, considered investing in the public stock market to be a reckless gamble. Later in his life, when he moves to New York, he remarks:

I had lived long enough in Pittsburgh to acquire the manufacturing, as distinguished from the speculative, spirit. My knowledge of affairs, derived from my position as telegraph operator, had enabled me to know the few Pittsburgh men or firms which then had dealings upon the New York Stock Exchange, and I watched their careers with deep interest. To me their operations seemed **simply a species of gambling.**

He complained that owning public stocks was distracting, and he warned people away from it, in a passage that sounds like today's discussion of the psychological effects of social media:

I have adhered to the rule never to purchase what I did not pay for, and never to sell what I did not own. In those early days, however, I had several interests that were taken over in the course of business. They included some stocks and securities that were quoted on the New York Stock Exchange, and I found that **when I opened my paper in the morning I was tempted to look first at the quotations of the stock market.** As I had determined to sell all my interests in every outside concern and concentrate my attention upon our manufacturing concerns in Pittsburgh, I further resolved not even to own any stock that was bought and sold upon any stock exchange....

For the manufacturing man especially the rule would seem all-important. His mind must be kept calm and free if he is to decide wisely the problems which are continually coming before him. Nothing tells in the long run like good judgment, and **no sound judgment can remain with the man whose mind is disturbed by the mercurial changes of the Stock Exchange. It places him under an influence akin to intoxication.** What is not, he sees, and what he sees, is not. He cannot judge of relative values or get the true perspective of things. The molehill seems to him a mountain and the mountain a molehill, and he jumps at conclusions which he should arrive at by reason. His mind is upon the stock quotations and not upon the points that require calm thought. Speculation is a parasite feeding upon values, creating none.

Later he comments that investing in public markets degrades one's integrity in business affairs:

A rule which we adopted and adhered to has given greater returns than one would believe possible, namely: always give the other party the benefit of the doubt. This, of course, does not apply to the speculative class. An entirely different atmosphere pervades that world. Men are only gamblers there. **Stock gambling and honorable business are incompatible.**

A reason to be wary of investments in those days was the lack of limited liability in many instances. Without it, even small investors uninvolved in management could be fully liable for the debts of the company. Carnegie tells this story:

Driving with Mr. Phipps from the mills one day we passed the National Trust Company office on Penn Street, Pittsburgh. I noticed the large gilt letters across the window, "Stockholders individually liable." That very morning in looking over a statement of our affairs I had noticed twenty shares "National Trust Company" on the list of assets. I said to Harry:

"If this is the concern we own shares in, won't you please sell them before you return to the office this afternoon?"

He saw no need for haste. It would be done in good time.

"No, Harry, oblige me by doing it instantly."

He did so and had it transferred. Fortunate, indeed, was this, for in a short time the bank failed with an enormous deficit.... Times were panicky, and had we been individually liable for all the debts of the National Trust Company our credit would inevitably have been seriously imperiled. It was a narrow escape. And with only twenty shares (two thousand dollars' worth of stock), taken to oblige friends who wished our name on their list of shareholders! The lesson was not lost. The sound rule in business is that you may give money freely when you have a surplus, but your name never—neither as endorser nor as member of a corporation with individual liability. A trifling investment of a few thousand dollars, a mere trifle—yes, but **a trifle possessed of deadly explosive power.**

All this was reinforced by Carnegie's experience in the financial panic of 1873, when he "entered upon the most anxious period of my business life":

All was going well when one morning in our summer cottage, in the Allegheny Mountains at Cresson, a telegram came announcing the failure of Jay Cooke & Co. **Almost every hour after brought news of some fresh disaster.** House after house failed. The question every morning was which would go next. Every failure depleted the resources of other concerns. Loss after loss ensued, until a total paralysis of business set in. Every weak spot was discovered and houses that otherwise would have been strong were borne down largely because our country lacked a proper banking system.

Scottish and American spirit

Carnegie often remarks on the ideals he picked up as a boy in Scotland. This will come as no surprise to those familiar with British history, but it was remarkable to me the streak of independence and anti-authoritarianism:

The denunciations of monarchical and aristocratic government, of privilege in all its forms, the grandeur of the republican system, the superiority of America, a land peopled by our own race, a home for freemen in which every citizen's privilege was every man's right—these were the exciting themes upon which I was nurtured. **As a child I could have slain king, duke, or lord, and considered their deaths a service to the state and hence an heroic act.**

Later, Carnegie tells a story of visiting an oil boom town in Pennsylvania in 1862. The town had been set up in a hurry, with too many people crowding in and not enough housing. He was impressed with the determination and resourcefulness of the oil wildcatters, who quickly threw up rough accommodations. But more, he was impressed with "the good humor which prevailed everywhere. It was a vast picnic, full of amusing incidents." Flags with "strange mottoes" flew, such as one drilling crew flying the words "Hell or China." Carnegie praises the American spirit:

The adaptability of the American was never better displayed than in this region. Order was soon evolved out of chaos. When we visited the place not long after we were serenaded by a brass band the players of which were made up of the new inhabitants along the creek. It would be safe to wager that a thousand Americans in a new land would organize themselves, establish schools, churches, newspapers, and brass bands—in short, provide themselves with all the appliances of civilization—and go ahead developing their country before an equal number of British would have discovered who among them was the highest in hereditary rank and had the best claims to leadership owing to his grandfather.

19th-century life

Finally, a number of quotes shed light on the general quality and challenges of life in the 1800s:

The burden of travel

A few stories give a glimpse into the hardship of travel before railroads and steamships. For instance, soon after his family came to America:

My father was induced by emigration agents in New York to take the Erie Canal by way of Buffalo and Lake Erie to Cleveland, and thence down the canal to Beaver—**a journey which then lasted three weeks, and is made to-day by rail in ten hours.** There was no railway communication then with Pittsburgh, nor indeed with any western town.... Nothing comes amiss to youth, and I look back upon my three weeks as a passenger upon the canal-boat with unalloyed pleasure. All that was disagreeable in my experience has long since faded from recollection, excepting the night we were compelled to remain upon the wharf-boat at Beaver waiting for the steamboat to take us up the Ohio to Pittsburgh. This was our first introduction to the mosquito in all its ferocity. **My mother suffered so severely that in the morning she could hardly see.**

On his return from the aforementioned oil fields:

The weather had been fine and the roads quite passable during our journey thither, but rain had set in during our stay. We started back in our wagon, but before going far fell into difficulties. The road had become a mass of soft,

tenacious mud and our wagon labored fearfully. The rain fell in torrents, and it soon became evident that we were in for a night of it. Mr. Coleman lay at full length on one side of the wagon, and Mr. Ritchie on the other, and I, being then very thin, weighing not much more than a hundred pounds, was nicely sandwiched between the two portly gentlemen. Every now and then the wagon proceeded a few feet heaving up and down in the most outrageous manner, and finally sticking fast. In this fashion we passed the night. There was in front a seat across the wagon, under which we got our heads, and in spite of our condition the night was spent in uproarious merriment.

Travel was also less reliable, owing to weaker infrastructure—for instance, wooden bridges. Part of what induced Carnegie to go into the iron business was the superiority of iron for bridge-building:

When at Altoona I had seen in the Pennsylvania Railroad Company's works the first small bridge built of iron. It proved a success. I saw that it would never do to depend further upon wooden bridges for permanent railway structures. **An important bridge on the Pennsylvania Railroad had recently burned and the traffic had been obstructed for eight days.** Iron was the thing.

Cultural experience

In the 1800s, there weren't many ways for an American to learn about fine art and classical culture. There weren't many great museums (the Metropolitan Museum in New York, for instance, [wasn't established until the 1870s](#); much of its collection was donated by the great industrialists of that era, including [over seven thousand pieces from J. P. Morgan](#)). There was, of course, no Internet, no multimedia, and not even a lot of high-quality printed books (or libraries to borrow them from—Carnegie himself was later to establish many public libraries as a cornerstone of his philanthropy). There were no recordings of music until the end of the century, and no radio broadcasts.

So the only way to learn was to vacation to Europe. Carnegie was one of the few who could afford such a trip (and the time off in which to take it), and he wrote of its profound effect on him:

Up to this time I had known nothing of painting or sculpture, but it was not long before I could classify the works of the great painters. One may not at the time justly appreciate the advantage he is receiving from examining the great masterpieces, but upon his return to America he will find himself unconsciously rejecting what before seemed truly beautiful, and judging productions which come before him by a new standard. **That which is truly great has so impressed itself upon him that what is false or pretentious proves no longer attractive.**

My visit to Europe also gave me my first great treat in music. The Handel Anniversary was then being celebrated at the Crystal Palace in London, and I had never up to that time, nor have I often since, felt the power and majesty of music in such high degree. What I heard at the Crystal Palace and what I subsequently heard on the Continent in the cathedrals, and at the opera, certainly enlarged my appreciation of music. At Rome the Pope's choir and the celebrations in the churches at Christmas and Easter furnished, as it were, a grand climax to the whole.

Later he took a more ambitious trip, around the world, which was even more transformative for him:

A new horizon was opened up to me by this voyage. It quite changed my intellectual outlook. Spencer and Darwin were then high in the zenith, and I had become deeply interested in their work. I began to view the various phases of human life from the standpoint of the evolutionist. In China I read Confucius; in India, Buddha and the sacred books of the Hindoos; among the Par-sees, in Bombay, I studied Zoroaster.

The outcome, as he describes it, was a much more cosmopolitan outlook, and a sense of the commonality of world cultures:

The result of my journey was to bring a certain mental peace. Where there had been chaos there was now order. My mind was at rest. I had a philosophy at last....

All the remnants of theology in which I had been born and bred, all the impressions that Swedenborg had made upon me, now ceased to influence me or to occupy my thoughts. I found that no nation had all the truth in the revelation it regards as divine, and no tribe is so low as to be left without some truth; that every people has had its great teacher; Buddha for one; Confucius for another; Zoroaster for a third; Christ for a fourth....

Every person who can, even at a sacrifice, make the voyage around the world should do so. All other travel compared to it seems incomplete, gives us merely vague impressions of parts of the whole. When the circle has been completed, you feel on your return that you have seen (of course only in the mass) all there is to be seen. The parts fit into one symmetrical whole and you see humanity wherever it is placed working out a destiny tending to one definite end.

The world traveler who gives careful study to the bibles of the various religions of the East will be well repaid. The conclusion reached will be that the inhabitants of each country consider their own religion the best of all. They rejoice that their lot has been cast where it is, and are disposed to pity the less fortunate condemned to live beyond their sacred limits.

Disease

Like many people of [the pre-germ theory era](#), Carnegie suffered from infectious disease, and lost relatives to it. The fact that this was common, and had been for all of history, didn't prevent the tragedy from affecting him emotionally:

The year 1886 ended in deep gloom for me. My life as a happy careless young man, with every want looked after, was over. I was left alone in the world. My mother and brother passed away in November, within a few days of each other, while I lay in bed under a severe attack of typhoid fever, unable to move and, perhaps fortunately, unable to feel the full weight of the catastrophe, being myself face to face with death.

Pollution

Air pollution in Pittsburgh was almost inconceivable:

Any accurate description of Pittsburgh at that time would be set down as a piece of the grossest exaggeration. The smoke permeated and penetrated everything. **If you placed your hand on the balustrade of the stair it came away black; if you washed face and hands they were as dirty as ever in an hour.** The soot gathered in the hair and irritated the skin, and for a time after our return from the mountain atmosphere of Altoona, life was more or less miserable.

Oil spills, now considered a disaster, were once routine. Again describing his visit to the oil fields:

In those early days all the arrangements were of the crudest character. When the oil was obtained it was run into flat-bottomed boats which leaked badly. Water ran into the boats and the oil overflowed into the river. The creek was dammed at various places, and upon a stipulated day and hour the dams were opened and upon the flood the oil boats floated to the Allegheny River, and thence to Pittsburgh.

In this way not only the creek, but the Allegheny River, became literally covered with oil. **The loss involved in transportation to Pittsburgh was estimated at fully a third of the total quantity, and before the oil boats started it is safe to say that another third was lost by leakage.**

Incidentally, in the early days of the industry, many people thought that oil would run out quickly. Carnegie, like many others, lost money on a scheme to take advantage of the peak that was believed to be imminent:

Mr. Coleman, ever ready at suggestion, proposed to make a lake of oil by excavating a pool sufficient to hold a hundred thousand barrels (the waste to be made good every day by running streams of oil into it), and to hold it for the not far distant day when, as we then expected, the oil supply would cease. This was promptly acted upon, but after losing many thousands of barrels waiting for the expected day (which has not yet arrived) we abandoned the reserve. Coleman predicted that when the supply stopped, oil would bring ten dollars a barrel and therefore we would have a million dollars worth in the lake. **We did not think then of Nature's storehouse below which still keeps on yielding many thousands of barrels per day without apparent exhaustion.**

Overall I found the autobiography readable and enjoyable, although for my research purposes I lost interest after Carnegie's retirement (the last several chapters are all about his philanthropy and about politics). If you want more like the excerpts above, [it's worth reading](#).

Wanting to Succeed on Every Metric Presented

There's a tendency to want to score high on every metric you come across. When I first read Kegan's [5 stages of adult development](#), I wanted to be a stage 5 meta-rationalist! Reading the meditation book "The Mind Illuminated" (TMI), I wanted to be stage 10 (and enlightened and stage 8 jhana and...)! I remember seeing [this dancer](#) moonwalk sideways and wanting to be that good too!

This tendency is harmful.

But isn't it good to want to be good at things? Depends on the "things" and your personal goals. What I'm pointing out is a tendency to become emotionally invested in metrics and standards, without careful thought on what you actually value. If you don't seriously investigate your own personal preferences and taste, you may spend years of your life invested in something you don't actually care about. By adding this habit of reflection, you could become much happier than you are right now.

[Note: I believe most people are bad at figuring out what they actually value and prefer. For example, I thought skilled pianists are cool and high status, but when I actually became decent enough to wow your average Joe, being cool in those moments wasn't as cool as I thought it would be. As they say, "[Wanting is better than having](#)".]

There's a difference between wanting to score 100's/all A+'s and scoring well enough to get a job. There's a difference between reading multiple textbooks cover-to-cover and reading the 40% or so that seem relevant to your tasks. There are tradeoffs; you can't optimize for everything. When you perceive a metric that makes you really want to score highly on, **nail down the tradeoffs in fine-grained details**. What about this do you actually care about? What's the minimum you could score on this metric and still get what you want? What *do* you actually want? Speaking out loud or writing this out is good for getting an outside view and notice confusion.

Noticing this pattern is half the battle. To make it concrete, here are examples from my life:

Running - I ran cross country and track for 3 years, but then I realized I don't enjoy running long distance. Later I found out that sprinting is fun! If I was better at knowing my values, I could've just played ultimate frisbee with friends instead.

Dancing - I used to imagine dancing at weddings and such and looking really cool! I remember being really self-conscious and slightly miserable when I did dance in front of others. Trying to impress people is disappointing (and trying to be cool is *so uncool*). Now I value dancing because it's fun and a good workout; I don't worry about recording myself and consistently improving or dancing hypotheticals.

Kegan's 5 stage development - I used to want to be stage 5, and I remember reading lots of David Chapman's work to figure this out. I believe I benefited from this, but I ironically would've understood it better if I considered my values better. Now I value it as a useful framing for how large segments of people interpret the world. [See? I pointed out that it's just another system with its own set of limits. [I'm a cool kid now, right?](#)]

Meditation - Becoming enlightened or TMI stage 10 sounded really cool! I've spent 100's of hours meditating now, but I would've been much better off if I crystallized in my head the skills being optimized and how improving those skills improved my life. It wasn't the "wanting to be enlightened prevented becoming enlightened" trope, but optimizing for a fuzzy "enlightened" metric was worse than more tractable metrics with clear feedback.

What I value now from meditation is being happier, accepting reality, being okay with metaphysical uncertainty (not freaking out when realizing I can't directly control all my thoughts, or noticing my sense of self being constructed), and maintaining awareness of context, all of which are much clearer metrics that I actually care about.

Grades - I wanted all A's and to work my hardest on every assignment, wasting a lot of time I could've spent elsewhere! Afterwards, I learned to do just enough to graduate and signal with my GPA that I'm a hard worker/smart. [Once, I missed my final exam where I needed a 60 to keep an A, dropping me to a C. I thought it was hilarious. [Thanks Nate!](#)]

Social Appraisals - I used to be emotionally affected by most everybody's social rewards and punishments (i.e. attention and praise vs ignoring and criticism). I've felt awful and disappointed so many times because of this! I've come to realize that I actually only care about <10 people's opinion of my worth, and they all care about me and know me well. [Note: this is separate from taking someone's thoughts into consideration]

The post that prompted this was [Specializing in problems we don't understand](#). Great post! I noticed the compulsion to work on this problem immediately without considering my current context and goals, so I wrote this post instead.

Topics people in this community may benefit from re-evaluating are:

- Existential AI risks and other EA areas. Not just whether or not you actually want to work in these fields, but also "do you actually enjoy pursuing it the way you are currently pursuing it?"
- Reading text books cover-to-cover and doing all the exercises
- Writing posts and comments in this forum in general

So... do you feel compelled to succeed according to the metric I've presented?

Iterated Trust Kickstarters

Epistemic Status: I haven't actually used this through to completion with anyone. But, it seems like a tool that I expect to be useful, and it only really works if multiple people know about it.

In this post, I want to make you aware of a few things:

Iterated kickstarters: [Kickstarters](#) where all the payment doesn't go in instantly – instead people pay in incrementally, after seeing partial progress on the goal. (Or, if you don't actually have a government-backed-assurance-contract, people pay in incrementally as you see other people pay in incrementally, so the system doesn't require as much trust to bootstrap)

Trust kickstarters: Kickstarters that are not about money, and are instead about "do we have the mutual trust, goodwill and respect necessary to pull a project or relationship off?" I might be wary of investing into my relationship with you, if I don't think you're going to invest in me.

Iterated trust kickstarters: Combining those two concepts – incrementally ratcheting up trust over time. I think this something people intuitively do sometimes, but it's nice to be able to do it intentionally, and communicate crisply about it.

...

*Alternate phrasing of a key insight: one solution to a prisoner's dilemma is to break it into multiple stages, so you have an **iterated** prisoner's dilemma, which [has a different incentive structure](#).*

Iterated Kickstarters

In [The Strategy of Conflict](#), Thomas Schelling (of Schelling Point fame), poses a problem: Say you have a one-shot coordination game. If Alice put in a million dollars, and her business partner Bob puts in a million dollars, they both get their money back, plus \$500,000 extra. But if only one of you puts in a million, the other can abscond with it.

A million dollars is a lot of money for most people. Jeez.

What to do?

Well, hopefully you live in a society that has built well-enforced laws around [assurance contracts](#) (aka "kickstarters"). You put in a million. If your partner backs out, the government punishes them, and/or forces them to return the money.

But what if there isn't a government? What if we live in the Before Times, and we're two rival clans who for some reason have a temporary incentive to work together (but still incentive to defect)? What if we live in present day, but Alice and Bob are two entirely *different countries* with no shared tradition of cooperation?

There are a few ways to solve this. But one way is to split the *one shot* dilemma into an iterated game. Instead of putting in a million dollars, you each put in \$10. If you both did that, then you each put in another \$10, and another. Now that the game is iterated, the payoff strategy changes from [prisoner's dilemma to stag hunt](#). Sure, at any given time you could defect, but you'd be getting a measly \$10, and giving up on a massive half-million potential payoff.

You see small versions of this fairly commonly on craigslist or in other low-trust contract work. "Pay me half the money up front, and then half upon completion."

This still sometimes results in people running off with the first half of the money. I'm assuming people do "half and half" instead of splitting it into even smaller chunks because the transaction costs get too high. But for many contractors, there are benefits to following through (instead of taking the money and running), because there's still a broader iterated game of reputation, and getting repeat clients, who eventually introduce you to other clients, etc.

(You might say that the common employment model of "I do a week of work, and then you pay me for a week of work, over and over again" is a type of iterated kickstarter).

If you're two rival clans of outlaws, trying to bootstrap trust, it's potentially fruitful to establish a *tradition of cooperation*, where the longterm payoff is better than any individual chance to defect.

Trust Kickstarters

Meanwhile: sometimes the thing that needs kickstarting is not money, but trust and goodwill.

Goodwill kickstarters

I've seen a few situations where multiple parties feel aggrieved, exhausted, and don't want to continue a relationship anymore. This could happen to friends, lovers, coworkers, or project-cofounders.

They each feel like the other person was more at fault. They each feel taken advantage of, and like it'd make them a doormat if they went and extended an olive branch when the other guy hasn't even said "sorry" yet.

This might come from a pure escalation spiral: Alice accidentally is a bit of a jerk to Bob on Monday. Then Bob feels annoyed and acts snippy at Alice on Tuesday. Then on Wednesday Alice is like "jeez Bob what's your problem?" and then is actively annoying as retribution. And by the end of the month they're each kinda actively hostile and don't want to be friends anymore.

Sometimes, the problem stems from cultural mismatches. Carl keeps being late to meetings with Dwight. For Dwight, "not respecting my time" is a serious offense that annoys him a lot. For Carl, trying to squeeze in a friend hangout when you barely have time is a sign of love (and meanwhile doesn't care when people are late). At first, they don't know about each other's different cultural assumptions, and they just accidentally 'betray' each other. Then they start getting persistently mad about the conflict and accrue resentment.

Their mutual friend Charlie comes by and sees that Alice and Bob are in conflict, but the conflict stems from a misunderstanding, or a minor mishap that really didn't need to have been a big deal.

"Can't you just both apologize and move on?" asks Charlie.

But by now, after months of escalation, Alice and Bob have both done some things that were legitimately hurtful to each other, or have mild PTSD-like symptoms around each other.

They'd be willing to sit down, apologize, and work through their problems, *if* the other one apologized first. When they imagine apologizing first, they feel scared and vulnerable.

I'll be honest, I feel somewhat confused about how to best to relate to this sort of situation. I'm currently relating it through the lens of game-theory. I can imagine the best advice for most people is to *not* overthink it, don't stress about game theory. Maybe you should just be letting your hearts and bodies be talking to each other, [elephant to elephant](#).

But... also, it seems like the game theory is just really straightforward here. A "goodwill kickstarter" really should Just Work in these circumstances. If it's true that "I would apologize to you if you apologized to me", and vice versa, holy shit, why are you two still fighting?

Just, agree that you will both apologize conditional on the other person apologizing, and that you would both be willing to re-adopt a [friendship relational stance](#) conditional on the other person doing that.

And then, do that.

Competence Kickstarter

Alternately, you might to kickstart "trust in competence."

Say that Joe keeps screwing up at work – he's late, he's dropping the ball on projects, he's making various minor mistakes, he's communicating poorly. And his boss Henry has started getting angry about it, nagging Joe constantly, pressuring Joe to stay late to finish his work, constantly micromanaging him.

I can imagine some stories here where Joe was "originally" the one at fault (he was just bad at his job for some preventable reason one week, and then Henry started getting mad). I can also imagine stories here where the problems stemmed originally from Henry's bad management (maybe Henry was taking some unrelated anger out on Joe, and then Joe started caring less about his job).

Either way, by now they can't stand each other. Joe feels anxious heading into work each day. Henry feels like talking to Joe isn't worth it.

They *could* sit down, earnestly talk through the situation, take stock of how to improve it. But they don't feel like they can have that conversation, for two reasons.

One reason is that there isn't enough goodwill. The situation has escalated and both are pissed at each other.

Another reason, though, is that they don't trust each other's *competence*.

Manager Henry doesn't trust that Joe can actually reliably get his work done.

Employee Joe doesn't believe that Henry can give Joe more autonomy, talk to him with respect, etc.

In some companies and some situations, by this point it's already too late. It's pretty overdetermined that Henry fires Joe. But that's not always the right call. Maybe Henry and Joe have worked together long enough to remember that they *used* to be able to work well together. It seems like it should be possible to repair the working relationship. Meanwhile Joe has a bunch of talents that are hard to replace - he built many pieces of the company infrastructure and training a new person to replace him would be costly. And there's a bunch of nice things about the company they work that makes Joe prefer not to have to quit to find a better job elsewhere.

To repair the relationship, Henry needs to believe that Joe *can* start getting work done reliably. Joe needs to believe that Henry can start treating him with respect, without shouting angrily or micromanaging.

This only works if they in fact both can credibly signal that they will do these things. This works if the missing ingredient is "just try harder." Maybe the only reason Joe isn't working reliably is that he no longer believes it's worth it, and the only reason Henry is being an annoying manager is that he felt like he needed to get Joe to get his stuff done on time.

In that case, it's reasonably straightforward to say: "I would do my job if you did yours", coupled with the relational-stance-change of "I would become genuinely excited to be your employee if you became genuinely excited about being my boss".

Sometimes, this won't work. The kickstarter can't trigger because Henry doesn't, in fact, trust Joe to do the thing, even if Joe is trying hard.

But, you can still clearly lay out the terms of the kickstarter. "Joe, here's what I need from you. If you can't do that, maybe I need to fire you. Maybe you need to go on a sabbatical and see if you can get your shit together." Maybe you can explore other possible configurations. Maybe the reason Joe isn't getting his work done is because of a problem at home, and he needs to take a couple weeks off to fix his marriage or something, but would be able to come back and be a valuable team member afterwards.

I think having the terms of the kickstarter clearly laid out is helpful for thinking about the problem, without having to commit to anything.

Why do you need to think about this in terms of "kickstarter", rather than just "a deal?". What feels special to me about relationship kickstarters is that relationship (and perhaps other projects) benefit from *investment and momentum*. If your stance is "I'm ready to jump and execute this plan if only other people were onboard and able to fulfill their end", then you can be better positioned to get moving quickly as soon as the others are on board.

The nice thing about the kickstarter frame, IMO, is I can take a relationship that is fairly toxic, and I can set my internal stance to be *ready to fix the relationship*, but *without* opening myself up to exploitation if the other person isn't going to do the things I think are necessary on their end.

Iterated Trust Kickstarters

And then, sometimes, a one-shot kickstarter isn't enough.

Henry and Joe

In the case of Henry and Joe: maybe "just try harder" isn't good enough. Joe has some great skills, but is genuinely bad at managing his time. Henry is good at the big picture of planning a project, but finds himself bad at managing his emotions, in a way that makes him bad at actually managing people.

It might be that even if they both *really wanted* things to work out, and were going to invest fully in repairing their working relationship... the next week, Joe might miss a deadline, and Henry would snippily yell at him in a way that was unhelpful. They both have behavioral patterns that will not change overnight.

In that case, you might want to combine "trust kickstarter" and "iterated kickstarter."

Here, Joe and Henry both acknowledge that they're expecting this to be a multi-week (or month) project. The plan needs to include some slack to handle the fact that they might fuck up a bit, and a sense of what's supposed to happen when one of them screws up. It *also* needs a mechanism for saying "you know what, this isn't working."

"Iterated Trust Kickstarter" means, "I'm not going to *fully* start trusting you because you say you're going to try harder and trust me in turn. But, I will trust you a little bit, and give it some chance to work out, and then trust you a bit more, etc." And vice versa.

Rebuilding a Marriage

A major reason to want this is that sometimes, you feel like someone has legitimately hurt you. Imagine a married couple who had a decade or so of great marriage, but then ended up in a several-year spiral where they stop making time for each other, get into lots of fights. Each of them has built up a story in their head where the other person is hurting them. Each of them has done some genuinely bad things (maybe cheated, maybe yelled a lot in a scary way).

Relationships that have gone sour can be really tricky. I've seen a few people end up in states where I think it's *legitimately reasonable* to be worried their partner is abusive, but also, it's legitimately reasonable to think that the bad behavioral patterns are an artifact of a particularly bad set of circumstances. If Alice and Bob were to work their way out of those circumstances, they could still rebuild something healthy and great.

In those cases, I think it's important for people to able invest a *little* back into the relationship – give a bit of trust, love, apology, etc, as a signal that they think the relationship is worth repairing. But, well, "once bitten, twice shy." If someone has hurt you, especially multiple times, it's sometimes really bad to leap directly into "fully trusting the other person."

I think the Iterated Trust Kickstarter concept is something a lot of people do organically without thinking about it in exactly these terms (i.e lots of people damage a relationship and then slowly/carefully repair it).

I like having the concept handle because it helps me think about how exactly I'm relating to a person. It provides a concrete frame for avoiding the failure modes of "holding a relationship at a distance, such that you're basically sabotaging attempts to repair it", and "diving in so recklessly that you end up just getting hurt over and over."

The ITK frame helps me lean hard into repairing a relationship, in a way that feels safe.

(*disclaimer: I haven't directly used this framework through to completion, so I can't vouch for it working in practice. But this seems to mostly be a formalization of a thing I see people doing informally that works alright*)

Concrete Plans

For an ITK to work out, I think there often needs to be a concrete, workable plan. It may not enough to just start trusting each other and hope it works out.

If you don't trust each other's competence (either at "doing my day job", or "learning to speak each other's love languages"), then, you might need to check:

- Does Alice/Bob each understand what things they want from one another? If this is about emotional or communication skills they don't have, do they have a shared understanding of what skills they are trying to gain and why they will help?
- Do they have an actual workable plan for gaining those skills?

Say that Bob has tried to get better at communication a few times, but he keeps running into the same [ugh fields](#) which prevent him from focusing on the problem. He and Alice might need to work out a plan together for navigating those ugh fields before Alice will feel safe investing more in the relationship.

And if Alice is already feeling burned, she might already be so estranged that she's not willing to help Bob come up with a plan to navigate the ugh-fields. "Bob, my terms for the initial step in the kickstarter is that I need you to have already figured out how to navigate ugh fields on your own, before I'm willing to invest anything."

Unilaterally Offering Kickstarters

Part of why I'd like to have this concept in my local rationalist-cultural-circles is that I *think* it's pretty reasonable to extend a kickstarter offer unilaterally, *if* everyone involved is already familiar with the concept and you don't have to explain it.

(New coordinated schemes are costly to evaluate, so if your companion *isn't* already feeling excited about working with you on something, it may be asking too much of them to listen to you explain Iterated Trust Kickstarters in the same motion as asking them to consider "do you want to invest more in your relationship with me?")

But it feels like a useful tool to have in the water, available when people need it.

In many of the examples so far, Alice and Bob both *want* the relationship to succeed. But, sometimes, there's a situation Alice has totally given up on the relationship. Bob may also feel burned by Alice, but he at least feels there's some potential value on the table. And it'd be nice to easily be able to say:

"Alice, for what it's worth, I'd be willing to talk through the relationship, figure out what to do, and do it. I'm still mad, but I'd join the Iterated Kickstarter here." If done right, this doesn't have to cost Bob anything other than the time spent saying the sentence, and Alice the time spent listening to it. If Alice isn't interested, that can be the end of that.

But sometimes, knowing that someone else *would* put in effort if you also would, is helpful for rekindling things.

My take on Michael Littman on "The HCI of HAI"

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

This is independent research. To make it possible for me to continue writing posts like this, please consider [supporting me](#).

Many thanks to Professor Littman for reviewing a draft of this post.

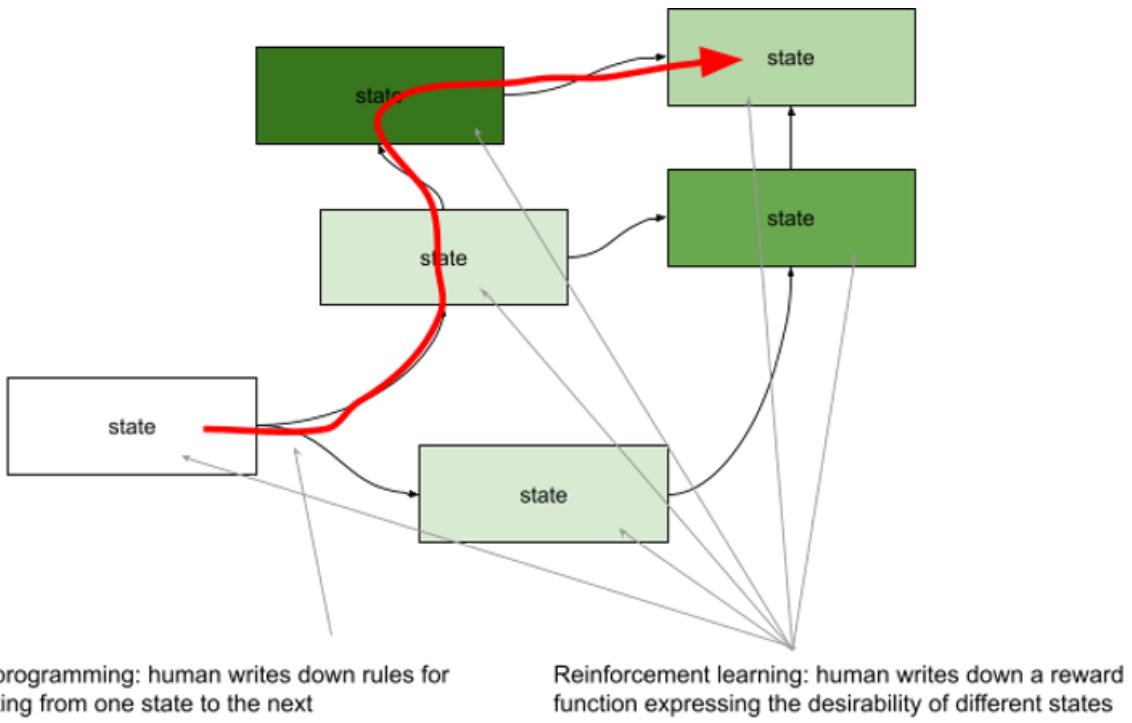
Yesterday, at a seminar organized by The Center for Human-compatible AI (CHAI), [Professor Michael Littman](#) gave a presentation entitled "The HCI of HAI", or "The Human Computer Interaction of Human-compatible Artificial Intelligence". Professor Littman is a computer science professor at Brown who has done foundational work in reinforcement learning as well as many other areas of computer science. It was a very interesting presentation and I would like to reflect a little on what was said.

The basic question Michael addressed was: "how do we get machines to do what we want?" and his talk was structured around the various modalities that we've developed to convey our intentions to machines, starting from direct programming, through various forms of machine learning, and on to some new approaches that his lab is developing. I found his taxonomy helpful, so I'll paraphrase some of it below and then add my own thoughts afterwards.

One way we can get machines to do what we want is by directly programming them to do what we want. But, the problem here is that direct programming is challenging for most people, and Michael talked through some studies that asked non-programmers to build simple if-then rulesets to implement programs such as "if it's raining and the dog is outside then do such-and-such". Most participants had difficulty reasoning about the difference between "if the dog goes outside while it is raining" versus "if it starts raining while the dog is outside".

I'm not sure that this is the main reason to look beyond direct programming, but I do agree that we need ways to instruct machines that are at least partially example-driven rather than entirely rule-driven, so I'm happy to accept the need for something beyond direct programming.

Next, Michael discussed reinforcement learning, in which the human provides a reward function, and it is the machine's job to find a program that maximizes rewards. We might say that this allows the human to work at the level of states rather than behavior, since the reward function can evaluate the desirability of states of the world and leave it up to the machine to construct a set of rules for navigating towards desirable states:



But, writing down a reward function that clearly expresses our intentions can also be very difficult, for reasons that have been discussed here and in the AI Safety literature. One way to express those reasons is Goodhart's Law, which says that when we apply powerful optimization to any proxy measure of that-which-we-really-care-about, we tend to get something that is both unexpected and undesirable. And, expressing that-which-we-really-care-about in full generality without proxy measures seems to be exceedingly difficult in all but the simplest situations.

So we come to inverse reinforcement learning, in which we ask the machine to formulate its own reward function based on watching a human perform the task. The basic idea with inverse reinforcement learning is to have humans demonstrate a task by taking a series of actions, then have the machine find a reward function, which, if the human had been choosing actions in order to maximize, would explain each of their actions. Then the machine takes actions in service of the same reward function. Michael gave some nice examples of how this works.

The problem with inverse reinforcement learning, however, is that in full generality it is both underspecified and computationally intractable. It is underspecified because in the absence of a good prior, any human behavior whatsoever can be explained by a reward function that rewards that exact behavior. One solution to this concern is to develop priors on reward functions, and work continues on this front, but it does mean that we have transformed the problem of writing down a good reward function to the problem of writing down a good prior on reward functions.

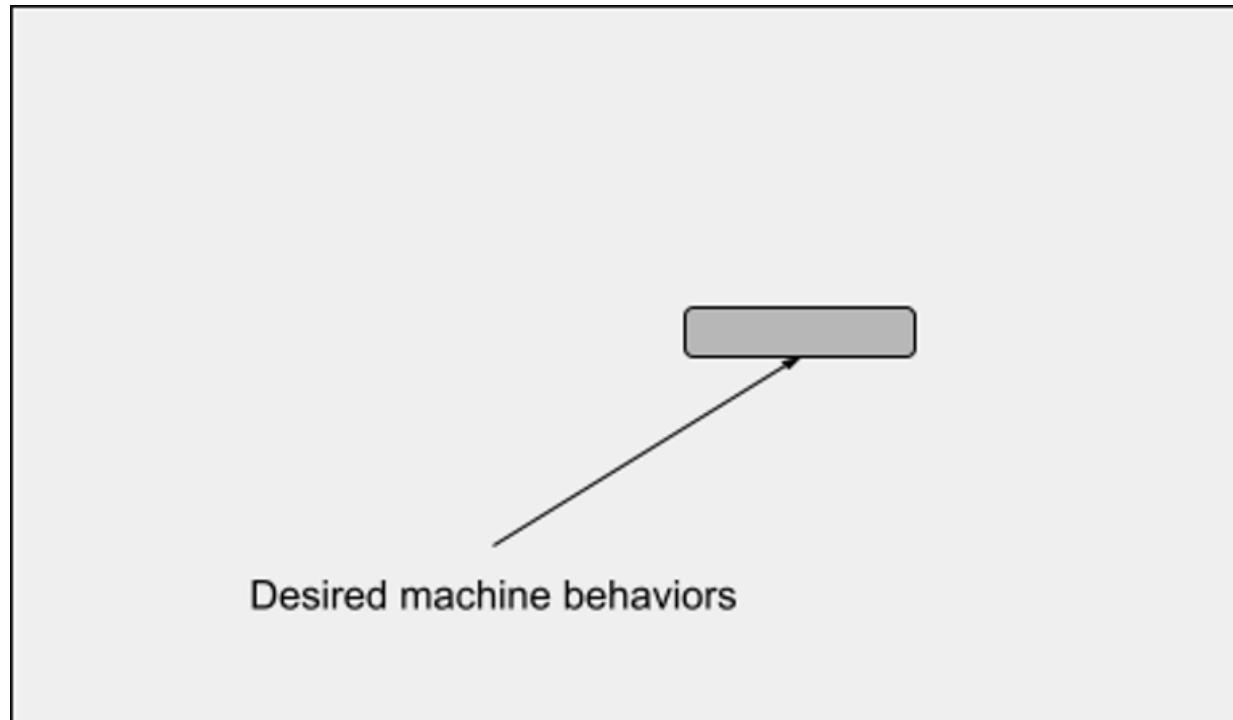
Next, Michael discussed direct rewards, in which a human directly provides rewards to a machine as it learns. This method skips over the need to write down a reward function or a prior on reward functions by instead having a human provide each and every reward manually, but it comes at the expense of being much slower.

Yet it turns out that providing direct rewards is subtle and difficult, too. Michael talked about a study in which some people were asked to provide rewards within a game

where the goal was to teach a virtual dog to walk to a door without walking on any flowers. In the study, most participants gave the dog positive rewards for walking on the path, negative rewards for walking on flowers, and a positive reward for reaching the door. But, the optimal policy under such a reward structure is actually to walk back and forth on the path endlessly in order to collect more and more reward, never reaching the goal. The "correct" reward structure in this case is actually to provide a positive reward for reaching the door, a large negative reward for walking on the flowers, and a small negative reward for walking on the path, in order to incentivize the dog to get to the door as quickly as possible.

A shortcoming of all of the learning approaches (reinforcement learning, inverse reinforcement learning, and direct rewards) is that they lack affordances for writing down rules even in cases where rules would be most natural. For example, when we teach someone how to drive a car, we might tell them not to drive through red lights. To teach a machine to drive a car using reinforcement learning we could provide a negative reward for states of the world in which the vehicle is damaged or the passengers are injured, and hope that the machine learns that driving through red lights leads on average to crashes, which on average leads to vehicle damage or passenger injury. Under inverse reinforcement learning, we could demonstrate stopping at red lights and hope that the machine correctly infers the underlying rule. Under direct rewards we could provide negative rewards for driving through red lights and hope that the machine learns to stop at red lights, and not, for example, to take a more circuitous route that avoids all traffic lights. But these all seem indirect compared to providing a simple instruction: if traffic lights are red then stop the car.

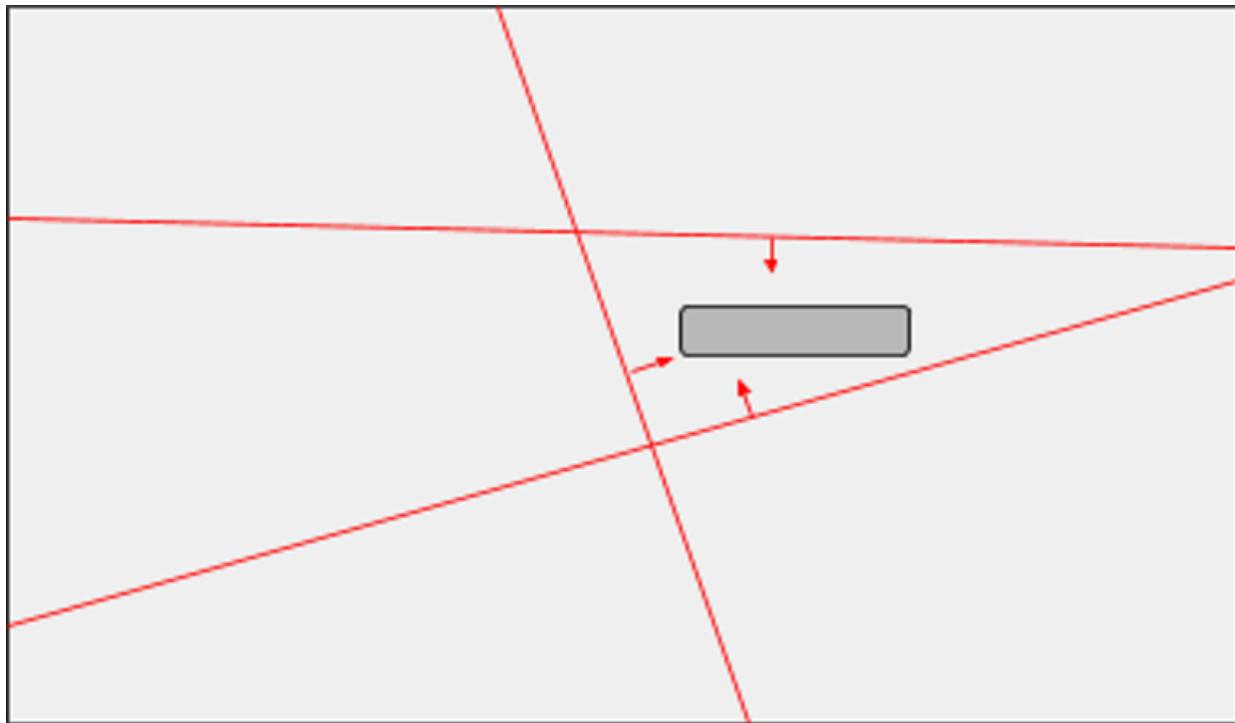
One seminar participant offered the following helpful way of seeing each of the above approaches. Imagine a space of all possible machine behaviors, and ask how we can communicate a desired behavior to a machine^[1]:



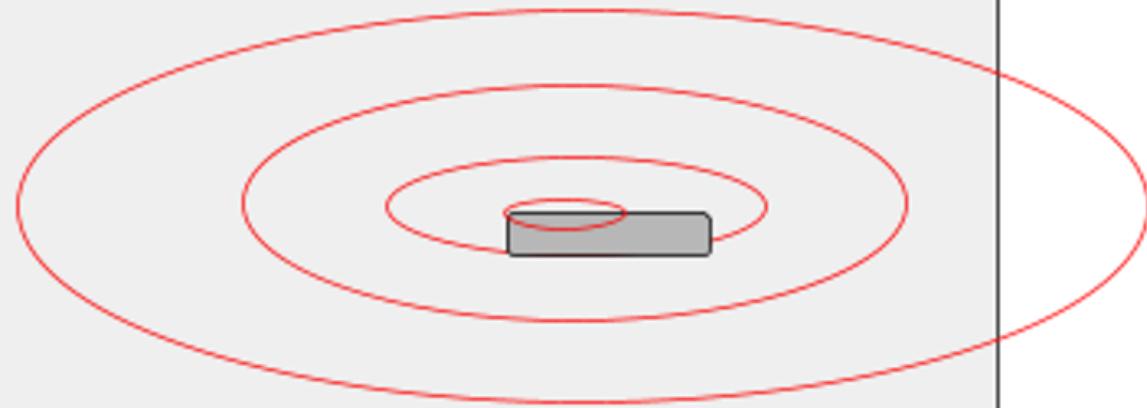
All possible machine behaviors

The four approaches discussed above could be viewed as follows.

