

# Best of LessWrong: June 2021

1. [Taboo "Outside View"](#)
2. [Intentionally Making Close Friends](#)
3. [Precognition](#)
4. [Two non-obvious lessons from microcovid.org](#)
5. [Social behavior curves, equilibria, and radicalism](#)
6. [Power dynamics as a blind spot or blurry spot in our collective world-modeling, especially around AI](#)
7. [The Point of Trade](#)
8. [The Apprentice Experiment](#)
9. [The Apprentice Thread](#)
10. [Irrational Modesty](#)
11. [Alcohol, health, and the ruthless logic of the Asian flush](#)
12. [Selection Has A Quality Ceiling](#)
13. [On the limits of idealized values](#)
14. [An Intuitive Guide to Garrabrant Induction](#)
15. [Reward Is Not Enough](#)
16. [Why did we wait so long for the threshing machine?](#)
17. [Bad names make you open the box](#)
18. [Frequent arguments about alignment](#)
19. [Environmental Structure Can Cause Instrumental Convergence](#)
20. [Book review: "Feeling Great" by David Burns](#)
21. [Attributions, Karma and better discoverability for wiki/tag features](#)
22. [Discussion: Objective Robustness and Inner Alignment Terminology](#)
23. [Scaling Networks of Trust](#)
24. [Covid 6/17: One Last Scare](#)
25. [AXRP Episode 9 - Finite Factored Sets with Scott Garrabrant](#)
26. [Empirical Observations of Objective Robustness Failures](#)
27. [Thoughts on the Alignment Implications of Scaling Language Models](#)
28. [Five Suggestions For Rationality Research and Development](#)
29. [Rogue AGI Embodies Valuable Intellectual Property](#)
30. [We need a standard set of community advice for how to financially prepare for AGI](#)
31. [The value of low-conscientiousness people on teams](#)
32. [Search-in-Territory vs Search-in-Map](#)
33. [The Language of Bird](#)
34. [Reinforcing Habits](#)
35. [Parameter counts in Machine Learning](#)
36. [Experiments with a random clock](#)
37. [The Moon is Down; I have not heard the clock](#)
38. [Announcing the Replacing Guilt audiobook](#)
39. [How do the ivermectin meta-reviews come to so different conclusions?](#)
40. [Notes on War: Grand Strategy](#)
41. [Survey on AI existential risk scenarios](#)
42. [Intro to Debt Crises](#)
43. [Avoid News, Part 2: What the Stock Market Taught Me about News](#)
44. [Changing my life in 2021, halfway through](#)
45. [Covid 6/10: Somebody Else's Problem](#)
46. [Why do patients in mental institutions get so little attention in the public discourse?](#)
47. [Big picture of phasic dopamine](#)
48. [The Generalized Product Rule](#)
49. [Covid 6/24: The Spanish Prisoner](#)
50. [Dangerous optimisation includes variance minimisation](#)

# Best of LessWrong: June 2021

1. [Taboo "Outside View"](#)
2. [Intentionally Making Close Friends](#)
3. [Precognition](#)
4. [Two non-obvious lessons from microcovid.org](#)
5. [Social behavior curves, equilibria, and radicalism](#)
6. [Power dynamics as a blind spot or blurry spot in our collective world-modeling, especially around AI](#)
7. [The Point of Trade](#)
8. [The Apprentice Experiment](#)
9. [The Apprentice Thread](#)
10. [Irrational Modesty](#)
11. [Alcohol, health, and the ruthless logic of the Asian flush](#)
12. [Selection Has A Quality Ceiling](#)
13. [On the limits of idealized values](#)
14. [An Intuitive Guide to Garrabant Induction](#)
15. [Reward Is Not Enough](#)
16. [Why did we wait so long for the threshing machine?](#)
17. [Bad names make you open the box](#)
18. [Frequent arguments about alignment](#)
19. [Environmental Structure Can Cause Instrumental Convergence](#)
20. [Book review: "Feeling Great" by David Burns](#)
21. [Attributions, Karma and better discoverability for wiki/tag features](#)
22. [Discussion: Objective Robustness and Inner Alignment Terminology](#)
23. [Scaling Networks of Trust](#)
24. [Covid 6/17: One Last Scare](#)
25. [AXRP Episode 9 - Finite Factored Sets with Scott Garrabant](#)
26. [Empirical Observations of Objective Robustness Failures](#)
27. [Thoughts on the Alignment Implications of Scaling Language Models](#)
28. [Five Suggestions For Rationality Research and Development](#)
29. [Rogue AGI Embodies Valuable Intellectual Property](#)
30. [We need a standard set of community advice for how to financially prepare for AGI](#)
31. [The value of low-conscientiousness people on teams](#)
32. [Search-in-Territory vs Search-in-Map](#)
33. [The Language of Bird](#)
34. [Reinforcing Habits](#)
35. [Parameter counts in Machine Learning](#)
36. [Experiments with a random clock](#)
37. [The Moon is Down; I have not heard the clock](#)
38. [Announcing the Replacing Guilt audiobook](#)
39. [How do the ivermectin meta-reviews come to so different conclusions?](#)
40. [Notes on War: Grand Strategy](#)
41. [Survey on AI existential risk scenarios](#)
42. [Intro to Debt Crises](#)
43. [Avoid News, Part 2: What the Stock Market Taught Me about News](#)
44. [Changing my life in 2021, halfway through](#)
45. [Covid 6/10: Somebody Else's Problem](#)
46. [Why do patients in mental institutions get so little attention in the public discourse?](#)

47. [Big picture of phasic dopamine](#)
48. [The Generalized Product Rule](#)
49. [Covid 6/24: The Spanish Prisoner](#)
50. [Dangerous optimisation includes variance minimisation](#)

# Taboo "Outside View"

No one has ever seen an AGI takeoff, so any attempt to understand it must use these outside view considerations.

—[Redacted for privacy]

What? That's exactly backwards. If we had lots of experience with past AGI takeoffs, using the outside view to predict the next one would be a lot more effective.

—My reaction

Two years ago I wrote a [deep-dive summary](#) of *Superforecasting*, and the associated scientific literature. I learned about the “Outside view” / “Inside view” distinction, and the evidence supporting it. At the time I was excited about the concept and wrote: “...*I think we should do our best to imitate these best-practices, and that means using the outside view far more than we would naturally be inclined.*”

Now that I have more experience, I think the concept is doing more harm than good in our community. The term is easily abused and its meaning has expanded too much. I recommend we permanently [taboo](#) “Outside view,” i.e. stop using the word and use more precise, less confused concepts instead. This post explains why.