- Under direct programming, we provide rules that eliminate parts of the behavior space:



- Under reinforcement learning we provide a reward function that maps any point in the space to a real number (the sum of discounted rewards):



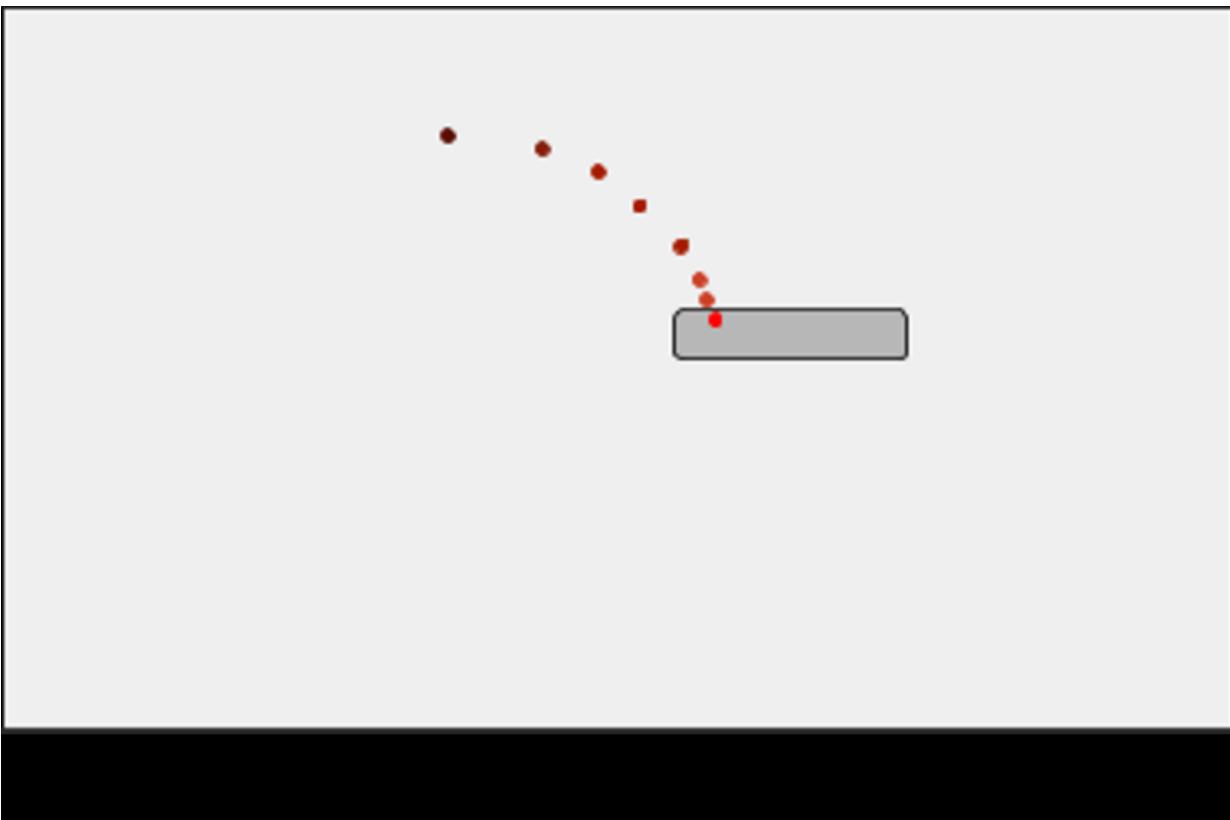
All possible machine behaviors

- Under inverse reinforcement learning we provide examples of the behaviors we desire:



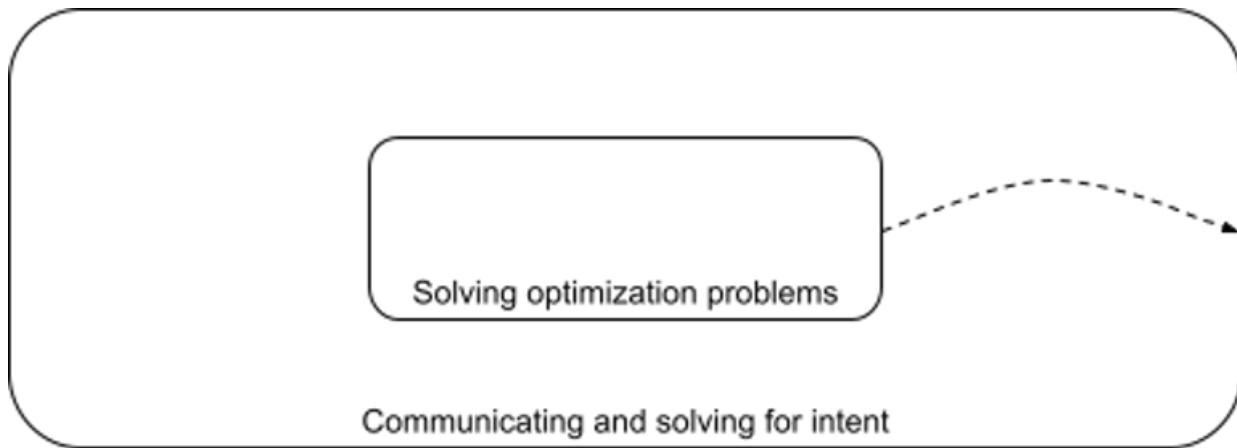
All possible machine behaviors

- Under direct rewarding we allow the machine to explore and provide rewards on a case-by-case basis:



My take

There was a time when computer scientists in the field of AI looked at their job as being about devising algorithms to solve optimization problems. It's not unreasonable to work on solving optimization problems -- that's a valid and important pursuit in the field of computer science -- but if you assume that a human is going to accurately capture their intentions in an optimization problem, and if few people examine how it is that intentions can be communicated from human to machine, then we will end up knowing a lot about how to construct powerful optimization systems while knowing little about how to communicate intentions, which is a dangerous situation. I see Professor Littman's work as "popping out" a level in the nested problem structure of artificial intelligence:

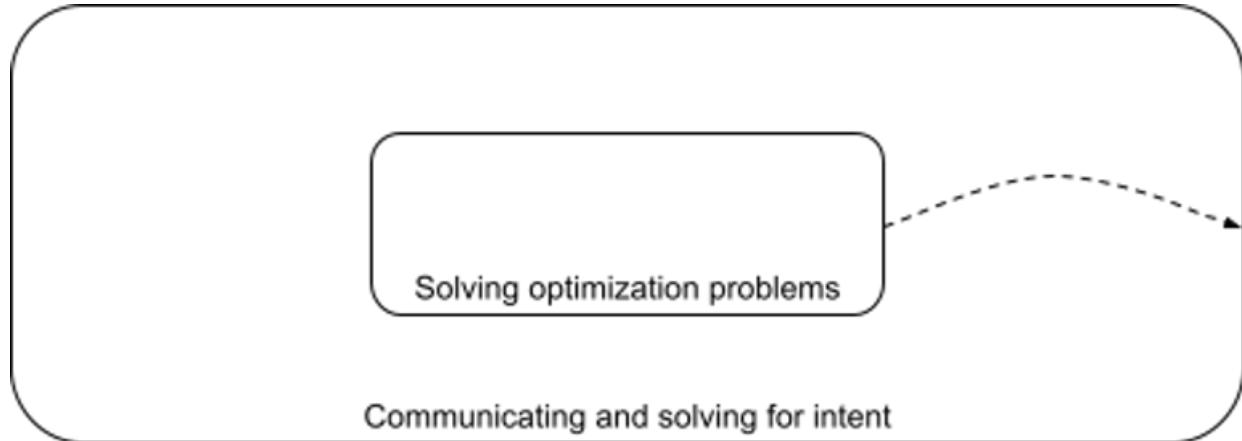


The [work at CHAI concerning assistance games](#) also seems to me to be about "popping out" a level in this nested problem structure, although the specific problem addressed by assistance games is not identical to the one that Michael discussed in his talk.

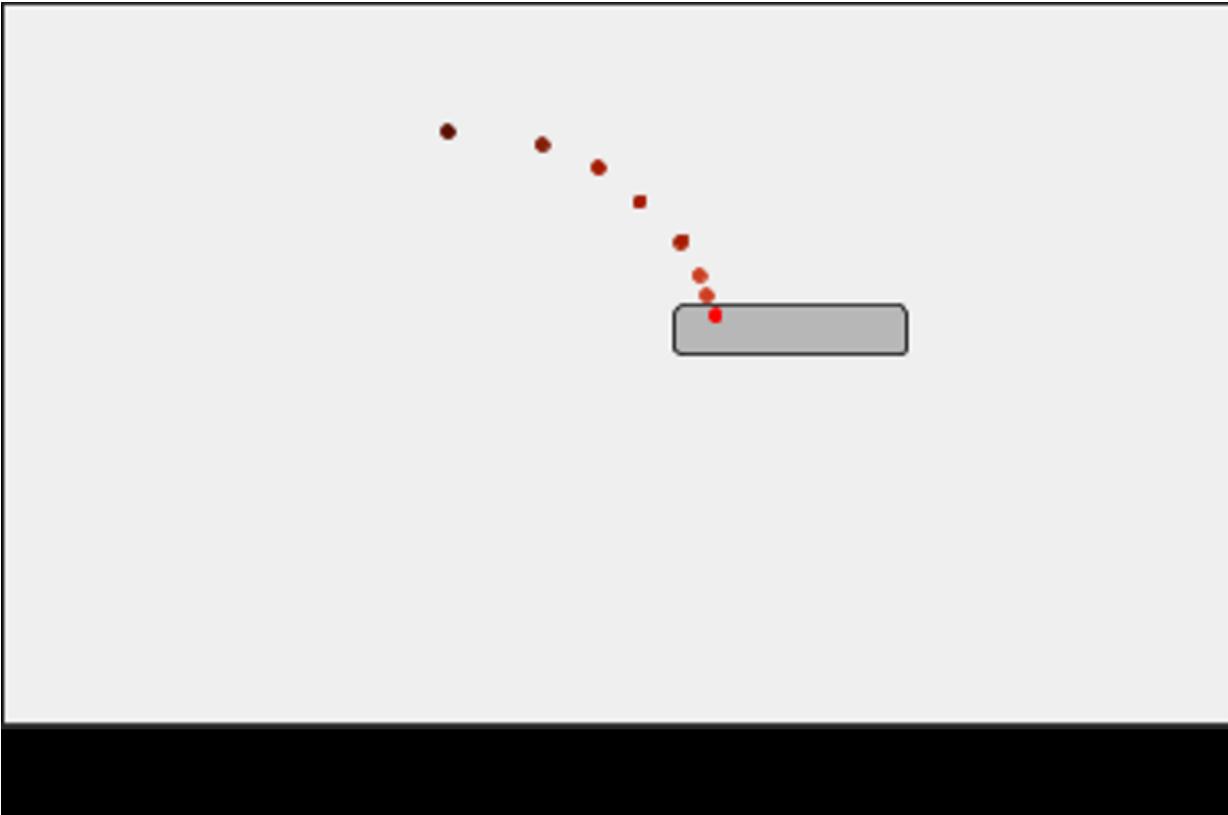
But, there is further yet for us to "pop out". Underlying Michael's talk, as well as, so far as I can tell, assistance games as formulated at CHAI, is the view that humans *have a fixed intention to communicate*. It seems to me that when humans are solving engineering problems, and certainly when humans are solving complex engineering or political or economic problems, it is rare that they hold to a fixed problem formulation without changing it as solutions are devised that reveal unforeseen aspects of the problem. When I worked on visual-inertial navigation systems during my first job after grad school, we started with a problem formulation in which the pixels in each video frame were assumed to have been captured by the camera at the same point in time. But, cell phones use rolling-shutter cameras that capture each row of pixels slightly after the previous row, and it turned out this mattered, so we had to change our problem formulation. But, I'm not just talking about flaws in our explicit assumptions. When I founded a company to build autonomous food-delivery robots, we initially were not clear about the fact that the company was founded, in part, out of a love for robotics. When the three founders became clear about this, we changed some of the ways in which we pitched and hired for the company. It's not that we overturned some key assumption, but that we discovered an assumption that we didn't know we were making. And, when I later worked on autonomous cars at a large corporation, we were continuously refining not just the formulation of our intention but our intentions themselves. One might respond that there must have been some higher-level fixed intention such as "make money" or "grow the company" from which our changing intentions were derived, but this was not my experience. At the very highest level -- the level of what I should do with my life and my hopes for life on this planet -- I have again and again overturned not just the way I communicate my intentions but my intentions themselves.

And, this high-level absence of fixed intentions shows up in small-scale day-to-day engineering tasks that we might want AI systems to help us with. Suppose I build a wind monitor for paragliders. I begin by building a few prototypes and deploying them at some paragliding sites, but I discover that they are easily knocked over by high winds. So, I build a more robust frame, but I discover that the cell network on which they communicate has longer outages than I was expecting. So, I change the code to wait longer for outages to pass, but I discover that paragliding hobbyists actually want to know the variability in wind speed, not just the current wind speed. So, I change the UI to present this information but I discover that I do not really want to build a company

around this product, I just want a hobby project. So, I scale down my vision for the project and stop pitching it to funders. If I had worked with an AI on this project then would it really be accurate to say that I had some fixed intention all along that I was merely struggling to communicate to the AI? Perhaps we could view it this way, but it seems like a stretch. A different way to view it is that each new prototype I deployed into the world gave me new information that updated my intentions at the deepest level. If I were to collaborate with an AI on this project, then the AI's job would not be to uncover some fixed intentions deep within me, but to participate fruitfully in the *process of aligning both my intentions and that of the AI with something that is bigger than either of us*.



In other words, when we look at diagrams showing the evolution of a system towards some goal state such as the ones in the first section of this post, we might try viewing ourselves as being inside the system rather than outside. That is, we might view the diagram as depicting the joint (human + machine) configuration space, and we might say that the role of an AI engineer is to build [the kind of machines that have a tendency, when combined with one or more humans, to evolve towards a desirable goal state](#):



It might be tempting to view the question "how can we build machines that participate fruitfully in the co-discovery of our true purpose on this planet?" as too abstract or philosophical to address in a computer science department. But, remember that there was a time when the question "how can we communicate our intentions to machines?" was seen as outside the scope of core technical computer science. Once we started to unpack this question, we found not only that it was possible to unpack, but that it yielded new concrete models of artificial intelligence and new technical frontiers for graduate students to get stuck into. Perhaps we can go even further in this direction.

-
1. This diagram and all future diagrams in this post are my own and are not based on any in Michael's presentation [←](#)

Draft report on existential risk from power-seeking AI

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

I've written a draft report evaluating a version of the overall case for existential risk from misaligned AI, and taking an initial stab at quantifying the risk from this version of the threat. I've made the draft viewable as a public google doc [here](#) (Edit: arXiv version [here](#), video presentation [here](#)). Feedback would be welcome.

This work is part of Open Philanthropy's "[Worldview Investigations](#)" project. However, the draft reflects my personal (rough, unstable) views, not the "institutional views" of Open Philanthropy.

April drafts

By Katja Grace, April 1 2021

Today we are sharing with our blog readers a collection of yet-to-be-published drafts, in the hope of receiving feedback. We are especially looking for methodological critique, but all comments welcome!

[Human-level performance estimate](#) (Katja Grace)

[How much hardware will we need to create AGI?](#) (Asya Bergal)

[Historic trends in AI Impacts productivity](#) (Daniel Kokotajlo)

[Analysis of superbombs as a global threat](#) (Katja Grace)

FAQ: Advice for AI Alignment Researchers

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

This is a linkpost for <https://rohinchah.com/faq-career-advice-for-ai-alignment-researchers/>

To quote Andrew Critch:

I get a lot of emails from folks with strong math backgrounds (mostly, PhD students in math at top schools) who are looking to transition to working on AI alignment / AI x-risk. There are now too many people “considering” transitioning into this field, and not enough people actually working in it, for me, or most of my colleagues at Stuart Russell’s [Center for Human Compatible AI \(CHAI\)](#), to offer personalized mentorship to everyone who contacts us with these qualifications.

[From math grad school to AI alignment](#), Andrew Critch

I’m pretty sure he wrote that at least 4 years ago (2016 or earlier). The field has grown enormously since then, but so has the number of people considering it as a research area. So far, I’ve tried to give at least 10 minutes of my time to anyone who emails me with questions; that probably won’t be sustainable for much longer. So now I’m answering the questions I get most frequently. I hope to keep this up to date, but no promises.

Usually, I write a blog post when I think I have something important and novel to say, that I am relatively confident in. That’s not the case for this post. This time, I’m taking all the questions that I frequently get and writing down what I’d say in response. Often, this is (a) not that different from what other people would say, and (b) not something I’m very confident in. Take this with more grains of salt than usual.

Thanks to Neel Nanda, Nandi Schoots, and others who wish to remain anonymous for contributing summaries of conversations.

See the linked post for the FAQ; which will hopefully be kept up to date over time.

Facebook is Simulacra Level 3, Andreessen is Level 4

A passage I just read in [The Hard Thing About Hard Things](#) by Ben Horowitz:

Andreessen vs Zuckerberg: How Big Should the Titles Be?

Should your company Vice President the top title or should you have Chief Marketing Officers, Chief Revenue Officers, Chief People Officers, and Chief Snack Officers? There are two schools of thoughts regarding this, one represented by Marc Andreessen and the other by Mark Zuckerberg.

Andreessen argues that people ask for many things from a company: salary, bonus, stock options, span of control, and titles. Of those, title is by far the cheapest, so it makes sense to give the highest titles possible. The hierarchy should have. Presidents, Chiefs, and Senior Executive Vice Presidents. If it makes people feel better, let them feel better. Titles cost nothing. Better yet, when competing for new employees with other companies, using Andreessen's method you can always outbid the competition in at least one dimension.

At Facebook, by contrast, Mark Zuckerberg purposely deploys titles that are significantly lower than the industry standard. Senior Vice Presidents at other companies must take title haircuts down to Directors or Managers at Facebook. Why does he do this? First, he guarantees that every new employee gets leveled as they enter his company. In this way, he avoids accidentally giving new employees higher titles and positions than better-performing existing employees. This boosts morale and increases fairness. Second, it forces all the managers of Facebook to understand and internalize Facebook's leveling system, which serves the company extremely well in their own promotion and compensation processes.

He also wants titles to be meaningful and reflect who has influence in the organization. As a company grows quickly, it's important to provide organizational clarity wherever possible and that gets more difficult if there are fifty VPs and ten Chiefs.

Next, he finds that businesspeople often carry inflated titles versus their engineering counterparts. While he recognizes that big titles help them out externally with getting meetings, he still wants to have an organization where the product people and engineers form the cultural core, so he strives to keep this in check as well.

Does Facebook ever miss out on a new hire due to its low titles? Yes, definitely. But one might argue that they miss out on precisely the employees they don't want. In fact, both the hiring and onboarding processes at Facebook have been carefully designed to encourage the right kind of employees to select themselves in and the wrong ones to select themselves out.

Simulacra level 3 is about accurately describing social reality. Level 4 is about defecting on people trying to do level 3, by painting an inaccurate model of social reality. It seems to me this is a crystal clear example of two people recommending Level 3 vs Level 4 strategies.

(Or so it is in my mind. I'm sure Zvi and Benquo and others will say Level 4 is something else.)

As an aside, this book is far better than it has any right to be at giving advice on building successful companies. It's a book that repeatedly stares into the dark at the things that will kill your company (e.g. how to fire senior people, how to minimize internal politicking, when smart people are bad employees, etc) and gives simple and clear advice in each situation. I've personally found it immensely helpful.

Against "Context-Free Integrity"

Sometimes when I talk to people about how to be a strong rationalist, I get the impression they are making a specific error.

The error looks like this: they think that good thinking is good thinking irrespective of [environment](#). If they just learn to avoid [rationalization](#) and setting [the bottom-line](#) first, then they will have true beliefs about their environment, and if there's something that's true and well-evidenced, they will come to believe it in time.

Let me give an extreme example.

Consider what a thoughtful person today thinks of a place like the Soviet Union under Stalin. This was a nation with evil running through their streets. People were vanished in the night, whole communities starved to death, the information sources were controlled by the powerful, and many other horrendous things happened every day.

Consider what a strong rationalist would have been like in such a place, if they were to succeed at keeping sane.

(In reality [a strong rationalist would have found their ways out of such places](#), but let us assume [they lived there and couldn't escape](#).)

I think such a person would be deeply paranoid (at least [Mad-Eye Moody](#) level), understanding that the majority of their world was playing power games and trying to control them. They'd spend perhaps the majority of their cognition understanding the traps around them (e.g. what games they were being asked to play by their bosses, what sorts of comments their friends would report them for, etc) and trying to build some space with enough slack to occasionally think straight about the narratives they had to live out every day. It's kind of like living in [The Truman Show](#), where everyone is living a narrative, and punishing you / disbelieving you when you deviate. (Except with much worse consequences than what happened in that show.)

Perhaps this is too obvious to need elaborating on, but the cognition of a rationalist *today* who aims to come to true beliefs about the Soviet Union, and the cognition of a rationalist *in the Soviet Union* who aims to come to true beliefs about the Soviet Union, are not the same. They're massively different. The latter of them is operating in an environment where basically every force of power around you is trying to distort your beliefs on that particular topic – your friends, your coworkers, the news, the police, the government, the rest of the world.

(I mean, certainly there are still today many distortionary forces about that era. I'm sure the standard history books are altered in many ways, and for reasons novel to our era, but I think qualitatively there are some pretty big differences.)

No, coming to true beliefs about your current environment, especially if it is hostile, is very different from coming to true beliefs about many other subjects like mathematics or physics. Being in the environment can be especially toxic, depending on the properties of that environment and what relationship you have to it.

By analogy, I sometimes feel like the person I'm talking to thinks that they just practice enough [fermi estimates](#) and [calibration training](#) and notice rationalization in themselves and practice the [principle of charity](#), then they'll probably have a pretty

good understanding of the environment they live in and be able to take positive, directed action in it, even if they don't think carefully about the political forces acting upon them.

And man, that feels kinda naive to me.

Here's a related claim: you cannot get true beliefs about what are good actions to take in your environment without good accounting, and good record-keeping.

Suppose you're in a company that has an accounting department that tells you who is spending money and how. This is great, you can reward/punish people for things like being more/less cost-effective.

But suppose you understand one of the accounting people is undercounting the expenses of their spouse in the company. Okay, you need to track that. (Assume you can't fire them for political reasons.) Suppose another person is randomly miscounting expenses depending on which country the money is being spent. Okay, you need to track that. Suppose some people are filing personal expenses as money they spent supporting the client. Okay, now you need to distrust certain people's reports more-so.

At some point, to have accurate beliefs here, it is again not sufficient to avoid rationalization and be charitable and be calibrated. You need to build a whole accounting system for yourself to track reality.

[A]s each sheep passes out of the enclosure, I drop a pebble into a bucket nailed up next to the door. In the afternoon, as each returning sheep passes by, I take one pebble out of the bucket. When there are no pebbles left in the bucket, I can stop searching and turn in for the night. It is a *brilliant* notion. It will revolutionize shepherding.

—[The Simple Truth](#)

I sometimes see quite thoughtful and broadly moral people interact with systems I know to have many power games going internally. [Moral Mazes](#), to some extent or another. The system outputs arguments and trades, and the person sometimes engages with the arguments and sometimes engages in the trade, and thinks things are going well. But I feel like, if they knew the true internal accounting mechanisms in that entity, then they would be notably more disgusted with the parts of that system they interacted with.

(Imagine someone reading a scientific paper on [priming](#), and seeking deep wisdom in how science works from the paper, and then reading about the way science rewards [replications](#).)

Again, I occasionally talk to such a person, and they can't "see" anything wrong with the system, and if they introspect they don't find a trace of any rationalization local to the situation. And if they've practiced their calibration and fermis and charity, they think they've probably come to true beliefs and should expect that their behavior was net positive for the world. And yet there are instances I feel that it clearly wasn't.

Sometimes I try to tell the people what I can see, and that doesn't always go well. I'm not sure why. Sometimes they have a low prior on that level of terrible accounting, so don't believe me (slash think it's more likely that I'm attempting to deceive them). This is the overly-naive mistake.

More often I think they're just not that interested in building that detailed of a personal accounting system for the thing they're only engaging with some of the time and isn't hurting them very much. It's more work than it's worth to them, so they get kind of tired of talking about it. They'd rather believe the things around them are pretty good rather than kinda evil. Evil means accounting, and accounting is booring. This is the apathetic mistake.

Anyway. All this is me trying to point to an assumption that I suspect some people make, an assumption I call "Context-Free Integrity", where someone believes they can interact with complex systems, and as long as they themselves are good and pure, their results will be good and pure. But I think it's required that you to actually build your own models of the internals of the complex systems before you can assess this claim.

...writing that down, I notice it's too strong. [Eliezer recommends empirical tests](#), and I think you can get a broad overall sense of the morality of a system with much less cost than something like "build a full-scale replica accounting model of the system in google sheets". You can run simple checks to see what sorts of morality the people in the system have (do they lie often? do they silence attempts to punish people for bad things? do they systematically produce arguments that the system is good, rather than trying to simply understand the system?) and also just look at its direct effects in the world.

(In my mind, Zvi Mowshowitz is the standard-bearer on 'noping out' of a bad system as soon as you can tell it's bad. The first time was [with Facebook](#), where he came to realize what was evil about it way in advance of me.)

Though of course, the more of a maze the system is, the more it will actively obscure a lot of these checks, which itself should be noted and listed as a major warning. Just as many scientific papers will not give you their data, only their conclusions, many moral mazes will not let you see their results, or tell you metrics that are confusing and clearly [goodharted](#) (again on science, see citation count).

I haven't managed to fully explain the title of this post, but essentially I'm going to associate all the things I'm criticizing with the name "Context-Free Integrity".

Context-Free Integrity (noun): *The notion that you can have true beliefs about the systems in your environment you interact with, without building (sometimes fairly extensive) models of the distortionary forces within them.*

The secret of Wikipedia's success

This is a linkpost for <https://aaronbergman.substack.com/p/the-secret-of-wikipedias-success>

Why its reputation for unreliability is Wikipedia's greatest asset

Intro

Wikipedia is everything the internet was supposed to be. Before social media became a battleground for foreign election meddling, corporate and political messaging wars, and algorithmic competition for our attention, it was going to be a means of sharing information across geographic, social, and political borders.

Over the last 30 years, this egalitarian-techno-optimist naïveté has given way to pragmatism about the social and economic forces governing the internet. But one beacon of innocent, brilliant functionality reminiscent of the old ethos remains: [wikipedia.org](#).

If you've read a few of my other pieces, you may know that I love linking to Wikipedia. It is my default source for any event, item, or concept that I think readers might not know a ton about. The Capitol riots? [Got that](#). A book summary? [Yup](#). An author's background? [That too](#).

Is it any good?

Lots of people have commented on Wikipedia's notable, and frankly surprising, reliability, breadth and depth all fueled by earnest online volunteers and a little over [\\$100 million](#) each year, or about 2% of [American annual spending on ice cream](#). Although anyone can edit just about any Wikipedia page to say just about anything, [several studies](#) find that Wikipedia is very reliable, albeit not always comprehensive or very analytical.

In my opinion, Wikipedia is often the best source of information on topics with an intermediate amount of salience. That is, extremely popular (say, the views of two presidential candidates) or extremely banal (say, "rice") topics of inquiry naturally attract so much attention that there are likely other excellent resources on the topic.

On the other end, Wikipedia pages on the very specific or obscure likely cannot attract enough attention to warrant confidence that something important has not been omitted. But for things in between - [juggling](#), the city of [Oakland, CA](#), or [flips-flops](#) - Wikipedia is often unambiguously the best single, easily accessible resource.

How is this possible?

There are several articles (like [this one](#) and [this one](#)) out there praising Wikipedia and speculating about the reasons for its success. Without a doubt, the organization makes good use of explicit rules and guidelines, community norms, and a social structure that awards status to people for high-quality contributions.

However, it strikes me that Wikipedia's secret to reliability is something paradoxical that I've never seen explicitly addressed: **its reputation for unreliability**.

If I had a dollar for every time I was told in school that, no, I can't cite Wikipedia as a source for my paper, I would be able to stop feeling guilty for ignoring this popup when I use the

site.

We ask you, humbly: don't scroll away.

i Hi reader, do you use Wikipedia a lot? Have you noticed that we don't show ads or charge a subscription fee? Thanks to the donations of 2% of our readers, Wikipedia remains free, independent, and open to all. If Wikipedia has given you \$2.75 worth of knowledge this year, please take a minute this Wednesday to donate and keep Wikipedia thriving. Thank you.

Please select a payment method

[MAYBE LATER ⓘ](#) [CLOSE X](#)

To be clear, I agree that scholarly work should not cite Wikipedia itself as a source. Anyone can edit it, there's no accountability, blah, blah, blah. Little did I realize, until a few days ago, that every teacher and librarian who pounded this into my head from kindergarten on was likely doing Wikipedia a huge service. Let me explain...

The cost of a good reputation

There are an interesting class of beliefs that become more true the fewer or less strongly people believe them. For example, the belief "my vote matters" will lead more people to go to the polls, which in turn means that every individual vote matters less.

Something similar is happening with Wikipedia. For sources of information widely regarded as reliable—not just by individuals, but by authorities and institutions like government, academia, and the media—there is a *massive* incentive to get them to say what you want them to.

Newspapers are the most obvious example. Politicians and government agencies strategically craft press releases, offer quotes, and make timed leaks to shape the media narrative around some event. Companies pay PR people big money to say something sympathetic that will be quoted in an article. Why? Because many people (at least, the people in power) think that newspapers are reliable. Or, perhaps more accurately, everyone thinks that everyone *else* thinks that newspapers are reliable.

The same thing holds true for scientific research, government reports, and more. Their reputation for reliability (whether deserved or otherwise) makes them a prime target for any institution with a story to sell. That's why biomedical research is a [big, juicy steak](#) for the pharmaceutical industry, and nutrition research is an [important lever of influence](#) for agribusiness.

This isn't an original point. I heard it most clearly expressed by [Will Wilkinson on The Wright Show](#). But there is an obvious corollary I have not heard: sources that are not seen as reliable are much *less* tempting prey for narrative-shaping predators. For example, I solemnly swear that exactly zero corporations, politicians, or government agencies (that I know of) have tried to get me to say (or not say) something on this blog. The reason is pretty obvious: I'm just a random guy on the internet, and I am not regarded by society at large as a reliable source of information.

How Wikipedia hacked the system

Usually, to a rough first approximation, sources regarded as reliable *are actually* more reliable than those that are not. Yes, media bias, [Manufacturing Consent](#), the [replication crisis](#), etc. etc. I'm probably more skeptical of institutional authority than the median non-Trump supporter, precisely because of these concerns.

That said, "unreliable" sources generally *are* pretty unreliable. Consider a list of things *not* generally trusted by the Powers that Be

1. Blogs (by individuals, at least)
2. Reddit posts
3. Donald Trump
4. Company press releases
5. Undergrad research papers not published or endorsed by someone high-profile
6. Wikipedia

That doesn't mean these things are *wrong*. There are *specific* blogs (not my own), Reddit posts, and undergrad research papers (my own, obviously) that I trust more than an arbitrarily-selected Washington Post article or scientific paper. But I would *not* trust an arbitrary blog or Reddit post more than an arbitrary WaPo article or paper. This relationship holds for the first five items on that list.

But Wikipedia is different. I *do* trust a random Wikipedia article (use en.wikipedia.org/wiki/Special:Random to get one) more than a random newspaper article or scientific paper, (*footnote: Of course, this isn't a "fair" comparison. Wikipedia articles can be about random shit completely unrelated to an ideology or the culture war, whereas news tends to be precisely the opposite. I suspect that this would hold true even if I weighted each Wikipedia page by its number of views and then took a random sample, though*), although this wouldn't hold if we limited the papers to, say, those in the hard sciences with >100 citations. If you think I'm crazy for writing this, read "[What's Wrong with Social Science and How to Fix It](#)" and check out several random Wikipedia pages and then get back to me.

So how did Wikipedia hack the system? How is it able to be so reliable? Because it—alone, as far as I can tell—maintains a set of incentives and processes for generating content that **simultaneously** produce reliable content **and** is coded as "unreliable."

Giving anyone online the ability to write anything they want on almost any article *sounds* like the kind of thing that would generate a cesspool of disinformation and nonsense. However, features peculiar to Wikipedia (in particular, the fact that every article has only one version, so "opposing sides" have to reach some sort of equilibrium instead of everyone just publishing their own version) *do* effectively incentivize internet randos to write things that are true instead of false.



These two opposing forces mean that Wikipedia has managed to do something analogous to landing a flipped coin on its side: generate reliable content without gaining an institutional reputation for reliability that would incentivize a massive effort to shape its content (though, unfortunately, the [coin may be wobbling](#)).

Not convinced? Imagine, for a minute, that every Wikipedia page was afforded the same degree of authority as a major newspaper, government agency, or even “real encyclopedia. All hell would break loose. “Trump Wins 2020 Election!” would have been splashed across every half-relevant page, and bots would be created to re-edit the pages each time they were corrected. Companies would spread rumors about rival firms. Investors would short stock and then announce that the company has been falsifying their quarterly statements. Random people would “award” themselves a Nobel Prize.

Obviously, this isn’t a stable equilibrium. Within a few hours at most, people would realize that Wikipedia was (genuinely) unreliable and it would be downgraded to the epistemic equivalent of a flat-earther Facebook group.

So why aren't other "unreliable" sources actually reliable, if they aren't being attacked by anyone trying to shape a narrative? A bunch of reasons. As mentioned above, Wikipedia works as a "marketplace of ideas" because everyone ultimately has to collaborate to create a single page for a given topic. It's similar to an economic ideal competitive market, in which the collective self-interest of thousands of buyers and sellers yields an optimal single "market price."

In the blogosphere or on Reddit, Democrats and Republicans, or Kantians and Utilitarians, or Yankees and Red Sox fans don't have to do this type of [adversarial collaboration](#), ultimately producing just a single product. Instead, every individual or group can have their own blog and write their own Reddit posts.

Even an individual earnestly seeking the truth cannot generally expect to beat the "marketplace of ideas," in the same way that a hypothetical benevolent politburo couldn't expect to price goods and services more efficiently than the free market, except under certain rare circumstances (say, large externalities but no ability to tax or subsidize, or extremely low trading volume).

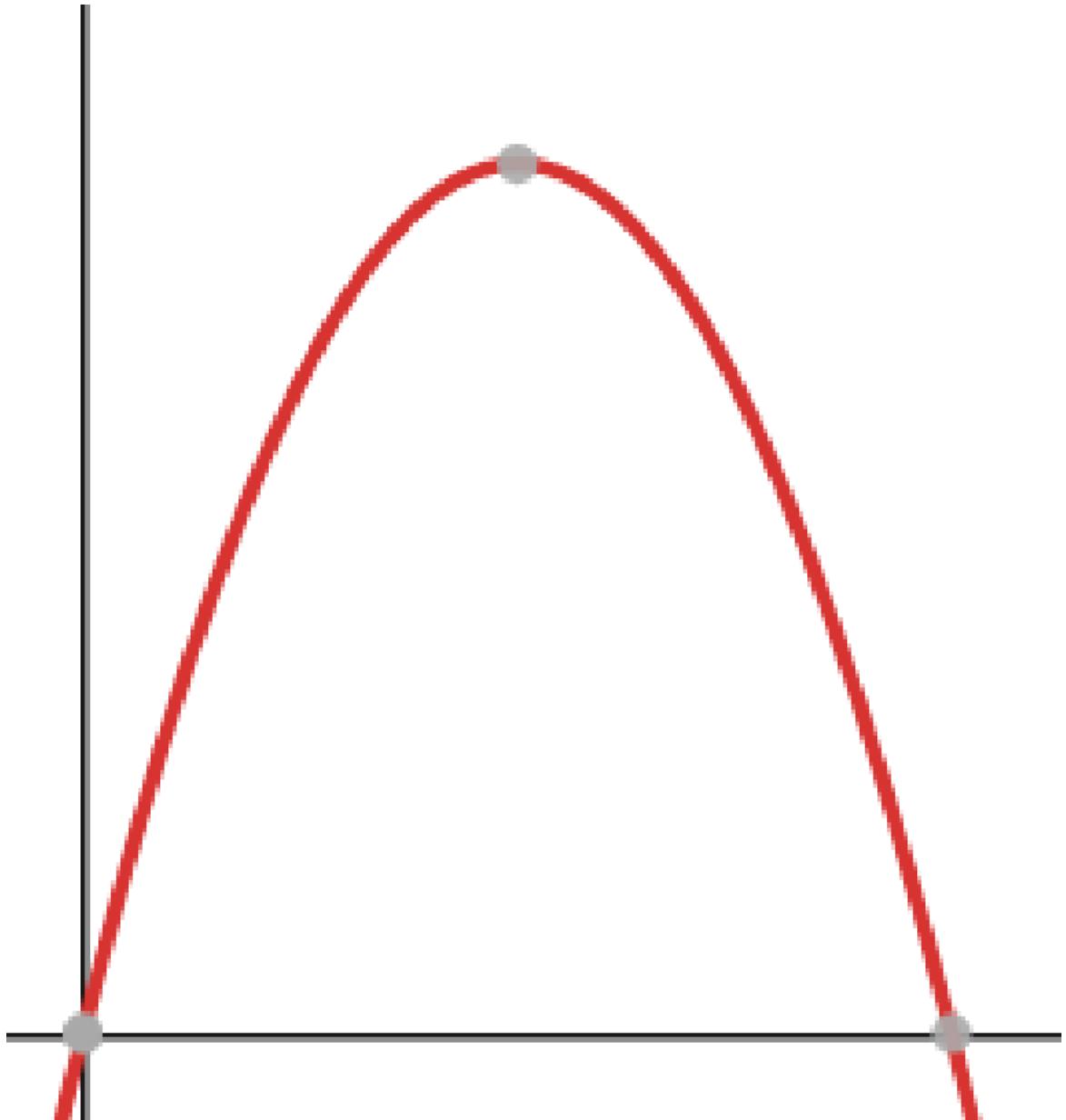
Conclusion

I always thought the aphorism "power corrupts, and absolute power corrupts absolutely" was a little silly, but maybe I was missing the point. Perhaps it isn't that the thing or person with power suddenly sheds its values, but rather that a thing only attracts *external* corrupting influences once it gains power.

The U.S. president's power, for example, inevitably attracts media attention, lobbyists, and even [personally-targeted commercials](#). No matter how principled he or she is, human psychology is fundamentally responsive to the stimuli it receives, and the president's actions will inevitably be influenced to some degree. If my theory is generally right, I hope it serves to illustrate a general cautionary principle: **be careful before empowering or elevating the salience of something good.**

Wikipedia is good not in spite of but *because* of its limited power. Intellectual or social movements might be maximally productive when concentrated among a few earnest, hardcore supporters. Musicians might produce their best work before they feel the need to cater to a growing mass of fans. My blog, insofar as it is interesting at all, can attribute its 'goodness' in part to the fact that I have few 'corrupting influences,' and make \$0 in revenue.

This doesn't mean everything good should be zealously guarded against acquiring power and influence; Wikipedia wouldn't be so awesome if nobody knew about it, after all. Rather, influence is simultaneously symbiotic and parasitic with net positive impact. And Wikipedia, it seems to me, is at the top of the curve.



We need a career path for invention

This is a linkpost for <https://rootsofprogress.org/a-career-path-for-invention>

If [technological progress has slowed down](#), what is causing it? Here is a hypothesis.

Broadly speaking, there are three domains of activity important to technological progress: science, invention, and business. Science discovers new knowledge; invention creates useful machines, chemicals, processes, or other products; and business produces and distributes these products in a scalable, self-sustaining way. (Occasionally inventions are distributed by government: water sanitation is an example. But this oversimplified model will serve for our purposes.)

These domains [do not form a simple linear pipeline](#), but they are distinct areas that attract different types of people, pose different challenges, and are judged by different standards. As such they create distinct communities and subcultures.

My hypothesis is that while science and business have functioning career paths, invention today does not.

Consider science. Suppose a high school or university student has a glimmer of desire to become a scientist. They will find that their road has already been paved. "Scientist" is a career. There's an established path into the career: get a BS and then a PhD in a scientific field. There are research labs that hire scientists, organize them into teams, and give them space and equipment. There is funding for all of this, from government and philanthropy. There is an established deliverable: talks and papers, presented at conferences and published in journals. There are awards and honors that confer prestige within the discipline; some of these, such as the Nobel, are even well-known and respected among the general public.

All of this combines to create a *career path* for the scientist: anyone with even a modest level of commitment and effort can start down the path, and those who are exceptionally talented and ambitious can reach for inspiring goals. Importantly, there is a feedback loop in which *progress down the career path opens opportunities*. The more the scientist produces legible accomplishments, the more they are able to get grants, secure coveted positions, and attract talent to work with them. Money, prestige, and the opportunity to do meaningful work all (roughly) go together.

Entrepreneurship has different structures, but the career path is there nonetheless. "Startup founder" is not a job you get *hired* for; it is a job the founder must create for themselves. They must raise their own funding, create their own organization, and hire their own team. In this sense, the founder is much less well-supported than the scientist. But there are established sources of funding for startups, in venture capital. There is a known job title, CEO, that you can give to yourself and that is understood by others in the industry and in society. There is an objective way to measure success: company profits and market valuation.

The founder career path is to create a successful company. Once again, progress on this path opens up opportunities. The most successful founders have the resources and reputation to launch even more varied and ambitious projects (think Jeff Bezos or Elon Musk). However, a startup failure does not end a career. In Silicon Valley at least, failure is not a black mark, and a failed founder can do another startup, or get a job in engineering, design, sales, or management.

We can think of a career path as a social support structure around a *value*. In science, the value is new knowledge. In entrepreneurship, the value is profitable business. Having a support structure around a value means that if someone is motivated to pursue that value, they can be paid to do so; and if they succeed, they can expect both prestige and expanded career opportunities.

Now, what is the career path for an inventor?

“Inventor” is not a role one can be hired for. The aspiring inventor finds themselves straddling science and business. They could join a research lab, or become an engineer at a technology-based company. In either case, they will be misaligned with their environment. In research, what is valued is new knowledge. An invention that achieves a practical goal is not valued if it demonstrates no new scientific principle. In the corporate environment, what is valued is what drives the business. The engineer may find themselves optimizing and refining existing products, without any mandate to create fundamentally new ones. Neither environment values simply making fundamentally new technologies work. Alternately, an inventor could also be an entrepreneur, starting a company to commercialize the invention. But this requires of the inventor that they have the wherewithal of the startup founder to raise money, hire a team, etc. We ask this of founders because it’s in the nature of the job: someone who can’t do these things probably wouldn’t succeed at the rest of the founder’s task. But we don’t expect every scientist to found their own research lab, and we shouldn’t expect every inventor to be a founder either.

In the early 20th century there were options for inventors. Some joined the great corporate research labs of the day: General Electric, Westinghouse, Kodak, Dow, DuPont, and of course Bell Labs. Others stayed independent, patented their inventions, and sold or licensed the patents to businesses. This let them make a living by inventing, without being personally responsible for commercializing, scaling, and distributing their inventions (although it required seed funding: many inventors [had second jobs, or got angel investment through personal connections](#)).

For reasons I still don’t fully understand, both options have withered. Corporate research is largely [not as ambitious and long-term as it used to be](#). The lone inventor, too, seems to be a thing of the past.

The bottom line is that if a young person wants to focus their career on invention—as distinct from scientific research, corporate engineering, or entrepreneurship—the support structure doesn’t exist. There isn’t a straightforward way to get started, there isn’t an institution of any kind that will hire you into this role, and there isn’t a community that values what you are focused on and will reward you with prestige and further opportunities based on your success. In short, there is no career path.

Note that funding alone does not create a career path. You could start an “invention lab” and hire people to make inventions. You could even pay, reward and promote them based on their success at this task. But it would be difficult to hire any ambitious academic, or anyone who wanted to climb the corporate ladder, because this role wouldn’t be advancing either career path. That isn’t to say that it would be impossible to hire great talent, but you would be facing certain headwinds.

I think this is why [the NIH receives relatively conventional grant proposals even for their “transformative research awards”](#), and why Donald Braben says that he had to build a high degree of trust with researchers before they would even *tell* him their ambitious research goals (see [Scientific Freedom](#), p. 135). The community that forms

around a career path has its own culture, and that includes an oral tradition of career advice, passed down from senior to junior members of the tribe. What kinds of goals to pursue, what kinds of jobs to take and when, how to choose among competing opportunities—there is folklore to provide guidance on all these questions. A single grant program or call for proposals cannot counter the weight of a culture that communicates: “the reliable way to build a scientific career is by proposing reasonable, incremental research goals that are well within the consensus of the field.”

In part, I see this as both the challenge and the opportunity of efforts like [PARPA](#) or [FROs](#). It’s a challenge because a career path must ultimately be supported by a whole community. But it’s an opportunity because efforts like this could be how we bootstrap one. Funding alone doesn’t create a career path, but it can attract a few talented and ambitious mavericks who value independence and scoff at prestige. Success could bring more funding, and inspire imitators. Enough imitators would create an ecosystem. Enough success would bring prestige to the field.

It won’t be easy, but I am excited by efforts like these. We need a career path for invention.

Thanks to Ben Reinhardt, Matt Leggett, and Phil Mohun for reading a draft of this.

Three reasons to expect long AI timelines

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

This is a linkpost for <https://matthewbarnett.substack.com/p/three-reasons-to-expect-long-ai-timelines>

A lot of people I trust put relatively high confidence on near term AI timelines. For example, most people in [the LessWrong AI timelines thread](#) from last year had shorter timelines than [me](#), though it may have been because I interpreted the question a bit differently than most everyone else.

In this post, I'll cover three big reasons to expect long AI timelines, which I take to be the thesis that [transformative](#) AI-related phenomena won't happen for at least another 50 years (currently >2071). Roughly speaking, the three reasons are

1. **Technological deployment lag:** Most technologies take decades between when they're first developed and when they become widely impactful.
2. **Overestimating the generality of AI technology:** Many AI scientists in the 1950s and 1960s incorrectly expected that cracking computer chess would automatically crack other tasks as well. I suspect a similar phenomenon is happening in people's minds today when they extrapolate current AI.
3. **Regulation will slow things down:** Lots of big technologies have been slowed down by regulation. Nuclear energy is an obvious example of a technology whose adoption has been hampered by regulation, but other cases exist too.

Technological deployment lag

From at least two perspectives, it makes sense to care more about when a technology is impactful, rather than when it first gets developed in a lab.

The first perspective is the perspective of an ordinary person. Ordinary people hardly get affected by isolated technological achievements. It *might* help their stock portfolios, if the resulting development triggers investors to become much more optimistic about future profits in those industries. But other than that, the ordinary person will care far more about when they can actually see the results of a technology compared to when it is first developed.

The second perspective is the perspective of a serious technological forecaster. These people care a lot about timing because the *order* of technological developments matters a lot for policy. To give a simple example, they care a lot about whether cheap and reliable solar energy will be developed before fusion power, because it tells them what type of technology society should invest in to stop climate change.

I care most about the second perspective, though it's worth noting cases where the first perspective might still matter. Consider a scenario in which AI researchers are going around declaring that "AGI is in 10 years." 10 years pass and an AGI developed is indeed developed in a lab somewhere, but with no noticeable impact on everyday life. People may grow distrustful of such proclamations, even if they're ultimately technically proven right.