## What does “Outside view” mean now?

Over the past two years I’ve noticed people (including myself!) do lots of different things in the name of the Outside View. I’ve compiled the following lists based on fuzzy memory of hundreds of conversations with dozens of people:

### Big List O’ Things People Describe As Outside View:

- **Reference class forecasting**, the practice of computing a probability of an event by looking at the frequency with which similar events occurred in similar situations. Also called comparison class forecasting. [EDIT: [Eliezer rightly points out](#) that sometimes **reasoning by analogy** is undeservedly called reference class forecasting; reference classes are supposed to be held to a much higher standard, in which your sample size is larger and the analogy is especially tight.]
- **Trend extrapolation**, e.g. “AGI implies insane GWP growth; let’s forecast AGI timelines by extrapolating GWP trends.”
- **Foxy aggregation**, the practice of using multiple methods to compute an answer and then making your final forecast be some intuition-weighted average of those methods.
- **Bias correction, in others or in oneself**, e.g. “There’s a selection effect in our community for people who think AI is a big deal, and one reason to think AI is a big deal is if you have short timelines, so I’m going to bump my timelines estimate longer to correct for this.”
- **Deference to wisdom of the many**, e.g. expert surveys, or appeals to the efficient market hypothesis, or to conventional wisdom in some fairly large group of people such as the EA community or Western academia.
- **Anti-weirdness heuristic**, e.g. “How sure are we about all this AI stuff? It’s pretty wild, it sounds like science fiction or doomsday cult material.”
- **Priors**, e.g. “This sort of thing seems like a really rare, surprising sort of event; I guess I’m saying the prior is low / the outside view says it’s unlikely.” Note that I’ve heard this

said even in cases where the prior is *not* generated by a reference class, but rather from raw intuition.

- **Ajeya's timelines model** ([transcript of interview](#), [link to model](#))
- ... and probably many more I don't remember

## Big List O' Things People Describe As Inside View:

- **Having a gears-level model**, e.g. "Language data contains enough structure to learn human-level general intelligence with the right architecture and training setup; GPT-3 + recent theory papers indicate that this should be possible with X more data and compute..."
- **Having any model at all**, e.g. "I model AI progress as a function of compute and clock time, with the probability distribution over how much compute is needed shifting 2 OOMs lower each decade..."
- **Deference to wisdom of the few**, e.g. "the people I trust most on this matter seem to think..."
- **Intuition-based-on-detailed-imagining**, e.g. "When I imagine scaling up current AI architectures by 12 OOMs, I can see them continuing to get better at various tasks but they still wouldn't be capable of taking over the world."
- **Trend extrapolation combined with an argument for why that particular trend is the one to extrapolate**, e.g. "Your timelines rely on extrapolating compute trends, but I don't share your inside view that compute is the main driver of AI progress."
- **Drawing on subject matter expertise**, e.g. "my inside view, based on my experience in computational neuroscience, is that we are only a decade away from being able to replicate the core principles of the brain."
- **Ajeya's timelines model** (Yes, this is on both lists!)
- ... and probably many more I don't remember

## What did “Outside view” mean originally?

As far as I can tell, *it basically meant reference class forecasting*. Kaj Sotala tells me the original source of the concept (cited by [the Overcoming Bias post](#) that brought it to our community) was [this paper](#). Relevant quote: "The outside view is ... essentially ignores the details of the case at hand, and involves no attempt at detailed forecasting of the future history of the project. Instead, it focuses on the statistics of a class of cases chosen to be similar in relevant respects to the present one." If you look at the text of *Superforecasting*, the "it basically means reference class forecasting" interpretation holds up. Also, "Outside view" redirects to "[reference class forecasting](#)" in Wikipedia.

To head off an anticipated objection: I am not claiming that there is no underlying pattern to the new, expanded meanings of "outside view" and "inside view." I even have a few ideas about what the pattern is. For example, priors are sometimes based on reference classes, and even when they are instead based on intuition, that too can be thought of as reference class forecasting in the sense that intuition is often just unconscious, fuzzy pattern-matching, and pattern-matching is arguably a sort of reference class forecasting. And Ajeya's model can be thought of as inside view relative to e.g. GDP extrapolations, while also outside view relative to e.g. deferring to Dario Amodei.

However, it's easy to see patterns everywhere if you squint. These lists are still pretty diverse. I could print out all the items on both lists and then mix-and-match to create new lists/distinctions, and I bet I could come up with several at least as principled as this one.

# This expansion of meaning is bad

When people use “outside view” or “inside view” without clarifying which of the things on the above lists they mean, I am left ignorant of what exactly they are doing and how well-justified it is. People say “On the outside view, X seems unlikely to me.” I then ask them what they mean, and sometimes it turns out they are using some reference class, complete with a dataset. (Example: [Tom Davidson’s four reference classes for TAI](#)). Other times it turns out they are just using the anti-weirdness heuristic. Good thing I asked for elaboration!

Separately, various people seem to think that the appropriate way to make forecasts is to (1) use some outside-view methods, (2) use some inside-view methods, but only if you feel like you are an expert in the subject, and then (3) do a weighted sum of them all using your intuition to pick the weights. This is *not* Tetlock’s advice, nor is it the lesson from the forecasting tournaments, especially if we use the nebulous modern definition of “outside view” instead of the original definition. (For my understanding of his advice and those lessons, see [this post](#), part 5. For an entire book written by Yudkowsky on why the aforementioned forecasting method is bogus, see *Inadequate Equilibria*, especially [this chapter](#). Also, I wish to emphasize that I myself was one of these people, at least sometimes, up until recently when I noticed what I was doing!)

Finally, I think that too often the good epistemic standing of reference class forecasting is illicitly transferred to the other things in the list above. I already gave the example of the anti-weirdness heuristic; my second example will be bias correction: I sometimes see people go “There’s a bias towards X, so in accordance with the outside view I’m going to bump my estimate away from X.” *But this is a different sort of bias correction.* To see this, notice how they used *intuition* to decide how much to bump their estimate, and they [didn’t consider other biases towards or away from X](#). The original lesson was that biases could be corrected by [using reference classes](#). Bias correction via intuition may be a valid technique, but it shouldn’t be called the outside view.

I feel like it’s gotten to the point where, like, only 20% of uses of the term “outside view” involve reference classes. It seems to me that “outside view” has become an [applause light](#) and a smokescreen for over-reliance on intuition, the anti-weirdness heuristic, deference to crowd wisdom, correcting for biases in a way that is itself a gateway to more bias...

**MAINSTREAM OPINION IS CRAZY!**

**FROM THE OUTSIDE VIEW THE RATIONALISTS ARE CRAZY**

**YOU WERE SUPPOSED TO DESTROY  
INTUITION-WIELDING PUNDITS, NOT JOIN THEM!**

I considered advocating for a return to the original meaning of “outside view,” i.e. reference class forecasting. But instead I say:

# Taboo Outside View; use this list of words instead

I'm *not* recommending that we stop using reference classes! I love reference classes! I also love trend extrapolation! In fact, *for literally every tool on both lists above, I think there are situations where it is appropriate to use that tool.* Even the anti-weirdness heuristic.

What I ask is that we stop using the *words* "outside view" and "inside view." I encourage everyone to instead be more specific. Here is a big list of more specific words that I'd love to see, along with examples of how to use them:

- **Reference class forecasting**
  - "I feel like the best reference classes for AGI make it seem pretty far away in expectation."
  - "I don't think there are any good reference classes for AGI, so I think we should use other methods instead."
- **Analogy**
  - Analogy is like a reference class but with lower standards; sample size can be small and the similarities can be weaker.
  - "I'm torn between thinking of AI as a technology vs. as a new intelligent species, but I lean towards the latter."
- **Trend extrapolation**
  - "The GWP trend seems pretty relevant and we have good data on it"
  - "I claim that GPT performance trends are a better guide to AI timelines than compute or GWP or anything else, because they are more directly related."
- **Foxy aggregation (a.k.a. multiple methods)**
  - "OK that model is pretty compelling, but to stay foxy I'm only assigning it 50% weight."
- **Bias correction**
  - "I feel like things generally take longer than people expect, so I'm going to bump my timelines estimate to correct for this. How much? Eh, 2x longer seems good enough for now, but I really should look for data on this."
- **Deference**
  - "I'm deferring to the markets on this one."
  - "I think we should defer to the people building AI."
- **Anti-weirdness heuristic**
  - "How sure are we about all this AI stuff? The anti-weirdness heuristic is screaming at me here."
- **Priors**
  - "This just seems pretty implausible to me, on priors."
  - "(Ideally, say whether your prior comes from intuition or a reference class or a model. Jia points out "on priors" has similar problems as "on the outside view.")
- **Independent impression**
  - i.e. what your view would be if you weren't deferring to anyone.
  - "My independent impression is that AGI is super far away, but a lot of people I respect disagree."
- **"It seems to me that..."**
  - i.e. what your view would be if you weren't deferring to anyone or trying to correct for your own biases.
  - "It seems to me that AGI is just around the corner, but I know I'm probably getting caught up in the hype."
  - Alternatively: "I feel like..."
  - Feel free to end the sentence with "...but I am not super confident" or "...but I may be wrong."
- **Subject matter expertise**
  - "My experience with X suggests..."

- **Models**
  - “The best model, IMO, suggests that...” and “My model is...”
  - (Though beware, I sometimes hear people say “my model is...” when all they really mean is “I think...”)
- **Wild guess (a.k.a. Ass-number)**
  - “When I said 50%, that was just a wild guess, I’d probably have said something different if you asked me yesterday.”
- **Intuition**
  - “It’s not just an ass-number, it’s an intuition! Lol. But seriously though I have thought a lot about this and my intuition seems stable.”

## Conclusion

Whenever you notice yourself saying “outside view” or “inside view,” imagine a tiny Daniel Kokotajlo hopping up and down on your shoulder chirping *“Taboo outside view.”*

*Many thanks to the many people who gave comments on a draft: Vojta, Jia, Anthony, Max, Kaj, Steve, and Mark. Also thanks to various people I ran the ideas by earlier.*

# Intentionally Making Close Friends

This is a linkpost for <https://www.neelnanda.io/blog/43-making-friends>

## Introduction

One of the greatest sources of joy in my life are my close friends. People who bring excitement and novelty into my life. Who expose me to new experiences, and ways of seeing the world. Who help me learn, point out my blind spots, and correct me when I am wrong. Who I can lean on when I need support, and who lean on me in turn. Friends who help me grow more into the kind of person I want to be.

I am especially grateful for this, because up until about 4 years ago, I didn't have *any* close friends in my life. I had friends, but struggled to form real emotional connections. Moreover, it didn't even occur to me that I *could* try to do this. It wasn't that I knew how to form close friends but was too anxious to try, rather, 'try to form close friendships' was a **non-standard action**, something that never even crossed my mind. And one of my most life-changing experiments was realising that this was something I wanted, and actually trying to intentionally form close friends.

It's easy to slip into a passive mindset here, to think of emotional connections as 'something that take time' or 'need to happen naturally'. That to be intentional about things is 'inauthentic'. I think this mindset is absolutely crazy. My close friendships are one of the most important components of my life happiness. Leaving it up to chance feels like passing up an incredible opportunity. As with all important things in life, [this can be optimised](#) - further, if done right, this adds a massive amount to the lives of me and of my future close friends.

The first half of this post is the story of how I approached intentionally forming close friends, and the second half is an attempt to distill the lessons I learned from this. As such, this post is more autobiographical than most. Feel free to skip to the advice section if you don't want that. Further, what you value in close friendships is highly personal - this post will focus on what *I* want in friendships and how I try to get it, but you should adapt this to your own situation, values, and what feels missing in your life!

**Exercise:** Think about your closest friends, and how these friendships happened. What needs are you fulfilling in each other's lives? Are you happy with this state of affairs, or is something missing? What could be better?

## My story

## The Problem

Back when I was in school, I never had close friends. I had *friends*, people I liked, people I spent time with, whose company I genuinely enjoyed. But I was pretty terrible at being vulnerable and forming emotional connections. These friendships rarely went beyond the surface level. In hindsight, I expect these *could* have been far richer (and

I've formed much stronger friendships with some of these friends since!), but I never really tried.

I find it hard to introspect on exactly what the internal experience of past Neel was like, but I think the core was that trying wasn't **available** as a possible action. That I spent much of my life doing what felt socially conventional, normal and expected, for the role I saw myself in. And 'go out of your way to form emotional connections' wasn't part of that. It wasn't an action I considered, weighed up the costs and benefits, and decided against - it never even occurred to me to try. It didn't feel like a void missing from my life - things just felt normal. It was like playing a video game, and having a list of actions to choose from, like 'ask about their day', 'complain about a shared experience' or 'discuss something cool I learned recently'; but this list contained nothing about 'intentionally form an emotional connection'. It wasn't in my reference class of things I could do.

One of the core parts of my life philosophy now is the skill of **agency**, of [actually doing things](#). The skill of going out of your way to *make* opportunities. To identify what's missing in my life, and in the world. Finding the actions that I don't *need* to take, that no one else will make me take, or do for me, and deciding to take them anyway. Fixing that which is broken. Finding that which is *not* broken, and deciding to make it better anyway. Exploring and trying new things. Challenging my self-image and growing. Fundamentally escaping the mindset of needing permission, and breaking past the [illusion of doing nothing](#). I think this is one of the most valuable skills anyone can learn, and one I cherish, though I am far from perfect at it. And this experience is a large part of why I value it. Not realising I could make close friends was a **failure of agency**, an **unknown-unknown** that cut out a massive amount of potential happiness, without even realising it.

## The solution

Despite all this talk of agency, I stumbled my way out of this problem pretty much by accident. When I was 18, in my final year in school, I ended up in a long-term romantic relationship (with a girl who, thankfully, was far better at taking initiative than me!). And this was one of my first times really feeling a deep, emotional connection with someone. And, surprisingly, found that this was great, and added a ton to my life! And further, got a bunch of surface area on what emotional connections actually felt like, and how they formed.

That relationship ended in my first year of university. And as part of trying to move on and recover, I did a lot of introspection on how the relationship had changed me, and what now felt missing from my life. And one of the biggest things missing was having a deep emotional connection with someone. So I decided to fix this.

The obvious next question was, what to actually do? In full 19-year-old-Neel fashion, I took a pretty reductionist approach to this. I made a list of all the people I considered close or close-ish friends, and tried to figure out how we became close friends. And in each case, I identified the main shifts in our relationship after intense, 1-1 conversations, where we were both being emotionally vulnerable and authentic, and talking about personal things. So, to make *more* close friends, all I needed to do was engineer more of these 1-1 conversations!

This was also inspired by a time when I was 17, and at a [rationality camp for high-schoolers](#). We were doing a workshop on Comfort Zone Expansion (CoZE), where the

intention was to identify something we were uncomfortable with but wanted to explore and try in a safe environment. I and another participant noticed we were both uncomfortable with being vulnerable and authentic, and tended to use humour to deflect from anything personal. So, we decided to find a private place and spend two hours having a fully authentic conversation, with no deflection allowed. This was kinda terrifying, but also a *really* great comfort zone expansion experience, and I felt much closer to him afterwards.