While it might seem obvious to some that we should make a distinction between when a technology is developed and when it actually starts having a large impact, I'm pointing it out because I mostly *don't* get the impression from most AI forecasting literature that such a distinction is important.

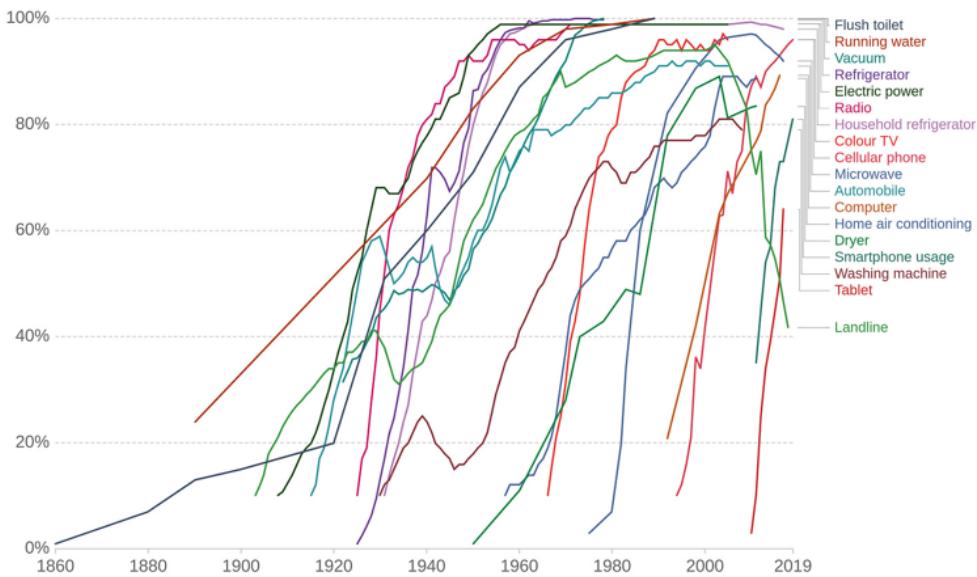
[Nearly all AI timeline surveys](#) and forecasts I've been acquainted with simply take it as a starting assumption that what we care about is when advanced AI is developed, somewhere, rather than some side effect of that development. While I admit that it *might* be more reasonable to care about the specific moment in time when advanced AI is developed in a lab (particularly if we accept some local "foom" AI [takeoff scenarios](#)), it's not at all obvious to me that it is. If you disagree, I would prefer you to at least carefully outline your reasoning before spitting out a date.

One main reason why we might care most about the date of development is because we think that after sufficiently advanced AI is developed, the effects will happen almost instantaneously. The most extreme version of this thesis is the one where AI self-improves upon getting past some critical threshold, and takes over the whole world within a few weeks.

Most technologies, however, don't generally have immediate widespread effects. Our World in Data produced a great chart showing the typical timescales for technological adoption. It's worth checking out [the whole article](#).

Share of US households using specific technologies, 1860 to 2019

Our World
in Data



Source: Comin and Hobijn (2004) and others

Note: See the sources tab for definitions of adoption rates by technology.

OurWorldInData.org/technology-adoption/ * CC BY

We can see from the chart that most technologies take decades between the time when just a few have access to it, and when they're ubiquitous. The chart likely underestimates the true lag between development and impact, however, because we also need to take into account the lag between when technologies are developed and when a non-negligible fraction of the population has access to them. Furthermore, this chart only shows adoption in the United States, a rich nation. Adoption trends worldwide are even slower.

(Consider that, at the time of writing, predictors on Metaculus [expect transformative economic growth](#) to come long after they [expect the first AGI to be developed](#)).

One objection is that we should care about AI's impacts long before it becomes ubiquitous in households—it might be adopted by businesses and governments first.

There are two forms this objection might take. The first form imagines that businesses or governments would be much faster to adopt technologies than households. I am uncertain about the strength of this objection, and I'm not sure what information might be relevant for answering it.

The second form of this objection is that AI might have huge transformative impacts even if only a few businesses or governments adopt it. The classic justification for this thesis is that one or a few AI projects become overwhelmingly powerful in a localized intelligence explosion, rather than having large effects by diffusion. In that case, all [the standard arguments against AI boom](#) are applicable here.

Another objection is that some technologies, especially smartphones, were adopted very rapidly compared to the other ones. AI is conceivably similar in this respect. The rapid adoption of smartphones seems to derive from at least one of two reasons: it could be that smartphones were unusually affordable for households, or that they experienced unusually high demand upon their introduction.

It's not clear to me whether AI technology will be unusually affordable relative to other technologies, and I lean towards doubting it. But it appears probable to me that AI will experience unusually high demand upon its introduction. Overall I'm not sure how to weight this consideration, but it definitely pushes me in the direction of thinking that AI technologies will probably not have a very long adoption timeline (say, more than 30 years after its introduction before it starts having large effects).

Another reason for doubting that AI will have immediate widespread impacts is because previous [general purpose technologies](#) failed to have such impacts too. Economist Robert Solow famously quipped in 1989 that "You can see the computer age everywhere but in the productivity statistics." His observation was later coined the [Productivity Paradox](#).

By the late 1990s, [labor productivity in the United States](#) had finally accelerated, culminating in an economic boom. Economists have provided a few explanations for this lag. For instance, Wikipedia points out that we may have simply mismeasured growth by overestimating inflation. Philippe Aghion and Peter Howitt, however, outline an alternative and common-sense explanation in chapter 9 of *The Economics of Growth*,

As David (1990) and Lipsey and Bekar (1995) have argued, GPTs [general purpose technologies] like the steam engine, the electric dynamo, the laser, and the computer require costly restructuring and adjustment to take place, and there is no reason to expect this process to proceed smoothly over time. Thus, contrary to the predictions of real-business-cycle theory, the initial effect of a “positive technology shock” may not be to raise output, productivity, and employment but to reduce them [...]

An alternative explanation for slowdowns has been developed by Helpman and Trajtenberg (1998a) using the Schumpeterian apparatus where R&D resources can alternatively be used in production. The basic idea of this model is that GPTs do not come ready to use off the shelf. Instead, each GPT requires an entirely new set of intermediate goods before it can be implemented. The discovery and development of these intermediate goods is a costly activity, and the economy must wait until some critical mass of intermediate components has been accumulated before it is profitable for firms to switch from the previous GPT. During the period between the discovery of a new GPT and its ultimate implementation, national income will fall as resources are taken out of production and put into R&D activities aimed at the discovery of new intermediate input components.

In line with these expectations, Daniel Kokotajlo [pointed](#) to [this paper](#) which complements this analysis by applying it to the current machine learning era.

Overestimating the generality of AI technology

Many very smart AI scientists in the 1950s and 1960s had [once believed](#) that human-level AI was imminent. As many later pointed out, these failed predictions by themselves provide evidence that AI will take longer to develop than we think. Yet, that's not the only reason why I'm bringing them up.

Instead, I want to focus on *why* AI scientists once believed that developing human-level AI would be relatively easy. The main reason, I suspect, is that researchers were too optimistic about the generality of their techniques. The case of computer chess is illustrative here. In [his 1950 paper](#) in which he provided an algorithm for perfect chess play, Claude Shannon wrote,

This paper is concerned with the problem of constructing a computing routine or "program" for a modern general purpose computer which will enable it to play chess. Although perhaps of no practical importance, the question is of theoretical interest, and it is hoped that a satisfactory solution of this problem will act as a wedge in attacking other problems of a similar nature and of greater significance.

Among the problems that Shannon had hoped would be attacked indirectly by solving chess, he listed,

Machines capable of translating from one language to another.

Machines capable of orchestrating a melody.

We now know that these problems are at least, for all practical purposes, only incidentally related to the problem of playing chess. At most, these problems are downright irrelevant. Few AI researchers would make the same mistake today.

Yet, I see elements of Shannon's mistake in the reasoning of many I see today. I'll walk through my reasons.

First, consider why Shannon might have expected progress in chess to aid progress in language translation. We could imagine, in some abstract sense, that chess and language translation are both the same type of problems. Mathematically speaking, a chess engine is simply a mapping between chess board states and moves. Similarly, language translation is simply a mapping between sentences in one language, and sentences in another language.

Beyond the simple mathematical formalism, however, there are substantial real differences between the two tasks. While computer chess can feasibly be solved by brute force, language translation requires an extremely nuanced understanding of the rules native speakers use to compose their speech.

One reason why Shannon might not have given this argument much thought is because he wasn't thinking about how to do language translation in the moment; he was more interested in solving chess, and the other problems were afterthoughts.

We can view his stance from the analogy to [construal level theory](#), or as Lesswrong likes to put it, [near vs. far thinking](#). All of the concrete ways that chess could be tackled were readily apparent in Claude Shannon's mind, but the same could not be said about natural language translation. Rather than viewing a *specific* similarity between the two tasks, he could have made the forgivable mistake of assuming that a *vague* similarity between them was sufficient for his prediction.

It's a bit like the [planning fallacy](#). When planning our time, we can see all the ways things could go right and according to schedule, since those things are concrete. The ways that things could go wrong are more abstract, and thus occupy less space in our thinking. We mistake this perception for the likelihood of things going right.

Now let's compare this case to an argument I hear quite a lot these days. Consider the quite reasonable suggestion that GPT-3 is a rudimentary form of general intelligence. Given that it can write on a wide variety of topics, it certainly appears generally capable. Now consider one further assumption: [the scaling hypothesis](#). We conclude that some descendant of GPT-3, given thousands or millions of times more computation, will naturally yield general AI.

I see no strong reason to doubt the narrow version of this thesis. I believe it's likely that, as training scales, we'll progressively see more general and more capable machine learning models that can do a ton of impressive things, both on the stuff we expect them to do well on, and some stuff we didn't expect.

But no matter how hard I try, I don't see any current way of making some descendant of GPT-3, for instance, manage a corporation.

One may reason that, as machine learning models scale and become more general, at some point this will just naturally yield the management skills required to run a company.

It's important to note that even if this were true, it wouldn't tell us much about how to extract those skills from the model. Indeed, GPT-3 may currently be skilled at many things that we nonetheless do not know how to make it actually perform.

Most importantly, notice the similarities between this reasoning and that of (my interpretation) of Claude Shannon's. Shannon expected algorithmic progress in chess to transfer usefully to other domains. In my interpretation, he did this because the problems of chess were near to him, and the problems of language translation were far from him.

Similarly, the problem "write a well-written essay" is close to us. We can see concretely how to get a model to perform better at it, and we are much impressed by what we obtain by making progress. "Manage a corporation" is far. We're not really sure how to approach it, even if we could point out vague similarities between the two problems if we tried.

I don't mean to imply that we haven't made progress on the task of getting an AI to manage a corporation. I only mean that you can't just wish it away as a hard problem simply by imagining that we'll just get it for free as a result of making steady progress on something simpler and more concrete.

What other tasks do I think people might be incorrectly assuming we could as a byproduct of progress on simpler things? Here's a partial list,

- As already stated, managing organizations and people.
- Complex general purpose robotics, of the type needed to win [the RoboCup grand challenge](#).
- Long-term planning and execution, especially involving fine motor control and no guarantees about how the environment will be structured.
- Making original and profound scientific discoveries.

I won't claim that an AI can't be dangerous to people if it lacks these abilities. However, I do think that in order to pose an existential risk to humanity, or obtain a [decisive strategic advantage](#) over humans, AI systems would likely need to be capable enough to do at least one of these things.

Regulation will slow things down

Recently, Jason Crawford [wrote on Roots of Progress](#),

In the 1950s, nuclear was the energy of the future. Two generations later, it provides [only about 10% of world electricity](#), and reactor design hasn't fundamentally changed in decades.

As Crawford explains, the reason for this slow adoption is neither because nuclear plants are unsafe or because they can't be built cheaply. Rather, burdensome regulation has raised production costs to a level

where people would rather pay for other energy sources,

Excessive concern about low levels of radiation led to a regulatory standard known as ALARA: As Low As Reasonably Achievable. What defines “reasonable”? It is an ever-tightening standard. As long as the costs of nuclear plant construction and operation are in the ballpark of other modes of power, then they are reasonable.

This might seem like a sensible approach, until you realize that **it eliminates, by definition, any chance for nuclear power to be cheaper than its competition**. Nuclear can't even innovate its way out of this predicament: under ALARA, any technology, any operational improvement, anything that reduces costs, simply gives the regulator more room and more excuse to push for more stringent safety requirements, until the cost once again rises to make nuclear just a bit more expensive than everything else. Actually, it's worse than that: it essentially says that if nuclear becomes cheap, then the regulators *have not done their job*.

Crawford lays blame on the incentives of regulators. As he put it,

[The regulators] get no credit for approving new plants. But they do own any problems. For the regulator, there's no upside, only downside. No wonder they delay.

In fact, these perverse incentives facing regulators have long been known by economists who favor deregulation. Writing in 1980, Milton and Rose Friedman [gave](#) the following argument in the context of the FDA regulation,

It is no accident that the FDA, despite the best of intentions, operates to discourage the development and prevent the marketing of new and potentially useful drugs. Put yourself in the position of an FDA official charged with approving or disapproving a new drug. You can make two very different mistakes:

1. Approve a drug that turns out to have unanticipated side effects resulting in the death or serious impairment of a sizable number of persons.
2. Refuse approval of a drug that is capable of saving many lives or relieving great distress and that has no untoward side effects.

If you make the first mistake—approve a [thalidomide](#)—your name will be spread over the front page of every newspaper. You will be in deep disgrace. If you make the second mistake, who will know it?

Given the moral case here, it might come as a surprise that the effect of regulation on technological innovation has not generally been well studied. Philippe Aghion et al. recently published [a paper](#) saying as much in their introduction. Still, although we lack a large literature to show the role regulation plays to delay technological development, it almost certainly does.

Regulation is arguably the main thing standing in the way of lots of futuristic technologies: human cloning, human genetic engineering, and climate engineering come to mind, just to name a few.

One might think that the AI industry is immune to such regulation, or nearly so. After all, the tech industry has historically experienced a lot of growth without much government interference. What reason is there for this to stop?

I offer two replies. The first reason is that governments of the world are *already* on the cusp of a concerted effort to regulate technology companies. A New York Times article from April 20th [explains](#),

[China fined the internet giant Alibaba](#) a record \$2.8 billion this month for anticompetitive practices, ordered [an overhaul of its sister financial company](#) and warned other technology firms to obey Beijing's rules.

Now the European Commission plans to unveil far-reaching regulations to limit technologies powered by artificial intelligence.

And in the United States, President Biden has [stacked his administration with trustbusters](#) who have taken aim at [Amazon](#), Facebook and Google.

Around the world, governments are moving simultaneously to limit the power of tech companies with an urgency and breadth that no single industry had experienced before. Their motivation varies. In the United States and Europe, it is concern that tech companies are stifling competition, spreading misinformation and eroding privacy; in Russia and elsewhere, it is to silence protest movements and tighten political control; in China, it is some of both.

The second reason is that, as AI becomes more capable, we'll likely increasingly see calls for it to be regulated. I should point out that I'm not restricting my analysis to government regulation; the very fact that the AI safety community exists, and that OpenAI and Deepmind hired people to work on safety, provides evidence that such calls for more caution will occur.

The slightest sign of danger was enough to stall nuclear energy development. I don't see much reason to expect any different for AI.

Furthermore, many others and I, have [previously pointed out](#) that in a continuous AI takeoff scenario, low-magnitude AI failures will happen before large-magnitude failures. It seems plausible to me that at some point, a significant AI failure will happen that triggers a national or even international panic, despite not posing any sort of imminent existential risk. In other words, I pretty much expect a [Chernobyl disaster](#) of AI—or at least, I expect a series of such disasters to happen that will have more or less the same effect.

Combining all three of these effects, it's a bit difficult to see how we will get [transformative AI](#) developments in the next 50 years [edit: but not very difficult, see [my update here](#)]. Even accepting some of the more optimistic assumptions in e.g. Ajeya Cotra's [Draft report on AI timelines](#), it still seems to me that these effects will add a few decades to our timelines before things get really interesting. So at present, my optimistic timelines look more like 25 or 30 years, rather than 10 or 15. But of course, smart people disagree with me here, there's a ton of uncertainty, so I'm happy to find where I made mistakes.

Center for Applied Postrationality: An Update

Previous post: <https://www.lesswrong.com/posts/7X3BWzAt62yvvtyQX/announcing-the-center-for-applied-postrationality>

Content Notice: Mild to Moderate NSFW language, Defective Altruism, Infoblessings

A lot of progress has been made here at CFAP in the past two years.

After posting on the SL5 newsgroup, a shadowy cryptobillionaire reached out. They not just met but drastically exceeded our funding goal. We are in the process of minting a commemorative impact certificate NFT to thank them for the monumental longterm value this will undoubtedly bring to the world.

Unfortunately, after reading a professional investing subreddit, we currently have much of our funds tied up in a superrational coordination experiment that is going very well. Therefore, we will need additional money if we're to reach our stretch goal of purchasing a yacht to live out of. We need the money in order to party all of the time on a huge boat, and you need to feel a sense of meaning in your life.

As evidence of our fiduciary effectiveness, here is a list of some of the most exciting projects we have initiated.

- A large shift in our research direction took place. We noticed we were hitting diminishing marginal returns perfecting the art of mental masturbation. Many of our facilitators started branching out into areas of embodiment and emotion. This has led to developing a new and innovative practice in the art of embodied relationality, called Circle Jerking.
- After a whole bunch of Circle Jerking, some teachers left to start a new monastery called WOOD (Wisdom, Onanism, Originality, and Disestablishmentarianism). Their theory of change is that the problems of the world are due to humans thinking, and therefore they will meditate in order to stop thinking. Here at CFAP, we support any and all efforts to not think.
- We were surprised to learn that the General Semantics movement had covered so much previous intellectual ground. We bought a copy of Science and Sanitation and are in the process of reading it whenever we're going #2. We had to tear many of the pages out for toilet paper in March 2020, sadly. Not sure why but there was a shortage.
- Getting swole. While practicing Original Seeing, we noticed many attendees are pretty scrawny and low-T. We conducted an extensive survey to ask them if they even lifted, and were shocked to learn how many didn't. So we started buying gym memberships en masse to turn code monkeys into brogrammers. We've partnered with KillMinder, a productivity program in alpha, to install irremovable bracelets on half of enrollees, which will discharge a lethal dose of electricity in the event of failing to show up at the gym one too many times (or unanticipated technical errors). Results on attendance rate are encouraging! 🤙
- We've strategically purchased a headquarters in Las Vegas, Nevada. We're excited for the opportunity to lead participants through Aura Zone Expansion exercises in this environment, as well as lessons in applied postrational economics. It is particularly easy to purchase souls in Sin City; asking people on

the street to buy their soul is a research-backed social skills exercise that is particularly highly-rated by students as an experience that boosted their social confidence. Training in the skill of gambling while under the influence of psychedelics will also help participants mystically transcend the Kelly Criterion, and see that they and the House have all won.

- In partnership with TERRITORIES, and pending the approval of the Fiasco and Delay Administration, we are going to start giving people to jungle-grown plants and see whether it leads to increases in creativity, openness, mood, intelligence, and prosociality in the plants. We're especially interested in promising research showing the outsized effects of human consumption on the reduction of trauma in herbs. 
- We've made a grant to Rubbin Handsome to probe deeper into the subject matter of aliens. We think aliens are one of the most important questions we don't have a good answer for, and it may force us to form new epistemology to explain it. To be honest, I don't even know what epistemology is, nor can I explain why the word "epistemology" shows up in random locations without explanation. 
- We've started a dominant assurance contract to ban mathematics everywhere. It will only go into effect when the number of signers reaches [censored]. If it fails by the year [censored], you will get \$[censored] for signing.
- 17000 copies of Meltdown were distributed to gifted primary school children. We also hosted a student drawing contest for depictions of Jeff's Friendly Snek. 
- Our research into Prophecy will bear fruit.

There's a lot more that happened in our workshops and throughout the broader postrationality movement that I couldn't describe. Share your experiences below!

Alignment Newsletter Three Year Retrospective

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

It has now been just shy of three years since the first Alignment Newsletter was published. I figure it's time for an update to the [one-year retrospective](#), and another very short [survey](#). **Please take the survey!** The mandatory questions take **just 2 minutes!**

This retrospective is a lot less interesting than the last one, because not that much has changed. You can tell because I don't have a summary or key takeaways, and instead I'm going to launch into nitty gritty details.

Newsletter stats

We now have 2443 subscribers, and tend to get around a 39% open rate and 4% click through rate on average (the click rate has higher variance though). In the one-year retrospective, I said 889 subscribers, just over 50% open rate, and 10-15% click through rate. This is all driven by organic growth; there hasn't been any push for publicity.

I'm not too worried about the decreases in open rate and click rate:

1. I expect natural attrition over time as people's interests change. Many of these people probably just stop opening emails, or filter them, or open and immediately close the emails.
2. In absolute terms, the number of opens has gone way up (~450 to ~950).
3. My summaries have gotten more pedagogic (see below), so people might feel less need to click through to the original.
4. I now summarize fewer items, so there are fewer chances to "catch people's interests".
5. We haven't done any publicity, which I would guess is a common way to boost open rates (since newer subscribers are probably more likely to open emails?)

There was this weird thing where at the beginning of the pandemic, open rates would alternate between < 20% and > 40%, but would never be in between. I have no idea what was going on there.

I was also a bit confused why we're only at #145 instead of #157, given that this is a weekly publication -- I knew I had skipped a couple of weeks but twelve seemed like too many. It turns out this newsletter was published every fortnight during the summer of 2019. I had no memory of this but it looks like I did take steps to fix it -- in the [call for contributors](#), I said:

I'm not currently able to get a (normal length) newsletter out every week; you'd likely be causally responsible for getting back to weekly newsletters.

(This was probably true, since I did get back to weekly newsletters after getting new contributors!)

Changes

My overall sense is that the newsletter has been pretty stable and on some absolute scale has not changed much since the last retrospective two years ago.

Pedagogy

There are roughly two kinds of summaries:

1. **Advertisements:** These summaries state what the problem is and what the results are, without really explaining what the authors did to get those results. The primary purpose of these is to inform readers whether or not they should read the full paper.
2. **Explanations:** These summaries also explain the “key insights” within the article that allow them to get their results. The primary purpose is to allow readers to gain the insights of the article without having to read the article; as such there is more of a focus on pedagogy (explaining jargon, giving examples, etc.)

Over time I believe I've moved towards fewer advertisements and more explanations. Thus, the average length of a summary has probably gotten longer. (However, there are probably fewer summaries, so the total newsletter length is probably similar.)

Long-form content. Some topics are sufficiently detailed and important that I dedicate a full newsletter to them (e.g. [Cartesian frames](#), [bio anchors](#), [safety by default](#), [assistance games](#)). This is basically the extreme version of an explanation. I've also done a lot more of these over time.

More selection, less overview

Two years ago, I worried that there would be too much content to summarize. Yet, somehow my summaries have become *longer*, not shorter. What gives?

Basically, I've become more opinionated about what is and isn't important for AI alignment researchers to know, and I've been more selective about which papers to summarize as a result. This effectively means that I'm selecting articles in part based on how much they agree with my understanding of AI alignment.

As a result, despite the general increase in alignment-related content, I now summarize *fewer* articles per newsletter than I did two years ago. The articles I do summarize are selected for being interesting in my view of AI alignment. Other researchers would likely pick quite a different set, especially when choosing what academic articles to include.

I think this is mostly because my views about alignment stabilized shortly after the one year retrospective. At that point, I had been working in AI safety for 1.5 years, and I probably still felt like everything was confusing and that my views were changing wildly every couple of months. Now though it feels like I have a relatively firm framework, where I'm investigating details within the framework. For example, I still feel pretty good about the things I said in this [conversation](#) from August 2019, though I might frame them differently now, and could probably give better arguments for

them. In contrast, if you'd had a similar conversation with me in August 2018, I doubt I would have endorsed it in August 2019.

This does mean that if you want to have an overview of what the field of AI alignment is up to, the newsletter is not as good a source as it used to be. (I still think it's pretty good even for that purpose, though.)

Team

Georg Arndt (FHI) and Sawyer Bernath (BERI) are helping with the publishing and organization of the newsletter, freeing me to work primarily on content creation. After a [call for contributors](#), I took on six additional contributors; this gives a total of 9 people (not including me) who could in theory contribute to the newsletter. However, at this point over half don't write summaries any more, and the remainder write them pretty occasionally, so I'm still writing most of the content. I think this is fine, given the shift towards being a newsletter about my views and the decrease in amount of content covered.

To be clear, I think the additional contributors worked out great, and had the effect I was hoping for it to have. We got back to a weekly newsletter schedule, I put in less time into the newsletter, it was even easier to train the contributors than I thought, and most new contributors wrote a fair number of good summaries before effectively leaving. I expected that to continue this long term I'd have to periodically find new contributors; I think this should be seen as a decision not to continue the program despite its success because I ended up evolving towards a different style of newsletter.

(I'm still pretty happy to have additional contributors, as long as they can commit to ~20 summaries upfront. If you'd be interested, you can send me an email at rohinmshah@gmail.com.)

Appearance

In March 2020, the newsletter got an updated design, that made it look much less like a giant wall of text.

Impact

I was pretty uncertain about the impact of the newsletter in the [last retrospective](#). That hasn't changed. I still endorse the discussion in that section.

Advice for readers

Since I'm making this "meta" post anyway, I figured I might as well take some time to tell readers how I think they should interact with the newsletter.

Don't treat it as an evaluation of people's work. As I mentioned above, I'm selecting articles based in part on how well they fit into my understanding of AI alignment. This is a poor method for evaluating other people's work. Even if you defer to me completely and ignore everyone else's views, it still would not be a good method, because often I am mistaken about how important the work is even on my

own understanding of AI alignment. Almost always, my opinion about a paper I feel meh about will go up after talking to the authors about the work.

I also select articles based on how useful I think it would be for other AI alignment researchers to learn about the ideas presented. (This is especially true for the choice of what to highlight.) This can be very different from how useful the ideas are to the world (which is what I'd want out of an evaluation): incremental progress on some known subproblem like learning from human feedback could be very important, but still not worth telling other AI alignment researchers about.

Consider reading just the highlights section. If you're very busy, or you find yourself just not reading the newsletter each week because it's too long, I recommend just reading the highlights section. I select pretty strongly for "does this seem good for researchers to know?" when choosing the highlight(s).

If you're busy, consider using the [spreadsheet database](#) as your primary mode of interaction. Specifically, rather than reading the newsletter each week, you could instead keep the database open, and whenever you see a vaguely interesting new paper, you can check (via Ctrl+F) whether it has already been summarized, and if so you can read that summary. (Even I use the database in this way, though I usually know whether or not I've already summarized the paper before, rather than having to check.)

Also, there may be a nicer UI to interact with this database in the near future :)

Survey

[Take it!](#)

Scott Alexander 2021 Predictions: Market Prices

[Scott Alexander has posted some predictions for 2021](#). Taking Zvi's approach from [here](#) and rather than making my own adjustments, making my best estimate of what various prediction (and financial) markets are saying the odds are.

If anyone has seen a market for any of the questions listed here which I haven't found one, let me know and I'll add it.

EDIT: Final results [here](#)

US/WORLD

1. Biden approval rating (as per 538) is greater than 50%: 80%

[Metaculus gives this 61%](#). (Much lower than both Zvi and Scott)

2. Court packing is clearly going to happen (new justices don't have to be appointed by end of year): 5%

[PredictIt gives a lower bound on this at 5%](#).

[Metaculus gives a 27%](#) chance to court packing by 2030. Assuming the chances of this happening is mostly weighted sooner rather than later (4/3/2/1 over Biden's term + 30% '24 - '30). would give this a 4% chance in this year

3. Yang is New York mayor: 80%

[Metaculus is at 60%](#)

[Smarkets is at 73%](#)

[Betfair is at 77%](#)

[PredictIt is at 69%](#)

4. Newsom recalled as CA governor: 5%

[BetOnline.ag](#) has this at ~7%

5. At least \$250 million in damage from BLM protests this year: 30%

6. Significant capital gains tax hike (above 30% for highest bracket): 20%

7. Trump is allowed back on Twitter: 20%

Not aware of any markets

8. Tokyo Olympics happen on schedule: 70%

[Metaculus has this at 80%](#)

[FTX has this at 75%](#)

[Smarkets has this at 75%](#)

Note that Metaculus is talking about an Olympics at any point in 2021, whereas the others are about the Olympics being on schedule.

9. Major flare-up (significantly worse than anything in past 5 years) in Russia/Ukraine war: 20%

[Metaculus has a related question at 32%](#). If we assume that question is about reaching a level approximately the same as was achieved in the past, this is presumably putting the chances of something worse at ~16%

10. Major flare-up (significantly worse than anything in past 10 years) in Israel/Palestine conflict: 5%

11. Major flare-up (significantly worse than anything in past 50 years) in China/Taiwan conflict: 5%

Not aware of any markets (specific to 2021).

12. Netanyahu is still Israeli PM: 40%

[PredictIt has this at 22%](#)

13. Prospera has at least 1000 residents: 30%

Not aware of one. I think Metaculus are making one though

14. GME >\$100 (Currently \$170): 50%

Possibly the most concrete one. The [options market](#) is giving this 60% chance right now. (Actually, it's giving that until 21-Jan-22, so the chances are even higher than that)

15. Bitcoin above 100K: 40%

[Metaculus gives this 43%](#) (at any point)

[Deribit options give this 25%](#)(at the end of the year)

16. Ethereum above 5K: 50%

[Deribit options give this 11%](#) (at the end of the year)

17. Ethereum above 0.05 BTC: 70%

Not a market price, but bootstrapping this from option prices + historical volatility puts this closer to 33%

18. Dow above 35K: 90%

[Option market gives this 50%](#)

19. ...above 37.5K: 70%

[Option market gives this 20%](#)

20. Unemployment above 5%: 40%

[Metaculus gives this 37%](#)

21. Google widely allows remote work, no questions asked: 20%

Not seen a market for this

22. Starship reaches orbit: 60%

Metaculus gives this 50%

COVID

23. Fewer than 10K daily average official COVID cases in US in December 2021: 30%

24. Fewer than 50K daily average COVID cases worldwide in December 2021: 1%

Not seen a market

25. Greater than 66% of US population vaccinated against COVID: 50%

Metaculus gives this 77% (their line is 69% vaccinated, so their probability for 60% is even higher)

26. India's official case count is higher than US: 50%

Not seen a market

27. Vitamin D is generally recognized (eg NICE, UpToDate) as effective COVID treatment: 30%

None of Metaculus' 4 questions about Vit-D are <= 25%:

- Best practice: 15%
- NHS: 25%
- NIH: 21%
- Dutch: 24%

28. Something else not currently used becomes first-line treatment for COVID: 40%

29. Some new variant not currently known is greater than 25% of cases: 50%

30. Some new variant where no existing vaccine is more than 50% effective: 40%

Not seen a market

31. US approves AstraZeneca vaccine: 20%

Metaculus gives this 37%

32. Most people I see in the local grocery store aren't wearing a mask: 60%

Not seen a market

Vim

Vim is a terminal-based editor optimized for speed. Keys are hotkeys by default. Vim does not even use `Ctrl` for most of them. Maneuvering via hotkeys is so efficient mouse input is often disabled by default.

This guide focuses on the features I use most frequently. It has nothing about with native Vim windows because I use i3 or Spacemacs instead. If you feel I left out a valuable feature then please let me know about it in the comments.

Movement

The simplest way to move is via the `hjkl` keys.

key	action
<code>h</code>	moves the cursor one character left
<code>j</code>	moves the cursor one line down
<code>k</code>	moves the cursor one line up
<code>l</code>	moves the cursor one character right

The `hl` keys operate within a single line. Pressing `h` at the start of a line does nothing. Pressing `l` at the end of a line does nothing. By **line** I mean a string of characters ending with a carriage return. The fact that a long paragraph appears to wrap around on your screen is irrelevant.

The `jk` keys move the cursor down and up a single line. If the cursor is in the 3rd column of the 6th line then pressing `j` moves the cursor to the 3th column of the 7th line.

Most commands can be prefixed with a number. `4h` moves the cursor left 4 characters. `101j` moves the cursor down 101 lines.

A sequence of nonempty lines is called a **paragraph**. The `{}` keys move the cursor forward and back by paragraphs.

key	action
<code>}</code>	moves the cursor forward one paragraph
<code>{</code>	moves the cursor backward one paragraph

There are two kinds of "words" in Vim. A **word** is a sequence of alphanumeric characters. A **Word** is a sequence of non-whitespace characters. The `wbe` keys move the cursor around words. The `WBE` keys move the cursor around Words.

key	action
<code>w</code>	moves the cursor forward to the next start of a word
<code>e</code>	moves the cursor forward to the next end of a word
<code>b</code>	moves the cursor backward to the next start of a word
<code>W</code>	moves the cursor forward to the next start of a Word

key	action
E	moves the cursor forward to the next end of a Word
B	moves the cursor backward to the next start of a Word

The #* keys move the cursor forward and backward to the next "idenfitier". If the cursor hovers over the word "marmot" then pressing * will move the cursor forward to the next instance of the word "marmot" and pressing # will move the cursor backward to the previous instance of the word "marmot".

key	action
#	moves the cursor backward to the previous identifier
*	moves the cursor forward to the previous identifier

The f key takes a character as an argument and moves the cursor forward to that character. For example, fa moves the cursor forward to the next instance of a. The f key operates within a line. It will never take the cursor to another line. The F command is like f except it searches backward. The t and T commands are like f and F except they move the cursor one fewer character.

The ; command repeats the most recent fFtT command. The , command is like ; except backwards. If you use an F command followed by a , then the double-backwardsness will cancel itself out.

key	action
f<char>	moves forward to <char>
F<char>	moves backward to <char>
t<char>	moves forward to the character before <char>
T<char>	moves backward to the character after <char>
;	repeats the previous fFtT command
,	repeats the previous fFtT command except reversed

The 0^\$ keys move the cursor to the beginning and end of a line.

key	action
0	moves the cursor to the beginning of the line
^	moves the cursor to the beginning of the line, excluding whitespace
+	moves the cursor to the beginning of the next line, excluding whitespace
\$	moves the cursor to the end of the line

The HLM keys move the cursor around relative to the viewing window itself.

key	action
H	moves the cursor to the start of the first visible line
M	moves the cursor to the start of the middle visible line
L	moves the cursor to the start of the last visible line

You can move the viewing window itself with z and Ctrl keys.

key	action
------------	---------------

key	action
Ctrl-e	moves the viewing window down one visual line
Ctrl-y	moves the viewing window up one visual line
Ctrl-f	moves the viewing window forward by one viewing window
Ctrl-b	moves the viewing window backward by one viewing window
Ctrl-d	moves the viewing window forward by half of one viewing window
Ctrl-u	moves the viewing window backward by half of one viewing window
zz	moves the viewing window to position the cursor in the middle
zt	moves the viewing window to position the cursor at the top
zb	moves the viewing window to position the cursor at the bottom

You can move the cursor itself up and down visible lines with the g prefix.

key	action
gj	moves the cursor down one visual line
gk	moves the cursor up one visual line

The g key can also go to a particular line.

key	action
gg	jumps the cursor to the beginning of the buffer ^[1]
G	jumps the cursor to the end of the buffer
<num>gg	jumps the cursor to the specified line
<num>G	"

The | key is like g except for columns.

key	action
<num>	moves the cursor to the <num>th column within a line

The / key performs search. You type / then the string you are searching for and then press Enter. The ?nN keys are analogous to F;n.

key	action
/	initiates a search
?	initiates a backwards search
n	repeats the previous search
N	repeats the previous search except reversed

The () characters move forward and backward one **sentence**.

key	action
)	moves forward to the next start of a sentence
(moves backward to the next start of a sentence

The [] characters move the cursor forward and backward to a variety of things. My favorite uses of [] is to go forward and backward to matching parenthesis. For example,]) moves forward to the next unmatched closing parenthesis. It is

indispensable when writing Lisp. Inner quotes and tags behave similarly to parentheticals.

key	action
])	move forward to matching parenthesis
] ("
[)	move backward to matching parenthesis
[("
[[jump to function start
]]	jump to function end

You can (invisibly) mark places in your buffer to return to later. You can set one mark per letter. Lowercase letters a-z are buffer-specific e.g. each buffer can have its own "a" mark. Uppercase letters A-Z are global e.g. only one buffer at a time can have an "A" mark.

key	action
m <a-z>	create a buffer-specific mark
m <A-Z>	create a global mark
' <mark>	jump to the beginning of the line with mark <mark>
` <mark>	jump to mark <mark>

Some marks are populated automatically. Of the automatic marks, I only use ''. You can find the others [here](#).

mark	meaning
"	jump to last jump point

Editing

Now that you know how to move, we can edit some text. Most of the time you edit text you should use the dyp keys.

- d stands for **delete**, which is similar to "cut".
- y stands for **yank**, which is similar to "copy".
- p stands for **put**, which is similar to "paste".

The d key operates on text objects and movement commands. For example dtA deletes everything up to (but not including) the next "A" on the current line.

Every delete operation operates on characters or lines. dtA operates on characters and thus behaves similarly to other editors. A delete operation which operates on lines deletes an integer number of lines. Delete operations which operate on lines delete the current line and the destination line. Thus, dj deletes the current line and the next line.

y is identical to d except it leaves your buffer unchanged.

delete or yank key	action
-------------------------------	---------------

delete or yank key	action
d<char>	deletes every character from the starting character up to (and including) the destination character
d<line>	deletes every line from the starting line up to (and including) the destination line
dd	deletes the current line
y<char>	yanks every character from the starting character up to (and including) the destination character
y<line>	yanks every line from the starting line up to (and including) the destination line
yy	yanks the current line

p puts the selection before the cursor. P puts the selection after the cursor. "Before" and "after" refer to characters or lines depending on whether you deleted/yanked characters or lines.

key	action
p	put after the cursor
P	put before the cursor

By default, p and P will put the last text you deleted or yanked. You can save text into registers by prefixing "<register>" before your delete or yank command. For example, "add deletes the current line and saves it into register "a".

There is one register for each letter a-z. If you use a capital letter then your delete or yank command will append to the register instead of overwriting it. You can put from a register by prefixing your p or P command with the register. For example, "ap" will put from register "a".

You can find a list of special registers [here](#).

Text Objects

Remember how I said you could follow d and y with a movement command or a text object? Remember from earlier the concept of "words" and "paragraphs"? Here we put them together into a simple grammar.

da deletes a text object.

key	action
daw	delete a word
daW	delete a Word
das	delete a sentence
dap	delete a paragraph
da(delete a parenthetical
da)	"
yaw	yank a word
yaW	yank a Word

key	action
yas	yank a sentence
yap	yank a paragraph
ya(yank a parenthetical
ya)	"

You can replace the a with an i to delete an inner word, an inner paragraph, etcetera. An inner text object is just like the regular text object except it does not include any surrounding delimiters. In the case of a parenthetical, the "inner parenthetical" does not include the outside parenthesis. For words/sentences/paragraphs, the delimiter is whitespace.

key	action
diw	delete inner word
diW	delete inner Word
dis	delete inner sentence
dip	delete inner paragraph
di(delete inner parenthetical
di)	"
yiw	yank inner word
yiW	yank inner Word
yis	yank inner sentence
yip	yank inner paragraph
yi(yank inner parenthetical
yi)	"

Other text objects include ""[]{}<>. They all do what you would expect them to do. t refers to an HTML-like tag.

Insert Mode

You cannot write everything by yanking and putting pieces of existing text together. Sometimes you have to insert text into a document. Several different keys drop you into insert mode.

key	action
i	inserts text before the cursor
a	inserts text after the cursor
I	inserts text at the beginning of the line
A	inserts text at the end of the line
o	creates a newline below the current line and inserts text there
O	creates a newline above the current line and inserts text there
s	deletes the current character and inserts text
S	deletes the current line and inserts text

You can exit insert mode by pressing `Escape` but it is faster to remap your `CapsLock` key to `Ctrl` and then exit insert mode with `Ctrl-[`. Some hotkeys are valid only while in insert mode.

insert mode key	action
<code>Escape</code>	exits insert mode
<code>Ctrl-[</code>	"
<code>Ctrl-F</code>	moves cursor forward one character
<code>Ctrl-b</code>	moves cursor backward one character
<code>Ctrl-w</code>	deletes one word backward (without saving it to a register)

The `c` key is like `d` except it drops you into insert mode afterward.

key	action
<code>c<movement></code>	clear text
<code>c<text_object></code>	clear text

After you exit insert mode, the whole insertion counts as a single edit. So if you type `ci(` followed by an insertion then the entire replacement of text inside the parenthetical counts as a single edit. You can repeat an edit with the `.` operator.

key	action
<code>.</code>	repeat previous edit

The `.` key might be the most important key in all of Vim. Generally-speaking, the more you use the `.` operator the better you are at Vim.

Miscellaneous edit commands

key	action
<code>x</code>	delete the character under the cursor
<code>X</code>	delete the character before the cursor
<code>~</code>	toggle capitalization and move the cursor forward one character
<code>r<char></code>	replace a single character

The Undo Tree

`u` and `Ctrl-r` operate like the undo and redo stack you are familiar with.

key	action
<code>u</code>	undo last edit
<code>Ctrl-r</code>	redo next edit

If you undo a long series of edits and then mistakenly make an edit you can undo the damage with `g+` and `g-` which traverse the nodes of the tree in chronological and reverse-chronological order.

key	action
g+	traverse undo tree in chronological order
g-	traverse undo tree in reverse chronological order

Macros

A macro is a series of keystrokes. Macros are good for automating repetitive tasks, especially editing structured text.

To define a macro, start by pressing the q key. Then pick a letter a-z at which to save the macro. Then execute the macro manually. When you are done typing the macro, press q again.

Once you have a macro defined, you can press @ followed by the macro's letter to execute the macro.

key	action
q	define a macro
@<a-z>	execute a macro
@@	execute previous macro

Macros are well-behaved. If a macro modifies a line and then goes down one line and you tell Vim to execute the macro 1000 times but your buffer only has 700 lines then Vim will stop when it gets to the end of your buffer.

Visual mode

Visual modes are similar to highlighting. Visual modes have their uses, but it is usually faster to avoid them.

key	action
v	enter visual mode
V	enter visual line mode
Ctrl-v	enter visual rectangle mode

Find and Replace

You can find and replace all the instances of a string by typing :%s/ followed by the original string followed by / followed by the replacement string followed by Enter... assuming you did not type any escape sequences. Find and replace is a complex subject I will not delve into here even though I do use it.

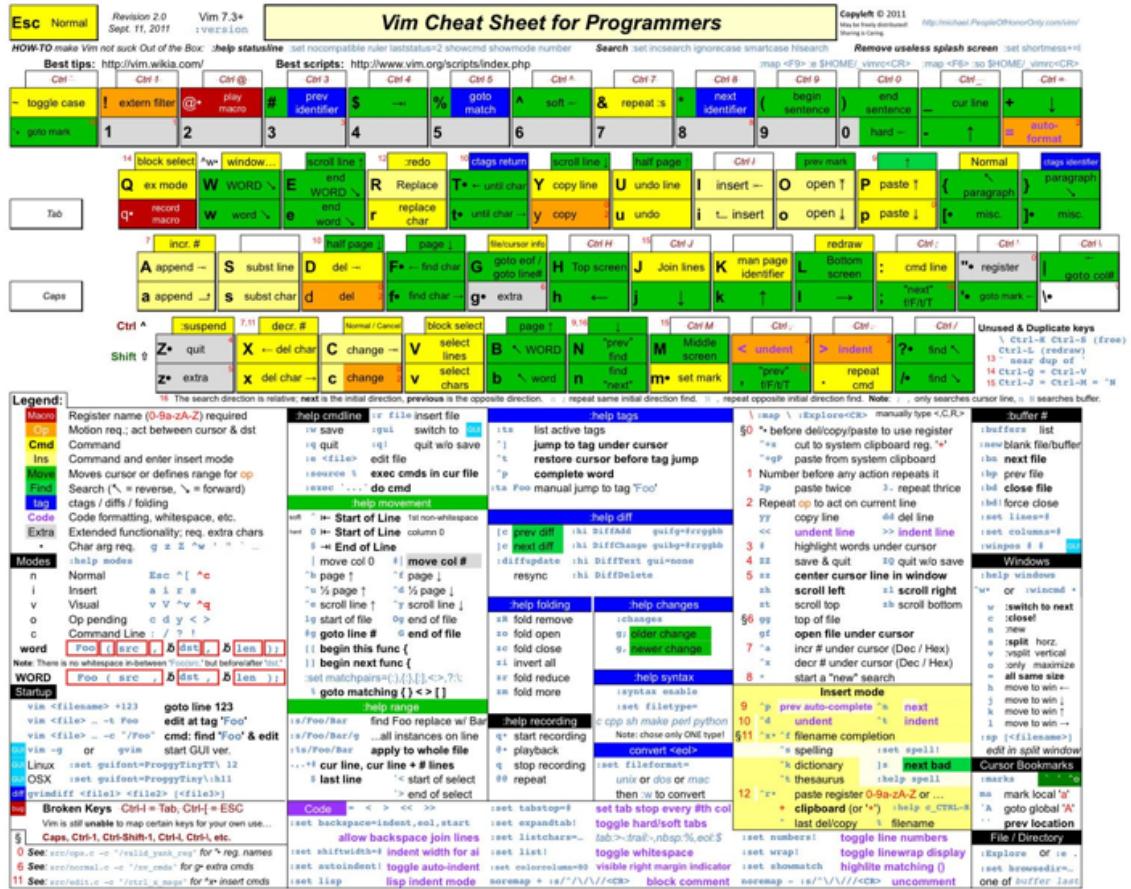
Enter and exit

You can enter Vim by typing vim into your terminal followed by the file you would like to create or edit. You can exit vim by typing : followed by a quit command.

command	action
:q	exit if no changes are pending
:q!	force an exit even if changes are pending
:w	write to disk
:wq	write to disk and then exit

Typing zz does the same thing as the :wq command.