One decent way of engineering an authentic 1-1 conversation is to go through a bunch of personal and vulnerability-inducing questions together, a la [36 Questions that Lead in Love](#) (after cutting the  $\frac{2}{3}$  of questions that I found dull). So I made a list of questions I considered interesting, which I expected to lead to authentic and vulnerable conversations. And then went up to the 10-20 people I felt most friendly with, explained the experiment, and asked if they'd be interested in blocking out a few hours, and going through the list together.

Somehow, this worked! About 80% of the people I asked said yes, and I felt much closer with about 50% of them afterwards. Some people were weirded out, but most of my friends were down to try the experiment. With some people the questions felt awkward, but with some people I *really* vibed. And some people were *extremely* enthusiastic about the idea from the start - I explained the idea to a guy I vaguely knew, he loved the idea and suggested doing it together, we hit it off immediately, and he's now probably my closest friend.

If you're interested in the questions, you can see the full list [here](#). Some of my favourites:

- What's the best way to get to know you as a person?
- What's your life story?
- What traits do you envy/value in those around you?
- What do you feel insecure about?
  - This one is higher variance - I don't recommend leading with it!
- What do you value in friendships? What are the best ways they add to your life?
- How, historically, have you become close to people?
- If you could design a personal set of social norms for how your friends interact with you, what would they be?
- How would other people describe you? How does this compare to how you want to be perceived?
- What in life do you get truly excited about?

## Retrospective

I am *incredibly* happy I ran this experiment. It has made my life massively better. And, in hindsight, I am still really surprised that the success rate was so high! If this idea sounds compelling, I would highly recommend people try it - it was an excellent growth experience.

That said, I still somewhat cringe looking back on that. I think having a literal list of questions made the interactions much more artificial. Since then, my conversational style has evolved to be a lot more natural, while trying to preserve the spirit. [I really like asking questions](#), and will often weave these questions into a conversation if appropriate. And strongly try to create an atmosphere where people are comfortable being honest and vulnerable, and where I show vulnerability in turn.

One of the main reasons I'd recommend others try this is that it **broke me out of a bad equilibria**. I was trapped in a 'normal' mode of conversation - making small talk, being inoffensive, feeling aversion to being weird, respecting where I *thought* other people's boundaries were. But when I tried something totally different and super weird, it often went great! Sometimes other people hate small talk too, and also seek a genuine connection. But by default, this would never have happened - it took one of us taking initiative to break past the trap of social norms. And only by having a bunch of unusual and scary conversations yet having them go great did I develop the courage to be weird. Respecting other people's *actual* boundaries is important, but the conventional approach assumes an un-negotiable, one-size-fits-all to boundaries. And empirically, this picture is often totally off - some people think / want them to act normally, but are open to a much wider range of conversational styles if I initiate it.

Another key lesson is that closeness isn't just about spending a lot of time being friends - intentional, authentic, 1-1 time together makes a big difference. Sometimes I feel closer to someone after a single *amazing* conversation, than to people I've considered friends for years. Emotional connections aren't something that just happen to me - they're something I need to actually try to form. And effort, intelligently applied, can really pay off.

## Advice

So, that was my story. Now, I want to try distilling some key lessons that I think might apply to other people's quests to form close friends.

A key caveat to everything that follows: I argue for being much more intentional about social things than normal. When doing this, it's easy to come across as cold and calculating. I think it's super important to try to remain authentic, and to *signal* authenticity. I find it helpful to generally be friendly, make jokes, be honest and transparent, and be willing to be vulnerable. Eg, being open about the strategies I'm running and why, if it ever comes up.

## Seek excitement

A key mindset I use when forming connections via conversation is: "**If we aren't both excited about this conversation, do something differently**". This applies especially when talking to someone I don't know well, and want to figure out whether we might become good friends.

Most social norms optimise for conversations that feel safe, not ones that feel exciting, so I need to do something differently! This means asking the other person questions. This means taking a genuine interest in what we're talking about - and if I *can't* take a genuine interest, then I am doing something wrong.

A tactic I find helpful here is what I call **recursive curiosity**. I lead by asking an open-ended question that invites a detailed answer. Then, I introspect and try to notice excitement, find the part of their answer I find most interesting, and ask a follow-up open-ended question about it. Then, I repeat this process on their new answer. After about 3-4 iterations, we've normally gotten somewhere that feels alive and novel, where we're both learning, rather than the same stale conversations they have all the time. The follow-up questions don't need to be thoughtful or elaborate, often just

'[specific detail] sounds interesting, tell me more' or '[specific detail] didn't really make sense to me, can you clarify? Did you mean [naive interpretation]?' are more than enough. Introspecting on confusion or curiosity also works well. Often, rather than having a clear purpose to my questions I try to **maximise surface area** - just asking questions that point at my confusions and try to maximise the new information I gain, and the amount that I learn. This tends to feel fairly reactive - just responding to whatever was most interesting in the last thing said.

Sometimes I feel trapped in a boring conversation direction because the structure of small talk feels hard to break out of. When this happens, I like to go meta, eg observing 'man, I feel like I keep having the same kinds of conversations at these places' or 'let's get the boring questions out of the way - [standard small talk done rapidly]'. If they seem to empathise, this is a good opener for a more fun question, eg: 'what kind of things do you get excited about?', 'what's something cool you learned recently?', or 'have you had any particularly memorable conversations in [this context]?'

I've found that practicing this skill has made me *much* better at forming instant connections when meeting new people, and is *much* more fun than small talk! Even beyond meeting new people or following concrete algorithms, the spirit of 'seek the most exciting thread of the conversation' makes talking to friends way more fun!

**Warning:** these tactics sometimes get the other person to monologue - I am fine with this, and most people enjoy talking to an engaged audience, but some people feel bad at one-sided conversations. If you apply these techniques, I recommend being willing to monologue in turn if the other person seems interested - otherwise it can feel like you're being insincere.

## Being vulnerable

For me, vulnerability, and especially shared vulnerability, are really core to forming emotional connections. But this is difficult to manage because vulnerability, by definition, is hard. Different people have very different boundaries and comfort levels, and respecting boundaries is *super* important here. But, conversely, successfully creating shared vulnerability is really valuable and worth striving for.

My main approach is to create a space in which it feels safe to be vulnerable, but try to avoid creating obligations. I try to be honest and vulnerable myself, and freely share things that feel authentic throughout the conversation. I prefer to express lots of small vulnerabilities throughout the conversation rather than sharing something major and making it feel like a big deal - the latter tends to create an obligation/expectation of reciprocation, while the former better establishes a 'I consider this fine and normal' norm. I also find that both are effective for breaking people's social scripts/default ways of acting by being weird and unexpected - I find this is often a good first step to actually having a meaningful conversation. I find that sharing anxieties and insecurities can work particularly well here - almost everyone has them, it feels stigmatised to discuss them but people tend to respect you when you do, and they're often much more common and relatable than people think. I've had a bunch of these conversations, and still find it exciting (and sad) when I meet someone with really similar problems to me!

The 'without obligations' point is particularly important here, and hard to thread - sometimes people would enjoy sharing something vulnerable, but fear that it would

make me uncomfortable. I like to ask questions that invite a vulnerable response, but to give the person an 'out', some kind of reasonable excuse they could use to deflect without losing face. And by gauging their reactions, and seeing how much further to probe.

Overall, this is pretty hard to gauge and balance. It definitely takes a lot of practice, and I'm far from perfect at it. But I find it very worthwhile to practice.

Personally, I tend to be very anxious about whether I'm making other people uncomfortable, so I find this technique pretty aversive at times. My main approach to motivation here is internalising that I want to be [a person who actually does things](#) - that being vulnerable and welcoming vulnerability are skills I find uncomfortable but valuable, and I am growing as a person if I cultivate them. And, empirically, I've found this *really* useful to practice. I find this often leads to really awesome interactions, often in my first interaction with someone.

## **Hits-based Befriending**

*Alternate title: Making friends like an [r-strategist](#)*

Another key insight about friendship is that it's [all about upside risk](#). I will meet many, many more people in my life than I could ever sustain friendships with, let alone close friendships. Thus, if I am meeting new people and want to find potential close friends, I want to **filter fast** for compatible people. Further, compatibility is heavy-tailed - I won't really vibe with most people, but some people are *awesome*. I want to explore and optimise for information. This pushed towards high-variance strategies. If I meet 100 people, and want to pursue a friendship with just a handful, this is great!

This is a very, very different mindset from standard social norms, which push me towards being bland and inoffensive, and minimising the probability of bad interactions. A bad interaction (so long as it doesn't damage my reputation) is just as useless as a mediocre interaction for finding potential friends. Instead I want to maximise the probability that, if someone *is* compatible with me, we have an *awesome* interaction. This is a key part of why I push for excitement and vulnerability - many people won't vibe with that, but it makes it much more likely that I hit it off with the right kind of person.

Further, I am not constrained by the number of people I could meet - there are a lot of interesting people in the world. This means it's OK (but sad) if some people I *could* be compatible with don't vibe with my approach. Some people are pretty closed at first, and take a while to warm up to new people, but are awesome once this happens - my strategies around eg minimising small talk work much less well on this kind of person, which is sad. But the ability to filter *fast* is crucial. (Note: The trade-off between efficiency and precision depends on your situation, and I expect I'm further towards efficiency than most readers)

An important part of this is that a good filter is something that identifies people I'm compatible with *and* convinces them that they're compatible with me - if it feels like I'm coldly analysing or interviewing them, this is unlikely to go well. This is another part of why I am excited about approaches centred on excitement + vulnerability, those tend to go well if reciprocated.

**Warning:** This logic does not apply with people who I will need to interact with regularly anyway, eg co-workers/classmates. Social norms around minimising weirdness/potential for bad outcomes make much more sense in those situations, since downside risk is much higher. These mindsets work best when eg meeting people at a party or meetup or friends of friends, where I won't necessarily interact with them again.

Another key part of hits-based befriending is meeting lots of people, and exposing myself to lots of possible hits. Some of my favourite approaches:

- Going to meetups
- Going to events that will attract people with similar interests
- Talking to people around me in talks/lectures
- Asking my friends for intros to their friends - both generically ('do you know anyone I might get on with?') and specifically ('can you introduce me to [specific person]?')
- Proactively reaching out to people who seem interesting
  - I find this pretty anxiety-inducing, but it has a surprisingly high success rate. Most people are flattered!
- Having public forms on my website for people who want to have a [chat](#) or go on a [date](#)

**Exercise:** What traits do you value in your friends? What kind of person would you love to be friends with? How could you identify these traits in someone in a first meeting?

## Take Social Initiative

See longer form thoughts on this in [Taking Social Initiative](#)

A key second step to the hits-based befriending mindset is to **follow-up** once I identify someone cool! I try to make sure I get their contact info, and reach out shortly after meeting them trying to arrange a call/meetup. I find that many people are too socially anxious to do this, but **this is a really useful skill to practice**. The vast majority of my current friendships would not exist if I was bad at reaching out. Most people find this great and flattering, don't overthink it.

If you feel convinced of this logic, but still feel anxious about it, my main advice is to **practice**. Find some safe-ish ways to try it at first, eg with people you *really* hit it off with, or who seem incredibly friendly, or who you feel really comfortable around. If you're overthinking it, talk it through with a trusted friend and let them talk you into it. If you're concerned you won't know what to talk about, do some research on the person and make an agenda: a list of possible topics or questions to ask them.

Initially, it takes a lot of willpower and effort, and may feel super anxiety inducing - this is normal. But after I did it a few times and it went well, my mind started to update, and it now feels like a habit. I find that a similar strategy works for most forms of comfort zone expansion.

Another tactic for overcoming anxiety is to **other-ise**. Imagine you met someone at a party, *they* thought you were cool, and messaged you afterwards asking to meet up again. I don't know about you, but I'd find that pretty flattering. Or, imagine a specific friend coming to you with an analogous situation, asking whether they should follow-up. What advice would you give to the friend? And, if it differs from your internal

thoughts about your situation, why? Personally, I have yet to find a situation where the advice I give to the friend is worse than the advice I give myself.

A related and important skill is keeping in touch. Most people are really bad at keeping in touch, especially without a structure like university that keeps you in frequent contact. This means that many friendships fizzle without this structure. And this is really sad! I find a common mindset is 'if friend X really valued this friendship, I wouldn't be the one to always reach out'. But, empirically, I am confident many of my friends value our friendship, but also suck at reaching out. Being conscientious and organised is just hard, and varies a lot between people. Some people are very organised and keep in touch with everyone with ease, others easily lose track of close friends. 'Does this friend reach out to me?' is an *incredibly* noisy signal for how much they like you, and I consider it to convey approximately no information. I want to be good at keeping in touch, because I want to be able to remain friends with less conscientious people. And if I want to know if a friend actually likes me, there are much more direct ways of asking them.

But, fundamentally, keeping in touch should *not* be that hard - you just need to regularly reach out to arrange a call/meetup. This *can* be solved by being highly conscientious and having a good memory, but if you're lazy like me, [the correct way to solve this is with systems](#). I have a pretty barebones spreadsheet ([see template](#)) that lets me set an interval of N days to reach out to each friend, and reminds me to reach out N days after our last call, which has completely solved this problem. Other systems, such as [Calendly](#), are great for streamlining both following-up and keeping in touch, by making it trivial to schedule things.

For me, much of the anxiety around following-up and keeping in touch centre on being a burden, and bothering other people. A mindset I find helpful is reframing it all as **providing a public good**. Taking social initiative is *hard* and most people aren't very good at it. But most people *do* value fun social interactions. And, for some reason, people often enjoy interacting with me. This means that by taking social initiative, I am creating more opportunities for both of our lives to be better, which *is* something I find deeply motivating. 'I want to be a person who creates win-win situations' is fairly core to my identity. Whether it's following-up, keeping in touch, organising parties, suggesting group activities, etc, I want there to be more people in the world who do this. So I want to cultivate this skill myself.