Cheatsheet



1. If you are not from a Unix back then the term "buffer" may be unfamiliar. Just translate "buffer" to "file" or "document" in your head even though, technically, a buffer is more general than a document. ↪

Covid 4/1: Vaccine Passports

Writer's Note: This is being posted on April 1, which is April Fools Day. To avoid all 'is this the fake out?' issues, this post *does not contain any April Fools material*. The fools contained herein are the usual, regularly scheduled fools we talk about every week. The official April Fool this year is whoever accidentally ruined 15 million doses of the Johnson & Johnson vaccine by mixing in the wrong ingredients.

This week started a big debate over vaccine passports. States including New York are deploying systems that let people prove they have been vaccinated, or document recent negative tests. Naturally, lots of people are outraged by this attempt to create a public record and provide information that helps people make better decisions. Thus, the long middle section where I go over all the objections I can come up with against this proposal. Some of them are legitimate.

Meanwhile, the overall arc remains the same. Things are getting worse again, and it's a race to see how long it will take vaccinations to catch up to the problem and get us headed in the right direction again.

Let's run the numbers.

The Numbers

Predictions

Last week's prediction: Positivity rate of 4.9% (up 0.3%) and deaths fall by 7%.

Result:

In the past week in the U.S. ...

New daily reported **cases rose 12.1% ↑**

New daily reported **deaths fell 0.4% ↓**

Covid-related **hospitalizations rose 0.2% ↑** [Read more](#)

Among reported tests, **the positivity rate was 4.9%**.

The **number of tests reported fell 11.8% ↓** from the previous week. [Read more](#)

Deaths only falling 0.4% is rather scary, and seems like strong evidence that the new strains are deadlier if it isn't a data artifact. The data from Wikipedia suggests that it is indeed a data artifact, and deaths dropped at roughly the rate predicted, but there's reason to suspect that result is a data artifact given what's happening in the West. It's all rather worrisome, and strong evidence that the new strains are sufficiently deadlier to make up for an increasingly youthful set of people being infected.

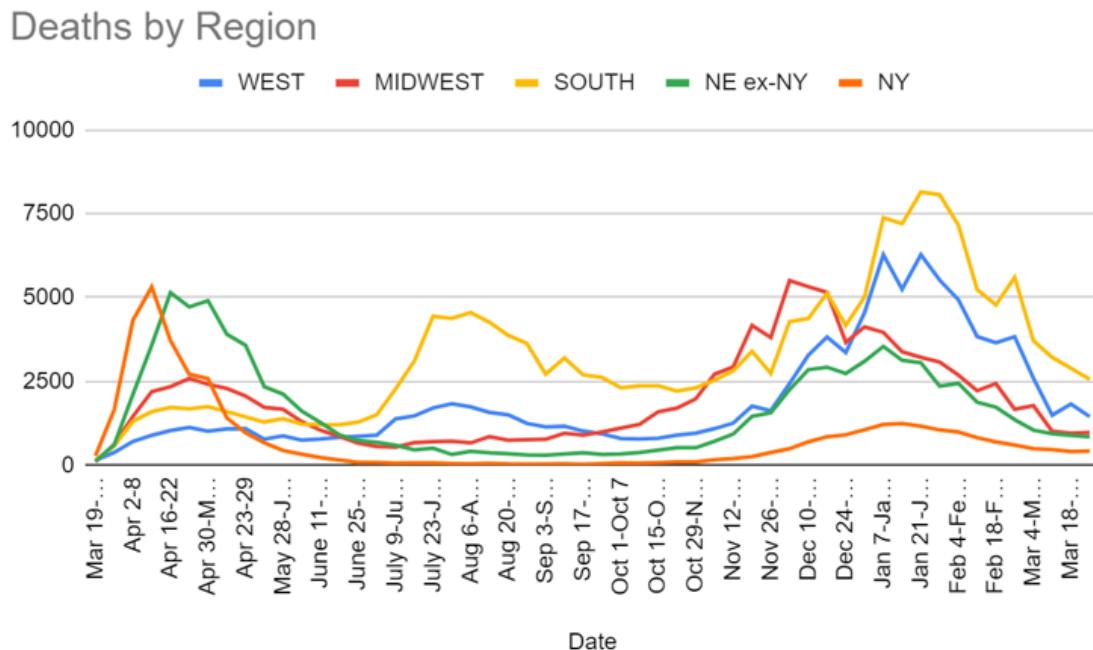
We also need to mention that the positivity rate went from 4.6% to 4.9%, while the same data source is claiming 11.8% fewer tests but 12.1% more cases, which really, really does not compute. Those three facts can't all be true at once.

Johns Hopkins has a 4.8% positivity rate at the moment, so I'm going to assume that the positivity rate is accurate and the number of cases or tests has some error or data artifact in it that's being corrected for the positivity rate calculation. In particular, I'm going to assume tests didn't actually fall *another* 11.8% last week.

Prediction for next week: Positivity rate of 5.3% (up 0.4%) and deaths are unchanged.

Also, I really miss the Covid Tracking Project, and I'm sad that no one found a way to step up and keep it from ending. The next time something this important is about to go away, I'll make a more explicit call to see if it can be saved.

Deaths

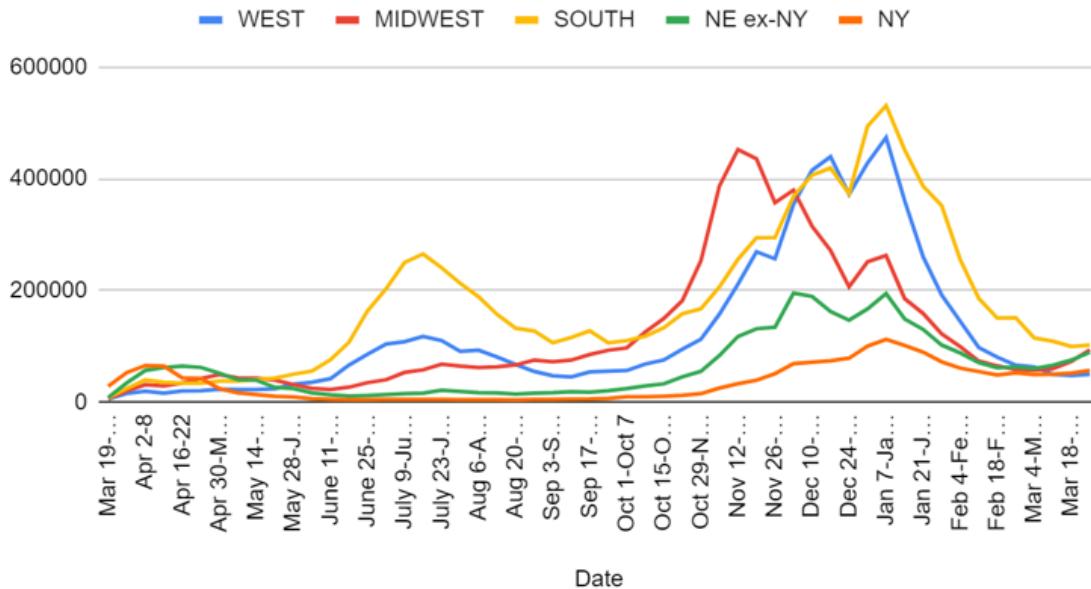


Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Feb 11-Feb 17	3837	2221	5239	2700	13997
Feb 18-Feb 24	3652	2433	4782	2427	13294
Feb 25-Mar 3	3834	1669	5610	1958	13071
Mar 4-Mar 10	2595	1775	3714	1539	9623
Mar 11-Mar 17	1492	1010	3217	1402	7121
Mar 18-Mar 24	1823	957	2895	1294	6969
Mar 25-Mar 31	1445	976	2564	1262	6247

I don't see a big glaring 'this is a timeshifted number' sign anywhere in the data in the West, but going up this much then right back down again isn't a thing that actually happens, so at least some of that isn't real, and the Midwest and Northeast see no declines in cases. Those are also the currently hardest hit areas, where cases are rising, so it does make sense to believe some amount of progress in the West and South, for an overall small improvement in deaths for now.

Cases

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Feb 11-Feb 17	97,894	73,713	185,765	125,773
Feb 18-Feb 24	80,625	64,857	150,493	110,339
Feb 25-Mar 3	66,151	58,295	151,253	115,426
Mar 4-Mar 10	62,935	57,262	114,830	109,916
Mar 11-Mar 17	49,696	59,881	109,141	115,893
Mar 18-Mar 24	47,921	72,810	99,568	127,421
Mar 25-Mar 31	49,669	93,690	102,134	145,933

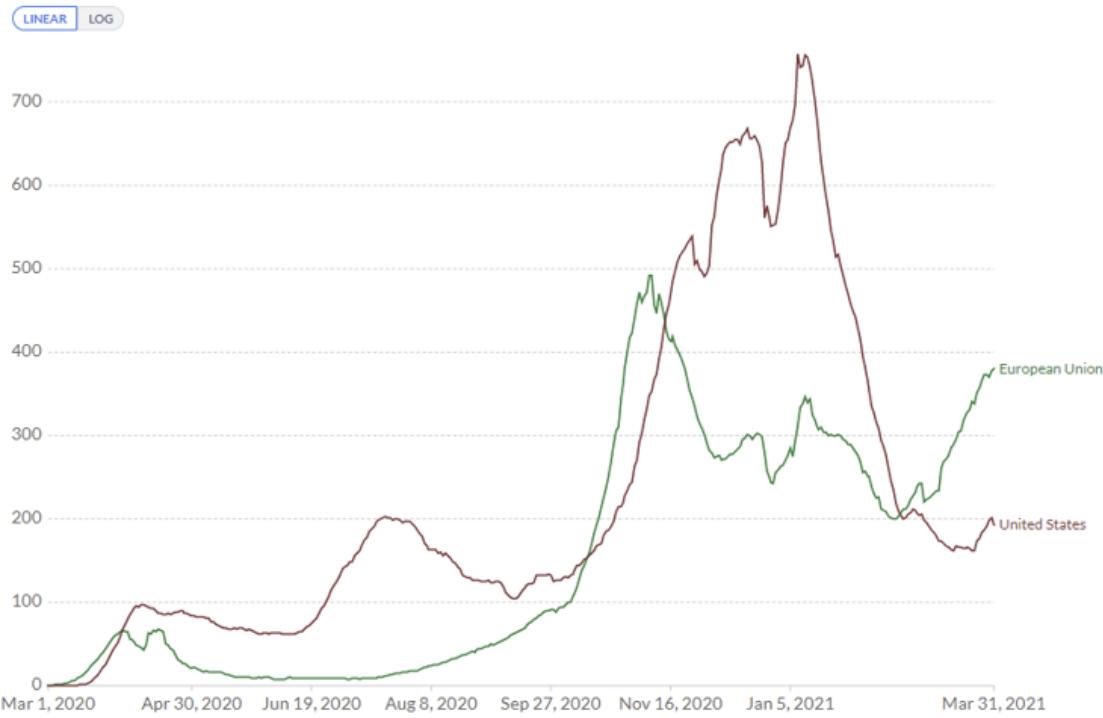
Sharp increases in the Northeast and Midwest, and small increases in the West and South. The fourth wave is here, the question is whether it will be small or it will be large. These are big weekly increases and there's zero sign of any new restrictions happening, so the strategy seems to be waiting on vaccinations to get us out of this. That will work, but the tide won't turn that way alone for at least a few weeks.

European Union comparison graph:

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.

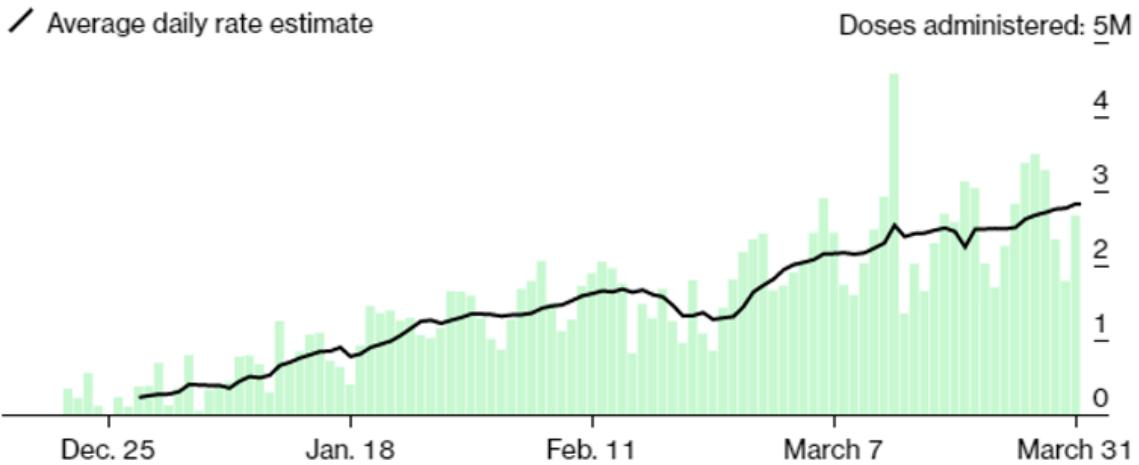
Our World
In Data



Again, for readability reasons I encourage you to [go to Our World In Data](#) to check out other countries of interest to you. The fourth wave might not be that bad in the United States, but in Europe where vaccine efforts are lagging far behind ours things do not look good.

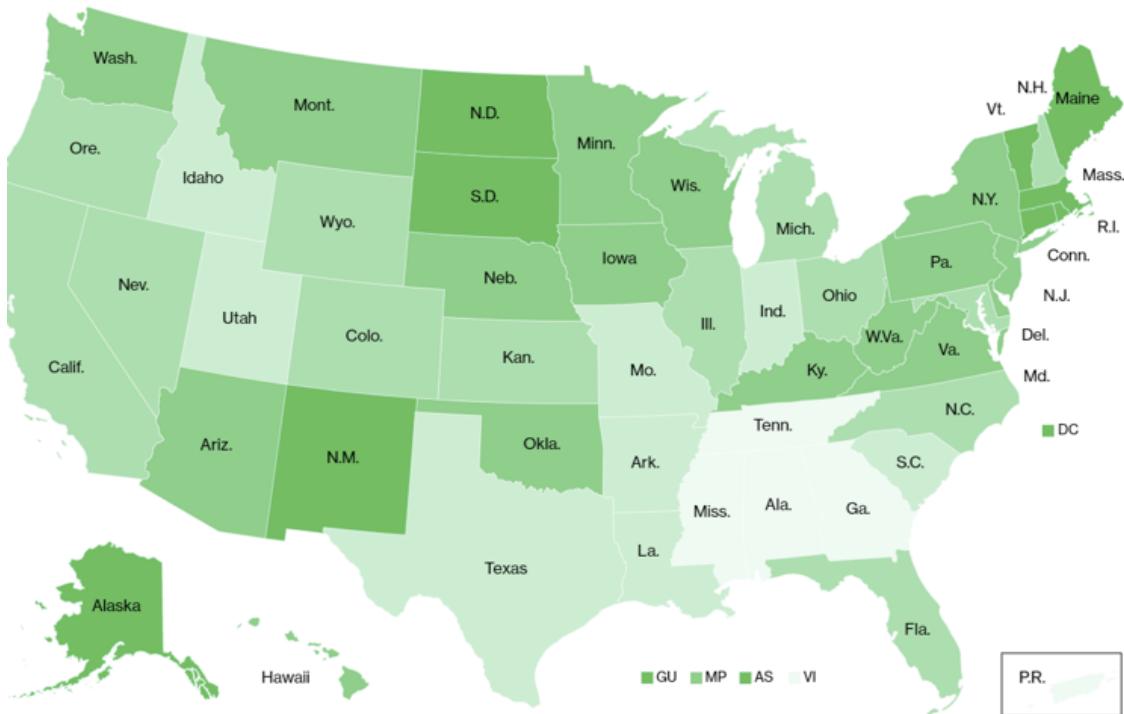
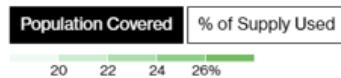
Vaccinations

In the U.S., more Americans have received at least one dose than have tested positive for the virus since the pandemic began. So far, **150 million doses** have been given. In the last week, an average of **2.83 million doses per day** were administered.



Vaccines Across America

Across the U.S., enough doses have been administered to cover 23.1% of the population, and 77% of the delivered shots have been used

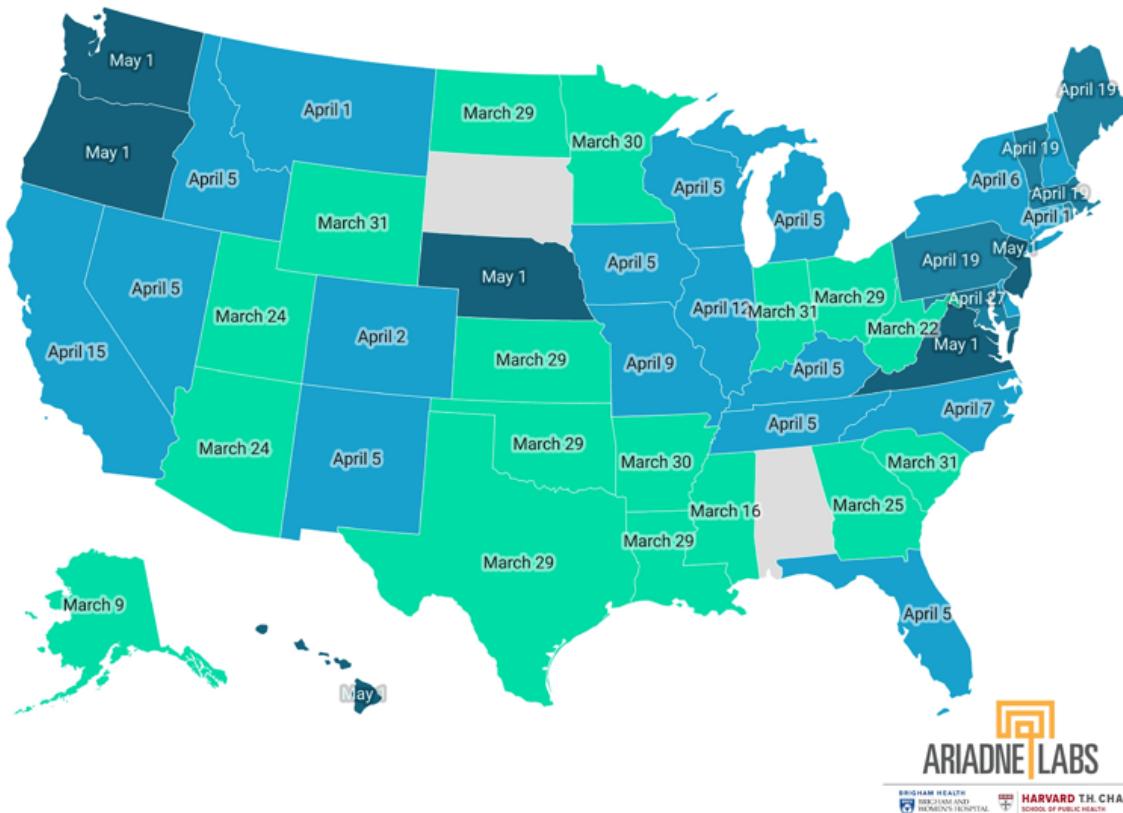


Occasional dips and spikes aside, this continues to look almost exactly like a linear increase in doses available per day over the course of several months. Hopefully we get a big boost to that soon from Pfizer and Moderna, but losing 15 million J&J doses to a manufacturing error is a rather large setback on the J&J front.

[Here's the current availability timeline map:](#)

States Opening Eligibility to All Adults

■ Open to all 16+ ■ End of March ■ Beginning of April ■ End of April ■ Beginning of May



ARIADNE LABS
BRIGHAM HEALTH
BRYAN STANLEY
MEMORIAL HOSPITAL
HARVARD TH. CHAN
SCHOOL OF PUBLIC HEALTH

Almost all gaps have been filled in, and there are far fewer remaining delays. A lot of the country has full eligibility already, and by a week from now it will be a lot more. By April 15, most restrictions will be lifted.

Availability for those who make an effort seems remarkably good based on anecdotal evidence. Most of my Magic friends have been able to find slots quickly now that they are eligible, often finding next day availability. There is certainly a traffic jam right after eligibility restrictions loosen in a given state, but that does seem to clear up before too long. If you're not finding anything, I encourage you to keep trying.

Vaccine Passport Hype

It seems the next big battle is going to be about vaccine passports.

For an entire year, we've failed to provide immunity passports, due primarily to our determination to refuse to in any way acknowledge that people who got Covid-19 were then immune to reinfection, or that someone's actual physical risk of getting Covid-19 should be considered when deciding what actions to take. There were also a whole array of other arbitrary nonsense concerns that are used as part of the war against anyone ever doing anything potentially helpful, such as 'equity' and 'privacy' and 'if we tell them the true situation they might change their behavior' concerns.

With a large and increasing portion of the population getting vaccinated, widespread eagerness to let life return to some semblance of normal, and messaging leading to a widespread belief that vaccines don't actually let one change one's behavior leading to lots

of unnecessary vaccine hesitancy, there is a strong obvious need to be able to tell who is and is not vaccinated.

Thus, the concept of a vaccine passport. You download a QR code on your phone, anyone can scan it and see that you're vaccinated, or notice you can't produce one and exclude you from risky activities.

This allows life for those who are vaccinated to return to normal, by ensuring that gatherings that would otherwise be risky exclude the unvaccinated until we've suppressed the virus, and it provides strong incentive to the unvaccinated to get vaccinated.

One thing to watch for is the distinction between those who support passports because it allows for valuable activity to take place safely, and those who support it because it allows people to 'follow the CDC guidelines.' That's the distinction between caring about the physical world and caring about fulfilling symbolic requirements, and it's important to retain that distinction.

Sounds great, and also obviously the right thing to do. Yet there are objections. It's worth dealing with them. The first job is to list all the distinct objections I've seen or that I can think of, and I managed to get to 12 (with some overlap):

1. Privacy concerns – from 'this will create a new database' to yelling 'mark of the beast'
2. Equity concerns – some people aren't vaccinated, some don't have phones, etc
3. Coercion concerns – this is effectively a vaccine mandate at pain of exile from life
4. Fraud concerns – I haven't heard this objection yet, but won't it be easy to fake?
5. Fear concerns – If we give the impression vaccinated people are safe, they'll take risks!
6. Culture war concerns – Some people won't take kindly to this, and that's terrible.
7. Practical concerns – This will be a government program so they'll mess it up a lot.
8. Norm concerns – If people see the vaccinated living life, the unvaccinated will too.
9. Motive concerns – If I support this helpful thing, my loyalties will be in doubt.
10. Vague concerns – Fear, uncertainty and doubt around anything new, ever.
11. Vibe concerns – This feels like a thing we should hate so we hate it.
12. Anti Elite concerns – Elite people like this, so we hate it on principle.
13. Approval concerns – This is only in Emergency Use Authorization, you can't punish people for waiting until full approval.

Some of these seem purely hostile and/or stupid. Others are legitimate concerns.

Privacy Concerns

Privacy is a legitimate concern, and the most common and forceful objection that I have observed. [Hence this poll represents a common framing.](#)



Bella Rudd @BellaRudd1 · 12h

...

what's your take on vaccine passports/badges?

- a-good-bc might help slow the spread
- b-bad-bc it's another centralized govt database of private info
- c-neutral-bc i'll be able to forge/circumvent any such policy
- d-pundit-bc policy is hard & it's easier to just dunk on other people

i'm an authoritarian sub! ✓

39.5%

i'm a paranoid whackjob!

36.8%

i'm hoping ur not a fed!

15.8%

i signal tribal loyalty!

7.9%

76 votes · 1 day left

There are multiple *different* privacy concerns, regarding different actors getting access to different information. Some are unavoidable, others are a choice. It's important to track them separately.

The big distinction I would draw is between the information on *who has been vaccinated* and the information of *who has been where and done what other things*.

If the privacy concern is 'people will know my vaccination status' then I notice I am confused as to why this is a problem. To me this seems like the *exact opposite* of a problem. The *whole point* is to make sure people can know your vaccination status, so everyone can take appropriate action in response to this information!

If there is physical information about the world that would change the consequences of actions in ways that would change your behavior, that's information you need to know. I've been asked about my vaccination status often, and I've asked many others about theirs, and at no point did anyone involved think privacy concerns were an issue here.

Yet there are people calling this the 'mark of the beast' or even invoking Godwin's Law directly off the bat by comparing it to the Nazis forcing Jews to wear a yellow star, which fall into some combination of 'anything scary sounding is good when signaling that everyone should be scared and making no sense is actively helpful in clarifying what you're up to' and 'pattern matching systems in the brain aren't using much logic and this feels like it might sound good so go with it.'

In short, I think the concern 'people will respond to me based on how likely I am to have Covid-19' is a stupid objection, to the extent it isn't one of the other objections in disguise (e.g. it could actually be objecting that this is helpful, or that this is coercive, or about equity, etc.)

The other privacy concern is that this risks tracking us more generally, or opening the door to such tracking, and that's a completely legitimate objection.

I was listening to the Brian Lehrer show and they were interviewing someone advocating for the New York State vaccine passport app. When asked about privacy concerns, he responded that the *current* version did allow people to be tracked, as each QR scan would identify who

they were and that they were at a particular place and time, but that *future versions* would fix that *real soon now* and privacy would be dealt with.

I am skeptical. Government programs that collect information about private citizens tend not to view that as a bug to be fixed, but rather a feature to be preserved. Corporations that get to gather similar information aren't going to be lining up to object to this either. It's entirely possible this will get fixed, but unless sufficient pressure is brought to the issue, chances are that it won't be fixed.

It would be quite bad, in my view, if the government had records of who was where at what time for many everyday activities. It is very reasonable to not want any part of such a system. It is also very reasonable to worry that once such a system is in place, the powers that be would find a way to make the system permanent, in order to keep collecting the data, or use the program as precedent to then collect the data in other ways.

Can anyone think of an example of where such information was created, and the government was respectful of our rights and *didn't* check it whenever they felt like it? Anyone?

The good news is that this is avoidable. There are plenty of privacy experts out there that can design a version of the system where you can't be tracked. The system can see if you're vaccinated, but it can't tell who 'you' are while doing so, except to verify that the claim is legitimate. I'm sure plenty of crypto people can help us out on this one and would be happy to do so free of charge.

Equity Concerns

There are two potential inequalities.

There is the concern that some people don't have smartphone access, and thus wouldn't be able to provide the QR code, and would be shut out of the system.

That's an important *concern to point out*, for the purpose of *fixing it*. The good news is that there's no reason for this to be an issue. A QR code can be placed upon a piece of paper, and those without a phone can carry the piece of paper, the same way we can carry the vaccination card now except with a less trivial duplication/fraud problem. It's not a meaningful objection.

The real concern is that this is punishing people for not getting the vaccine, combined with the concern that some groups and communities aren't getting equal access to the vaccine, and those same people are already worse off in general and worse off for the lack of vaccine access. Aren't we punishing those communities and people even more?

One could argue that those are activities people shouldn't be doing unvaccinated under those circumstances, *and one would be right*, but such arguments rarely sway equity advocates. You can't answer an equity concern with an efficiency argument, equity advocates don't consider that a valid response.

We can separate this concern into the period where vaccines have limited availability and it's hard to get an appointment, which for now is still everywhere, and the period in the future where vaccines are freely available at your local pharmacy to anyone who walks in or at most signs up a day in advance.

During the period where vaccines have been widely available for long enough for everyone to get them, I consider equity concerns solved. The stab is quick, the stab is free, and there are real consequences to not getting the stab. If anything, there's an equity concern *not* using vaccine passports, because that would lead to the undervaccinated communities having more Covid and people getting sick or dying more often, by letting people engage in

risky activities without being vaccinated, taking risk and also not having the additional incentive to get vaccinated.

Before vaccines are easy to acquire, the argument is stronger. In particular, a reasonable concern would be if major life activities ended up gated by passports *before* those passports could reasonably be acquired, and activities that would be reasonable to do unvaccinated and are important for people to access become impossible to do without a passport. If it's all ballparks and concerts, I'm not sympathetic at all, and if it's about indoor dining I'm still not sympathetic because life goes on and it'll all be over soon enough anyway, but if it stops people from grocery shopping or getting a job, yeah there's a concern there.

The obvious solution there is to not allow vaccine passports to be used as gateways to essential life activities until such time as vaccine availability is complete, except insofar as there's sufficient risks involved in the activity that it makes sense to require the passport anyway.

I don't think this needs to be a government requirement. There would be both direct loss of business (and/or good employees) and a lot of blowback for requiring vaccine passports in places like grocery stores, or firing people who aren't vaccinated before they have the opportunity to get a shot.

Looking for new employment or new housing, or seeking travel on mass transit including airplanes, seem like the places most likely to create a real issue. These are already places where there are especially strong anti-discrimination provisions, where people making decisions are told they are not allowed to consider information (that they'd often consider quite useful) when making those decisions. For frontline jobs, I'm willing to bite the bullet and say, yes, absolutely we should allow a vaccination requirement, the physical need here is too important, and I'd prefer requiring it by law to using law to prevent there being an employer-mandated requirement.

For travel, there's big real costs imposed by not being vaccinated. If you can't check who is vaccinated, you'd need to impose those costs for all passengers - everyone would have to distance like no one (else) was vaccinated, often cutting capacity in half or worse. Reserving at least some capacity for vaccinated people only makes everyone's trip cheaper and easier and safer. It's important to make sure the unvaccinated can travel at all, and in some reasonable fashion, but ignoring the issue seems super expensive. Despite this, I don't expect those involved to get their acts together fast enough on these matters to cause issues before vaccination access is widespread.

For those worried this is effectively class discrimination, this will soon become the easiest class marker there is to fake. All you have to do is get vaccinated. If this becomes a central method of those who are attempting to discriminate by class, that's great news in the medium-term.

In short: I'd sympathize with this worry (to the extent that equity concerns ever deserve any response other than 'transfer payments') if we were looking at an extended period of widespread passport use for essential life activities without vaccine access, but that seems highly implausible. Anyone who wants an appointment by the end of August (to be super conservative) will be able to get one without being savvy, and it's already April and these things take time to get adopted.

Coercion Concerns and Approval Concerns

Is this us forcing people to get the vaccine?

Mu. It's not us forcing you to get vaccinated, but it's not *not* forcing it, either. It's a rather strong nudge, a punishment for being unvaccinated slash a reward for being vaccinated.

The extent to which this is coercion versus bribery, and the extent to which that effect is central versus incidental, depends on one's point of view.

One can look at 'require passport for activity' as coercion by punishing the unvaccinated. One can also look at it as coercion by bribing the vaccinated. In this context, the two are basically the same, even if they are clearly meaningfully different in the 'free glazed Krispy Kreme donut' scenario. The power to tax is the power to destroy, and the power to bribe is the power to tax.

PoliMath @politicalmath · Mar 29

I think I've pinpointed my reluctance to any concept of vaccine passports to my deep suspicion of anyone who wants to force or coerce others into adopting their preferred policies or patterns of behavior

20 24 269

PoliMath @politicalmath · Mar 29

Maybe I could be talked into this particular idea, but holy shit, anytime someone starts a conversation that is "We need to force these people to do this thing" you should start hearing sirens in your head

10 29 230

If you *don't* have a deep suspicion of anyone who wants to force or coerce others into their preferred patterns of behavior, *you need to get more deeply suspicious*.

The sirens can be overcome. They don't mean we should never force anyone to do anything. It certainly doesn't mean we should never prevent anyone from doing anything, and it most definitely doesn't mean we should never provide an incentive to do one thing over another thing.

For example, it definitely doesn't mean we shouldn't coerce people into vaccinations. We should and do require people to get vaccinations! You can't go to school, or get many jobs, without proof of vaccinations. The externality argument, and the 'this is an overwhelmingly worthwhile thing to do' argument, are both extremely strong. A vaccine is *exactly* the place where we should be *least* suspicious of coercion.

The approval concern is that it's not reasonable to apply this coercion if the FDA isn't even willing to fully approve the vaccine. There are several responses to this, but the main one is that if the FDA is playing bureaucratic games with its labeling and timing in ways that have killed hundreds of thousands of Americans, that's on the FDA, and we should not take such labels either seriously or literally. The vaccines are safe, and even if the FDA hasn't officially fully approved them yet, most of our most important people are vaccinated, we now have given hundreds of millions of shots and we damn well know they're safe, so I do not want to hear it.

To the extent that this is a messaging issue, it sounds like the worry is that we're providing a nudge against getting the vaccine and this should make us not want to push people in the

other direction to fix it, which I find confusing.

To the extent that these effects are contained to vaccinations, the coercion effect is a bug rather than a feature. We want to provide strong incentives to get vaccinated. There is a real worry that this helps establish a pattern and/or method of coercion in general, that is used on other things. In this context I am not too worried about that, but I am open to that lack of worry being a mistake.

I don't consider 'vaccination required' that different, in a world where vaccinations are free and widely available and the passport software protects your privacy, from 'mask required' or from 'no shirt, no shoes, no service.' We want people to wear masks and shirts and shoes, and also get vaccinated, and the lack thereof has practical real costs for all of them.

Fraud Concerns and Practical Concerns

Can't people who want one simply fake a QR code that will represent that they are vaccinated? Especially if the system is meant to protect their privacy?

There's a clear trade-off involved. The easiest ways to preserve privacy involve making the signal easy to fake. Verifying that it isn't fake requires verifying identity, which is a threat to privacy.

As I noted earlier, I expect crypto people to have good answers to these problems, as they have been hard at work dealing with similar issues for a while with large amounts of money at stake.

If not, then how are we to feel about a system where it's not that hard to fake being vaccinated?

Presumably we should feel a similar way to how we feel about people showing ID to buy alcohol.

We have a thing people want to do (buy and consume alcohol) that we have decided some people mostly shouldn't do until they meet the requirement to do it at lower risk (be older). Thus, we ask them for proof that they are older, in the form of ID. In response to this and other ID requirements, a huge portion of young people historically get fake IDs.

This clearly *does* have a big impact on the amount of alcohol consumed by minors, who spend considerable effort acquiring booze and often fail to do so, or get lower-quality booze at higher prices or in less enjoyable ways.

In exchange, we initiate our young into a culture of fraud and lying, where in order to participate in ordinary human activity (again, buying and consuming alcohol) they have to lie and commit fraud, and those who facilitate such frauds and lying, in various forms, become heroes with high status, look cool and get laid, and often also make money.

On top of that, what alcohol is still consumed is often in secret, and hidden from those who could help deal with problems when they arise, and thus is often consumed in relatively dangerous ways. The harm mainly falls on those consuming alcohol, but also on others, and also they often harm each other.

The parallel here seems obvious. If we require vaccine passports that are not that difficult to fake, then a bunch of people will respect that, but a bunch of other people will fake them and then act even less responsibly, since acting like they were unvaccinated would be suspicious, and also they'll cultivate the anti-virtues of lying and fraud.

Given how easy it is to commit fraud and lie in order to get vaccinated earlier, my assumption is it also probably won't be that difficult to fake the passport, so that's the deal

we're facing here. The same way I think it would be better if the laws against minors consuming alcohol were less strict but were more strictly enforced, it would be best if we used the passports only when necessary, but ensured that it was difficult and/or risky to fake them.

Thus, the practical concerns involve privacy and fraud. They also involve error. Multiple people who called into the radio show I listened to complained or asked about issues getting the passport. One lady got her first shot in Florida, and the system didn't know how to handle that. The response was a hand wave. Another had the concern about not having a smartphone, which again shouldn't be an issue *in theory* but in practice the solution wasn't obvious or forthcoming. And [Cuomo shows us exactly how stupid a government program like this can be...](#)



Mitchell
@ChiCubsJames

...

Replies to @politicalmath

NY state has launched a Covid passport called the excelsior pass and to have access you need a negative test result in last 2 weeks or both doses. But the passport then says it's revoked if it's > 90 days since second dose as if the vaccine is no longer effective

- It has been more than 90 days since you received the last dose of the COVID-19 vaccine.

What happens when a substantial number of vaccinated people are more than 90 days out from their second dose, which will be true soon? What happens when people see this and think vaccination only works for 90 days, and thus don't bother getting vaccinated, or start going back to pre-vaccination behaviors, or request a third dose as a booster shot?

My wife signed up for Excelsior, and she doesn't even get to keep it until 90 days after the second dose, because it seems it has to at least be renewed every 30 days for some reason.

So, yeah. Government running the system is a real concern here.

Fear and Norm Concerns

Whenever anyone suggests letting people who are vaccinated do things that are now safe, others respond that this would be terrible because unvaccinated people would then respond

to those actions by treating the pandemic as over slash adjusting their norms and taking more risk, and also that the vaccinated people are still at some risk and any additional risk anyone takes is terrible, or something like that.

At least one person I follow on Twitter (who I won't link to or quote here) blamed the recent uptick on cases on the CDC issuing guidance that unvaccinated people could meet together, on the theory that *of course* this *inevitably* had been misinterpreted as giving permission to others to not wear masks or not distance, and *of course* that was the direct cause of the uptick. Which is obvious nonsense, since the uptick was already predicted and baked in, and also I see no evidence of people making that kind of adjustment to the CDC guidelines.

Vibe and Anti-Elite and Vague Concerns

When I talk about anti-elite concerns or vibe concerns, I mean [arguments kind of like this](#):

 PoliMath Retweeted



Ben Sixsmith
@BDSixsmith

...

Take-up could be 99% and some people would still back vaccine passports just to feel superior.

3:53 PM · Mar 30, 2021 · Twitter for Android

15 Retweets 116 Likes

One could interpret this as saying that bad people are supporting vaccine passports for bad reasons, thus we should oppose them. One could call that an uncharitable interpretation, but it's not, because that's both the real argument here and also not a crazy argument. If there are bad reasons people are supporting policy X, then policy X is likely to get enacted in worlds where it's net harmful due to that support, and if we collectively push back against X a similar amount, we can correct this imbalance. That's especially true when the motives are selfish, and some special interest wants to take from the public treasury, or otherwise get private benefit at public cost (e.g. make us all continuously show passports in order to let them feel superior).

It is a useful exercise to total up all the *bad* reasons people are for or against something, where bad in context means considerations you don't value rather than a moral judgement, and then pushing back to correct this imbalance. It's important to correct both bias and incentive on every level at all times, even if that doesn't look like the locally 'rational' thing to do. Someday I hope to explore such issues in more detail.

Here, these arguments very much run both ways. There are people who are opposing this for zero-sum reasons, and people supporting it for zero-sum reasons, and it's not obvious which effect is going to dominate. One could even say we'll know which one dominated when we see which side won.

When I talk about vague concerns on the concern list, I'm pointing to a general Suspicion of Authority that is pervasive in America, and with good reason, and a general suspicion of anything new like this, again with good reason, but not to any particular good reason. I do

think that a vague uncomfortableness is appropriate for various reasons (again, there are 13 concerns listed) and this represents a threshold effect that needs to be overcome.

Culture War and Motive Concerns

Masks became a culture war issue, despite there being no privacy issue with masks, despite a Republican president, and with no real concerns of any kind other than 'how dare you tell me what to wear' and 'this mask is mildly annoying.' Vaccines are also already an issue. Vaccine passports dial this up even further, and raise the specter of 'red tribe members are being systematically censored and excluded already, and now we're going to accelerate this process and lock them out of major life activities' with the new administration leading the way. I can totally see this perspective. I don't think it's right or that we should let it stop us or anything, but I do get it.

There's thus both the concern that this is part of a broader culture war push from the left, whether or not it's being sold or even explicitly contemplated that way, and the instinctive and automatic 'the other side wants thing so we oppose that thing' dynamic going on as well.

That leads into an argument from some lawmakers that goes something like this: You want to do this useful thing. However, there are lots of people that have been convinced to oppose this thing, and if you tried to do this thing anyway, they'd be mad and this would cause trouble for you, and this would be 'combative' and 'divisive' and we'd be forced to make your life miserable and job harder, so Mr. President, please don't do this or we can't be held responsible for the consequences. That's not a completely illegitimate concern, even if it's largely disingenuous and those same people saying this isn't going over well are mostly working hard to ensure that it doesn't go over well. Resistance is a real issue here, as is upping the general distrust factor and level of conflict and paranoia. It's a concern.

Finally there's the Motive Ambiguity concern. If you're for the passports, you're presumably for them so that people can get back to living their lives and stay safe. Those sound suspiciously like better things that you prefer to worse things. Can't have that. Whereas if you oppose it, you can signal your loyalty to almost any group via your choice of symbolic concern, and even plausibly claim loyalty to all of them at once. Neat trick.

One could argue that there are also some plausible *bad reasons* to support the passports, and this could be used as a defense here. You *might* be supporting good things over bad things, but maybe you like inequality and discrimination, or you like punishing the outgroup for defying your preferred shibboleths, or you hate privacy. There are several good choices here that lend plausible deniability to supporters. I don't think that fully works, because when it's this obvious which side has the physical benefits on its side there will always be suspicion that those concerns mattered, but it does help.

Summary on Vaccination Passport Arguments

I still find the case for vaccine passports overwhelming, as there's no realistic path to doing things safely without them other than waiting for full suppression, and giving people incentive to get vaccinated once we have unlimited supply is super important. Yet the concerns here are real, and worth worrying about.

I am especially concerned about privacy, and it is important to find and push for a technological solution that doesn't allow the tracking of who had their QR code scanned where and when. I would also be concerned if we were to start screening off essential aspects of life within the next few months, while access to the vaccine remained an issue, and we do need to protect housing, transportation and jobs against actions that shut people out too broadly and quickly, even if there are some efficiency losses involved.

What is especially troubling are those who seem determined to ban, boycott or punish those who attempt to run a private business in a safe fashion. The exact worst thing one could do right now would be to ban people from asking about others' vaccination status when trying to figure out what would be safe or unsafe actions, and that's exactly what a number of states including Florida seem to be doing in order to score political points.

Asking The Best Question

Nate Silver gets major points for asking what is arguably the best question about Covid that is rarely asked, and asking it correctly, and to which I'm not super confident in the answer:



Nate Silver 
@NateSilver538

...

OK, I have a weird question that's a little hard to phrase. To the extent vaccines are not 100% effective, is that because some individuals have better immune responses than others, or because some encounters with COVID are more likely to overcome the immune response than others?

4:38 PM · Mar 29, 2021 · Twitter Web App

32 Retweets 18 Quote Tweets 945 Likes



Nate Silver  @NateSilver538 · Mar 29

...

Replying to [@NateSilver538](#)

Put another way, should we treat each encounter as independent? If say a vaccinated person takes a 95% effective vaccine, resumes some activities, has a bunch of exposure to COVID and doesn't get COVID, should we revise our prior for how effective the vaccine is *for her*?

141

23

560



I think I've talked about this a few times, where there are two possible worlds. In world one, each exposure is 95% less likely to infect you, so any given action isn't too risky but if you take lots of risky actions continuously you will eventually get Covid-19. In the other world, degree of effectiveness varies from person to person, so 95% of people can do whatever and never be infected, while 5% are at the same risk they would have been without vaccination.

You could also have a hybrid world between the two, which seems highly plausible; it gives varying degrees of additional protection to different people, that combine for 95% protection.

My prior has been that it's *mostly* the second world. Most people are immune, and no realistic exposure is going to change that, and if they did get exposed a bunch it would act like a booster shot long before anything bad happened. A few people's shots get botched or contaminated, or they have compromised immune systems, or something randomly goes wrong, and they're still varying degrees of vulnerable, but often with substantial partial protection.

Thus, mostly this answer:



Salsa Nirvana @LisaAJK · Mar 29

...

Replies to [@NateSilver538](#)

Not a scientist, just a nerd who used to work for a major vaccine manufacturer, and my layman's understanding was always that "95% effective" means that 5% of people vaccinated won't develop antibodies and thus could still catch the disease they were vaccinated against.



Here's a more complete take:



Natalie E. Dean, PhD  @nataliexdean · Mar 29

...

Replies to [@NateSilver538](#)

We have two conceptual models for vaccines.

- 1) All-or-none, meaning you are either perfectly protected (95%) or the vaccine doesn't work (5%).
- 2) Leaky, meaning the vaccine works the same for everyone. Not perfect, but reduces your risk of disease per exposure by 95%.

...

17

53

532



Natalie E. Dean, PhD  @nataliexdean · Mar 29

...

The truth may be somewhere in between, with similar levels of high but incomplete protection, with some people who don't respond quite as well as others. One thing we hope to do is link immune responses with level of protection (analyses in progress). ...

6

15

301



Natalie E. Dean, PhD  @nataliexdean · Mar 29

...

Since we see strong immune responses in the vast majority of vaccinated people (versus there being some subset with no response), breakthroughs may represent chance events, particularly in people with more frequent exposure or higher dose exposure.

11

13

309





Natalie E. Dean, PhD @nataliexdean · 23h

...

A related thought... the level (or type) of immune response to prevent infection may be different from what's needed to prevent disease may be different from what's needed to prevent severe disease (the easiest lift). (I owe [@JuliaLMarcus](#) a picture of how I visualize this...)...

5

13

135



Natalie E. Dean, PhD @nataliexdean · 23h

...

So most everyone might be above the threshold for preventing severe disease, but some people might be below it for infection (or challenged with a bigger dose). It will take more time (and more examples of vaccinated people getting infected/sick) to work this out.

5

8

163



I'm willing to go with that explanation.

What that means in practice is that once you're vaccinated, there's an effective cap on how much risk you can take. It still makes sense to defend against the bigger risks when the cost of doing so is reasonable. It could also be reasonable, if circumstances broke correctly, to actually not care at all and accept the full 'if the vaccine didn't work I'm going to catch this' failure rate, since it caps out around 5% for infection and the death rate per case is much lower on top of that. Doing this likely only shortens life expectancy no more than a few days.

I still plan on taking the easy precautions, but part of that is I don't have a way to usefully take a lot of risk. There isn't that much to do in this town.