## Deepening Friendships

See my post on [Friendships](#) for a deeper dive into what my ideal friendship looks like

My guess is that a surprising amount of the variance in friendship quality comes from finding the right people, and that things can often flow easily from there. But I think that it is also clearly valuable to practice the skill of deepening existing friendships. I have less time actively optimising this skill, and feedback loops are harder, but here are some thoughts:

- All my thoughts so far on vulnerability, excitement and authenticity also work for deepening friendships - I find that having regular authentic and meaningful conversations really help me feel closer to people
- Spending time together
  - In particular, searching for unusually fun/fulfilling ways to spend time together. Pay attention to their interests, experiment, and take social

- initiative!
- Relatively, actually *make* time to spend together. [Protect your Slack](#), and spend it on the people you care about. It's easy to make the mistake of considering social stuff the low priority thing to cut when I'm busy.
  - *Don't* feel constrained by fear of seeming weird, or social norms
    - Some social norms are good, some are bad. But if I really care about someone, I want our relationship to be optimised for *their* preferences. And this means figuring out what *they* want, and setting explicit boundaries and norms with each other
      - Eg, do they value honesty? Politeness? Bluntness? Proactivity? Compliments? Affection?
  - Relatedly, have good and clear communication about what we both want out of the friendship, and how the other person adds to our life
    - People often find this hard - it's hard to eg communicate about ways the other person annoys you without harming them. For me, a key is creating clear and explicit common knowledge that we both value this friendship and are invested in it. This frames all clear communication as **being on the same team** - we're trying to work together and share information, so we can both forge a better friendship
  - [Seek positive externalities](#) - be a pleasant person to be around, and find ways to add joy to the lives of those around you
  - Try to form a coherent model of how they add value to my life and vice versa. Once identified, try to actively optimise for the ways I add value.
    - [Obvious caveats](#): Make sure to remain authentic, account for uncertainty in the model, check for consent, etc
  - Practice relevant skills:
    - As above, good communication
    - **Love languages** - different people have very different ways of expressing affection, and it's easy to [typical mind fallacy](#). Some people really value appreciation, others value your time, others value gifts, others value physical affection, etc. I find it hard to empathise with people with different love languages, so it's important to explicitly notice and account for this, and understand what my friends want
    - **Emotional support/debugging** - one of the most valuable parts friends play in my life is providing emotional support, and help solving my problems. And I want to be able to provide this in turn. But this is really hard, and definitely a practicable skill!
      - My main tip is to avoid jumping to conclusions about the problem and what is needed, and instead to explore the problem more than feels necessary. I often explicitly ask 'what kind of help are you looking for?' - sometimes they want a solution, sometimes they just want to vent
      - See [this post](#) for more advice

But remember, these are just my takes, according to my friends and my values. You should experiment, try things, and figure out what works best for you!

**Exercise:** Make a list of your good friends. Which of them do you feel closest to, and what has led to this? What blocks are there to being closer to some, and what are you going to do about it?

## Conclusion

Close friends are very important to my life. And solving the problem of intentionally making close friends has added a *ton* of value, helped me make many friendships I cherish. Further, I've managed to add a lot of value to *their* lives. Friendship is one of the best mutually beneficial trades I've ever made.

The ideas in this post are all specialised to my tastes, and my experiences. I am a massive fan of 1-1 conversations, and of shared vulnerability. But your mileage may vary! You should experiment, try things, and forge your own vision of what intentionally forming close friendships looks like for you.

A point I've made throughout is that **this is a skill**. These are all things you can practice, experiment, iterate and get better at. It's easy to think of friendships as just something that happens to you. But I assert that, for most people, the cap on how awesome their friendships can be is *far* higher than where they are right now - it would be weird if that wasn't the case! Relationships are complex.

Finally, my key lesson from all this is that I need to take **agency**. It's easy to go through life never solving this problem, never even noticing the lack. Nothing will go visibly wrong. Nobody will stop you, or solve this for you. If you need permission from someone to do something differently, let this post be it - if you want your social life to be better, the only one who will fix this is you.

If the ideas in this post have resonated, I encourage you to take a moment to stop and reflect on your social life: Are you happy with how things are going? Do you feel happy with your ability to emotionally connect? Could it be better, and if so, how? What have you tried to do about this so far? And looking forwards, what are you *going* to do about this?

**Exercise:** Set a 5 minute timer, and list as many concrete ideas as you can for experiments to run, and things to try doing differently.

**Bonus exercise:** [Actually do something about it.](#)

# Precognition

This is a linkpost for <https://jasoncrawford.org/precognition>

It's almost impossible to predict the future. But it's also unnecessary, because *most people are living in the past*. All you have to do is see the present before everyone else does.

To be less pithy, but more clear: Most people are slow to notice and accept change. If you can just be faster than most people at seeing what's going on, updating your model of the world, and reacting accordingly, it's almost as good as seeing the future.

We see this in the US with covid: The same people who didn't realize that we all should be wearing masks, when they were life-saving, are now slow to realize/admit that [we can stop wearing them](#).

For a dramatic historical example (from [The Making of the Atomic Bomb](#)), take Leo Szilard's observations of 1930s Germany:

Adolf Hitler was appointed Chancellor of Germany on January 30, 1933. ... In late March, Jewish judges and lawyers in Prussia and Bavaria were dismissed from practice. On the weekend of April 1, Julius Streicher directed a national boycott of Jewish businesses and Jews were beaten in the streets. "I took a train from Berlin to Vienna on a certain date, close to the first of April, 1933," Szilard writes. "The train was empty. The same train the next day was overcrowded, was stopped at the frontier, the people had to get out, and everybody was interrogated by the Nazis. This just goes to show that if you want to succeed in this world you don't have to be much cleverer than other people, **you just have to be one day earlier.**"

## How to be earlier

**1. Independent thinking.** If you only believe things that are accepted by the majority of people, then by definition you'll always be behind the curve in a changing world.

**2. Listen to other independent thinkers.** You can't pay attention to everything at once or evaluate every area. You can only be the *first* to realize something in a narrow domain in which you are an expert. But if you tune your intellectual radar to other independent thinkers, you can be in the first ~1% of people to realize a new fact. Seek them out, find them, and follow them.

I was taking covid precautions in late February 2020, about three weeks ahead of official "lockdown" measures—but only because I was tuned in to the people who were six weeks ahead.

But:

**3. Distinguish independent thinkers from crackpots.** Both are "contrarian"; only one has any hope of being right. This is an art, honed over decades. Pay attention to both the source's evidence and their logic. Credentials are relevant, but they are neither necessary nor sufficient.

**4. Read broadly;** seek out and adopt concepts and frameworks that help you understand the world (e.g.: exponential growth, network effects, efficient frontiers).

Finally:

**5. Learn how to make decisions in the face of uncertainty.** Even when you see the present earlier, you won't see it with full clarity, nor will you be able to predict the future. You'll just have a set of probabilities that are closer to reality than most people's.

To return to the covid example: in January/February 2020, even the people farthest ahead of the curve weren't certain whether there would be a pandemic or how bad it would be. They just knew that the chances were double-digit percent, before it was even on most people's radar.

Find low-cost ways to avoid extreme downside, and low-investment opportunities for extreme upside. For example, when a pandemic *might* be starting, it makes sense to stock up on supplies, move meetings to phone calls, etc.—these are cheap insurance.

---

In some fantasy worlds, there are superheroes with “pre-cognition”, able to see the immediate future. They’re always one step ahead. But since most people are a few steps *behind* reality, you don’t need pre-cognition—just independent thinking.

# **Two non-obvious lessons from microcovid.org**

At the 2021 Summer Solstice, Elizabeth Van Nostrand made a brief speech thanking the organizers of [microcovid.org](https://microcovid.org), which I found very heartwarming and meaningful.

I wish I had thought in advance about taking the opportunity to make a few public remarks about [microcovid.org](https://microcovid.org): things I wish the community knew, that weren't obvious.

Here's what I would've said, in terms of my lessons from this project:

1. Build your own oxygen mask. Next, share it with others.
2. Connect and collaborate with non-rationalists.

## **1) Build your own oxygen mask. Next, share it with others.**

We didn't *start out* trying to create a resource for our whole community, let alone a website with many thousands of users. All we wanted was to save our own asses. We looked at the precipitous "autonomy crunch" we were facing, and said "oh shit, our house is going to explode if we don't fix this."

So, we built a spreadsheet — *for ourselves* first. Other group houses asked about it, and the momentum snowballed inexorably from there. Each broadening of project scope was compelled by a commensurate rise in demand, and corresponding deeply felt motivation.

I think many people who have altruistic or worldsaving ambitions could stand to have more focus on first making their own lives not suck. Fixing huge problems in your own life — and then later making an extra effort to share and export them — is one important path to altruistic impact.

## **2) Connect and collaborate with non-rationalists**

To my knowledge, I'm the only project member out of the top dozen or so top contributors who self-identifies as a rationalist.

The "core idea" is an extremely, extremely rationalist idea. But the implementation took writers, copyeditors, web developers, backend developers, UX designers, a medical doctor whose patients were among our first users, and many more. These folks had to understand the core idea and know how to use it, but did not have to be skilled enough at quantitative risk thinking to have designed it in the first place.

The final product had a vastly more scalable reach because many people, who had very little identity-level commitment to epistemics, looked at it and said things like "I

need to access this on my phone, I won't ever use a spreadsheet" or "This has too much jargon, move all these details to the appendix."

Thank you again everyone for the gratitude and recognition; and for using the system to make your own lives suck a little less!

# Social behavior curves, equilibria, and radicalism

This is a linkpost for <https://ericneyman.wordpress.com/2021/06/05/social-behavior-curves-equilibria-and-radicalism/>

## I.

Here are some hypotheticals to consider, with a common theme. Note that in each case I'm asking what you *would* do, rather than what you *should* do.

- In the fall, COVID cases drop to 10% of their current level. You're back to working/studying in person. You're vaccinated, as is everyone else. Mask-wearing isn't required, but 25% of your co-workers wear one anyway. Would you wear a mask too? What if 75% of your co-workers wear a mask? What if it's literally everyone else?
- You're having dinner with a party of 10 at a Chinese restaurant. Everyone else is using chop sticks. You know how to use chop sticks but prefer a fork. Do you ask for a fork? What if two other people are using a fork?
- (Inspired by [Scout Mindset](#)) For anyone who chose to have kids, or wants kids: if only 30% of adults had kids, would you still choose to have a kid? What if only 2% did? For anyone who doesn't want kids: if 90% of adults had kids, would you? What if it were 99%?
- You join a group video call with 20 co-workers/classmates. Everyone except the presenter/teacher has their video turned off. Do you turn yours off too? What if everyone has their video on? Typically, what fraction of other participants need to have their video on for you to keep yours on?

Your answers to these questions depend on your **personal circumstances**. How uncomfortable is wearing a mask? Do you find a fork *slightly* or *significantly* easier to use? How strong is your preference to have kids? Is there a Justin Bieber poster in your background?

But — unless your answer to each question didn't depend on the specific percentage — your decision also depends on **others' behavior**. If everyone else has their video on, you'll probably feel obligated to keep yours on too. Maybe you'll move the poster first. This applies not just to behavior, but also to preferences, beliefs, and opinions.

# WIKI FRIENDS:

I REALLY LIKED  
THAT MOVIE.

I HATED  
THAT MOVIE.

ME TOO.



Figure 1: [xkcd #185](#)

So far, everything I've said is pretty obvious. But let's throw a model at this observation and see if we can discover anything interesting.

Let's take our example with 20 people in a video call and line all of the participants up, from bottom to top, based on **how many other people need to have their video on in order for them to choose to keep their video on**.

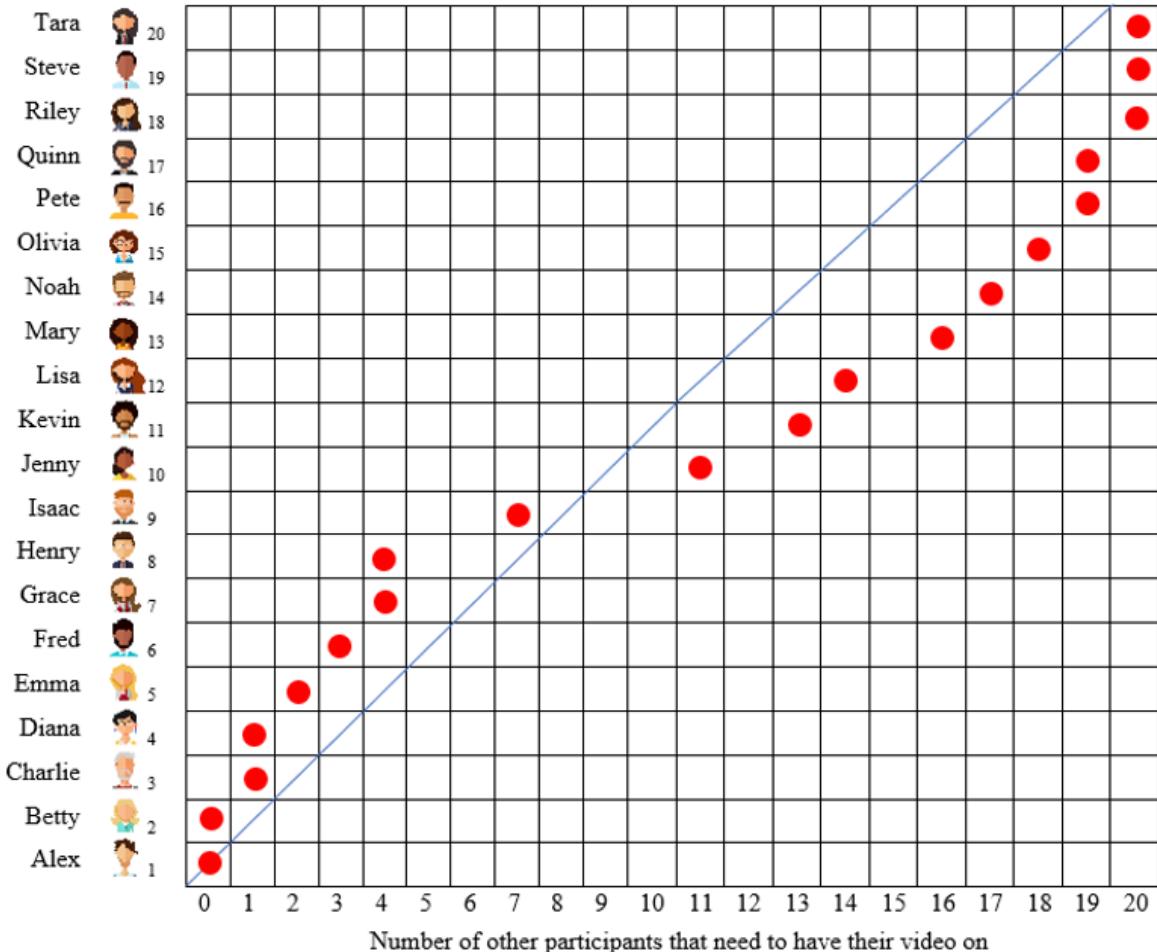


Figure 2: video chat, scenario 1

In this example, Alex and Betty will keep their video on no matter what; on the other hand, Riley, Steve, and Tara will turn their video off no matter what. Most others are somewhere in between: Isaac, for instance, will keep his video on if at least 7 of the other 19 participants have their video on.