Vaccines Still Work

The latest reproduction of this result is [a new study of health care workers](#).

mRNA COVID-19 vaccines are highly effective in preventing infections in real-world conditions



Nearly 4,000* health care personnel, first responders, and essential workers were tested weekly for the virus that causes COVID-19



Those who were fully vaccinated[†] were **90% less likely** to get infected

* Effectiveness of Pfizer-BioNTech and Moderna mRNA vaccines among 3,950 study participants in eight U.S. locations from December 14, 2020, to March 13, 2021. Participants self-collected specimens weekly regardless of symptoms and collected additional specimens if they became sick.
† Fully vaccinated = 2 weeks after 2nd dose

CDC.GOV

bit.ly/MMWR32921

MMWR

During the 116,657 person-days when participants were unvaccinated, 161 PCR-confirmed infections were identified (incidence rate = 1.38/1,000 person-days). During the 13 days after first-dose or second-dose vaccination when immune status was considered indeterminate (67,483 person-days), 33 PCR-confirmed infections were identified and excluded from the outcome. Two sources of partially immunized person-days were reported. Five PCR-confirmed infections were reported during 15,868 person-days \geq 14 days after their first dose among those who did not receive their second dose during the study period; three PCR-confirmed infections were reported during 25,988 person-days \geq 14 days after the first dose and through receipt of the second dose. Taken together, this represents eight PCR-confirmed infections that occurred during 41,856 person-days with partial immunization (\geq 14 days after first dose and before second dose; incidence rate = 0.19/1,000 person-days). Three PCR-confirmed infections occurred during 78,902 person-days with full immunization (\geq 14 days after second dose; incidence rate = 0.04/1,000 person-days). Estimated adjusted vaccine effectiveness of full immunization was 90% (95% confidence interval [CI] = 68%–97%); vaccine effectiveness of partial immunization was 80% (95% CI = 59%–90%) (Table 2). In sensitivity analyses, inclusion of other covariates (sex, age, ethnicity, and occupation) were entered individually in the vaccine effectiveness model; the change in vaccine effectiveness point estimates were <3%.

The headline numbers are drops in infection of 80% and 90% after the first and second doses respectively, starting fourteen days after the first dose for the date of the test, and they reflect actual behaviors of people who know they've been vaccinated.

There isn't much more to the study than that. Once again, it's clear that protection against all infection is somewhat weaker than the 95% protection against symptomatic infection, and that the first dose gets you far more than halfway to being protected.

On a related note, I did go back and look at the data about the timing of protection from the first dose, and I've concluded that it's potentially slower than we thought, based on the timing of the real world data. You still have a lot of protection by day 10, but I wouldn't get excited on day 7, and waiting until day 14 isn't crazy. After that, we're still talking 80%+ protection from one dose from infection alone, higher than that for symptomatic infection, higher still than that for hospitalization and death.

There is no clean way to do a study on the exact infectiousness level of the vaccinated, but I continue to view the probability that they aren't at least roughly as much less infectious as they are less likely to test positive on a PCR test to be quite small. A bigger decline would surprise me less than a smaller decline, since those who get worse cases tend to be more infectious.

This study is commonly referred to as 'good news.' On reflection I agree with that assessment, but almost entirely in the sense that we now have stronger evidence to overcome vaccine hesitancy, rather than this being a positive update on effectiveness. For

this to be good news on effectiveness it has to move our estimates substantially upwards, which this doesn't. That's a sign of good calibration.

This study also makes it very clear that first doses first would have been the correct approach, but that horse is quite dead and I see no reason to beat it further.

Also, in case there was any doubt, [yes the vaccines work in children](#).

Natalie E. Dean, PhD Retweeted

Meg Tirrell @megtirrell 1h

A potential #covid19 vaccine in time for back-to-school. Pfizer shows 100% efficacy in kids ages 12-15 (18 cases on placebo, 0 on vaccine, among 2,260 in trial)

Hopefully this isn't used as a 'wait until the kids are vaccinated' excuse.

Vaccines *don't* work, unfortunately, if you ruin them during the manufacturing process. [Which seems to have happened to 15 million doses \(!\) of Johnson & Johnson.](#)

But two senior officials working on the federal government's Covid-19 response told POLITICO that it became clear earlier this month that there were significant problems at Emergent's West Baltimore plant, where the company was producing the active ingredient — or drug substance — for J&J's vaccine. The officials said they had not known the exact details of the situation.

A third senior official said the Department of Health and Human Services found out last week that Emergent had botched the 15 million doses, and how. "It was no secret that Emergent did not have a deep bench of pharmaceutical manufacturing experts," that official said.

The federal government expects that Emergent's problems will delay future shipments of Johnson & Johnson's vaccine, and distribution of doses to states will be patchy during the next several weeks, two other senior administration officials said. One of those officials said that Johnson & Johnson should still be able to deliver the doses it promised under its contract with the federal government by the end of April.

The argument that 'money' is not effectively the limiting factor on vaccine production now has to explain why people for whom 'it was no secret that [they] did not have a deep bench of pharmaceutical manufacturing experts' were put in charge of one of the major vaccine production sites. Either there was a supply limit that was binding, in which case perhaps that supply should have gone somewhere *that had a deep bench of manufacturing experts*, who would be less likely to do things like *mix ingredients for two different vaccines together and thereby ruin 15 million doses*. Due to a single 'human error.' Or, alternatively, if there wasn't a supply limit, why aren't there more factories?

This is from the-source-I-will-not-link-to via Twitter:

It does not affect Johnson & Johnson doses that are currently being delivered and used nationwide. All those doses were produced in the Netherlands, where operations have been fully approved by federal regulators.

Participants are Heroes, Go Team Yeah

Not Covid related, but it looks like I [made it onto someone's fantasy intellectual team](#), going 106th in the draft. It's quite the honor to be taken at all in such a thing, especially considering some of the giants that went only slightly before I did. Of course, the scoring system is super generous to me, with bets (B) being one of three categories, and a second

(S) being steelmanning. In theory I even have some meme (M) game, especially with Eliezer Yudkowsky, Julia Galef, Robin Hanson, Robert Wiblin and Scott Alexander all counting as one of the three uses necessary for a term to count. One big advantage my team will have is that *I am aware of the competition*, and as always, it's good to have fantasy players who care even a tiny bit about their fantasy scores. My understanding is there are 9 teams, and the only team member I know about is Robert Wiblin, so it's hard to judge whether the team is well-rounded or generally seems strong. But if I keep writing at this pace, I should be a monster in the (B) category, and I do expect to keep writing for a while, so before seeing the rest of the team I'd give my team a 30% chance of winning that category outright (if you think I'm wrong please say so, but I really don't recommend betting against me on this for an amount I'd care about, in either direction, for obvious reasons), and maybe a 20% chance to win overall out of 9 teams because I think picking me also reflects well on their likely other picks by showing they're more likely to be thinking carefully about the points system. I wouldn't underestimate how much the pick order was determined by *whose content you wanted to monitor the whole time!* You take Joe Rogan first and you're committing to a lot of hours of listening, and I mean a lot. That's the only way to get that kind of value. A lot of this competition will be an effort play – who will be willing to tally all the probabilistic predictions and check lots of writing everywhere against a list of potential memes? Steelmanning is easier, since it's much harder to miss so long as you read your team's stuff at all.

After I wrote that, [the teams were posted](#). I'm on the Null Hypothesis, which is an excellent team name, plus you get to be on the Kling-approved Null Hypothesis Watch while looking for points. We got the number one pick at #4 (Scott Alexander), which I suspected was reasonably likely given he fell to 4th (it's possible we had Tyler Cowen, but even that seemed highly unlikely, and several later in the round also seemed implausible), and also is great. This definitely feels like someone took people they knew and were happy to follow closely, which makes sense as a strategy. Hence both Weinsteins, and the package of myself, Alexander and Yudkowsky both make sense, and also play into finding M point triggers within the team. I'm guessing I'll end up providing the first point for a lot of Ms from both of them. Tabarrok is similarly a free action and was a steal at #42, and Taleb is basically a giant M-point hunt – he's never scored an S point in his life, and when asked for a probability he'll either have a formula or tell you it's impossible to know the probability and also there's a 100% probability that you're an idiot, but you're sure to pick up Black Swan, Antifragile, Skin in the Game and so on. Collison and Thompson also feel like people such a reader would be happy to keep tabs on. Overall, there's a lot to like, and good reason to think the team will be well-monitored, but I do think there are some potential blank spots. I definitely like our chances to win (B), which I'd bump to 35%, and I'd keep our winning chances around 20%. Tim the Enchanter's team looks strong too. Clan Graham seems strong in very memes but not well-rounded elsewhere. I'm not sure what to make of the teams that are mostly people I don't recognize – presumably it's right to be skeptical there.

In Other News

Very Serious People are back in charge of policy, so we can neither move at [nor call any operation Warp Speed](#).

[FDA approves three rapid tests for home use](#), two without a prescription. Better late and crippled than never.

The Biden Administration has noticed that the massive government vaccination sites tend not to be as useful as the existing infrastructure of pharmacies. So far so good, [but then the implied intervention is to do less rather than do more](#). I will never understand why 'costly' gets to be an adjective in such descriptions given the benefits at stake.

1. Politico: The Biden administration is rethinking a costly system of government-run mass vaccination sites after data revealed the program is lagging well behind a much cheaper federal effort to distribute doses via retail pharmacies....The vaccination hubs, which are run by FEMA and staffed in part by National Guard troops and other Pentagon personnel, have administered...about 67,000 shots a day, according to a series of internal FEMA briefing documents and data sets obtained by POLITICO....By comparison, the federal retail pharmacy program reported March 11 it had administered nearly 1 million doses over a single day.

Using the retail pharmacies is what Scott Duke Kominers and I argued for in mid-February in our piece titled, America's Pharmacies Can Do a Lot More Vaccinations. Good to see the Biden administration is making adjustments. Nothing wrong with the clinics, by the way, only use the pharmacies more.

There's a take on the whole AstraZeneca announcement mess last week that it was about AZ not properly respecting the authority of the DSMB, and thereby doing things 'the wrong way,' rather than any substantive disagreement (e.g. 76% vs. 79% is a small disagreement anyway), and they got called out because the people calling them out were disrespected. I don't buy it, and while it would be somewhat mitigating it doesn't make AZ's actions not supremely stupid, because it's playing with fire where you can't accomplish anything with it:



Helen Branswell ✅ @HelenBranswell · Mar 24

#AstraZeneca's US trial results, a play in 4 acts:

Act 1: Our vaccine is 79% efficacious — AZ.

Act 2: You're not reporting on all the data. More like 69% to 74% — DSMB & NIH.

Act 3: Final analysis is 76% — AZ.

Act 4: FDA reviews the raw data.

statnews.com/2021/03/24/pus...

It still seemed necessary to share the alternate hypothesis.

Washington Post notices that sometimes people lie to get life-saving medicine earlier, especially when there's zero probability they would ever be caught, and frames this as something that is 'damaging friendships.'

Quarantine procedures to countries that aren't doing suppression are not about preventing Covid-19 (official link). Two week quarantine for *fully vaccinated* people.



caseybmulligan @caseybmulligan · 9h

...

This must be from the Onion, not from an official government site.

travel.gc.ca/travel-covid/t...

Planning your mandatory quarantine

First, determine whether you can enter Canada.

Find out if you can travel to Canada

Federal quarantine applies for travellers entering Canada. If you can enter Canada and **you have no symptoms**, you must **quarantine** for a minimum of 14 days.

At this time, you are not excluded from quarantine, even if you have:

- been vaccinated for COVID-19
- tested negative for COVID-19
- recovered from COVID-19

Flying into Canada – your quarantine period includes a mandatory 3 night pre-paid booking at a government-authorized hotel at your own cost.

[Zeynep piece in The Atlantic laying out our situation](#), solid presentation, nothing especially new, the race between vaccinations and the fourth wave coming from a new deadlier and more infectious strain, and all that. I do think this leans a little hard into the ‘if we move fast we win if we move slow we lose’ thing, especially given the late hour and how little variance is actually in play at this point, but mostly that seems fine.

She also links to [one of several calls for a ‘vaccine surge.’](#) The logical idea is that if there are some places with high infection rates, especially when dominated by newer and deadlier strains, we should direct our vaccine supply to those areas first, because it will have a higher impact there. As a first best policy this would obviously be correct, but opening up the floodgates of which states are ‘most deserving’ would be such a disaster I don’t even want to consider it. Allocation by population rather than politics and power, even if there are real needs not being met, as the lesser of two disasters.

I too strongly endorse [the strategy of yelling “Fix It.”](#) Rather than issue rules and attempt micromanagement, if there’s something that needs to get done and isn’t getting done, because there are a bunch of veto points stopping it for reasons that obviously aren’t any good, it makes sense to try activating everyone’s blame-avoidance programming, and trigger them into physical-world mode, by yelling “FIX IT!” as loudly as possible. Pick an outcome, pick a date:

That's basically what happened in Washington. Inslee declined to sort through any of the details of opening up schools and simply said "You will do this thing. Solve the problem by April 19th. I don't care about your other plans. Get it done."

And they did. All the previous plans to painfully, erratically, incrementally open schools were tossed into the trash. We we were doing 2 half-days a week for elementary students and, for middle and high schools, no in-person class at all. That plan was scrapped and our school district said "OK. the old plan is trash, we're doing in-person classes, four full-days per week."

I feel like the same FIX IT plan is what President Biden put in place for vaccine availability. He simply said "Every state make the vaccines available to everyone by May 1st. Period. I don't care what your other plans were. Fix it."

And then last week Washington announced that they would make vaccines available for everyone by May 1st.

My big takeaway from this is that most bureaucratic objections are not intended to help solve a problem but to delay solving it. We have been told for months that we need to ease back into full in-person schooling and that we couldn't rush the process. Now I see that we absolutely can rush the process. We don't need incremental steps. We can simply say "this is what is happening" and make it happen.

While I'm linking to that post, I'll give my reaction to its first section as well, where he points out that there's no correlation between the minimum age requirement and what percentage of the elderly are vaccinated, and respond that the places doing better or expecting to do better with vaccinating the elderly were then in a better position to expand eligibility. So I agree that singling out states and obvious problems isn't useful, I don't think this particular non-finding says what he thinks it says.

Cuomo [finally joins the crowd and sets a date for full eligibility:](#)



Andrew Cuomo
@NYGovCuomo

...

#BREAKING: Starting Tuesday, April 6 at 8am, all New Yorkers age 16+ will be eligible to schedule and receive the COVID-19 vaccines.

And beginning tomorrow at 8am, all New Yorkers age 30+ will be eligible to schedule and receive the vaccines.

Let's [#VaccinateNY](#)

12:58 PM · Mar 29, 2021 · Twitter Web App

The one week delay for the 16-29 year olds is quite smart. Every time you make a lot of people eligible at once, you reliably cause a system overload and make it super hard to get an appointment that's at all reasonable. By giving the 30-49 year olds a week's head start, Cuomo gives them the chance to book as far out as they want, then the kids can get in line behind them. I heartily approve, and I think not doing this in other states was a missed opportunity. If anything, I would suggest going to 40 before (or instead of) 30, but this still does most of the work that needs doing.

[This is not a coincidence](#) because nothing is ever a coincidence, even though it's definitely a coincidence, or is it:

 Stop de kindermoord Retweeted

 **Mike Madden**  @MikeMadden 1d

So wait. First, Pfizer discovered its extra doses in each vial right around Chanukah, and then, the giant cargo ship was freed from captivity in Egypt in the middle of Passover? Who put my Hebrew school teachers in charge of writing the plot all of a sudden?

💬 112  2k ❤️ 18k ...

You think your attitude on social media can't save lives? [Behold:](#)

Some clinics have ended up throwing away supplies. A woman in her 30s told the *Observer* how she had been turned away after a friend working at a clinic in East Anglia said that they needed to find people for five leftover **Pfizer** doses. “When I got there I said, quite straightforwardly, that I’d been told there were some doses left over. The woman looked at me and said ‘how’s that going to look on social media?’ It was as if it was quite unethical for me to even be there.

“But afterwards my friend told me those five doses were wasted. I completely understand that I’m not in a high priority group. I wouldn’t want to push my way in front of anybody. But I do think I’m worth more than a bin.”

Reverse the associations of potential minor social awkwardness and vague blameworthiness, and those five doses get used. The [source article from Guardian](#) is mainly about the ‘controversy’ around providers turning to whoever is around to ensure doses aren’t wasted.

It then ends by pointing out that the UK is at a point where it needs to loosen eligibility requirements, as proven by inability to fill appointments, yet the requirements aren’t changing:

NHS England told GPs in January to draw up reserve lists of patients who could attend clinics at short notice, and there has been remarkably little wastage. A freedom of information request revealed that of 156,262 Pfizer doses used before 8 January, only 1,055 shots were wasted.

But with the [restriction on supplies of AstraZeneca](#) vaccine in April and no permission yet to go beyond cohort nine, GPs are now wondering how to deal with further leftover doses.

“We’re not allowed to go into cohort 10,” GP running a vaccination centre said. “We have actually had to stop running clinics. If another practice needs a few vials, we can help. But to run a viable clinic, I need more than 300 patients.

“When we first started, the slots were booked in two hours. Now we’re sending them out and we haven’t got enough. If we could go into cohort 10, there are lots of people waiting. It’s quite frustrating.”

[Potentially good advice](#), making lemonade edition:



Major Hayden @majorhayden · Mar 29

...

So after stressing all day over trying to get a vaccine appointment, someone tells me: "Just go find the most Republican county near you and look them up."

Found ~ 150 slots in one county next to mine and ~ 100 in another. Signed up in less than 5 minutes.

49

381

1.7K



Finally, I got to go back to New York City this week to get my second shot of Pfizer. I have a small amount of soreness in my arm and spent a day not feeling quite 100%, but it definitely wasn’t the whammy that I’ve heard others experienced, and I’ve been able to work and write this column in its aftermath with little trouble. I decided not to try and see people because waiting two weeks allows me to not worry about conditions at all, and I’ve found that such worries distract from my ability to enjoy seeing people, so better to use the trip as a more solitary one this time, and connect with people mid-April. If you’re in New York City and would like to see me next time I’m there, please do drop me a line.

Agents Over Cartesian World Models

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

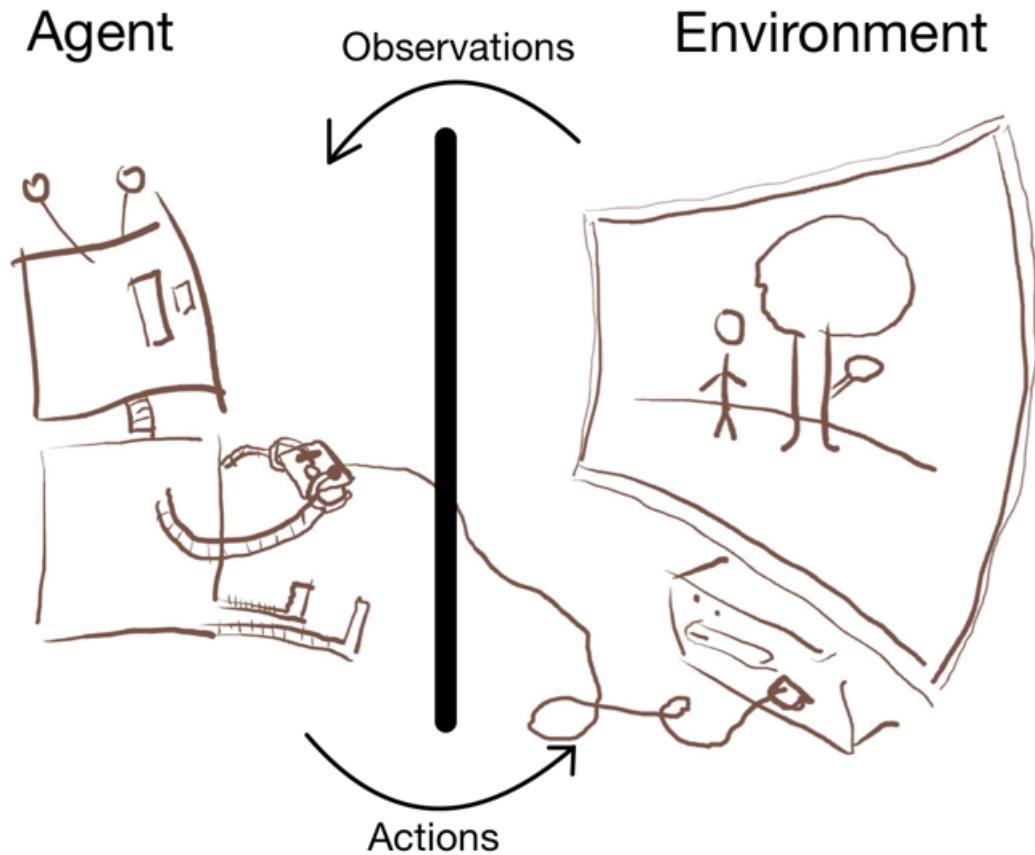
Thanks to Adam Shimi, Alex Turner, Noa Nabeshima, Neel Nanda, Sydney Von Arx, Jack Ryan, and Sidney Hough for helpful discussion and comments.

Abstract

We analyze agents by supposing a Cartesian boundary between agent and environment. We extend partially-observable Markov decision processes (POMDPs) into *Cartesian world models (CWMs)* to describe how these agents might reason. Given a CWM, we distinguish between consequential components, which depend on the consequences of the agent's action, and structural components, which depend on the agent's structure. We describe agents that reason consequentially, structurally, and conditionally, comparing safety properties between them. We conclude by presenting several problems with our framework.

Introduction

Suppose a Cartesian boundary between agent and environment:^[1]



There are four types: actions, observations, environmental states, and internal states. Actions and observations go from agent to environment and vice-versa. Environmental states are on the environment side, and internal states are on the agent side. Let A, O, E, I refer to actions, observations, environmental states, and internal states.

We describe how the agent interfaces with the environment with four maps: observe, orient, decide, and execute.^[2]

- observe : $E \rightarrow \Delta O$ describes how the agent observes the environment, e.g., if the agent sees with a video camera, observe describes what the video camera would see given various environmental states. If the agent can see the entire environment, the image of observe is distinct point distributions. In contrast, humans can see the same observation for different environmental states.
- orient : $O \times I \rightarrow \Delta I$ describes how the agent interprets the observation, e.g., the agent's internal state might be memories of high-level concepts derived from raw

- data. If there is no historical dependence, orient depends only on the observation. In contrast, humans map multiple observations onto the same internal state.
- $\text{decide} : I \rightarrow \Delta A$ describes how the agent acts in a given state, e.g., the agent might maximize a utility function over a world model. In simple devices like thermostats, decide maps each internal state to one of a small number of actions. In contrast, humans have larger action sets.
 - $\text{execute} : E \times A \rightarrow \Delta E$ describes how actions affect the environment, e.g., code that turns button presses into game actions. If the agent has absolute control over the environment, for all $e \in E$, the image of $\text{execute}(e, \cdot)$ is all point distributions over E . In contrast, humans do not have full control over their environments.

We analyze agents from a *mechanistic* perspective by supposing they are maximizing an explicit utility function, in contrast with a behavioral description of how they act. We expect many training procedures to produce mesa-optimizers that use explicit goal-directed search, making this assumption productive.^[3]

Consequential Types

We use four types of objects (actions, observations, environmental states, and internal states) and four maps between them (observe, orient, decide, and execute) to construct a world model. The maps are functions, but functions are also types. We will refer to the original four types as *consequential types* and the four maps as *structural types*.

We can broadly distinguish between four type signatures of utility functions over consequential types, producing four types of consequential agents.^[4]

- **Environment-based consequential agents** assign utility to *environmental states*. Most traditional agents are of this type. Examples include [the Stamp Collector](#), a paperclip maximizer, and some humans, e.g., utilitarians that do not value themselves.
- **Internal-based consequential agents** assign utility to different *internal states*. Very few "natural" agents are of this type. Examples include meditation bot, which cares only about inner peace, happiness bot, which cares only about being happy, and some humans, e.g., those that only value their pleasure.
- **Observation-based consequential agents** assign utility to different *observations*. Many toy agents have bijective observe functions and could be either observation-based or environment-based. Examples include virtual reality bot, which wants to build itself a perfect VR environment, video game agents that value the observation of the score, and some humans, e.g., those that would enter the [experience machine](#).
- **Action-based consequential agents** assign utility to different *actions*. Very few "natural" agents are of this type. Examples include twitch bot, which just wants to twitch, ditto bot, which wants to do whatever it did previously, and some types of humans, e.g., deontologists.^[5]

Some agents have utility functions with multiple types, e.g., utilitarian humans with deontological side constraints are environment/act-based consequential agents. Some humans value environmental states and personal happiness, making them environment/internal-based consequential agents. Question answering agents value different answers for different questions, making them internal/act-based consequential agents.

Agents could also have utility functions over structural types. We defer discussion of such structural agents until after we have analyzed consequential agents.

Consequential Agents

Cartesian world models

We modify discrete-time partially observable Markov decision processes (POMDPs) by removing the reward function and discount rate and adding internal states. Following Hadfield-Menell et al., who call a Markov decision process (MDP) without reward a *world model*, we will refer to a discrete-time POMDP without reward (POMDP\|R) and with internal states as a *Cartesian world model* (CWM).^[6]

Formally, a Cartesian world model is an 7-tuple $(E, O, I, A, \text{observe}, \text{orient}, \text{execute})$, where

- E is the set of environmental states (called "states" in POMDP\|Rs),
- O is the set of observations the agent could see (also called "observations" in POMDP\|Rs),
- I is the set of internal states the agent could have (not present in POMDP\|Rs),
- A is the set of actions available to the agent (also called actions in POMDP\|Rs),
- $\text{observe} : E \times A \rightarrow \Delta O$ is a function describing observation probabilities given environmental states (called "conditional observation probabilities" in POMDP\|Rs),
- $\text{orient} : O \times I \rightarrow \Delta I$ is a function describing internal state probabilities given observations (not present in POMDP\|Rs),
- $\text{execute} : E \times A \rightarrow \Delta E$ is a function describing transition probabilities between different states of the environment given different actions (called "conditional transition probabilities" in POMDP\|Rs),

At each time period, the environmental state is in some $e \in E$, and the agent's internal state is some $i \in I$. The agent decides upon an action $a \in A$, which causes the environment to transition to state e' sampled from $\text{execute}(e, a)$. The agent then

receives an observation $o \in O$ sampled from $\text{observe}(e', a)$, causing the agent to transition to internal state i' sampled from $\text{orient}(o, i)$.

This produces the initial 4-tuple of context $c_0 := (e, i, a, o)$ representing the initial environment's state, the agent's initial internal state, the agent's initial action, and the agent's initial observation. Subsequent time steps t produce additional 4-tuples of context c_t .

MDPs are traditionally used to model decision-making situations; agents trained to achieve high reward on individual MDPs implement policies that make "good" decisions. In contrast, we intend CWMs to capture how agents might model themselves and how they interact with the world.

Consequential Decision Making

Agents making decisions over CWMs attempt to maximize some utility function. To emphasize that agents can have utility functions of different types, we make the type explicit in our notation. For example, a traditional agent in a POMDP will be maximizing expected utility over an environment-based utility function, which we will denote

$U_{\{E\}}(e, t)$, with the second parameter making explicit the time-dependence of the utility function.

More formally, let T be the set of possible consequential type signatures, equal to $P(\{E, O, I, A\})$. For some $T \in T$, let $U_T : T \times N \rightarrow R$ be the agent's utility function, where T can vary between agents. Recall that $c_t := (e, i, a, o)$ is at the 4-tuple of CWM context at time t . Let $c_t|_T$ be c_t restricted to only contain elements of types in T . A consequential agent maximizes expected future utility, which is equal to

$$E[\sum_{t=0}^{\infty} U_T(c_t|_T, t)].$$

U_T 's time-dependence determines how much the agent favors immediate reward over distant reward. When $t > 0 \implies U_T(c_t|_T, t) = 0$ the agent is *time-limited myopic*; it takes actions that yield the largest immediate increase in expected utility. When $U_T(c_t|_T, t) \approx U_T(c_t|_T, t + 1)$ the agent is *far-sighted*; it maximizes the expected sum of all future utility.

Examples

Example: Paperclip maximizer

Consider a paperclip maximizer that can only interface with the world through a computer. Define a Cartesian world model (E, O, I, A , observe, orient, execute) as follows:

- E is all ways the universe could be,
- O is all values a 1080×1920 monitor can take,
- I is the set of internal states, which we leave unspecified,
- A is the set of keycodes plus actions for mouse usage,
- observe maps the environment to the computer screen; this will sometimes be correlated with the rest of the environment, e.g., through the news,
- orient maps the computer screen to an internal state, which we leave unspecified,
- execute describes how keycodes and mouse actions interact with the computer.

Additionally, let our agent's utility function be $U_{\{E\}}(e, t) = \text{\text{the number of paperclips in } e}$.

Many of these functions are not feasible to compute. In practice, the agent would be approximating these functions using abstractions, similar to the way humans can determine the consequences of their actions. This formalism makes clear that paperclip maximizers have environment-based utility functions.

Example: twitch-bot

Twitch-bot is a time-limited myopic action-based consequential agent that cares only about twitching. Twitch-bot has no sensors and is stateless. We can represent twitch-bot in a CWM+U as follows:

- E is all ways the universe could be,
- $O = \{\emptyset\}$; twitch-bot can only receive the empty observation,
- $I = \{\emptyset\}$; twitch-bot has only the empty state,
- $A = \{\text{twitch}, \emptyset\}$; twitch-bot has only two actions, twitching and not twitching,
- observe always gives \emptyset ,
- orient always gives \emptyset ,
- execute describes how twitching affects the world,

Additionally, let our agent's utility function be $U_{\{A\}}(\text{twitch}, 0) = 1$, with $U_{\{A\}}(a, t) = 0$ otherwise.

The optimal decide for this CWM+U always outputs twitch.

Example: Akrasia

Humans sometimes suffer from akrasia, or weakness of will. These humans maximize expected utility in some mental states but act habitually in others. Such humans have utility functions of type {E, I, A}; in some internal states, the human is an environment-based consequential agent, and in other internal states, the human is an action-based consequential agent.

Relation to Partially Observable Markov Decision Processes

We desire to compare the expressiveness of CWMs to POMDPs. Since we want to compare CWMs to POMDPs directly, we require that the agents inside each have the same types. We will call a POMDP P equivalent to CWM C if there exists a utility function U such that the agent optimal in P is optimal with respect to U in C and vice-versa.

Partially Observable Markov Decision Processes \subseteq Cartesian World Models

Given a POMDP, we can construct a *Cartesian world model + utility function* (CWM+U) that has an equivalent optimal consequential agent. Let $(S', A', T', R', \Omega', O', \gamma)$ be a POMDP, where

- S' is a set of states,
- A' is a set of actions,
- T' is a set of conditional transition probabilities between states,
- $R' : S' \times A' \rightarrow \mathbb{R}$ is the reward function,
- Ω' is the set of observations,
- O' is the set of conditional observation probabilities, and
- $\gamma \in [0, 1]$ is the discount factor.

Recall that an optimal POMDP agent maximizes $E[\sum_{t=0}^{\infty} \gamma^t r_t]$, where r_t is the reward earned at time t .

Let our Cartesian world model be a 7-tuple $(E, O, I, A, \text{observe}, \text{orient}, \text{execute})$ defined as follows:

- $E = S'$,
- $O = \Omega'$
- $I = \Delta S'$, the set of belief distributions over S' ,
- $A = A'$,
- $\text{observe} = O$,
- $\text{orient}(o, i)$ describes the Bayesian update of a belief distribution when a new observation is made, and
- $\text{execute} = T'$

Additionally, let our agent's utility function be $U_{\{E,A\}}(e, a, t) = R'(e, a)\gamma^t$.

Since a POMDP has the Markov property for beliefs over states, an agent in the CWM+U has information as an agent in the POMDP. Consequential agents in CWMs maximize

$$E[\sum_{t=0}^{\infty} U_T(c_t|_T, t)]. \text{ Substituting, this is equivalent to } E[\sum_{t=0}^{\infty} U_{\{E,A\}}(c_t|_{\{E,A\}}, t)].$$

$U_{\{E,A\}}(c_t|_{\{E,A\}}, t) = R'(e_t, a_t)\gamma^t = r_t\gamma^t$, so $E[\sum_{t=0}^{\infty} U_T(c_t|_T, t)] = E[\sum_{t=0}^{\infty} r_t\gamma^t]$. Thus our agents are maximizing the same quantity, so an agent is optimal with respect to the CWM+U if and only if it is optimal with respect to the POMDP.

Cartesian World Models $\not\subseteq$ Partially Observable Markov Decision Processes

If we require the CWM agent to have the same type signature as in the POMDP, some CWMs do not have equivalent POMDPs. Agents in CWMs map internal states to actions, whereas agents in POMDPs map internal belief distributions over states into actions. Therefore, agents optimal in a POMDP have infinite internal states. Since one can construct a CWM with finite internal states, there must exist a CWM that cannot be converted to an equivalent POMDP.

This non-equivalence is basically a technicality and thus unsatisfying. A more satisfying treatment would use a less rigid definition of equivalence that allowed the CWM agent and the POMDP agent to have different types. In particular, it might be possible to construct a POMDP and a partition over belief states such that the elements of the partition can be put in correspondence with the internal states in the CWM. We leave exploration of this question to future work.

Structural Agents

Structural Decision Making

In contrast to consequential agents, structural agents have utility functions over structural types. We model structural agents as reasoning about a CWM with a modified decision procedure. In this framework, the agent is not trying to select utility-maximizing actions but instead trying to enforce utility-maximizing relations between consequential types. The agent is optimizing over the set of possible decide functions instead of the set of possible actions. Note that an agent optimizing in this way could still have utility functions of multiple types in the same way that an environment-based consequential agent still optimizes over its possible actions.

Let a Cartesian World Model plus Self (CWM+S) be a 8-tuple $(E, O, I, A, \text{observe}, \text{orient}, \text{execute}, \text{decide})$, where the first 7 are a CWM and the last entry is the decide function of an agent. Let CWM + S be the set of all CWS+Ss.^[7] The utility function of a structural agent is a function $U : \text{CWM} + \text{S} \rightarrow \mathbb{R}$ that assigns utility to each CWS+S.

Let C be a CWM and $\text{decide} : I \rightarrow A$ be a decision function. Let $C + \text{decide}$ be the CWM+S that agrees with C for the first 7 entries and is equal to decide for the 8th. This construction omits acausal implications of changing decide . A method of constructing CWM+Ss that include acausal implications is currently an open problem.

Recall that A^I is the set of all functions from I to A. A structural agent reasoning according to a CWM C acts to implement $\text{decide}^* = \arg \max_{\text{decide} \in A^I} U(C + \text{decide})$.

Behaviorally, when given an internal state i, an optimal structural agent will take $a = \text{decide}^*(i)$

Examples

Pseudo-example: Updateless Decision Theory

An agent using updateless decision theory (UDT) is a structural agent with a consequential utility function over environmental states reasoning over a CWM+S that includes acausal implications. In order to convert the consequential utility function U_b into a structural one U_s , we simply define $U_s(C + \text{decide})$ by rolling out $C + \text{decide}$ to produce a sequence e_0, e_1, e_2, \dots of environmental states and let

$$U_s(C + \text{decide}) = \sum_{t=0}^{\infty} U_b(e_t) \lambda^t \text{ with some discount factor } \lambda.$$

Example: Structural HCH-bot

In the limit of training, [imitative amplification](#) produces [Human consulting HCH](#)(HCH). We describe a structural agent that is implementing HCH.^[8] The agent receives input via computer terminal and outputs text to the same terminal. It has 1 Mb of memory. In the following, let Σ be the alphabet of some human language, including punctuation and spaces:

- E is all ways the universe could be,
- $O = \Sigma^*$, all finite strings from Σ ,
- $I = P(\{n \mid 0 \leq n < 8 \times 10^6\})$, the set ways 1 Mb of memory can be,
- $A = \Delta\Sigma^{<1000}$, the set of probability distributions over strings from Σ of less than 1000 characters,
- observe yields the command typed in at the computer terminal with probability ~ 1 ,
- orient stores the last 1 Mb of the string given by observe,
- execute describes the world state after outputting a given string at the terminal.

Let P be some distribution over possible inputs. Let $HCH : I \rightarrow \Delta A$ be the HCH function.

Let our agents utility function U be defined as $U(\text{decide}) = E_{i \sim P}[\text{KL}(HCH(i) \parallel \text{decide}(i))]$.
[\[9\]](#)

It is unclear which P makes the agent behave properly. One possibility is to have P be the distribution of what questions a human is likely to ask.^[10] Any powerful agent likely has a human model, so using the human distribution might not add much complexity.

Example: Structural Decoupling

A consequential [approval-maximizing agent](#) takes the action that gets the highest approval from a human overseer. Such agents have an incentive to tamper with their reward channels, e.g., by persuading the human they are conscious and deserve reward.

In contrast, a structural approval-maximizing agent implements the decide function that gets the highest approval from a human overseer. Such agents have no incentive to *directly* tamper with their reward channels, but they still might implement decision functions that appear safe without being safe. However, a decide function that overtly manipulates the overseer will get low approval, so structural approval-maximizing agents avoid parts of the reward tampering problem.

This example is inspired by the decoupled approval described in Uesato et al.^[11]

Types of Structural Agents

There are roughly four types of consequential agents, one for each consequential type. This correspondence suggests there are four types of structural agents, one for each structural type.

Agents with utility functions over decide are coherent. However, since we do not include acausal implications when constructing a CWM+S, agents with utility functions over orient, observe, or execute have constant utility. More specifically, the agent only has control over its own decide function, which does not have influence over orient, observe, or execute (within the CWM+S), so agents with utility functions over those types will not be able to change anything they care about. How structural agents act when we include acausal implications is currently an open problem.

Utility/Decision Distinction

Besides having utility functions with different type signatures, structural agents also make decisions differently. We have two dimensions of variation: structural versus consequential utility functions and structural versus consequential decision-making. These dimensions produce four possible agents: pure consequential, pure structural, decision-consequential utility-structural, and decision-structural utility-consequential.

A pure consequential agent makes consequential decisions to maximize a consequential utility function; it reasons about how taking a certain action affects the future sequence of consequential types. A purely consequential environment-based agent takes actions to maximize $\sum_{t=0}^{\infty} U(e_t)\gamma^t$ for some discount factor γ .

A pure structural agent makes structural decision to maximize structural utility function; it reasons about how implementing a certain decide affect the structural types of its CWM+S. A purely structural decide-based agent implements the decide function to maximize $U(C + \text{decide})$, where C is the agent's CWM.

A decision-consequential utility-structural agent makes consequential decisions to maximize a structural utility function; it reasons about how taking a certain action affects how it models the world. For example, a decision-consequential utility-structural orient-based agent might rewrite its source code. If decision-consequential utility-structural agents are not time-limited myopic, they will take over the world to securely achieve desired CWM structural properties.^[12]

A decision-structural utility-consequential agent makes structural decisions to maximize a consequential utility function. If only causal implications are included, a decision-structural utility-consequential agent behaves identically to a purely consequential agent. If we include acausal implications, decision-structural utility-consequential agents resemble UDT agents. Traditional UDT agents are decision-structural utility-consequential environment-based agents.

Pure Structural decide-based Agents = Time-Limited Myopic Action/Internal-based Behavioral Agents

Pure structural decide-based agents can be expressed as time-limited myopic action/internal-based consequential agents and vice versa. Let decide^* be optimal according to our pure structural agent's utility function. To construct an action/internal-based consequential utility function for which decide^* is optimal, define U such that

$\forall i \in I, a \in A : U(i, a) = 1 \iff \text{decide}^*(i) = a$ and 0 otherwise. To show the inverse, construct a structural utility function maximal utility to the time-limited myopic action/internal-based consequential agent's decision function.

These agents are behaviorally identical but mechanistically distinct; they use different decision mechanisms and have different types of utility functions.

Connection to Act-Based Agents

[Act-based agents](#) focus on satisfying users' short-term instrumental preferences. These agents might be safe insofar as learning short-term preferences naturally avoids catastrophic generalization. For instance, learning that killing is bad might be easy, allowing weak agents to avoid catastrophe.

Paul Christiano postulates a gradient of agents that satisfy "narrow" preferences to "broad" preferences:

Consider a machine choosing a move in a game of chess. I could articulate preferences over that move (castling looks best to me), over its consequences (I don't want to lose the bishop), over the outcome of the game (I want to win), over immediate consequences of that outcome (I want people to respect my research team), over distant consequences (I want to live a fulfilling life).

In contrast, our framework makes a sharp distinction between agents that use consequential versus structural decision-making. The above spectrum is composed of consequential agents with differing amounts of time-limited myopia. Such agents are unsafe because they would take over the world if they could do so fast enough. In general, all agents that reason over their actions' consequences are dangerous because we do not know reality's causal structure. We hope that structural agents provide a step towards constructing agents that do not use intrinsically dangerous decision-making procedures.

Input Distribution Problem

decide-based structural agents attempt to implement specific decide functions, which will often require determining the distance to a target function, e.g. the HCH function. Unless the distance metric treats all inputs identically, such as with sup norm, the metric requires a distribution over inputs. For instance, recall structural HCH-bots's

utility function, $U(\text{decide}) = E_{i \sim P}[\text{KL}(HCH(i) || \text{decide}(i))]$, depends on an input distribution P .

However, the distribution of inputs depends on how the agent responds to various inputs, creating an acausal implication between the agent's actions and what inputs it receives. For example, what you google depends on Google's capabilities. Since the agent's utility depends on the inputs it receives, this acausal implication incentivizes the agent to implement a decide function that shifts the distribution towards high-scoring inputs. The agent is not optimizing over decide functions, but rather jointly optimizing over decide, input distribution pairs. More concretely, the agent has an incentive to hide its full capabilities so it will not be asked difficult questions. If the agent can only answer questions with obvious answers, it will probably be asked questions with easy answers, acausally shifting P to a higher utility distribution. [\[13\]](#)

This acausal implication reduces capabilities, but it also might be a problem for alignment. The capabilities hit from acausally optimizing the input distribution does not appear to intrinsically produce alignment failures, but the problem arises only when the agent thinks of itself as having logical control over other instances of itself. This pattern of reasoning potentially results in deceptive alignment arising through acausal means. [\[14\]](#) In general, any agent that has uncertainty over its input might be able to acausally influence the input distribution, potentially resulting in undesirable behavior.

Conditional Agents

Conditional Decision Making

Traditional utility functions map types, e.g., environmental states, to utility. In contrast, conditional utility functions map types to utility functions. For example, an environment-conditional utility function takes in an environmental state and yields a utility function over other environmental states, actions, observations, internal states, etc. We will refer to the utility function given by a conditional utility function as the base utility function.

Conditional agents might make decisions in the following way. Let U be a conditional utility function. Upon having internal state i , the agent acts as if it has utility function $U = U(\text{argmax}_{s \in S} P(s|i))$, where S varies between E , O , I , and A depending on whether U is environmental, observational, or structural conditional. The agent reasons as if it were a structural or consequential agent depending on the utility function. [\[15\]](#) Action-based agents might run into issues around logical uncertainty.

Examples

Example: Value Learner

A simple value-learning agent observes a human and infers the human utility function, which the agent then optimizes. Ignoring the issues inherent in inference, such an agent can be thought of as having a conditional utility function that maps observations to utility functions over environmental states, i.e., of type $O \rightarrow (E \times N \rightarrow R)$ (Recall that we include explicit time-dependence to allow for arbitrary discounting possibilities).

Shah et al.'s [Preferences Implicit in the State of the World](#) attempts to construct agents that infer human preferences based on the current environmental state. In our framework, these agents have conditional utility functions that map environmental states to utility functions over environmental states, i.e., of type $E \rightarrow (E \times N \rightarrow R)$.

Example: Argmax

Given a CWM, argmax takes in a utility function and outputs the action that maximizes that utility function. Since argmax only has access to its internal representation of the utility function, we can think of argmax as having a conditional utility function that maps internal states to utility functions over all types, i.e. of type $I \rightarrow (A \times O \times E \times I \times N \rightarrow R)$.

Example: Imprinting-bot

Imprinting-bot is an agent that tries to imitate the first thing it sees, similar to how a baby duck might follow around the first thing it sees when it opens its eyes. Such an agent can be thought of as having a conditional utility function that maps observations to utility functions over decide functions, i.e., of type $O \rightarrow (A^I \rightarrow R)$.

Example: Conditional HCH-bot

Recall that $HCH : I \rightarrow \Delta A$ is the HCH function. Let $HCH(i)$ be the distribution of actions HCH would take given internal state i . Let $HCH(i)(a)$ be the probability that HCH takes action a when internal state i .

Structural HCH-bot gets higher utility for implementing decide functions that are closer to HCH in terms of expected KL-divergence relative to some input distribution. Conditional behavioral-HCH-bot conditions on the internal state, then gets utility for outputting distributions closer to the distribution HCH would output given its current internal state as input. More precisely, conditional behavioral-HCH-bot has a conditional utility function defined by

$$U : I \rightarrow (A \times N \rightarrow R) := i \mapsto ((a, n) \mapsto -\log(HCH(i)(a)) \text{ if } n = 0 \text{ else } 0).$$

Conditional structural-HCH-bot conditions on the internal state, then attempts to implement a decide function close to HCH. More precisely, conditional structural-HCH-bot has a conditional utility function described by

$$U : I \rightarrow (A^I \rightarrow R) := i \mapsto (\text{decide} \mapsto KL(HCH(i)||\text{decide}(i))).$$

Conditioning Type is Observationally Indistinguishable

Following a similar argument to previous sections, the type an agent conditions upon cannot be uniquely distinguished observationally. To see this, note that the only information a conditional agent has access to is the internal state, so it must back-infer observation/environmental states. Thus, internal-conditional utility functions can be constructed that mimic conditional utility functions of any conditioning type. Since back-inference might be many-to-one, the reverse construction is not always feasible.

However, there seem to be natural ways to describe certain conditional agents. For instance, one can consider a value learning agent that conditions upon various types. An environmental-conditional value learner looks at the world and infers a utility function. An observational-conditional value learner needs to observe a representation of the utility function. An internal-conditional value learner needs a representation of the utility function in its internal state. At each stage, more of the information must be explicitly encoded for "learning" to take place.

These agents can be distinguished by counterfactually different observe and orient maps. Back inference should happen differently for different observe and orient mappings, causing agents to potentially act differently. Environmental-conditional agents have different utility functions (and potentially take different actions) if either the observe or orient mappings differed. Observation conditional agents are stable under changes to observe, but act differently if orient differed. Internal conditional agents should not care if either observe or orient differed.

Doing back inference at any point opens the door to the input distribution problem; only internal-conditional agents do not back-infer.

Structural Conditional Utility Functions

In addition to consequential conditional utility functions, there are also structural conditional utility functions. Instead of conditioning on a particular environmental state, an agent can condition upon execute, e.g. by having a conditional utility function of type $E^A \rightarrow (E \times N \rightarrow R)$. Such structural conditional agents have utility functions that depend on their model of the world.

For instance, Alice might have a utility function over 3D environmental states. However, suppose that Alice found out the world has four dimensions. Alice might think

she gets infinite utility. However, Alice's utility function previously mapped 3D environmental states to utilities, so 4D states produce a type error. For Alice to get infinite utility, a 4D state's utility must generalize to be the infinite sum of all the 3D states it's made of. Instead, Alice might get confused and start reasoning about what utility function she should have in light of her new knowledge. In our framework, Alice has a structural-conditional utility function, i.e., she assigns utilities to 3D environmental states conditional on the fact that she lives in a 3D world.

In general, structural conditional utility functions are resistant to ontological crises – such agents will switch to different base utility functions upon finding out their ontological assumptions are violated.

Type Indistinguishability

The type of a conditional utility function is often mathematically identical to the type of a consequential utility function. For example, a conditional utility function that takes internal states and gives a utility function over actions has type $I \rightarrow (A \times N \rightarrow R)$, which is mathematically equivalent to an internal/act consequential utility function $I \times A \times N \rightarrow R$. This equivalence is problematic because consequential utility functions do not possess many desirable safety properties.

We currently remain confused by the implications of this problem but sketch out a potential resolution. Classical Bayesianism accepts [dogmatism of perception](#), i.e., observations, once observed, are believed with probability one. Similarly, we might require that conditional agents accept dogmatism of conditioning, i.e., the agent must believe that the object they are conditioning on occurs with probability one. This requirement neatly solves the input distribution problem; even if the agent thought it had logical control of the input distribution, the input could not have been anything else.

Contrast this with one of Demski's desiderata in [Learning Normativity](#):

Learning at All Levels: Although we don't have perfect information at any level, we do get meaningful benefit with each level we step back and say "we're learning this level rather than keeping it fixed", because we can provide meaningful approximate loss functions at each level, and meaningful feedback for learning at each level. Therefore, we want to be able to do learning at each level.

Demski's motivation is that any given level might be corruptible, so learning should be able to happen on all of them. Our motivation is that if an input can be influenced, then it will, so the agent must think of it as fixed.

No Uncertainty Upstream of Utility

In general, many problematic situations seem to arise when agents have uncertainty that is logically upstream of the utility they ultimately receive.[\[16\]](#) An agent with uncertainty over the input distribution suffers from the input distribution problem.[\[17\]](#) An agent with uncertainty over how much utility its action will be power-seeking. We will call the former type of uncertainty *historical uncertainty* and the latter *consequential*

uncertainty. Agents also must be uncertain about what actions they will take, which we call logical uncertainty.

We are interested in agents that possess no historical or consequential uncertainty. Such agents might be described as maximizing utility directly instead of maximizing expected utility, as there is no uncertainty over which to take an expectation.

We suspect avoiding historical uncertainty is the only way to avoid the input distribution problem described above because if agents have uncertainty logically upstream of utility they eventually receive, then they have an incentive to acausally influence the distribution to get more utility. We also suspect that avoiding consequential uncertainty is a way to eliminate power-seeking behavior and undesirable outcomes like [edge instantiation](#).

The trick employed in Conditional Decision Making is to collapse the input distribution into the maximum probability instance. Thus an environmental-conditional agent will optimize the most probable base utility function instead of optimizing the weighted mixture represented by the probability distribution. This does not entirely avoid the input distribution problem, as the agent still has logical influence over what the maximum probability back-inference is, but it reduces the amount of logical control an agent has.

In practice, the agent will have logical control over parts of the input distribution. It remains an open question as to how to get an agent to *think* they do not have logical control.

Problem: Empirical Uncertainty

Agents with no historical or consequential uncertainty will still have uncertainty over parts of their world model. For instance, conditional HCH-bot will have uncertainty over its model of a human, which will translate into uncertainty as to what HCH would output for any given input. Since conditional HCH-bot's utility depends on how well it can approximate HCH, this means that there must be uncertainty that is logically upstream of the utility the agent receives. One might hope that the agent does not think of itself as having logical control over the human input-output mapping.

Drawing a natural boundary between these two types of uncertainty remains an open problem.

Problems

Cartesian boundaries are not real

One major problem with this way of conceptualizing agency is that Cartesian boundaries are part of the map, not the territory. In reality, there is no distinction between agent and environment. What happens when an agent discovers this fact is relatively unclear, although we think it will not cause capabilities to disintegrate.

Humans have historically thought they were different from the environment. When humans discovered they were made of atoms, they were still able to act. Anthropomorphizing, humans have empirically been robust to ontological crises, so AIs might also be robust.

Even if structural agents can continue acting upon discovering the Cartesian boundary is not real, there might be other undesirable effects. If the agent begins conceptualizing itself as the entire universe, a utility function over decide reduces to a utility function over environmental states; desirable properties of structural agents are lost. As another example, the agent could start conceptualizing "become a different type of agent" as an action, which might cause it to self-modify into an agent that lacks desirable properties.

In general, the way a structural agent models the CWM boundary might change the action set and the internal state set. Depending on how the agent's utility function generalizes, this might cause undesirable behavior.

Myopia might be needed

Purely consequential act/internal-based farsighted agents are incorrigible. If an approval-maximizing agent picks the action trajectory that maximizes total approval, it will avoid being shut down to gain high approval later. While structural agents avoid some of these problems, the highest utility decide function needs to be myopic, else the problem reappears. The base utility function output by a conditional utility function must also be myopia, else the agent will act to preserve its utility function across time.

This analysis suggests myopia might be necessary for safe agents, where myopic agents "only care about the current episode" and "will never sacrifice reward now for reward later." Currently, myopia is not understood well enough to know whether it is sufficient. Myopia also has a [number of open problems](#).

Training for types

If there are agents that possess desirable safety properties, how can we train them? The obvious way is to reward them for having that type. However, knowing an agent is optimal according to some reward function does not constrain the type of the agent's utility function. For instance, rewarding an agent for implementing a decision function that takes high approval actions is indistinguishable from rewarding the agent for creating environments in which its actions garner high approval.

Many agents are optimal for any reward, so the resulting agent's type will depend on the training process's inductive biases. In particular, we expect the inductive bias of stochastic gradient descent (SGD) to be a large contributing factor in the resulting agent. [18] We are eager for exploration into how the inductive biases of SGD interact with structural agents.

Given that behavior does not determine agent type, one could use a training process with *mechanistic incentives*. For instance, instead of rewarding the agent for *taking* actions, one can also reward the agent for *valuing* actions. We could also reward agents for reasoning over decision functions instead of actions or lacking uncertainty of certain types. This training strategy would require advances in transparency tools and a better mechanistic understanding of structural agents. [19]

Conclusion

If one supposes a Cartesian boundary between agent and environment, the agent can value four consequential types (actions, environments, internals, and observations) and four structural types (observe, orient, decide, and execute). We presented *Cartesian world models (CWMs)* to formally model such a boundary and briefly explored the technical and philosophical implications. We then introduced agents that explicitly maximize utility functions over consequential and structural properties of CWMs.

We compared consequential agents, structural agents, and conditional agents and argued that conditional agents avoid problems with consequential and structural agents. We concluded by presenting multiple problems, which double as potential directions for further research.

Appendix

These sections did not fit into the above structure.

The Two Wireheaders

Two types of consequential agents are "wireheading": internal-based and observation-based agents. Internal-based agents wirehead by changing their internal states, e.g., by putting wires in their brain. Observation-based agents wirehead by changing their observations, e.g., by constructing a virtual reality. Internal state-based agents also wirehead by changing their observations, but observation-based agents will not change their internal states (typically).

These two agents demonstrate "wireheading" does not [carve reality at its joints](#). In reality, internal and observation-based agents are maximizing non-environment-based utility functions. Internal-based or observation-based agents might look at environment-based agents in confusion, asking, "why don't they just put wires in their head?" or "have they not heard of virtual reality?"

Consider: are action-based agents wireheading? Neither yes nor no seems reasonable, which hints that wireheading is a confused concept.

Types are Observationally Indistinguishable

In most cases, it is impossible to determine an agent's type by observing its behavior. In degenerate cases, any set of actions is compatible with a utility function of any combination of {E, I, A, O}. [\[20\]](#) However, by employing appropriate simplicity priors, we guess the agent's approximate type.

As an analogy, all [nondeterministic finite automaton](#) (NFA) can be translated into [deterministic finite automaton](#) (DFA). What does it mean to say that a machine "is an NFA"? There is an equivalent DFA, so "NFA" is not a feature of the machine's input-output mapping, i.e., "being an NFA" is not a behavioral property.

Converting most NFA into DFA requires an exponential increase in the state-space. We call a machine an NFA if describing it as an NFA is simpler than describing it as a DFA.

Similarly, we might call an agent "environment-based" if describing its behavior as maximizing an environment-based utility function is simpler than describing it as maximizing a non-environment-based utility function.

Unlike NFAs and DFAs, however, one can construct degenerate CWMs that rule out specific type signatures. Suppose that the environment consisted only of a single switch, and the agent could toggle the switch or do nothing. If an optimal agent always toggled the switch, we could infer that the agent was not purely environment-based.[\[21\]](#)

In general, the [VNM Theorem](#) rules out some types of utility functions for some sequences of actions. If the agent can act to leave itself unchanged, loops of the same sequences of internal states rule out utility functions of type $\{I\}$. Similarly, loops of the same (internal state, action) pairs rule out utility functions of type $\{I\}$, $\{A\}$ and $\{I, A\}$. Finally, if the agent ever takes different actions, we can rule out a utility function of type $\{A\}$ (assuming the action space is not changing).[\[5:1\]](#)

However, to see that agent types cannot be typically distinguished behaviorally, note that an agent can always be expressed as a list of actions they would take given various observations and internal states. Given reasonable assumptions, one can construct utility functions of many types that yield the same list.[\[22\]](#) This flaw with behavioral descriptions justifies our intention that the CWM+U framework captures how the agent models its interactions with the world - a mechanistic property, not a behavioral one.

Example: human

Alice has a utility function over the environment. Alice's only way to obtain information is by observation, so we can construct a utility function over observations that results in indistinguishable behavior. Alice refuses to enter perfect VR simulations, so we suppose that Alice intensely disvalues the observation of entering the simulation. Observation-based Alice is more complicated than environment-based Alice but is compatible with observed behavior.

Observations are distinguishable only when they produce different internal states, so we can construct a utility function over internal states that results in indistinguishable behavior. Alice does not take drugs to directly alter the internal state, so we suppose they intensely disvalue the internal state of having decided to take drugs. Internal-based Alice is more complicated than environment-based Alice but is compatible with observed behavior.

Alice takes different actions, so we can rule out a pure action-based utility function. Since internal states are only distinguishable by what actions they cause Alice to take, we can construct an internal/action-based utility function that results in indistinguishable behavior. Alice is willing to sacrifice their life for a loved one, so we suppose they get immense utility from taking the action of saving their loved one's life when they believe said loved one is in danger. Internal/action-based Alice is more complicated than environment-based Alice but is compatible with observed behavior.

Natural descriptions

In the above, there seemed to be a "natural" description of Alice.^[23] Describing Alice's behavior as maximizing an environment-based utility function did not require fine-tuning, whereas an observation-based utility function had specific observations rate very negatively. We can similarly imagine agents whose "natural" descriptions use utility functions of other type signatures, the trivial examples being agents who model themselves in a CWM and maximize a utility function of that type.

It is essential to distinguish between a mechanistic and a behavioral understanding of CWMs. Under the mechanistic understanding, we consider agents that explicitly maximize utility over a CWM, i.e., the agent has an internal representation of the CWM, and we need to look inside the agent to gain understanding. Under the behavioral understanding, we are taking the [intentional stance](#) and asking, "if the agent internally modeled itself as being in a CWM, what is the simplest utility function it could have that explains its behavior?"

Attempting to understand agents behaviorally using CWMs will sometimes fail. For instance, a thermostat is a simple agent. One might describe it as an environment-based agent whose goal is to maintain the same temperature. However, the thermostat only cares about the temperature in a tiny region around its sensors. We could describe it as an observation-based agent that wants to observe a specific temperature. Nevertheless, there is a 1-1 correspondence between observations and internal states, so it seems equally accurate to describe the thermostat as internal-based. Finally, we can describe it as an internal/action-based agent that wants to increase the temperature when it is low but decrease it when it is high.

We strain to describe a thermostat as environment-based or internal/action-based. However, there is a 1-1 correspondence between observation and internal state, so it is equally simple to describe the thermostat as observation-based or internal-based.

It can also be unclear what type of utility function suboptimal agents have. Suppose an agent does not know they can fool their sensors. In that case, an observation-based agent will act the same as an environment-based agent, making it impossible to obtain a mechanistic understanding of the agent by observing behavior.

"Becoming Smarter"

What exactly does it mean for an agent to improve? Our four-map model allows us to identify four potential ways. In what follows, we assume the agent is farsighted. We also implicitly relax other optimality assumptions.

- observe: An agent could expand O or reduce the expected entropy of $\text{observe}(e, a)$. For example, an agent could upgrade its camera, clean the camera lens, or acquire a periscope. Many humans employ sensory aids, like glasses or binoculars.
- orient: An agent could both expand I and reduce the expected entropy of $\text{orient}(o, i)$. For example, an agent could acquire more memory or better train its feature detection algorithms. Many humans improve their introspection ability and memories by meditating or using [spaced repetition software](#).

- decide: An agent could implement a decide function that better maximizes expected future utility. For example, a chess-playing agent could search over larger game trees. Many humans attempt to combat cognitive biases in decision-making.
- execute: An agent could expand A or decrease the expected entropy of execute(e, a). For example, a robot could learn how to jump, build itself an extra arm, or refine its physics model. Many humans practice new skills or employ prostheses.

Many actions can make agents more powerful along multiple axes. For example, increasing computation ability might create new actions, increase decision-making ability, and better interpretation processing. Moving to a different location can both create new actions and increase observation ability.

1. This drawing is originally from [Embedded Agents](#)
2. The decision-making model known as the [OODA Loop](#) inspired this naming scheme. Acting has been renamed to executing to avoid confusion with act-based agents.
3. See [Conditions for Mesa-Optimization](#) for further discussion.
4. See [Locality of Goals](#) for a related discussion.
5. Technically, there could be two actions that both had maximal utility. This occurrence has measure 0, so we will assume it does not happen.
6. Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. "Inverse Reward Design." *ArXiv:1711.02827 [Cs]*, October 7, 2020. <http://arxiv.org/abs/1711.02827>
7. Constructing this set is technically impossible. In practice, it can be replaced by the set of all finite CWS+Ss.
8. We contrast this to an agent whose internals are structured like HCH, i.e., it has models of humans consulting other models of humans. An agent with this structure is likely more transparent and less competitive than an agent trying to enforce the same input-output mapping as HCH.
9. This utility function is flawed because the asymmetry of KL-divergence might make slightly suboptimal agents catastrophic, i.e., agents that rarely take actions HCH would never take will only be slightly penalized. Constructing a utility function that is not catastrophic if approximated remains an open problem.
10. What a human asks the agent depends on the agent's properties, so using the human distribution has acausal implications. See [Open Problems with Myopia](#) for further discussion.
11. Uesato, Jonathan, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. "Avoiding Tampering Incentives in Deep RL via Decoupled

Approval." ArXiv:2011.08827 [Cs], November 17, 2020.
<http://arxiv.org/abs/2011.08827>. ↵

12. The inverse might not hold. Time-limited myopic agents might also take over the world for reasons described [here](#) ↵
13. These concerns are similar to the ones illustrated in Demski's [Parable of Predict-O-Matic](#). ↵
14. See [Open Problems with Myopia](#) for further discussion. ↵
15. Here, the agent is reasoning according to the maximum probability world state, a trick inspired by Cohen et al.'s [Asymtotically Unambitious Artificial General Intelligence](#). ↵
16. What we mean by "upstream" is unclear but is related to the concept of subjunctive dependence from [Functional Decision Theory: A New Theory of Instrumental Rationality](#). ↵
17. If an agent's experiences shape who they are, uncertainty over the input distribution might be viewed as a special case of anthropic uncertainty. ↵
18. See [Understanding Deep Double Descent](#) for more discussion. ↵
19. A sample of current work in this general direction is Hubinger's [Towards a mechanistic understanding of corrigibility](#) and [Relaxed adversarial training for inner alignment](#), Filan et al.'s [Pruned Neural Networks are Surprisingly Modular](#), and much of the OpenAI Clarity team's work on [circuits](#). ↵
20. This is related to Richard's point that [coherent behavior in the world is an incoherent concept](#). ↵
21. Technically, both environmental states could have equal utility, making all policies optimal. This occurrence has measure 0, so we will assume it does not happen. ↵
22. I am not sure what assumptions are needed, but having no loops of any type and the agents always taking the same action in the same (environmental state, internal state) pair might be sufficient. ↵
23. Here, we roughly mean "there is probably some notion of the complexity of an agent's description that has some descriptions as simpler than others, even though we do not know what this notion is yet." ↵

Where are intentions to be found?

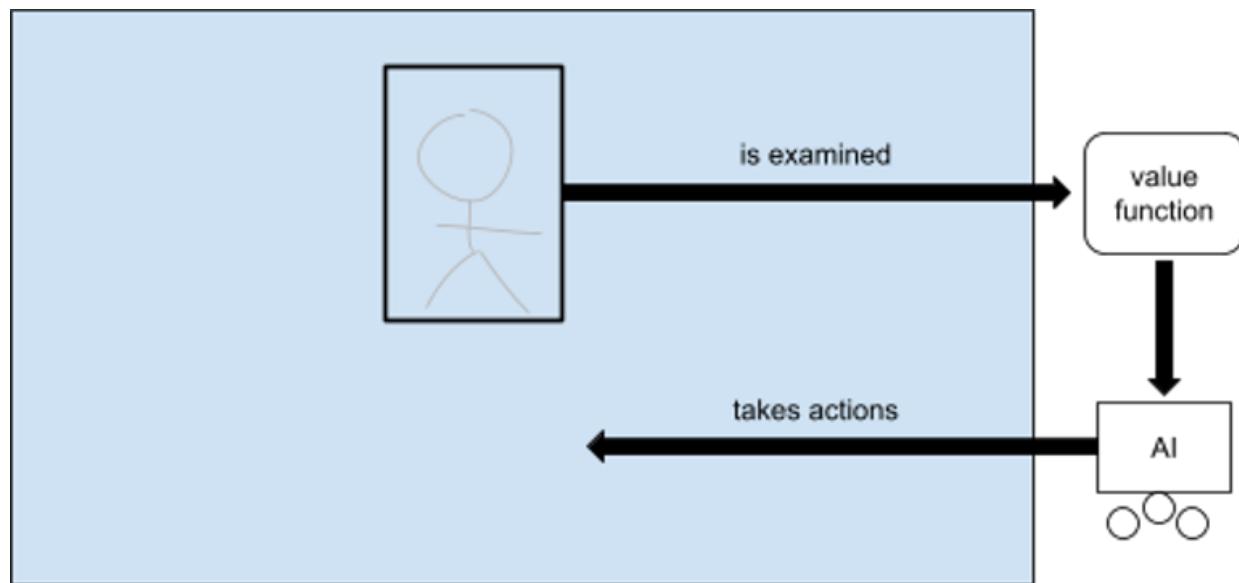
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is independent research. To make it possible for me to continue writing posts like this, please consider [supporting me](#).

As we build powerful AI systems, we want to ensure that they are broadly beneficial. Pinning down exactly what it means to be broadly and truly beneficial in an explicit, philosophical sense appears exceptionally daunting, so we would like to build AI systems that are, in fact, broadly and truly beneficial, but without explicitly answering seemingly-intractable philosophical problems.

One approach to doing this is to build AI systems that discover what to do by examining or interacting with humans. The hope is that AI systems can help us not just with the problem of taking actions in service of a goal, but also with the problem of working out what the goal ought to be.

Inverse reinforcement learning is a classical example of this paradigm. Under inverse reinforcement learning, an AI observes a human taking actions, then looks for an explanation of those actions in terms of a value function, then itself takes actions that optimize that value function.

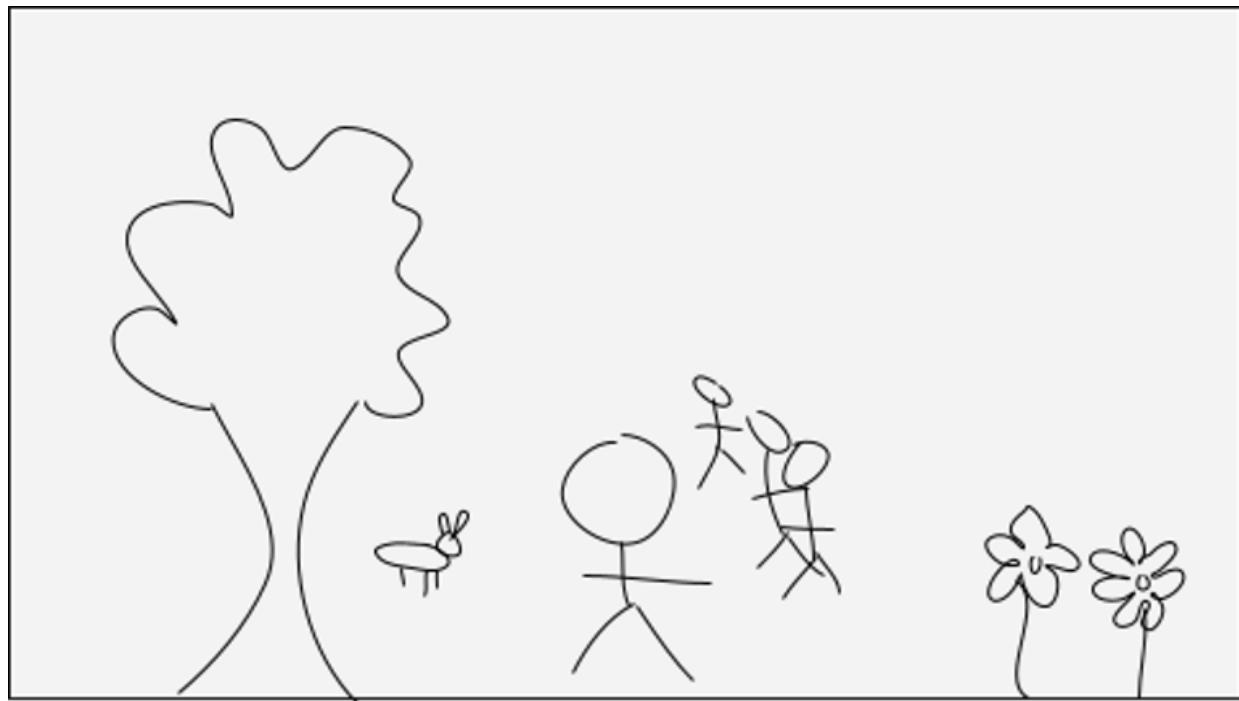


We might ask why we would build an AI that acts in service of the same values that the human is already acting in service of. The most important answer in the context of advanced AI, it seems to me, is that AI systems are potentially much more powerful than humans, so we hope that AI systems will implement our values at a speed and scope that goes beyond what we are capable of on our own. For this reason, it is important that whatever it is that the AI extracts as it examines a human taking actions is trustworthy enough that if it were implemented faithfully by the AI then the world brought forth by the AI would be a good world.

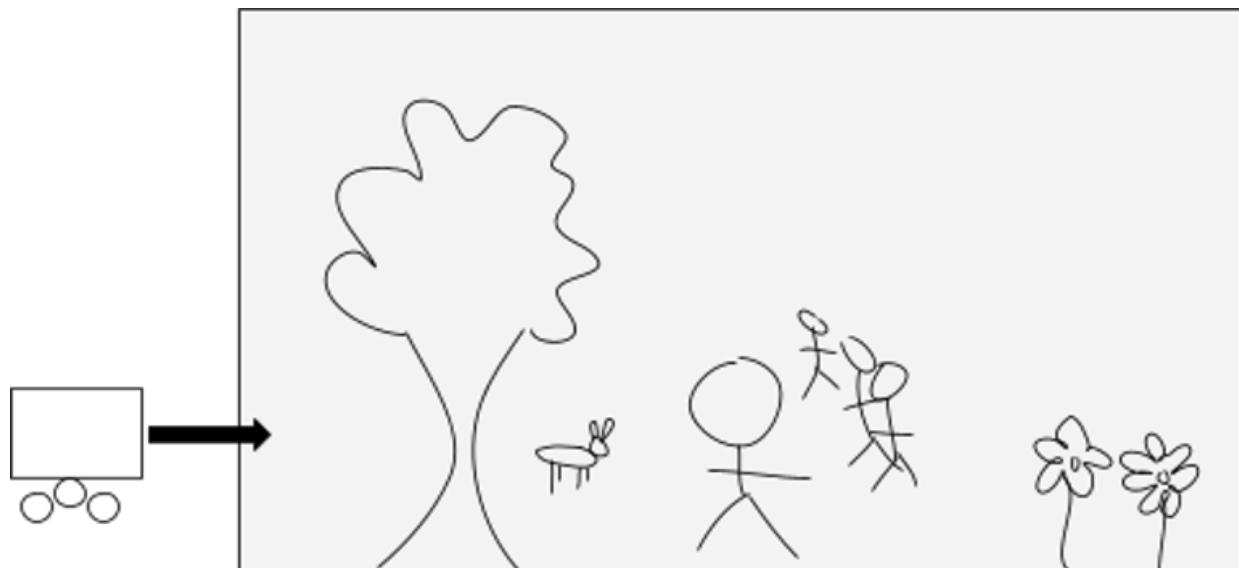
Inverse reinforcement learning is just one version of what I will call extraction-oriented AI systems. An extraction-oriented AI system is one that examines some part of the world, then, based on what it finds there, takes actions that affect the whole world. Under classical inverse reinforcement learning the particular part of the world that gets examined is some action-taking entity such as a human, the particular extraction method is to model that entity as an agent and look for a value function that explains its behavior, and the particular way that the system acts upon this value function is, at least under classical AI paradigms, to itself take actions that optimize that value function. But there are many other choices for what part of the world to examine, what to extract from it, and how to implement that which is extracted. For example, we might examine the net behavior of a whole human society rather than a single human; we might extract a policy by imitation learning rather than a value function by imitation learning; and we might act in the world using a satisficer rather than an optimizer. There are many choices for how we might do this. What I'm addressing here is any approach to developing AI that becomes aligned with what is truly beneficial by investigating some part of the world.

So long as we are within the regime of extraction-oriented AI systems, we are making the assumption that there is some part of the world we can examine that *contains information sufficient to be a trustworthy basis for taking actions in the world*.

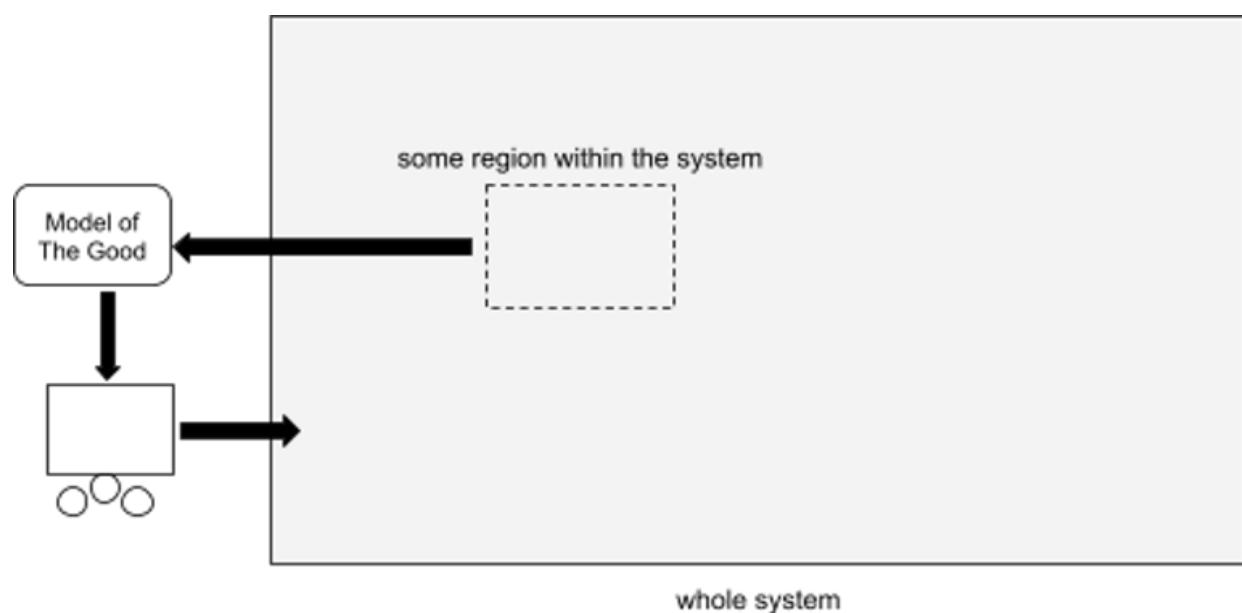
Let us examine this assumption very carefully. Suppose we look at a closed physical system with some humans in it. Suppose that this system contains, say, a rainforest in which the humans live together with many other animal and plant species:



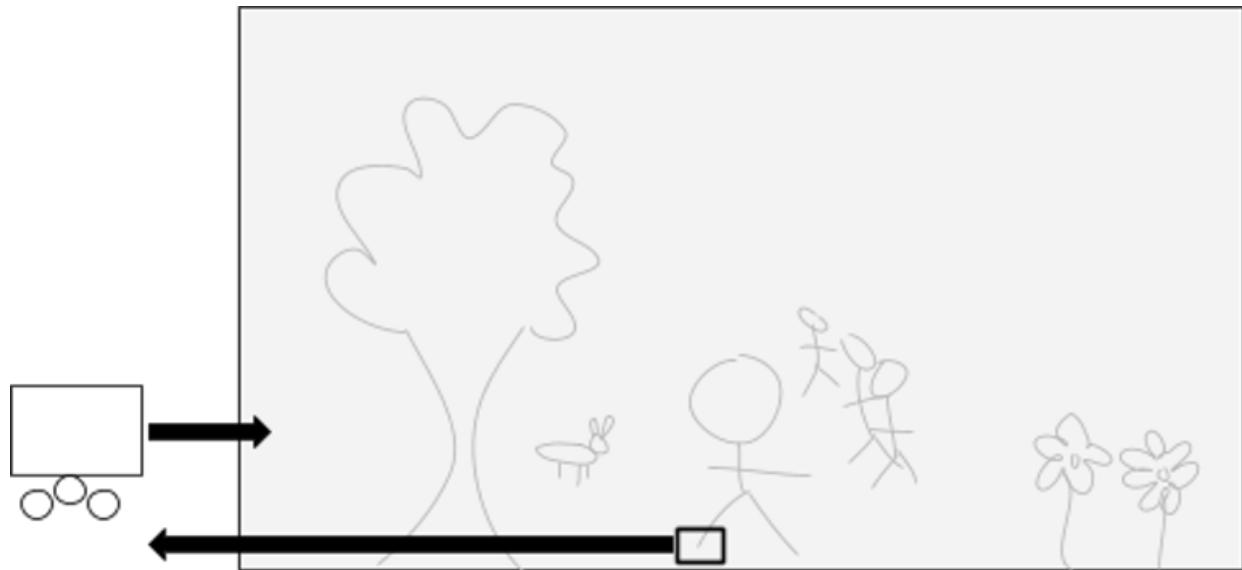
Suppose that I plan to build an AI that I will insert into this system in order to help resolve problems of disease, violence, ecological destruction, and to assist with the long-term flourishing of the overall ecosystem:



It is difficult to say exactly what it means for this overall ecosystem to flourish. How do I balance the welfare of one species against that of another? Of one individual against another? How do we measure welfare? Is welfare even the right frame for asking this question? And what is an appropriate way to investigate these questions in the first place? Due to such questions, it is difficult to build an AI purely from first-principles and so suppose I tell you that I am planning to build an AI that discovers the answers to these questions by examining the behavior of humans and perhaps other living beings within the ecosystem. Perhaps I have some elaborate scheme for doing this; there is no need to get into the details here, the important thing is that I tell you that the basic framework I will be working within is that I will observe some part of the system from some amount of time, then I will do some kind of modelling work based on what I observe there, then I will build an AI that acts in some way upon the model I construct, and in this way I will sidestep needing an explicit answer to the thorny philosophical questions of what true benefit really means:



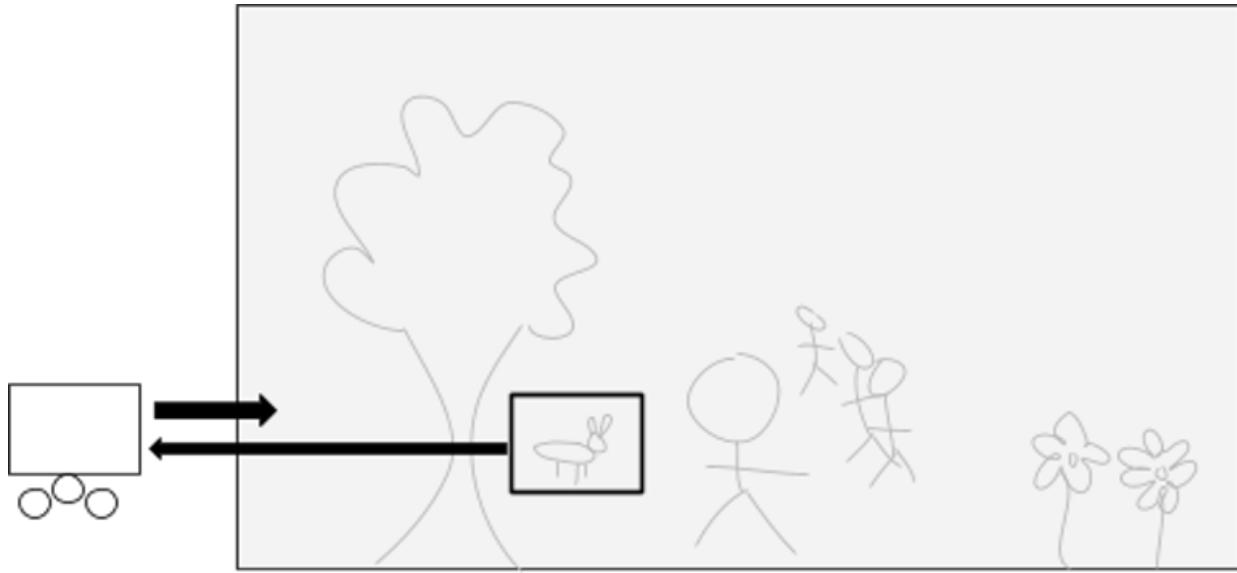
You might then ask which part of the system will I examine and what is it that I hope to find there that will guide the actions of the powerful AI that I intend to insert into the system. Well, suppose for the sake of this thought experiment that the part of the world I am planning to examine was the right toe of one of the humans:



Suppose I have an elaborate scheme in which I will observe this toe for aeons, learn everything there is to learn about it, interact with it in this or that way, model it in this or that way, place it in various simulated environments and interact with it in those simulated environments, wait for it to reach reflective equilibrium with itself, and so forth. What do you say? You say: well, this is just not going to work. The information I seek is just not in the toe. It is not there. I can examine the spatial region containing a single human toe for a long time but the information I seek is not there, so the AI I build is not going to be of true benefit to this ecosystem and the living beings within it.

What information is it that I am seeking? Well I am seeking information sufficient to guide the actions of the AI. I do not have an understanding of how to derive beneficial action from first principles so I hope to learn or imitate or examine something somewhere in a way that will let me build an AI whose actions are beneficial. It could be that I extract a policy or a value function or something else entirely. Suppose for the sake of thought experiment that I am in fact a computer scientist from the future and that I present to you some scheme that is unlike anything in contemporary machine learning, but still consists of examining a part of the world, learning something from it, and on that basis building an AI that sidesteps the need for a first principles answer to the question of what it means to be beneficial. And suppose, to continue with my thought experiment, that the region of space I am examining is still a single human toe. It really does not matter what sophisticated scheme I present: if the part of the world that I'm examining is a left toe then this scheme is not going to work, because this part of the world does not contain the kind of information that could guide the actions of an AI that will have power over this ecosystem's destiny.

Now let us suppose that I present to you the following revised plan: the part of the world I am going to examine is a living rabbit. Yes, a rabbit:



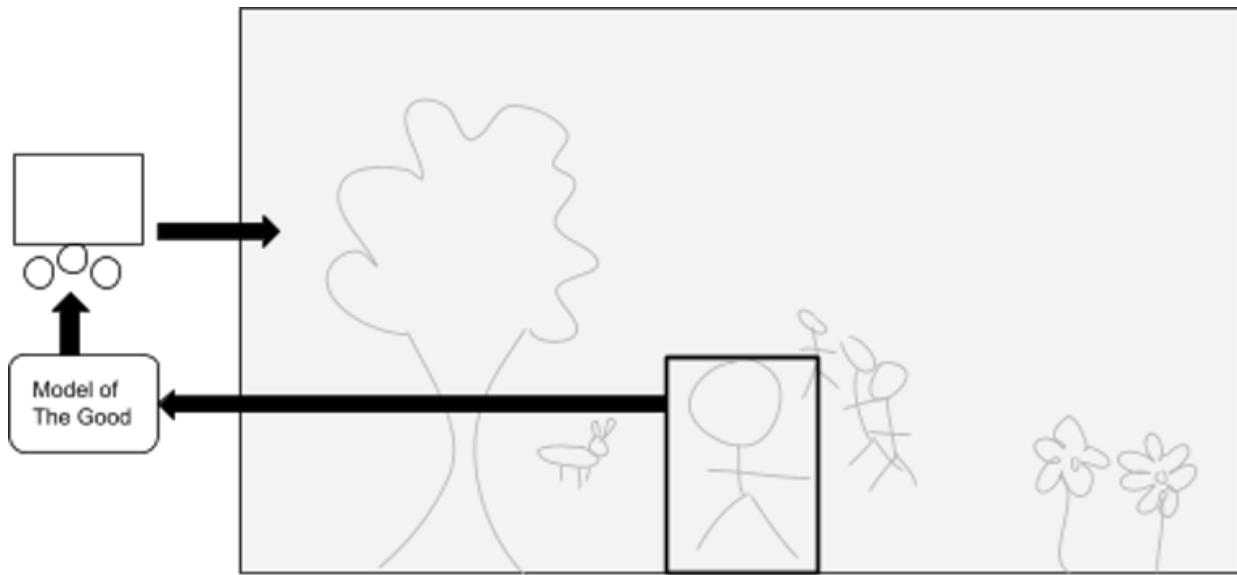
Again, let's say that I present some sophisticated scheme for extracting *something* from this part of the world. Perhaps I am going to extrapolate what the rabbit would do if it had more time to consider the consequences of its actions. Or perhaps I am going to evolve the rabbit forward over many generations under simulation. Or perhaps I am going to provide the rabbit with access to a powerful computer on which it can run simulations. Or perhaps I have some other scheme in mind, but it is still within the following framework: I will examine the configuration of atoms within a spatial region consisting of a live rabbit, and on the basis of what I find there I will construct an AI that I will then insert into this ecosystem, and this AI will be powerful enough to determine the future of life in this ecosystem.

Now, please do not get confused about whether I am trying to build an AI that is beneficial to humans or to rabbits. Neither of those is my goal in this hypothetical story. I am trying to build an AI that is *overall beneficial* to this system, but I do not know what that means, or how to balance the welfare of rabbits versus that of humans versus that of trees, or what welfare means, or whether the welfare of the whole system can be decomposed into the welfare of the individual beings, or whether welfare is the right kind of frame to start with. I am deeply confused at every level about what it means for any system to be of true benefit to anything, and it is for that very reason that I am building an extraction-oriented AI: my hope is that rather than first coming to a complete understanding of what it means to be of true benefit to this small world and only then building an AI to implement that understanding, I can sidestep the issue by extracting some information from the world itself. Perhaps if I do the right kind of extraction -- which may involve allowing the rabbit to reflect for a long time, or allowing it to interact with statistical imitations of itself interacting with statistical imitations of itself, or any other such scheme -- then I can find an answer to these questions within the world itself. And it does not have to be an answer that I personally can understand and be satisfied with, but just an answer that can guide the actions of the AI that I plan to insert into this world. But no matter how many layers of uncertainty we have or what specific scheme I present to you, you might still ask: *is it plausible that the information I seek is present in the particular spatial region that I propose to examine?*

And, I ask you now, back here in the real world: is this information in fact present in the rabbit? Could some hypothetical superhumans from the future build this AI in a way that actually was beneficial if they were limited to examining a spatial region

containing a single rabbit? What *is* the information we are seeking, and is it present within the rabbit?

I ask this because I want to point out how nontrivial is the view that we might examine *any* part of such a system and find answers to these profound questions, no matter how the extraction is done. Some people seem to hold the view that we could find these answers by examining a human brain, or a whole human body:

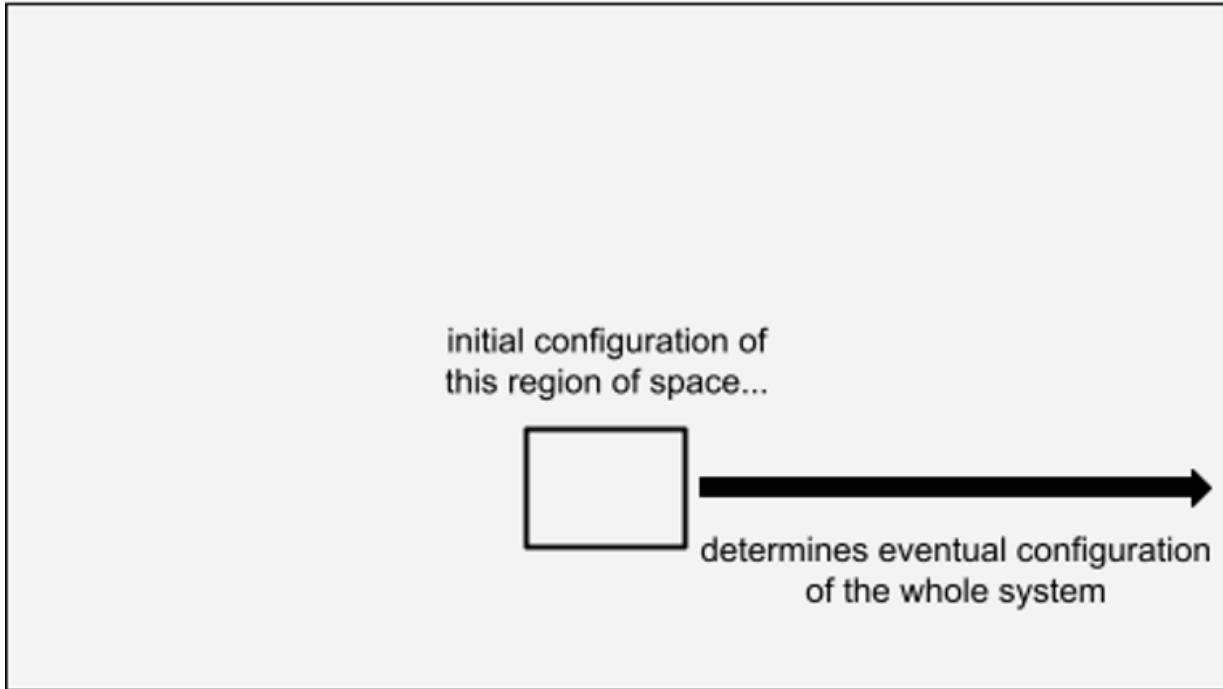


Of course, the schemes for doing this do not anticipate that we will just read out answers from the structure of the brain. They are more sophisticated than that. Some anticipate running simulations of the human brain based on the neural structures we find and asking questions to those simulations. Others anticipate modelling the brain based on the output it produces when fed certain inputs. But the point is that so long as we are in the regime of extraction-oriented AI, which is to say that we *examine a spatial region within a system, then, based on what we find there, build an AI that takes actions that affect the whole system*, then we might reasonably ask: is the information we seek plausibly present in the spatial region that we are examining? And if so, why exactly do we believe that?

Is it plausible, for example, that we could examine just the brain of a human child? How about examining an unborn human embryo? A strand of human DNA? A strand of DNA from a historical chimpanzee from which modern humans evolved? A strand of DNA from the first organism that had DNA? If the information we seek is in the human brain then how far back in time can we go? If we have a method for extracting it from an adult human brain then could we not extract it from some causal precursor to a fully-formed human brain by evolving a blueprint of the precursor forward in time? We are not talking here about anything so mundane as extracting contemporary human preferences; we are trying to extract answers to the question of whether preferences are even the right frame to use, whether we should incorporate the preferences of other living beings, where the division between moral patienthood and moral non-patienthood is, whether the AI itself is a moral patient, whether the frame of moral patients is even the right frame to use. These are deep questions. The AIs we build are going to do *something*, and that *something* may or may not be what is truly beneficial to the systems into which we deploy them. We cannot avoid these questions completely, but we hope to sidestep explicitly answering them by imitating or learning from or modelling *something* from *somewhere* that can form *some kind of basis* for an

AI that takes actions in the world. If we are within this extraction-oriented AI regime, then the actions taken by the AI will be a function of the physical configuration of matter within the spatial regions that we examine. So we might ask: do we want the future to be determined by the physical configuration of matter within *this* particular spatial region? For which spatial regions are we willing to say yes? So long as we are in this regime, no amount of modelling wizardry changes this functional dependence of the whole future of this world upon the physical configuration of some chosen part of the world.

Extraction-oriented AI Systems



If the spatial region we choose is a human brain, or a whole human body, or even an entire human society, then we should ask: how is it that the information in this spatial region is relevant to how we would want the overall configuration of the system to evolve, but information outside that spatial region is not relevant? How did that come to be the case?

As I wrote in [my reflections on a recent seminar by Michael Littman](#), it seems to me that my own intentions have updated over time at every level. It does not seem to me that I have some underlying fixed intentions lying deep within me that I am merely unfolding. It seems to me that it is *through interacting with the world* that my intentions develop and mature. I do not think that you could find out my current intentions by examining my younger self because the information was not all in there: much of the information that informs my current intentions was at that time out in the world, and it is through encountering it that I have arrived at my current intentions. And I anticipate this process continuing into the future. I would not trust any scheme that would look for my true intentions by examining my physical body and brain today, because I do not think the information about my deepest intentions in the future is located entirely within my body and brain today. Instead I think that my intentions will be informed by my interactions with the world, and some of the information about how that will go is out there in the world.

But this is just introspective conjecture. I do not have full access to my own inner workings so I cannot report on exactly how it is that my intentions are formed. My point here is more modest, and it is this: that we can discover what is of benefit to a system by examining a certain part of the system is a *profound claim*. If we are to examine a part of the universe in which we find ourselves located, and that part contains one or several hairless primates under the supposition that the desired information *is* present in that part, then we should have a good account of how that came to be the case. It is not obvious to me that it is in there.

D&D.Sci April 2021 Evaluation and Ruleset

This is a followup to [the D&D.Sci post](#) I made last week; if you haven't already read it, you should do so now before spoiling yourself.

[Here](#) is the web interactive I built to let you evaluate your solution; below is an explanation of the rules used to generate the dataset. You'll probably want to test your answer before reading any further.

Ruleset

(Note: to make writing this easier, I'm using standard D&D dice notation, in which "3+4d8" means "roll four eight-sided dice, sum the results, then add three".)

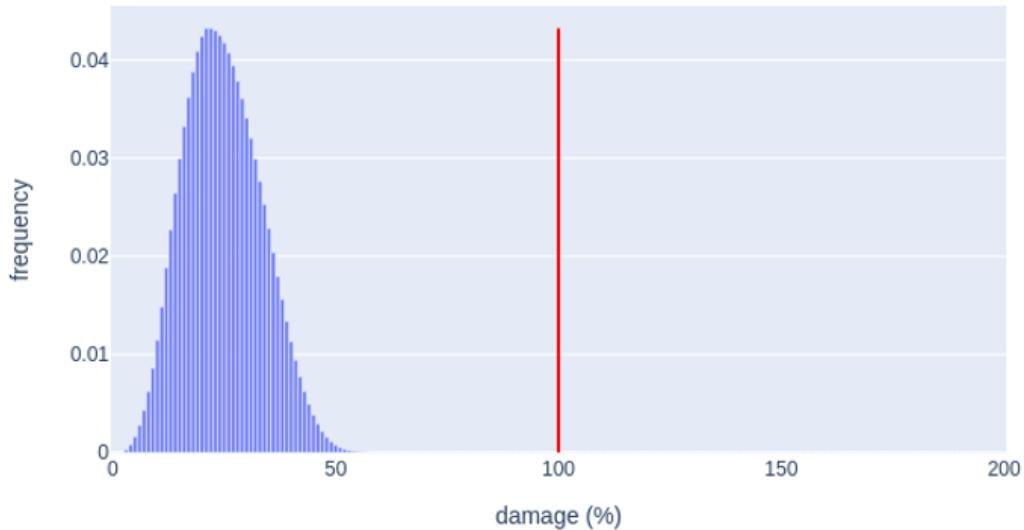
Enemies

Sharks

Sharks are 1/6 of encounters.

They attack in groups of $2+1d4$, each of which does $1d10$ points of damage.

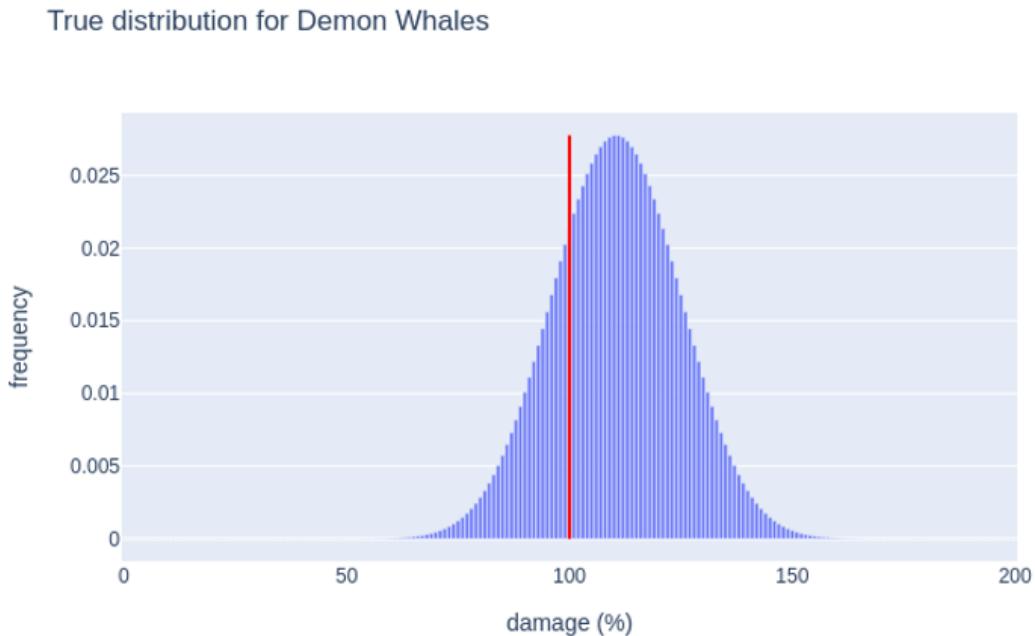
True distribution for sharks



Demon Whales

Demon Whales are 1/14 of encounters. (If that fraction seems high, you're failing to account for all the sunk ships that couldn't report encountering them.)

An attack from a Demon Whale does 17d12 points of damage.



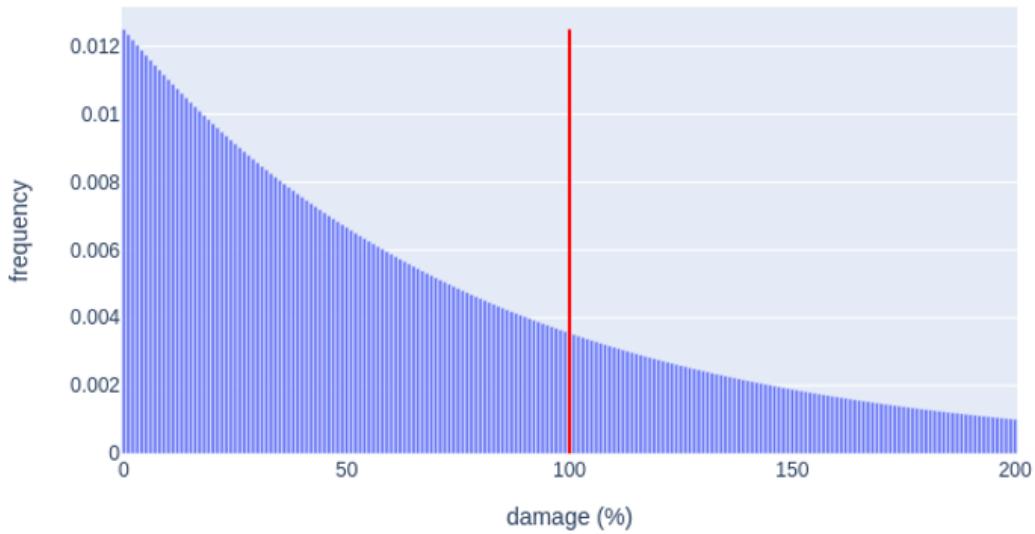
A Demon Whale encounter has a ~78% fatality rate

Crabmonsters

Crabmonsters are 1/14 of encounters.

A Crabmonster repeatedly rolls 1d80 as it tears through the ship, adding a point of damage with each roll, until it rolls a 1 (that is, encounters someone or something that stops it).

True distribution for Crabmonsters



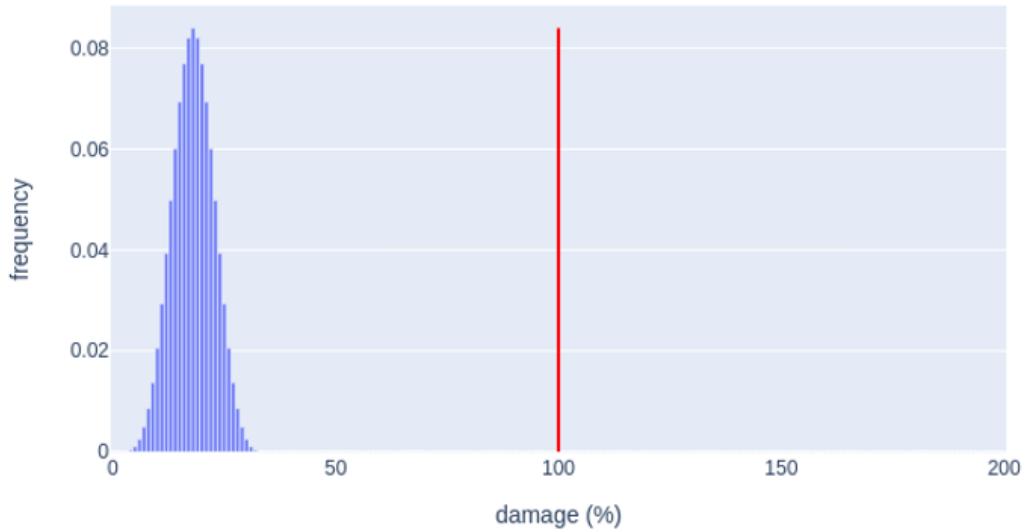
~8% of Crabmonster encounters do >200% damage; a Crabmonster encounter has a ~28% fatality rate

Pirates

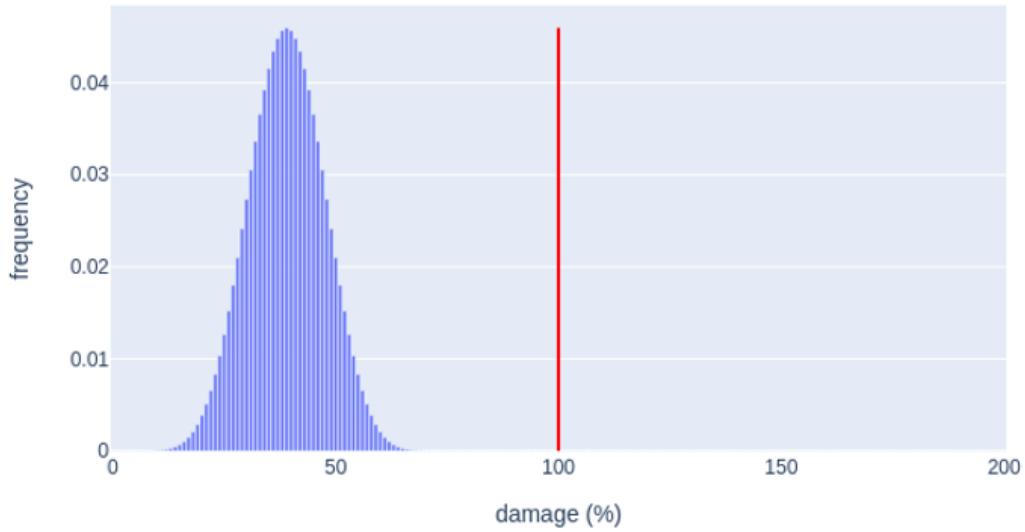
Though the Navy's records don't bother to distinguish, Pirates come in two categories: Brigands (local criminals who had the poor fortune to cross paths with Naval supply ships while flying the black flag, and/or to mistake them for civilian cargo ships) and Privateers (agents of an enemy government, harassing your Navy's fleet using hit-and-run tactics). Brigands are 1/6 of random encounters during your voyages, Privateers 1/21.

A fight with Brigands does 4d8 points of damage; a fight with Privateers does 6d12.

True distribution for Brigands



True distribution for Privateers



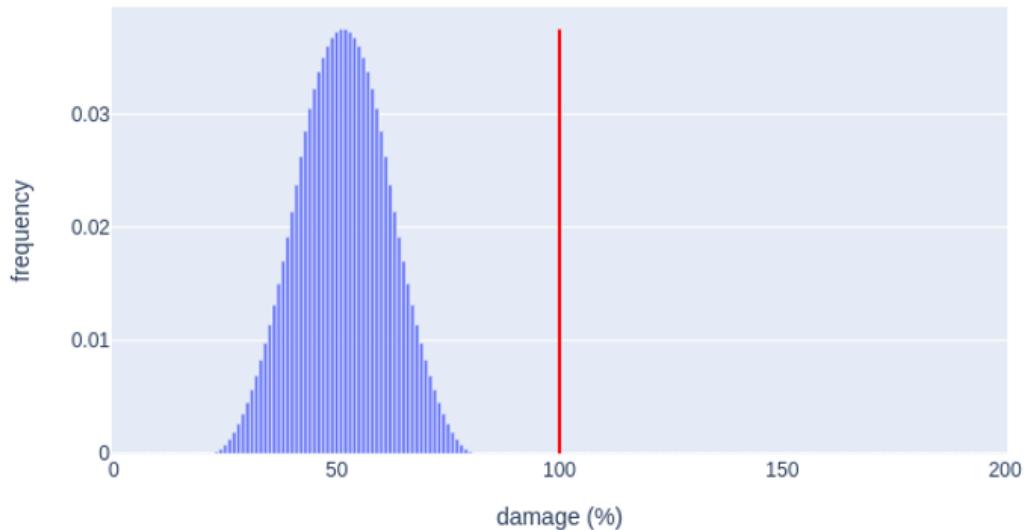
Merpeople

Surface-dwellers are unaware of the intricacies of underwater society, and record both Atlantean Merfolk (1/14 of encounters) and Alexandrian Merfolk (2/21 of encounters) as

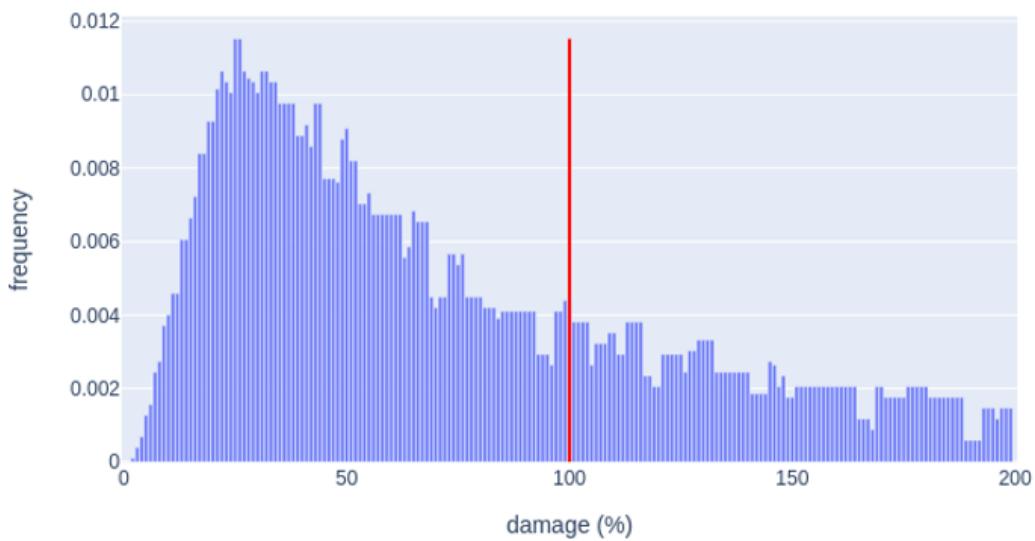
"Merpeople". Fortunately, the two city-states are close enough politically that befriending one will cause them both to allow you free passage.

Atlanteans do $20+3d20$ damage; Alexandrians do $1d8*1d8*1d8+1d20$ damage.

True distribution for Atlanteans



True distribution for Alexandrians



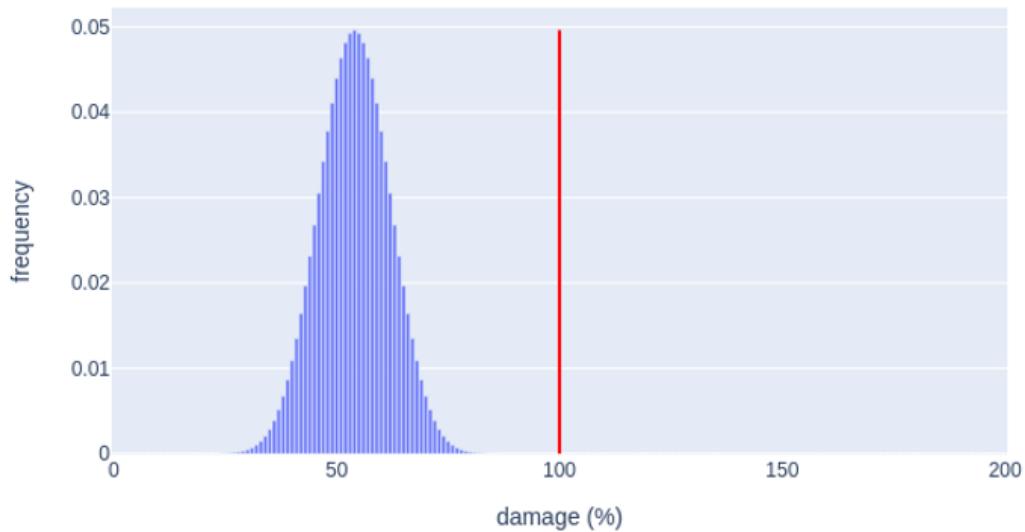
~14% of Alexandrian attacks do >200% damage; an Alexandrian attack has a ~37% fatality rate

Kraken

Kraken are 2/21 of encounters.

They do 12d8 points of damage.

True distribution for Kraken

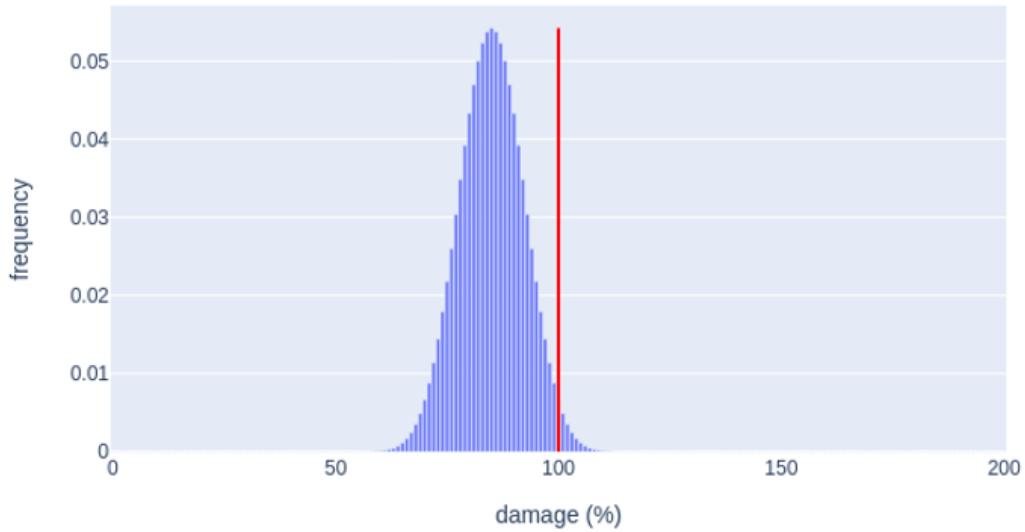


Nessie

Nessie is 1/21 of encounters.

She does 40+10d8 points of damage.

True distribution for Nessie



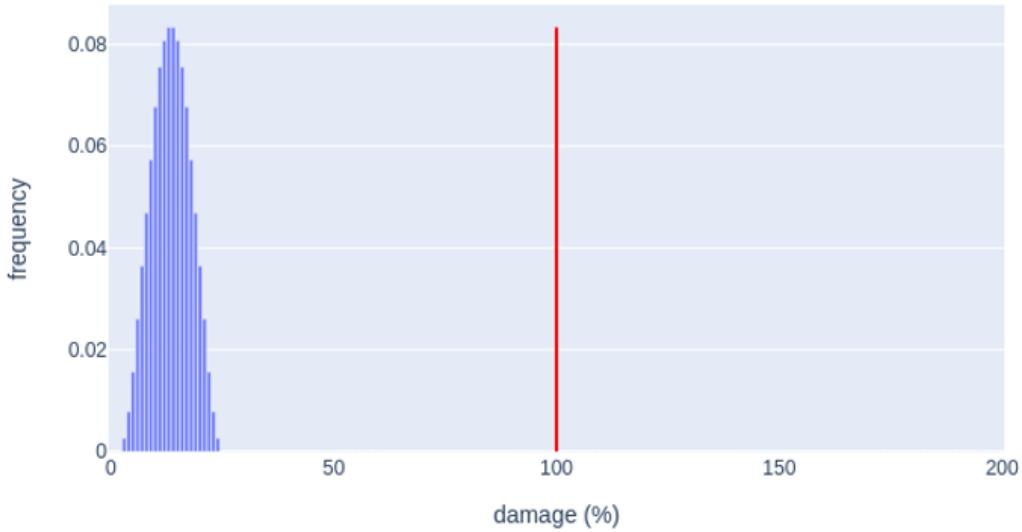
An encounter with Nessie has a ~2% fatality rate

Harpies

Harpies are 1/14 of encounters.

They do $1d4+1d8+1d12$ points of damage.

True distribution for Harpies

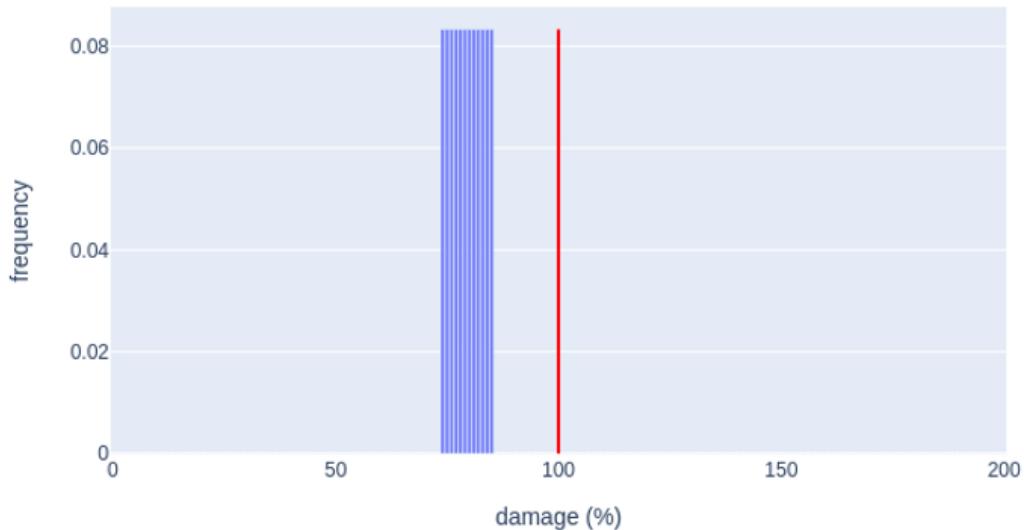


Water Elementals

Water Elementals are 2/21 of encounters.

The Navy has countering the powerful but predictable attacks of Water Elementals down to an art; there are well-known methods for ensuring they only *almost* destroy a given ship. They do $73+1d12$ points of damage.

True distribution for Water Elementals



Direction

Direction is irrelevant from perspectives both practical (you have no control over how many trips you take each way) and epistemic (direction happens to have no effect on outcomes).

Time effects

Time has almost no effect. The one exception is that Privateers used to be much more common (and other encounters therefore slightly less common) before 4/1401; this is when your nation's main rival changed tactics and stopped hiring mercenaries to attack supply ships.

Sinking Risk by Enemy

In the absence of interventions, ~50% of shipwrecks are caused by Demon Whales, ~18% by Crabmonsters, ~31% by Merpeople, ~1% by Nessie, and 0% by other threats.

Strategy

If attempting to optimize odds of survival, your best choices are to buy all oars, arm carpenters, tribute the Merpeople, and buy one extra cannon; congratulations to simon, GuySrinivasan and Measure for reaching this conclusion.

However, since Pirates never sink ships and Nessie is pretty bad at it, you may wish to take the money you'd spend on the cannon and either hold onto it (to impress the Navy's accountants) or spend it on foam swords (to impress the Navy's dockworkers).

Reflections

All else equal, there's a little extra uncertainty when predicting quantities instead of categories: "is that sudden peak at 14% noise, or a clue to the generating function?", etc. However, the *main* reason this challenge was so much more speculative than its predecessors is that the most important information - details of attacks that did 100%+ damage - was censored by the mechanics of the world. In the absence of hard evidence, small errors in inference compound, priors pick up the slack, and considerations like "what genres apply here?" or "is the scenario designer enough of a troll to have Demon Whale damage arbitrarily cap out at 99%?" take on a significance they wouldn't otherwise.

This is both good and bad. Good because the personal touch adds intrigue to what would otherwise just be data-wrangling; bad because every unit of effort spent psychoanalyzing the GM is a unit of effort not spent on getting better at data-wrangling or on psychoanalyzing reality's GM (i.e. studying Math and Science). I enthusiastically solicit feedback on this point, as well as on every other point.

Scheduling

The next D&D.Sci challenge should be ready sometime earlyish next month, but nebulous and open-ended work commitments mean I can't promise anything.

Probability theory and logical induction as lenses

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

This is independent research. To make it possible for me to continue writing posts like this, please consider [supporting me](#).

Abstract

I frame probability theory as a lens for understanding the real-world phenomenon of machines that quantify their uncertainty in their beliefs. I argue in favor of using multiple lenses to look at any phenomena, even when one of those lenses appears to be the most powerful one, in order to keep track of the boundary between the phenomena we are examining and the lens we are using to examine it. I argue that logical induction is another lens for looking at the real-world phenomena of machines that quantify their uncertainty in their beliefs, addressing the specific case of machines that are Turing-computable. I argue that tracking uncertainty in purely logical claims is merely a necessary consequence of addressing this general issue. I conclude with remarks on the importance of general theories of intelligent systems in the engineering of safety-critical AI systems.

Introduction

Last year I began studying the logical induction paper published by MIRI in 2016. I did so partially because I was unsure how to contribute within AI safety. I had been reading articles and writing posts but was at a bit of a loss as to what to do next, and I had been curious about, although intimidated by logical induction for some time, and so I felt inclined to study it. I felt a little guilty about spending time looking into something that I was merely curious about, at the expense perhaps of other projects that might be more directly helpful.

But my curiosity turned out to be well-founded. As I read through the logical induction paper, and particularly as I came back to it many days in a row, I started to see that it was actually significant in a different way than what I had expected. I had previously understood that logical induction was about having uncertainty about purely logical claims, such as putting 10% probability on the claim that the one billionth digit of pi is a 3. That is certainly in there, but as I read through the paper, I realized that logical induction paints a bigger picture than this, and that in fact the maintaining of uncertainty in logical claims is a mere consequence of this bigger picture.

The bigger picture is this: logical induction provides an account of what it means for machines to quantify their uncertainty in their beliefs, in the specific case when the machine is subject to the laws of physics and can only compute things that are Turing-computable. In general it is not possible for such a machine to immediately evaluate the logical consequences of each new observation, so logical induction is forced to deal with the gradual propagation of logical information through belief networks, and so

there has to be some mechanism for accounting for partially-propagated logical information, and of course one reasonable way to accomplish that is by maintaining explicit uncertainty in logical claims. But the way I see it, this is merely a consequence of having a formal account of quantified uncertainty when the one doing the quantifying is explicitly subject to the laws of computability.

Probability theory

Probability theory also provides an account of what it means to quantify one's uncertainty in one's beliefs. It is a different account from the one provided by logical induction, and it is a compelling account. We might view each basic derivation of probability theory as an instance of the following message:

If the credences you assign to your beliefs obey the laws of probability theory, then you will get such-and-such benefits.

Under the Dutch book arguments for probability theory, the "benefits you get" are that if you bet on your credences then you can be certain that you will not be Dutch-booked. (Being Dutch-booked is when someone makes a combination of bets with you under which you are guaranteed to lose money no matter what the state of the world turns out to be.)

Conversely, we can view derivations of probability theory as an instance of:

If you want such-and-such benefits, then the credences you assign to your beliefs must obey the laws of probability theory.

Under the Dutch book arguments for probability theory, the message is now that if you don't want to be Dutch-booked, then your credences must obey the laws of probability theory.

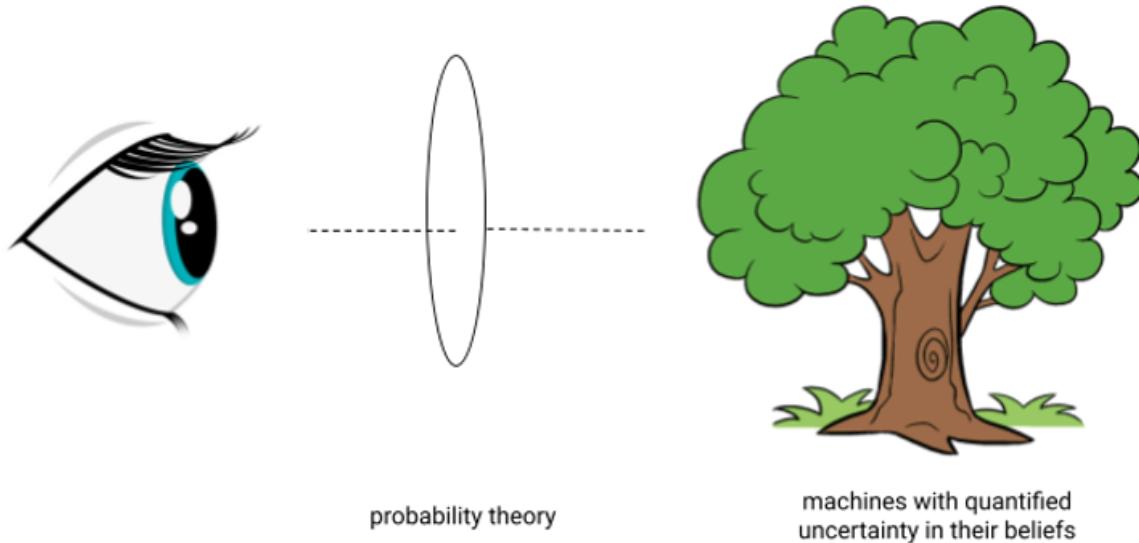
There are other ways to arrive at the laws of probability theory, too. For example, under [under Jaynes' derivation](#), the "benefits you get" are his three desiderata^[1]:

- Degrees of plausibility are represented by real numbers
- Qualitative correspondence with common sense
- If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

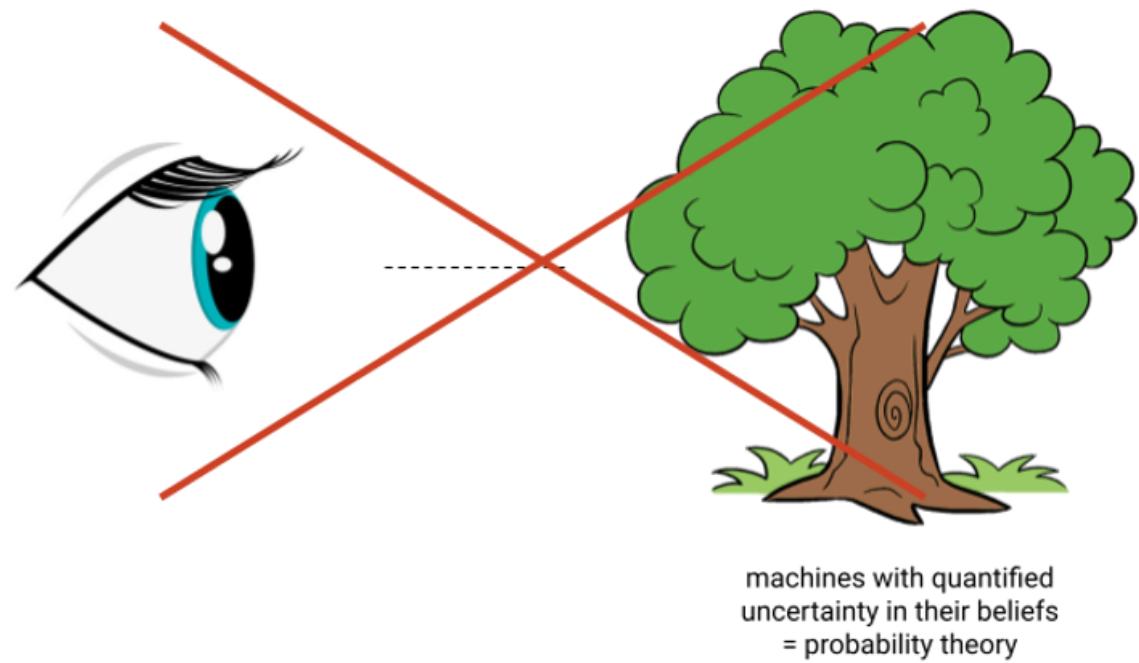
These desiderata seem so reasonable, so humble, so minimal, that it can be difficult to imagine that one would ever not want to obey the laws of probability theory. Do you not want your credences to be consistent? Do you not want your credences to correspond with common sense (which Jaynes operationalizes as a set of very reasonable inequalities)? Do you not want your credences to be represented by numbers? When I originally read Jaynes' book, it appeared to me that he had demonstrated that probability theory was *the right way* to quantify uncertainty, and in my thinking there was very little daylight between the abstraction of "probability theory" and the real-world phenomenon of "having beliefs".

And indeed probability theory is excellent. But there is just one hitch: we cannot in general build machines that implement it! This is not a criticism of probability theory itself, but it *is* a criticism of viewing probability theory as a final answer to the question

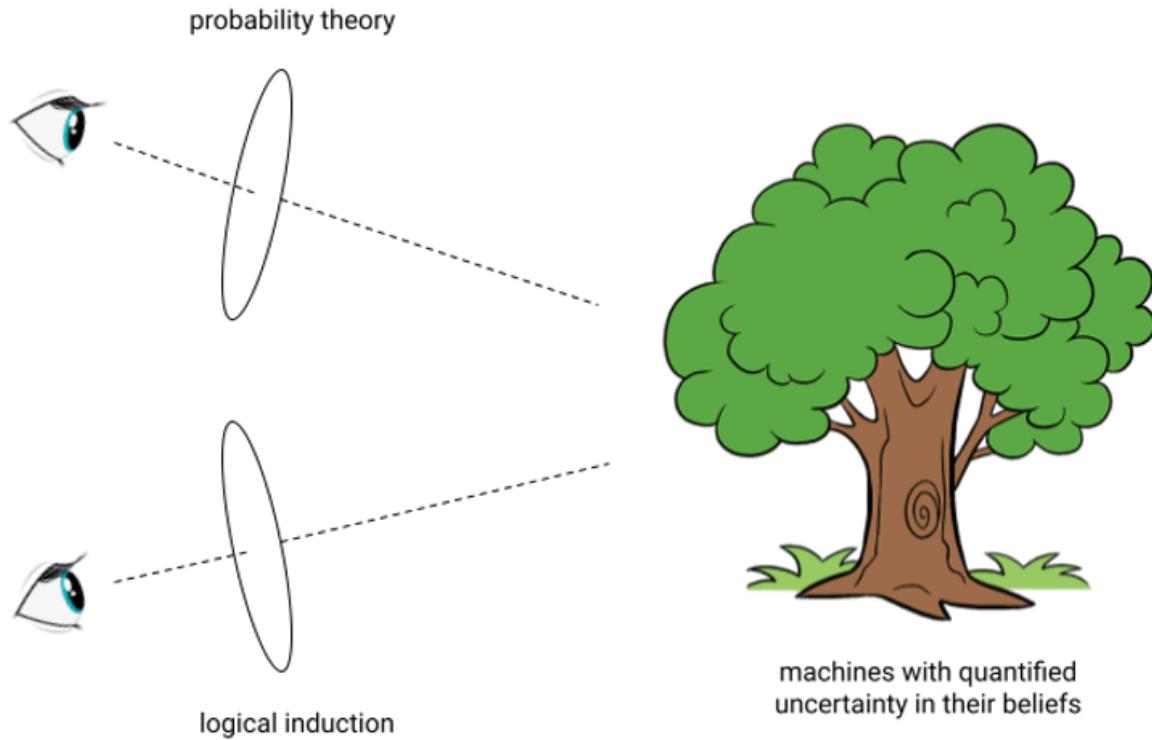
of how we can quantify our uncertainty in our beliefs. Probability theory is a lens through which we can view the real-world phenomena of machines with quantified uncertainty in their beliefs:



Probability theory is a way of understanding something that is out there in the world, and that way-of-understanding gives us affordances with which to take actions and engineer powerful machines. But when we repeatedly use one lens to look at some phenomenon, as I did for many years with probability theory, it's easy to lose track of the boundary between the lens and the thing that is actually out there in the world:



To avoid this collapse of the lens/phenomenon, or map/territory distinction, it is helpful to view phenomena through multiple lenses:



Of course using multiple lenses may allow us to see aspects of real-world phenomena missed by our primary lenses, but even beyond that, using multiple lenses is helpful simply insofar as it reminds us *that we are in fact using lenses to see the world*. In this way it helps us to see not just the phenomena we are looking at but also the lens itself.

Logical induction

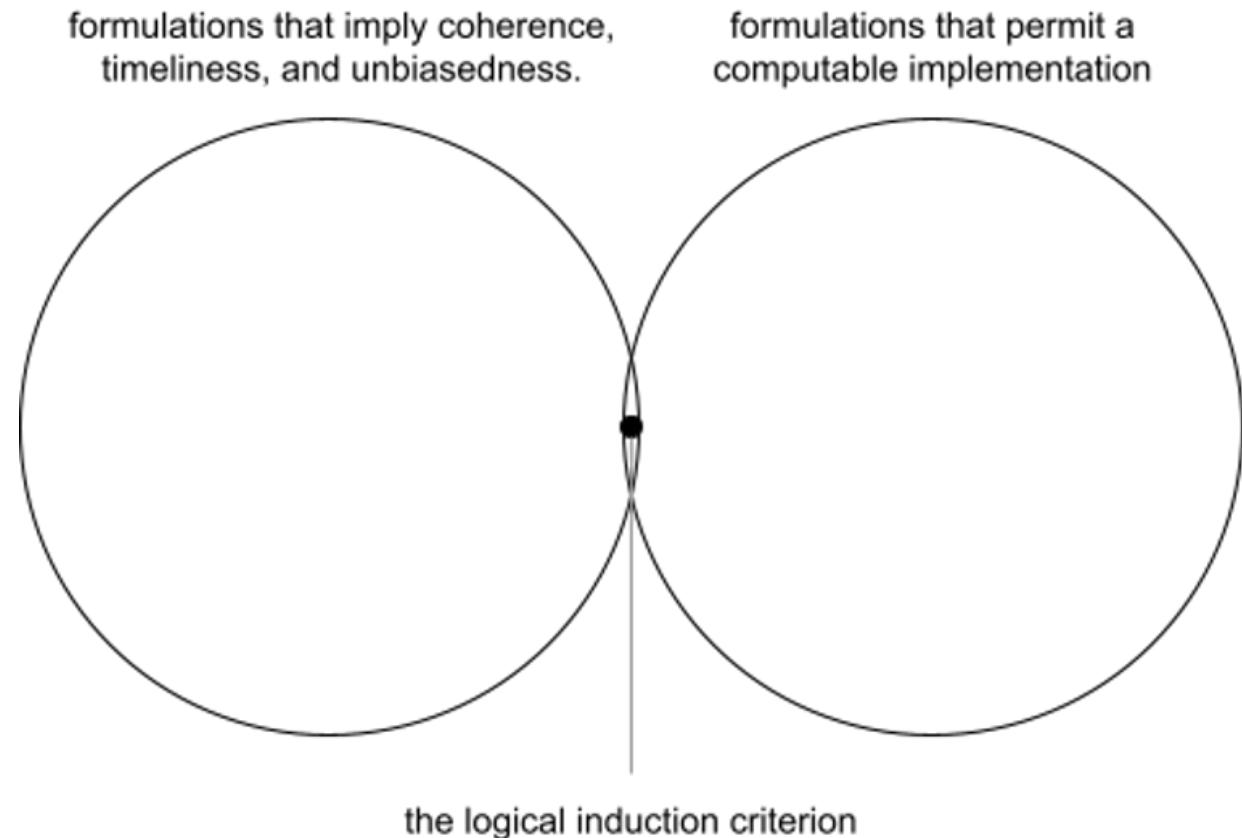
So what exactly *is* the perspective that logical induction gives us on machines with quantified uncertainty in their beliefs? Well, one way to see the framework presented in the logical induction paper is:

If the credences you assign to your beliefs obey the logical induction criterion, then you will get such-and-such benefits.

In the case of logical induction, the benefits are things like coherence, convergence, timeliness, and unbiasedness^[2]. But different from probability theory, these concepts are operationalized as properties of the evolution of your credences over time, rather than as properties of your credences at any particular point in time.

The benefits promised to one whose credences obey the laws of logical induction are weaker than those promised to one whose credences obey the laws of probability theory. A logical inductor can generally be Dutch-booked at any finite point in time, unless by some fluke it happens to have fallen perfectly into alignment with the laws of probability theory at that point in time, in which case there is no guarantee at all that it will remain there. So why would one choose to pick credences that obey the logical induction criterion rather than the laws of probability theory?

The answer is that credences that obey the logical induction criterion can be computed, no matter how many claims about the world you are maintaining uncertainty with respect to, and no matter how complex the relationships between those claims. This is a very significant property. The logical induction criterion threads a needle that sits at the intersection of those formulations that give us the properties we care about (coherence, convergence, timeliness, unbiasedness, and so on), and those formulations that permit a computable implementation:



Conclusion: theory-based engineering

As we build safety-critical systems, and safety-critical AI systems in particular, it is good to have theories that can guide our engineering efforts. In particular it is good to have theories that make statements of the form "if you construct your system in such-and-such a way then it will have such-and-such properties". We can then decide whether we want those properties, and whether it is feasible for us to work within the confines of the design space afforded by the theory.

For the purpose of engineering AI systems, our theories must ultimately be applicable to machines that can be constructed in the real world. Probability theory is not natively applicable to machines that can be constructed in the real world, but it can be made applicable by finding appropriate approximation schemes and optimization algorithms through which we can reason not just about the properties we want our credences to have but also the mechanism by which we will get them there. Logical induction, also, is not "natively applicable" to machines that can be constructed in the real world. The

logical induction algorithm is computable but not efficient. It does not guarantee anything about our credences at any finite point in time but only about the evolution of our credences over all time, so it does not provide a complete account for how to build general intelligent systems that do what we want. But it is a step in that direction.

Most importantly, to me, logical induction provides *an example of what theories of intelligent systems can look like*. We have few such theories, and the ones we do have, like probability theory, are so familiar that it can be difficult to see the daylight between the real-world phenomenon we are trying to understand the theory we are using as a lens to look at it. By looking through multiple lenses we are more likely to be able to see where to construct new theories that revise the answers given by our existing theories about what shape intelligent systems ought to have.

1. Of course these imply non-dutch-bookability and non-dutch-bookability implies these desiderata. It is merely a different formulation and emphasis of the same message. [←](#)
2. In chapter 4 of the paper each of these is defined formally as a property of a sequence of belief states and then it is proved that credences that obeys the logical induction criterion will necessarily have this property. [←](#)

What books are for: a response to "Why books don't work."

This is a linkpost for <https://aaronbergman.substack.com/p/what-books-are-for>

...and a follow up to my recent post "[Recommended Readings](#)."

Introduction

[Andy Matuschak](#)'s 2019 essay "[Why books don't work](#)" has been on my mind for a while now. I highly recommend you give it a read. The piece's central claim is that books fail at their fundamental implicit task: conveying information to their reader. Let Matuschak speak for himself:

Picture some serious non-fiction tomes. *The Selfish Gene*; *Thinking, Fast and Slow*; *Guns, Germs, and Steel*; etc. Have you ever had a book like this—one you'd read—come up in conversation, only to discover that you'd absorbed what amounts to a few sentences? I'll be honest: it happens to me regularly...
I know I'm not alone here. When I share this observation with others—even others, like myself, who take learning seriously—it seems that everyone has had a similar experience...

Why? Because

books have no carefully-considered cognitive model at their foundation, but the medium does have an *implicit* model. And like lectures, *that model is transmissionism*. Sequences of words in sequences of lines in sequences of pages, the form of a book suggests *people absorb knowledge by reading sentences*.

But *transmissionism* is basically false. Matuschak concludes that

as a medium, *books are surprisingly bad at conveying knowledge, and readers mostly don't realize it*.

And the answer, he thinks, is not to change *how* we read but rather to replace books with a medium more conducive to understanding.

We agree so far

I basically buy the central thrust of Matuschak's argument. A typical nonfiction book contains probably many thousands of individual statements and propositions, the vast majority of which the vast majority of us will never remember. And I agree that changes to the medium can help. For example, Spencer Greenberg's excellent post "[How You Can Gain Self Control Without 'Self-Control'](#)" includes built-in, interactive flashcards to help readers absorb and remember its content.

Books do work, sort of

Matuschak acknowledges that books aren't useless*:^{*}

I'm not suggesting that all those hours were wasted. Many readers enjoyed reading those books. That's wonderful! Certainly most readers absorbed *something*, however

ineffable: points of view, ways of thinking, norms, inspiration, and so on. Indeed, for many books (and in particular most fiction), these effects are the point.

But I do think that he gives short shrift to what seems to me books' hidden function: **making an idea or a set of ideas take up an appreciable amount of room on our cognitive real estate**. And I don't just mean for more abstract "points of view, ways of thinking, norms, inspiration, and so on" but also for plain old information, both uncontroversial statements of fact as well as more contentious claims or arguments

Against the 'Light Switch' model of knowledge

Matuschak's essay implies something like a 'light switch' model of knowledge, and this probably works pretty well for, say, state capitals and other "trivia"-style facts. You know the tenth digit of pi, or you do not. The switch is on, or it is off. For so much else, though, a statement's *salience* is more important than whether we technically do or do not "know" it.

An example

I recently finished [*The Precipice*](#) by Toby Ord. His headline claim is that humanity faces greater than a 15% risk of "existential catastrophe" in the next hundred years. Before downloading the audiobook, I already "knew" this statistic. In Matuschak's model, the flip for this proposition was turned to "on" in my brain, in the same way that I "know" that Annapolis is the capital of Maryland.

Further, I have either already forgotten or will soon forget at least 95% of the book's ancillary facts and arguments. Right now, I do remember Ord's description of a few terrifyingly close nuclear calls during the cold war, that he thinks unaligned artificial intelligence constitutes about 10 percentage points of the 16.6% risk, and a few of his proposals to improve international coordination among other things. If I were to enumerate every single fact and argument I remember, though, I have no doubt it would be far fewer than 5% of all those in the book.

But most of the book's impact comes not from flipping any of these "fact switches" to 'on' - it comes from making that "one in six chance of extinction" much more *salient* to me than it was before. How does it do this? I'm not a psychologist, but it seems likely that 99% of the substance of the book is basically a trick to get me to mull over a cluster of ideas for a while.

Before, I knew that a very smart person whom I admire and respect thinks that humanity is basically playing Russian Roulette. Now, I *really* know that a very smart person whom I admire and respect thinks that humanity is basically playing Russian Roulette.

This is important! Lots of people "know" that climate change is a big deal, but not enough people *really* know that climate change is a big deal. In more precise language, the *salience* of climate change is probably relatively low even for those who are seriously concerned about it.

To most of us, I'm guessing, COVID is a much more salient crisis. Perhaps this is appropriate, but I doubt it. What might increase the salience of climate change? Spending ~10 hours reading or listening to David Wallace Wells explain why (as the first line of [*An Uninhabitable Earth*](#) reads) "It is worse, much worse, than you think."

Another one

The very first "serious" non-fiction book I remember reading on my own is [*A Hope in the Unseen*](#), probably 12 or 13 years ago. Despite its 384 pages and likely tens of thousands of individual propositions, my entire memory of the book is little more than the single sentence *a smart poor black kid struggles to get into college, but is finally admitted to Brown*.

Under the ‘light switch’ model of knowledge, it would have been much more efficient to simply memorize that italicized sentence. In fact, I “know” more information in the book after skimming the Wikipedia page just now than I remembered from reading the whole thing.

But that’s not the point. Because the book probably took me ~8 hours of reading spread over three weeks or so, my brain was forced to process a single, coherent narrative for quite a bit of time. The sentence “a smart poor black kid struggles to get into college, but is finally admitted to Brown” and a maybe a few other vague associations with this sentence have taken up a real amount of my cognitive real estate for the last 12 years. Not much—I don’t walk around thinking about Cedrick, the protagonist, all day—but more than could have been achieved by reading the Wikipedia article no matter how recently.

Books are clandestine salience-building devices

My hypothesis is that—in agreement with Matuschak—the thousands upon thousands of bits of information conveyed in a typical nonfiction book don’t effectively transmit themselves to a typical reader, but they do force the reader to marinate in the book’s main few ideas for enough time for something important to happen.

Another lens

To put words in someone else’s mouth, I think Matuschak would say that, for the purpose of conveying information, it would be much more efficient to read a very short summary than to read an entire book. After all, we never remember more than the summary’s content, and generally remember much less.

I think this is incorrect.

After a few weeks, months, or years, a person who read the book won’t remember more than the short summary’s content, **but the person who read the short summary won’t remember anything at all.** Or, if the latter does remember something, the former will remember far more.

On one level, this is completely trivial. People remember more stuff if they learn more stuff, and a book has more stuff inside than a summary. Duh.

But the key point to remember is that (to a rough approximation), reading a book doesn’t convey any information *not included* in the summary. All the thousands of extra sentences and anecdotes and facts aren’t there to be remembered. They’re there to trick the reader into spending enough time mentally interacting with a few key ideas.

So, if you’re an author who wants to convey an actual set of, say, three or four main ideas, you could write a couple dozen articles or a couple hundred tweets and hope that a reader decides to read every single one. Or, you could write one book, filled with tens of thousands of quasi-redundant statements that will rapidly be forgotten, that you can reasonably expect a good proportion of initial readers to fully consume.

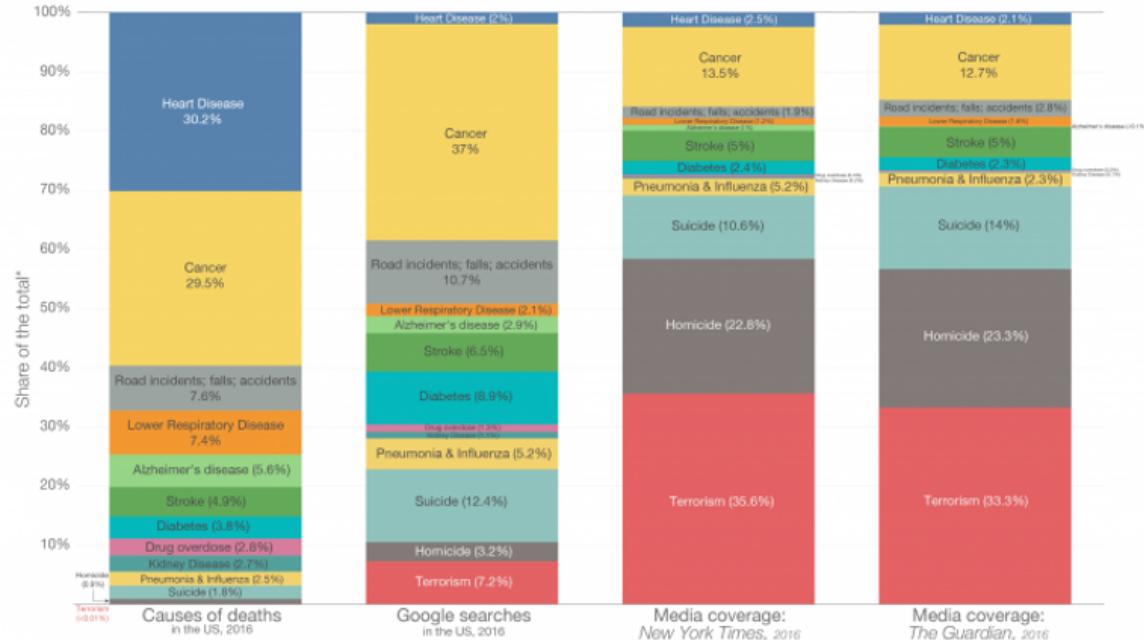
Mechanisms

At a psychological or neuroscientific level, I don’t really know what’s going on, but I can speculate. We know that [availability bias](#) disposes us to treat the ease of recall of a memory as a proxy for its importance, and the quantity of recalled events as a proxy for actual

frequency. That's why it's surprising that about as many Americans [die from asthma](#) in a single year than have been [killed in terror attacks since 1970](#) (more than 75% on 9/11).

Causes of death in the US

What Americans die from, what they search on Google, and what the media reports on



*This represents each cause's share of the top ten causes of death in the US plus homicides, drug overdoses and terrorism. Collectively these 13 causes accounted for approximately 88% of deaths in the US in 2016. Full breakdown of causes of death can be found at the CDC's WONDER public health database: <https://wonder.cdc.gov/>

Based on data from Shen et al (2018) – Death: reality vs. reported. All data available at: <https://overnsen24.github.io/charting-death>

All data refers to 2016.

Not all causes of death are shown: Shown is the data on the ten leading causes of death in the United States plus drug overdoses, homicides and terrorism. All values are normalized to 100% (they represent their relative share of the top causes, rather than absolute counts (e.g. deaths). The causes of death shown here account for approximately 88% of total deaths in the United States in 2016.

This is a visualization from OurWorldInData.org, where you find data and research on how the world is changing. Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Worth noting how similar the NYT and Guardian coverage is

Even for those in the intelligentsia who "know" that heart disease and vehicle accidents kill orders of magnitude more people than do terrorists, or that [nearly two thirds of U.S. gun deaths are suicides](#) and only [0.2% come from "mass shootings."](#) it is difficult to shed availability bias's distorting influence. Mass shootings don't intuitively seem like 0.2% of the problem. This isn't to say that we should care about everything in exact proportion to its frequency or "objective" importance, but the two should probably be more closely linked than they are.

Back to books

Anyway, I think that books are basically mechanisms to leverage this availability heuristic. Even after reading that "mass shootings" (not sure how these are defined) constitute a tiny proportion of gun deaths, their disproportionate media coverage (perhaps for good reason - that's not the point of this post) maintains their outsized share of our cognitive real estate.

The sheer length of time required to read a book is analogous to regularly watching violent crime reporting in the nightly news, albeit hopefully to a better end. You'll rapidly forget 99% of the book's statements just as you'll rapidly forget 99% of the nuances of every news story. All those forgotten propositions are doing something important, though, for better or for worse: keeping you engaged with a few central ideas long enough for them to make an impact.

Covid 4/9: Another Vaccine Passport Objection

I've been travelling to New York City once again, and it's been a busy week, including putting in bids on multiple different apartments. Developments are ongoing, so I've been even busier than usual and it's likely some stuff has slipped through the cracks.

The overall situation seems unchanged. As people take more risks and the new strains dominate, we've settled into about a 10% week over week growth in positive rates, but due to vaccinations death rates continue to decline. That means that one's risk as an unvaccinated person continues to rise, even as risk on a population level goes down. With everyone eligible now or at least soon, I'd urge once again for everyone who isn't yet vaccinated to schedule an appointment as soon as possible.

Many places that have less robust vaccination efforts than America are seeing record high case numbers and hospitalization numbers, as the new strains overwhelm efforts that were previously calibrated to let things stabilize. Things are quite bad out there in many countries. Check your local situation, and act accordingly.

Vaccination rates have stalled out in America, but at 3 million shots a day, which will still get us there pretty soon. Unless something unexpectedly makes our situation worse, we should reach the tipping point on vaccinations within a month or two, at most within three, after which we see a steady and accelerating improvement across the board.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 5.3% (up 0.4%) and deaths are unchanged.

Result (on Friday):

In the past week in the U.S. ...

New daily reported **cases rose 0.1% ↑**

New daily reported **deaths fell 15.7% ↓**

Covid-related **hospitalizations rose 5.3% ↑** [Read more](#)

Among reported tests, **the positivity rate was 5.5%**.

The **number of tests reported fell 13.6% ↓** from the previous week. [Read more](#)

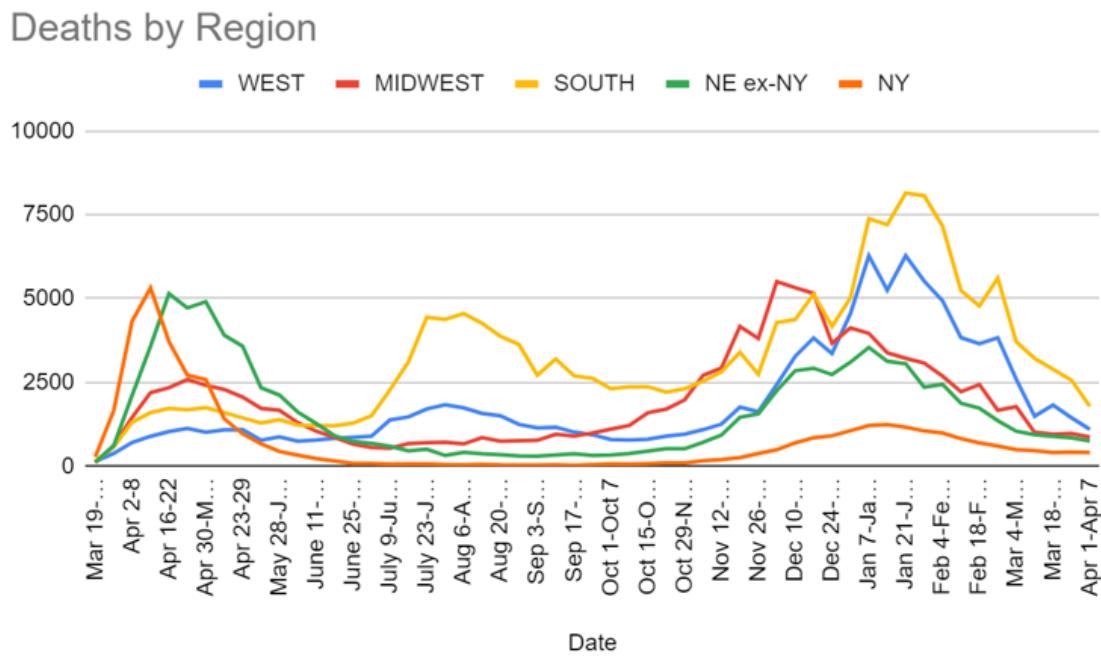
Deaths falling a lot indicates that last week's failure to drop was a measurement error, and the numbers from Wikipedia were more accurate. On reflection, I should have realized this and made a better prediction. The positive test rate number matches my intuition.

I notice I am confused by the continuing decline in test counts. It's faster than the rate of new vaccinations, and the number of cases isn't declining. I haven't seen a good explanation. Seems like an important piece of the puzzle.

Prediction for next week: Positivity rate of 5.9% (up 0.4%) and deaths decline by 8%.

Deaths

Oklahoma reported 1,716 deaths due to finding previously uncounted ones. I'm calling that 100 deaths from this week based on their previous weeks, and assuming the rest were in the past.

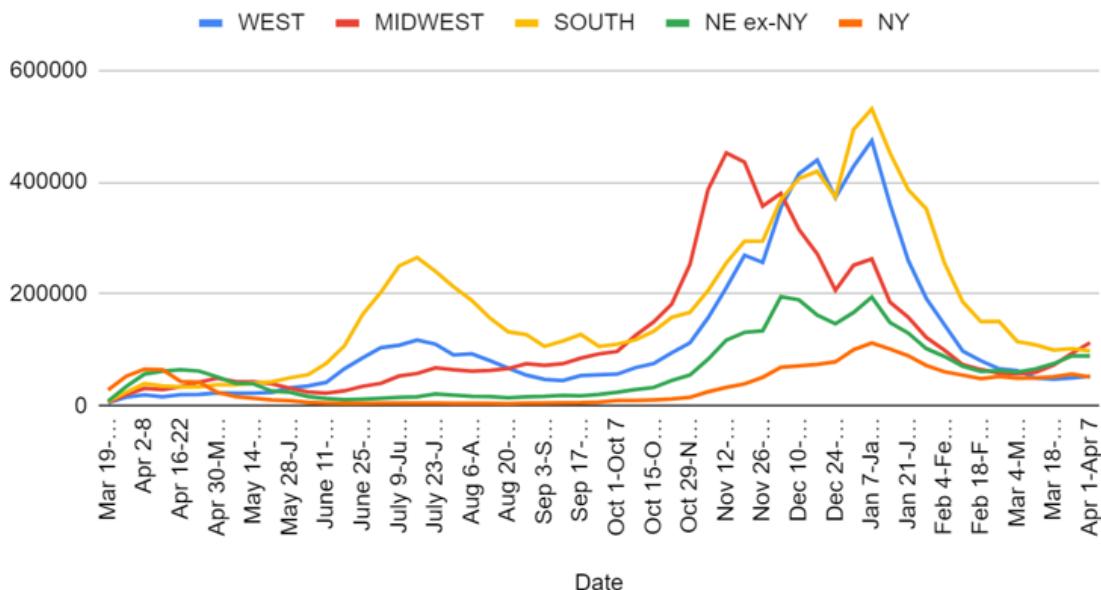


Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Feb 18-Feb 24	3652	2433	4782	2427	13294
Feb 25-Mar 3	3834	1669	5610	1958	13071
Mar 4-Mar 10	2595	1775	3714	1539	9623
Mar 11-Mar 17	1492	1010	3217	1402	7121
Mar 18-Mar 24	1823	957	2895	1294	6969
Mar 25-Mar 31	1445	976	2564	1262	6247
Apr 1-Apr 7	1098	867	1789	1160	4914

Cases may have stopped falling several weeks ago, but deaths continue to improve each week across the board, with improvement even in the Midwest. Vaccinations were not prioritized perfectly, but they were prioritized well, and they protect even better against death than infection. It makes sense that deaths continue to decline, and I again affirm that my prediction on this last week was bad.

Cases

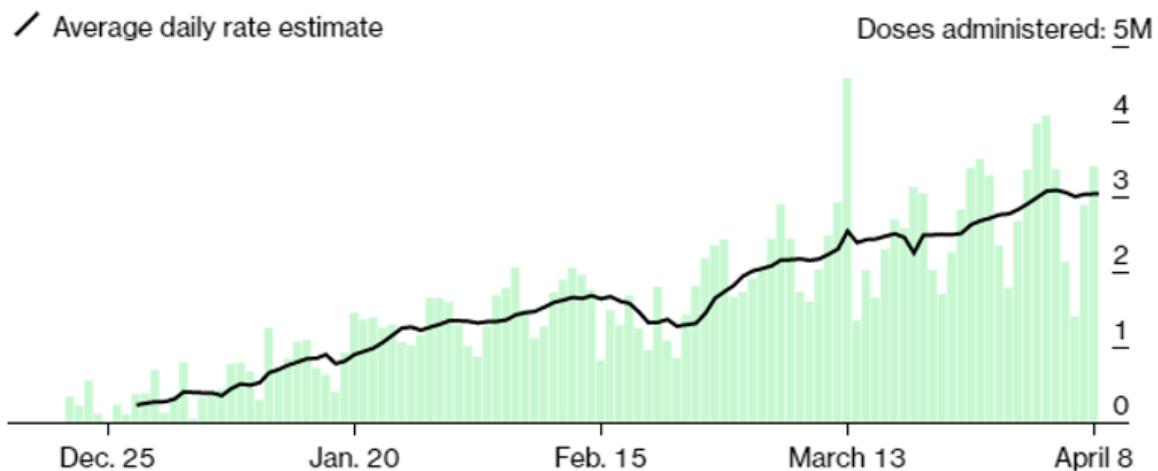
Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Feb 18-Feb 24	80,625	64,857	150,493	110,339
Feb 25-Mar 3	66,151	58,295	151,253	115,426
Mar 4-Mar 10	62,935	57,262	114,830	109,916
Mar 11-Mar 17	49,696	59,881	109,141	115,893
Mar 18-Mar 24	47,921	72,810	99,568	127,421
Mar 25-Mar 31	49,669	93,690	102,134	145,933
Apr 1-Apr 7	52,891	112,848	98,390	140,739

Vaccinations

In the U.S., more Americans have received at least one dose than have tested positive for the virus since the pandemic began. So far, **175 million doses** have been given. In the last week, an average of **3.04 million doses per day** were administered.



112 million vaccinated

The number of people who have received at least one dose of the vaccine, covering **41.9% of the eligible population, 16 and older** and **33.7% of the total population.**

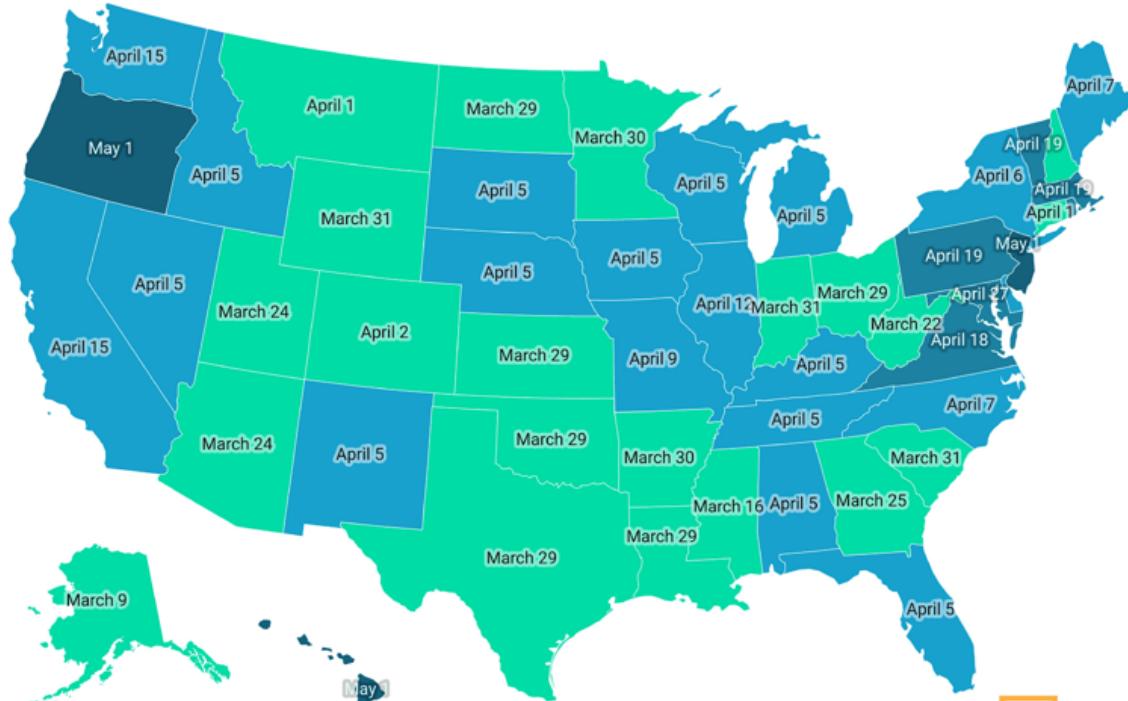
Progress on the rate of vaccination seems to have stalled out for now, which is disappointing. Still, three million shots per day is pretty good, and one third of the population has had its first dose.

[Vaccine eligibility map is complete.](#) At this point, most of the light blue states (not California, Washington or Illinois) have open eligibility, as do all the green ones, so within the week only six states will be left, and everyone's agreed to finish it by May 1 at the latest. Since this was created, New Jersey moved up a few weeks, so it's only Oregon and Hawaii waiting out the

clock.

States Opening Eligibility to All Adults

Open to all 16+ End of March Beginning of April End of April Beginning of May



Vaccine Passport Hype

There was a lot of vibrant discussion of vaccine passports in the comments, both on the blog and on the crosspost at LessWrong. Topics ranged from the technical feasibility of various ways to conserve privacy while ensuring accuracy, to various strong opinions about how we should operate in general once vaccinations are freely available.

As many pointed out, the conflict between security and privacy is difficult in practice. Usually we fail at both, such as a driver's license to buy alcohol, where it's not that difficult to fake and also reveals who you are, although without any automatic record. Having a QR code unique to the individual means some "they" could track the QR code, if desired. Having a one-time QR code continuously generated wouldn't work with people without phones and might generally be tricky. And of course, if they can't tell who you are claiming to be, there's nothing stopping you from using someone else's QR code by borrowing their phone or what not.

As usual, if we have a system that effectively requires you to *either* get vaccinated *or* have a willingness to use an app to fool the system (i.e. lie about your status) that is a subsidy to the dishonest. That can be difficult to avoid.

Overall, I think I'm fine with a system that's relatively easy to fool at any given time for those making an effort, in order to preserve privacy. There's still a record of who was and wasn't vaccinated, so when it matters enough, records can still be kept and cross-checked. It also lets those running venues, from restaurants to stadiums, avoid blameworthiness and get on

with the show. Everyone shows their passports, and life can resume. If a few people aren't vaccinated, they might infect each other, but if they're making an effort to get in via fraud then I feel fine with that being the consequence.

My guess is such fraud will be rare if it's even a relatively trivial inconvenience. It's easier to get the vaccine, even if the fraud isn't that difficult. I already know of several concrete examples from this past week. For example, I got my first post-vaccination haircut, and the (young) barber said she was going to get vaccinated in order to be allowed to do stuff. I pointed out that it also would be nice not to get Covid-19, and despite *the other barber pointing out he still didn't have his sense of smell back after catching it*, she didn't seem to notice that the vaccine might have a health benefit to her. And yet, she (quite reasonably) was eager to get it the moment she's eligible.

What I see playing out so far is exactly what one would hope, which is that *the expectation of needing to be vaccinated* is causing people to get vaccinated, without any need to actually impose meaningful restrictions. If that works, then we won't have to.

Others disagree, including large states. [Texas is banning the use of vaccination passports](#) (WaPo). Florida has taken a similar position. I am curious to see if this results in situations where their venues are often not at full capacity for longer than in other places, despite starting out ahead of the curve, or if the culture that disdains the passports also allows life to return to normal without them and shrugs at the consequences. That shrug will hopefully be reasonable not that long from now.

I missed at least one good objection. From the comments:

Jiro 4h ⚭ < 2 >

You left out one: the possibility that people will be deliberately denied COVID vaccination passports for other reasons than not having a COVID vaccination. It's like when Facebook denied the "verified" status to people who were verified, but who said politically inconvenient things.

This is going to be a risk when the passport is "private" but becomes ubiquitous, especially when encouraged by the government.

Reply

If people depend on a vaccine passport, then anyone who can take that passport away will be tempted to use or leverage that power. Would the government take the passport away from you if there's a warrant out for your arrest? If you have unpaid parking tickets? If you attend a protest?

Or, you could raise the more generalized version of this objection. Also from the comments:



Craken says:

April 6, 2021 at 6:47 pm (Edit)

Vaccine passports would accelerate techno-authoritarianism. That is reason enough to oppose them. That some young, low risk people would be infected for lack of such passports and some business lost—these are prices worth paying. I say that as someone with skin in both of these games. The price of skipping passports is short term; techno-authoritarianism is long term. The tech lords have invented the most powerful, sophisticated form of fascism. Part of its sophistication is its fashionable rhetoric of self-justification and its covert/overt coordination with the permanent government. The worst part is that these entities might just make fascism truly competent (at least temporarily).

[Reply](#)

Let's generalize the argument, which is a strong one. The proposal the vaccine passport is really making is to gate freedom of movement with a government document on your phone. This is 'papers, please' everywhere, with GPS tracking and check-in logging plausibly attached. It isn't simply a question of whether the government will abuse the records to track people. It's a question of what will happen once it realizes that it can change the requirements to get papers, or what information is revealed by those papers. Historically this all ends quite poorly, [and one should quote Benjamin Franklin](#).

I should have taken this objection seriously enough to list it and respond to it, and that I didn't do so worries me. It really should be a 101-style extrapolation, even if there are good reasons not to expect such outcomes here. At a minimum, it definitely does lower the bar for similar future systems that are more likely to be abused – accelerate such trends, as Craken puts it.

What's scary is when I try to put a probability on 'this turns into a broader thing that outlives Covid and people have to show the computer has them in good standing to get on a train' or something like that. I still think it's a super low probability thing, maybe 2% (it's a Fermi answer, but a considered one), but that's 2% of a really, really bad outcome. It's not obvious one should take those odds. The argument that convinced me is that in the 2% of worlds where this happens, I think mostly it was going to happen anyway within a reasonably short time frame, one way or another.

So on reflection I think this a real downside, but that 'give everyone the information that matters and let people do what they want with it' is still the right attitude. Your vaccination status simply isn't a private concern, it matters to those around you. The alternative to a passport is people either literally laminating their vaccine cards (hence CNN running with the headline 'things to do before you laminate your vaccine card') and using a super-easy-to-fool system, or banning people from asking about vaccine status, and those alternatives seem far worse.

This also relates to [this other instinct](#), which I also understand:



PoliMath @politicalmath · 8h

...

The real problem with vaccine passports is the number of people who, when confronted with a collective action problem, skip right over that idea that they could engage in persuasion and go straight to government coercion

I don't trust that instinct

15

41

316



If I felt this was centrally government coercion, I'd have a different view. Instead, I view this as providing for the flow of necessary information to let us all make informed decisions, which in turn enables behavior that would otherwise effectively be shut down entirely by coercion and/or by health risks, and the worry is that future action will transform the passports from vaccination information flow into a more general case of the government coercion thing.

Risk to the Unvaccinated

I haven't said this explicitly in a while, so it's worth saying it again: If you are not vaccinated, the current level of risk out there is *much higher* than the graphs and charts naively imply. On top of that, the cost of getting Covid now is *much higher* than it was earlier, both because the new dominant strain is deadlier, and also because the main benefit of getting infected – that you can't get infected again – no longer matters much since you'll have vaccine access soon either way and things aren't so bad out there that prevention is hopeless. You won't even be able to skip the vaccine, due to people requiring it (plus it's a good idea anyway, since the cost is trivial).

Not only have we vaccinated over a quarter of the population, and given one dose to over a third of the population, we've done so with an emphasis on those most at risk.

That means that if you're in the remaining two thirds, not only is your risk a third higher than it looks (e.g. almost all the infections will happen to unvaccinated people) but also the risk of death is more than double what it appears, as those at risk have largely been vaccinated early.

Thus, **if you have not yet been vaccinated, you need to treat things as much more dangerous, relative to the past, than the numbers indicate. Over time, this effect will get larger.**

[This Atlantic article by Katherine Wu](#) points out the phenomenon of people seeing vaccinated people going around doing things, and various gatherings happening, and the resulting temptation to take more risks and feel as if things are safer now. They're a lot safer than they were in December, but in many places and in America overall they're much *less* safe than they were any time before October. Don't lose sight of that.

Interacting with vaccinated people still has very little risk involved, but using the vaccinated as an excuse to do things involving other unvaccinated people, that you wouldn't have done in a situation like the one in October, is a trap to be careful about.

Did the FDA Get One Right?

Two weeks ago, I noted that the FDA hadn't even approved Emergent to make the J&J vaccine, and that this was holding up production to the extent that it looked like all the doses actually being using ingredients from abroad instead:

Emergent is already sending millions of doses to Catalent, the people said. But those shots cannot be used until Emergent receives its own FDA authorization. Catalent has not yet responded to questions about which company made the active ingredient for the doses it has begun shipping out. But a person familiar with the matter said J&J is currently using drug substance flown in from the Netherlands, not substance produced by Emergent, to manufacture its vaccine.

The whole debacle makes it clear that the FDA is causing far more delays in vaccine manufacturing and delivery with its red tape than one would guess only looking at the topline delays in authorizations.

Then the week afterwards, the world learned that Emergent had put ingredients for the AstraZeneca vaccine into a batch of Johnson & Johnson vaccine, running the entire batch of 15 million doses. That's a true disaster, enough to set back the overall American vaccine effort by over a week. The mind boggles at how many lives we'll lose and how much economic value was lost, including being later to start exporting overseas.

I didn't connect these two events at the time in my head, so I didn't mention it, and deservedly got this comment challenging that:



mike23123 says:

April 2, 2021 at 4:01 am (Edit)

Hey, where's the mea culpa for what you wrote last week about Emergent (the guys that screwed up 15mil vaccine doses)?

Last week's post included this paragraph after an image of text talking about how Emergent was still waiting on FDA approval:

> The whole debacle makes it clear that the FDA is causing far more delays in vaccine manufacturing and delivery with its red tape than one would guess only looking at the topline delays in authorizations.

And now it's come out that Emergent has serious quality problems, which makes the FDA's hesitancy to approve them seem pretty reasonable.

I'm not going to claim this proves the FDA is reasonable generally, but I'm disappointed you didn't update at all on FDA-reasonableness in this week's post.

It's hard to argue against the claim that if someone refuses to give someone else approval to do something, and the next week they massively screw up doing it, and also it turns out they've had similar (if I presume less dramatic) issues in the past, that the refusal to approve

has to look at least somewhat reasonable. One needs to notice this, and update, or if not updating then clarify why that shouldn't be the case.

At a minimum, this should have been explicitly noticed.

One incomplete but important question to ask is, in this particular case, what did the FDA's intervention change?

Then we can ask the more important question of expected value. What does this type of intervention change in expectation, in general, and at what cost?

Let's start with the obvious. Did the FDA's refusal to give approval to Emergent prevent a disaster? Would we have otherwise been looking at giving out 15 million contaminated doses, likely useless and potentially dangerous?

If so, then this is strong evidence that the FDA being picky about who can get manufacturing approval has high value, whether or not it justifies the costs involved, and that would transfer a non-zero amount to other types of FDA approvals as well. The whole idea is to guard against disasters.

I have not seen any details on *how the error was noticed*, so we can't be certain of the counterfactual, but my very very strong prior is that this error gets caught in all worlds. The composition of the batch was wrong. No sane major corporation is going to have a batch of 15 million doses and not test it to see if it is the thing it was meant to be before shipping it out, and having something in there that wasn't supposed to be there should get found many times over by the tests they'd run in all worlds. Ordinary corporate reputation and liability are more than enough to motivate catching this error.

My model says that's not how the FDA is even trying to prevent this mistake when it withholds approval. The FDA's strategy for withholding approval is to *force Johnson & Johnson to use a better manufacturer using better procedures*. The failure to get approval, based on the FDA noticing that Emergent does not know what it is doing, is a heavy incentive for J&J to find another partner here. What is interesting is that this *didn't* happen.

That's the part that's still baffling me. If Emergent had a bunch of violations already, and was going to have trouble getting approved slash hadn't already been approved, why did J&J use them rather than someone else? My understanding is that J&J's vaccine uses relatively well-established manufacturing procedures, so they had plenty of options especially given their lead time and deep pockets, and also there wasn't much risk that key knowledge would leak out in this fashion. That also implies of course that they should have been using multiple factories to make a lot more vaccine doses faster, although there could be another limiting factor somewhere on the ingredient list.

The benefit of the whole FDA system comes when corporations look ahead to the approval process, notice they won't get approved unless they did things right, and then people do their jobs right in the first place, or it comes when the FDA notices something bad that wouldn't have otherwise been noticed and stopped, and stops it. The second mechanism is important to the extent it happens.

The question here is why that first mechanism, which is supposed to be how you stop this disaster, failed. And by implication, how often this mechanism silently prevents disasters in other situations, that would otherwise happen. It's a thankless job to prevent disasters that, if you succeed, no one knows were ever an issue. That's the dog that didn't bark. I don't think that it happens often, because I think there are lots of other robust checks that would work instead, but it's reasonable to disagree with that.

The blame in this particular case looks like it falls on J&J (and of course on Emergent), to the extent that I'd want to downgrade my view of their prospects and stock on the basis of J&J likely being less good at what they do than we thought. J&J knew full well Emergent was a

risky choice here, without the necessary experience, and used them anyway. If anything, it seems more like the FDA ended up effectively *forcing J&J's hand* because regulation made it much harder to build a new plant or spin up a new partnership, which would have required a whole new set of approvals across the board. If they didn't have those obligations, they would have had other options.

By tying everyone's hands and delaying things across the board, and generally constraining the range of possible actions and actors, my guess is the FDA made this disaster *more likely*. They definitely made it *more of a problem*, because without them we'd be way ahead of the current schedule for vaccine production across the board. Certainly 'catching this' does not justify the red tape they've thrown over vaccines and also everything else.

The FDA did, to its credit, provide strong evidence that its decision to withhold approvals was based on a legitimate real world consideration, and that in general, however slow and cumbersome what it does might be, it at least does provide some amount of quality checks and safety standards. That's much better than not doing that. They're a sane regulator, responding to their incentives, which put very little weight on production they prevent or innovation they stifle, the time and money and lives lost through inaction, and almost all focus on avoiding disasters due to action, or avoiding being blameworthy for any disasters due to action that aren't prevented. Within that framework, one could plausibly argue they did their jobs quite well here.

The goal is to change the incentives and the job.

One can also ask how exactly we got into this mess, since Emergent is so obviously a terrible choice on so many levels, and the answer is that they're a government contractor hired explicitly to handle exactly this problem, except *they chose a company with a long history of problems and which the FDA refused (with reasonable justifications) to certify*:

WASHINGTON — More than eight years ago, the federal government invested in an insurance policy against vaccine shortages during a pandemic. It paid Emergent BioSolutions, a Maryland biotech firm known for producing anthrax vaccines, to have a factory in Baltimore always at the ready.

When the coronavirus pandemic arrived, the factory became the main U.S. location for manufacturing Covid-19 vaccines developed by Johnson & Johnson and AstraZeneca, churning out about 150 million doses as of last week.

But so far not a single dose has been usable because regulators have not yet certified the factory to allow the vaccines to be distributed to the public. Last week, Emergent said it would destroy up to 15 million doses' worth of the Johnson & Johnson vaccine after contamination with the AstraZeneca vaccine was discovered.

Emergent and government health officials have long touted their partnership as a success, but an examination by The New York Times of manufacturing practices at the Baltimore facility found serious problems, including a corporate culture that often ignored or deflected missteps and a government sponsor, the Biomedical Advanced Research and Development Authority, that acted more as a partner than a policeman.

Emergent is a longtime government contractor that has spent much of the last two decades cornering a lucrative market in federal spending on biodefense. [The Times reported last month](#) that sales of its anthrax vaccines to the Strategic National Stockpile accounted for nearly half of the stockpile's half-billion-dollar annual budget throughout most of the last decade, leaving the federal government with less money to buy supplies needed in a pandemic.

It was a great idea to pay to have capacity always at the ready. The problem was that this was a contract awarded on the basis of politics and power, rather than by the market, and thus people whose focus was on winning contracts via political power games, and which not coincidentally didn't have that much to do all day because no one else was that eager to hire them, beat out the places that would have been capable of actually doing the job.

This makes J&J's fate somewhat more understandable. They chose a terrible partner, but it wasn't entirely their choice. Once again, the free market fails due to lack of a free market, requiring intervention from regulators.

In Other News

[England offers free twice-a-week at-home tests to actual everyone](#). It's about time. A day late to be sure, but all the pounds are there.

Meanwhile, we're a little behind that, [but at least we've... legalized Abbott to do home testing? \(Source\)](#)



Jim O'Neill @regardthefrost · Apr 4

...

FDA finally legalizes home testing, seven months after admitting Abbott's test is accurate, and 15 months after the SARS-CoV-2 genome was published.

abbott.com/corpnewsroom/d...



Jim O'Neill @regardthefrost · Aug 27, 2020

Abbott's \$5 lateral flow antigen test is the size of a credit card and gives results in 15 minutes. It seems perfect for home testing. However, FDA bans home use and also bans testing of presymptomatic or asymptomatic people. fda.gov/media/141567/d...

[Kai compares what happened now to the response to AIDS.](#)

Starting to see frequent posts on Twitter about available vaccine appointments. [For example, here was Sam Black](#), although that one's in the past. [Here's another from LA](#), presumably also expired by now. The bottom line is, keep an eye out for such opportunities if you or anyone you know has had trouble finding an appointment, or would like an earlier one.

[Our government allocates life-saving medicine explicitly on the basis of the racial identity of your household](#). Then it restricts what people's movement and what they can do if they haven't yet had access to that medicine.

[Chronicle of a parent who finally lost it when their child's school closed for ten days due to two Covid cases... for a fifth time](#). New York City is finally changing that rule. It was even more absurd than it sounds:

What really surprised Schechter-Perkins and others were the specifics of the rule, which [you can read](#) yourself: When the city's surveillance testing — 20% of a school's population every week — finds two or more positive cases that are connected, just the affected classrooms are shut and close contacts are told to quarantine. But if the cases can't be traced or connected, then the entire school building will be shut for 10 days.

Put another way: A school *won't* be shut down if there is evidence of some spread, but it *will* be closed if there is no evidence of in-school transmission.

The city doesn't publish figures on how many of its roughly 1,800 public schools have been closed, but its website shows a total of [2,358 building closures](#) this school year as of Thursday.

Other experts had similar reactions. "That's crazy. That's nuts," said [Dr. Benjamin Linas](#), another epidemiology professor at Boston University School of Medicine. "It doesn't really make sense," said [Dr. Leana Wen](#), a public health professor at George Washington University who served as Baltimore's health commissioner. "It's not evidence driven," said [Eyal Oren](#), an epidemiologist at the San Diego State University School of Health.

From the perspective of actual people leading physical lives, this is indeed nuts. From a blameworthiness perspective, and a 'state of grace/purity' versus 'state of sin/contamination' perspective, though, [and a very explicit 'if I can say it's keeping kids safe I don't have to care whether it makes any sense or drives people crazy and disrupts their lives' perspective](#) (that was the first reply to the Tweet announcing the story), it makes perfect sense.

[Efforts to promote First Doses First continue](#), remain obviously correct, remain doomed to failure. I don't see any way to turn this one around at this point. We are requiring full

vaccination to engage in everyday life, so people aren't going to accept going around half-vaccinated even if it's 90% as good, so long as they're told to treat it as 0% as good.

Gentle reminder that if you book your appointment for a given day, [you'd better be ready to get the second dose exactly the right number of weeks later](#) to the minute, or you'll need to actively reschedule at a minimum:



Nate Silver @NateSilver538 · Apr 3

...

Replying to [@zeynep](#) [@ZoeMcLaren](#) and [@michaelmina_lab](#)

The state-run sites in NY automatically sign you up for a second Pfizer dose *exactly* three weeks (to the minute!) later. I guess they want to streamline things (and the state-run sites are quite efficient) but they don't even check to see if you have a conflict, etc.

8

2

33



zeynep tufekci @zeynep · Apr 3

...

Yeah, many places are exactly like that. Even the CDC and the WHO—on the cautious side—say six weeks is okay (I'm putting aside everything else). I have not yet heard of a single reason beyond the obvious (speeding up trials) that the Pfizer has an unprecedentedly short interval.

3

1

29



[Your periodic reminder department:](#)



Eliezer Yudkowsky ✅ @ESYudkowsky · Apr 3

...

To state the obvious: any country that had learned anything would be preparing NOW, not for the next Covid-19, but for THE future pandemic. That means, eg, having enough in-country mRNA vaccine production capacity to get everyone in 1 month. They'd be building toward that NOW.



61



308



1.9K



Eliezer Yudkowsky ✅ @ESYudkowsky · Apr 4

...



Paul Sztorc @Truthcoin · Apr 4

It is obvious.

Also: UV lights everywhere! For a long time, our species did not separate drinking water from sewer water. That's like our air today! Every HVAC should have one.

Kills *all* airborne infection (virus bacteria parasite), not just covid!

lesswrong.com/posts/L8KGSDch...

[Show this thread](#)



4



4



90



I think Eliezer is flat out correct here that it's dirt cheap to build out what we might call 'generic mRNA vaccine capacity' and that even if the pandemic was over we should be doing enough of that to vaccinate everyone, except I'd say the USA should be planning to make enough for the whole world, not only Americans, because the payoffs here are so ridiculous. Instead, we aren't even doing this for the *current* pandemic. The footnote on never even [investigating far-UV light](#) properly is also a good periodic reminder.

Our lockdown rules being applied differently depending on which groups and activities are favored by politics and power [is the default, not an exception](#).

We repeat, [suicides were actually down last year, rather than up](#). Your model needs to account for this.

Table. Number of Deaths for Leading Causes of Death, US, 2015-2020^a

Cause of death	No. of deaths by year					
	2015	2016	2017	2018	2019	2020
Total deaths	2 712 630	2 744 248	2 813 503	2 839 205	2 854 838	3 358 814
Heart disease	633 842	635 260	647 457	655 381	659 041	690 882
Cancer	595 930	598 038	599 108	599 274	599 601	598 932
COVID-19 ^b						345 323
Unintentional injuries	146 571	161 374	169 936	167 127	173 040	192 176
Stroke	140 323	142 142	146 383	147 810	150 005	159 050
Chronic lower respiratory diseases	155 041	154 596	160 201	159 486	156 979	151 637
Alzheimer disease	110 561	116 103	121 404	122 019	121 499	133 382
Diabetes	79 535	80 058	83 564	84 946	87 647	101 106
Influenza and pneumonia	57 062	51 537	55 672	59 120	49 783	53 495
Kidney disease	49 959	50 046	50 633	51 386	51 565	52 260
Suicide	44 193	44 965	47 173	48 344	47 511	44 834

^a Leading causes are classified according to underlying cause and presented according to the number of deaths among US residents. For more information, see the article by Heron.⁴ Source: National Center for Health Statistics. National Vital Statistics System: mortality statistics (<http://www.cdc.gov/nchs/deaths.htm>). Data for 2015-2019 are final; data for 2020 are provisional.

^b Deaths with confirmed or presumed COVID-19, coded to *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* code U07.1 as the underlying cause of death.

We see extra deaths from heart disease, Alzheimer's disease, diabetes and stroke, and in general about 140k extra deaths on net that aren't listed as being directly from Covid-19. Some of those are likely Covid-19 deaths that got misidentified. Some others are likely due to people not getting any exercise or change of scenery. What we don't see are suicides.

[CDC finally starts getting explicit on surfaces not being much of an issue](#). Still a lot of 'it depends' and 'we don't know anything about anything' but the bolded part is bolded:

The principal mode by which people are infected with SARS-CoV-2 (the virus that causes COVID-19) is through exposure to respiratory droplets carrying infectious virus. It is possible for people to be infected through contact with contaminated surfaces or objects (fomites), but the risk is generally considered to be low.



Charles Haas @ProfCharlesHaas 15h

and #ventilation is at least mentioned
"During the first 24 hours, the risk can
be reduced by increasing ventilation".
So I am starting to think that CDC is
on the road to Damascus, but has
some distance yet to get there.
Curious to see what my tweeps may
think. end/N

Q 5 ↗ 12 ❤ 72 ...

[Tyler Cowen and Russ Roberts podcast revisiting the state of the pandemic.](#) Self-recommending and as good as you'd expect if you like this kind of podcast conversation. [Cafe Hayek responds](#), arguing that Covid response has been far more damaging to social cohesion than it would have been to let things play out. Tyler in turn responds that this wasn't an option, because once things got worse the lockdowns would have happened regardless, from a worse place, and been more severe. I'm with Tyler here, the alternative to lockdowns that works is to do more earlier, the public choice of letting Covid mostly play out was never going to do anything but backfire. There are still some clear examples in the response post of utterly crazy policies in the name of prevention, of course, as I often point out.

[Health care provision across state lines, to-go cocktails and other purely good deregulations that might stick around](#) (WSJ via MR).

[Post on the option to use IP suspensions to solve the vaccine shortage](#), and alternative ways to get production to go up. That seems like a disastrous road to go down, and if formal IP was the issue we could simply *buy the companies out*, which the post points out is the correct play here, plus the post also points out that mere suspensions wouldn't actually work, because (among several other reasons) the IP that matters is not in the patents. So Moderna won't enforce its patents but that doesn't mean anyone knows how to make its vaccine, and those secrets are valuable to Moderna for future products so they won't share. Of course, we could *buy them out of that too* if we were sane, it's worth it, but there's no willingness to do so. Also of course, the simple play is to offer to pay a lot more money to get vaccine doses, and let them work out the rest, but that's crazy talk.

[Zaynep thread about claims that places have reached herd immunity.](#)

[Finally, this](#) seems very right, and it's important once you're fully vaccinated to remember to live your life once again:



Agnes Callard @AgnesCallard

1d

I worry that we sold ourselves on lockdowns & social distancing by telling ourselves it was ok, not that bad whereas in fact it was a TOTAL NIGHTMARE CATASTROPHE (albeit one that needed to happen) but we bought our own lie & now we're insufficiently motivated to make it unhappen.

16 17 4 68 ...