Take a moment to think about what will happen in this call. As a hint, you might want to consider the diagonal line I've drawn on the chart.

The answer is that nine participants — Alex through Isaac — will have their video on, and the rest will have it off. Why? Well, if everyone has their video on at the start, then Riley, Steve and Tara will turn their video off right away. A cascade will follow: Quinn and Pete, who are only willing to have their video on if everyone else has theirs on, will turn their video off. And so on — up through Jenny. Now Alex through Isaac (but no one else) will have their video on. But at this point the cascade stops: Isaac is happy to keep his video on, as is everyone else.

This would also be the end state if everyone started with their video off (assuming there's no status quo bias). Alex and Betty would turn their video on, Charlie and Diana would follow, and so on, up through Isaac. Indeed, no matter who has their camera on at the start, this will be the end state.

Let's look at a different example.



Figure 3: video chat, scenario 2

Now, Alex is willing to keep his video on so long as at least one other participants does as well. At the other end of the spectrum, Tara is willing to keep her video on if at least 17 other participants do. Now what will happen?

This time, the answer depends on the starting state. If everyone starts with their video on, everyone will keep it on. If everyone has it off, everyone will keep it off. In fact, small changes in the starting state cause dramatic differences in the end state. If Alex through Henry start with their video on, then Fred, Grace, and Henry will turn their video off, and there will be a downward cascade (ending with no one having their video on). If Alex through Jenny start with their video on, there will instead be an *upward cascade*, with Kevin turning his video on, followed by Lisa, and so on. (Exercise for the reader: what happens if Alex through Isaac start with their video on?)

Drawing a 20 by 20 chart is very particular to our example. Let's abstract that away; instead of plotting dots on a 20 by 20 grid, we'll plot curves on a square. That'll look something like this.



Figure 4: [Many other person do thing, person do thing. No many, no do thing.](#)

Just like in the video call example, we're ordering people by how willing they are to do a thing. In the video call example, the thing was having their video on. But now we're abstracting away the number of people and instead want to think about the *percentage* of other people doing the thing that is necessary for someone to choose to do the thing. I'll be calling these curves **social behavior curves**.

Perhaps a concrete example would be helpful: let's say that in the above plot, the "thing" is wearing a mask. In the plot, the people who are most willing to wear a mask will wear one if at least 3% or so of people are wearing a mask (these are the people toward the bottom). On the other hand, 3% or so aren't willing to wear a mask no matter what (these are the people at the very top). In the middle we have the median person in terms of willingness to wear a mask, who will wear a mask if at least 35% of others are wearing one.

Armed with a social behavior curve, it's pretty easy to reason about the social equilibria of mask-wearing. For example, suppose that just the 50% most willing-to-wear-a-mask people wear masks. This is *not* an equilibrium. That's because the person who is just a little bit above the 50% mark (i.e. the person who's just slightly less willing than median to wear a mask) will put on a mask: after all, they're willing to wear a mask if at least 35% of others are wearing masks, which is the case. And the next most willing person will put on a mask, and so forth: we'll have an upward cascade of mask-wearing until... when?



Figure 5: social equilibria

Until the purple point near the middle (around 65%) is reached. And at that point, the cascade will stop because people above the 65% line will need more than 65% of people to wear a mask in order to themselves wear a mask. And similarly, if the starting state were that 75% of people wore masks, there would be a downward cascade until the 65% mark was reached. (In general, an upward cascade will happen if you're at a point on the y-axis where the red curve is to the left of the blue line. And if you're at a point where the red curve is to the right of the blue line, you'll get a downward cascade.)

In this sense, the 65% mark is a **social equilibrium**. More precisely, the state of the world where the 65% most willing-to-wear-a-mask people wear a mask, and the rest don't, is a social equilibrium. There are other social equilibria in this example. One is 0%: if no one is wearing a mask, no one will put on a mask. And there's another one around 96% or so.

What about the points in orange? They're kind of weird! Consider the bottom-left one, which is around 25%. If 24% of people are wearing a mask, there will be a *downward cascade*. But if 26% of people are wearing a mask, there will be an *upward cascade*. In that sense, this point is an *unstable social equilibrium*. If you're right at 25%, everyone wearing a mask is happy to continue wearing one and everyone not wearing one won't put one on. Go just above or below 25% and you get a cascade. The same is true of the orange point around the 90% mark.

To summarize, **points where the social behavior (red) curve crosses the blue line from left to right are stable social equilibria. Points where the curve crosses the blue line from bottom to top are unstable equilibria.**

As with all mathematical models of social behavior, this one is incomplete. You might want to take a minute to think about the various things this model fails to capture. Still, I wish to make the case that this model is useful for understanding phenomena such as persuasion, radicalism, and rapid cultural shifts. Let's dive in.

## II.

In our model, **persuasion is the act of shifting the social behavior curve horizontally**. That's because a persuasive argument in favor of doing X lowers the percentage of other people who need to be doing X in order for you to join in and start doing X yourself.

To see this, bring yourself back into the very early days of the pandemic, when the virus was spreading but no one was wearing a mask. Now suppose you read a compelling argument in favor of wearing a mask. This alone probably wouldn't be sufficient for you to start wearing a mask. (If this isn't true for you personally, consider the "you" to be generic.) Instead, it would lower your threshold for how many other people need to be wearing a mask in order for you to be willing to wear one. Maybe beforehand you would have started wearing a mask if 30% of people around you were wearing one, but now a 25% masking rate would be sufficient. In other words, your point on the social behavior curve used to have an x-value of 30%, but now it's 25%. Your point on the curve shifted leftward.

(In a more naïve model of persuasion, one where if you hear a persuasive argument in favor of X, you start doing/believing X. I think basically no one does that; we're all shaped not just by arguments but by the beliefs and behaviors of those around us.)

Now, if everyone reads the argument then the *entire curve* will shift to the left. And if some fraction of the population comes across the argument, then you can still model the curve as

shifting left — you just need to multiply the amount of the shift by the fraction of people who come across the argument.

(If the argument systematically affects people in different spots of the curve differently, then the leftward shift won't be uniform. But I'll be assuming a uniform shift to avoid overcomplicating the model.)

Typically, the effect of persuasion looks something like this:



Figure 6: The red and green curves are, respectively, the social behavior curves before and after people read the argument in favor of X. The purple and grey points are the equilibria before and after, respectively.

You come up with a really clever argument in favor of X — enough to shift the red curve leftward by 5 percentage points. People who previously needed 80% of people around them to do X in order to themselves do X now only need 75%, and so on. This causes the equilibrium to shift from the purple point... up just a few percentage points to the grey point. Congratulations: you've successfully disseminated your super persuasive argument, and 3% more people believe X.

That's what typically happens. But in some cases, an equally persuasive argument can have dramatic effects.

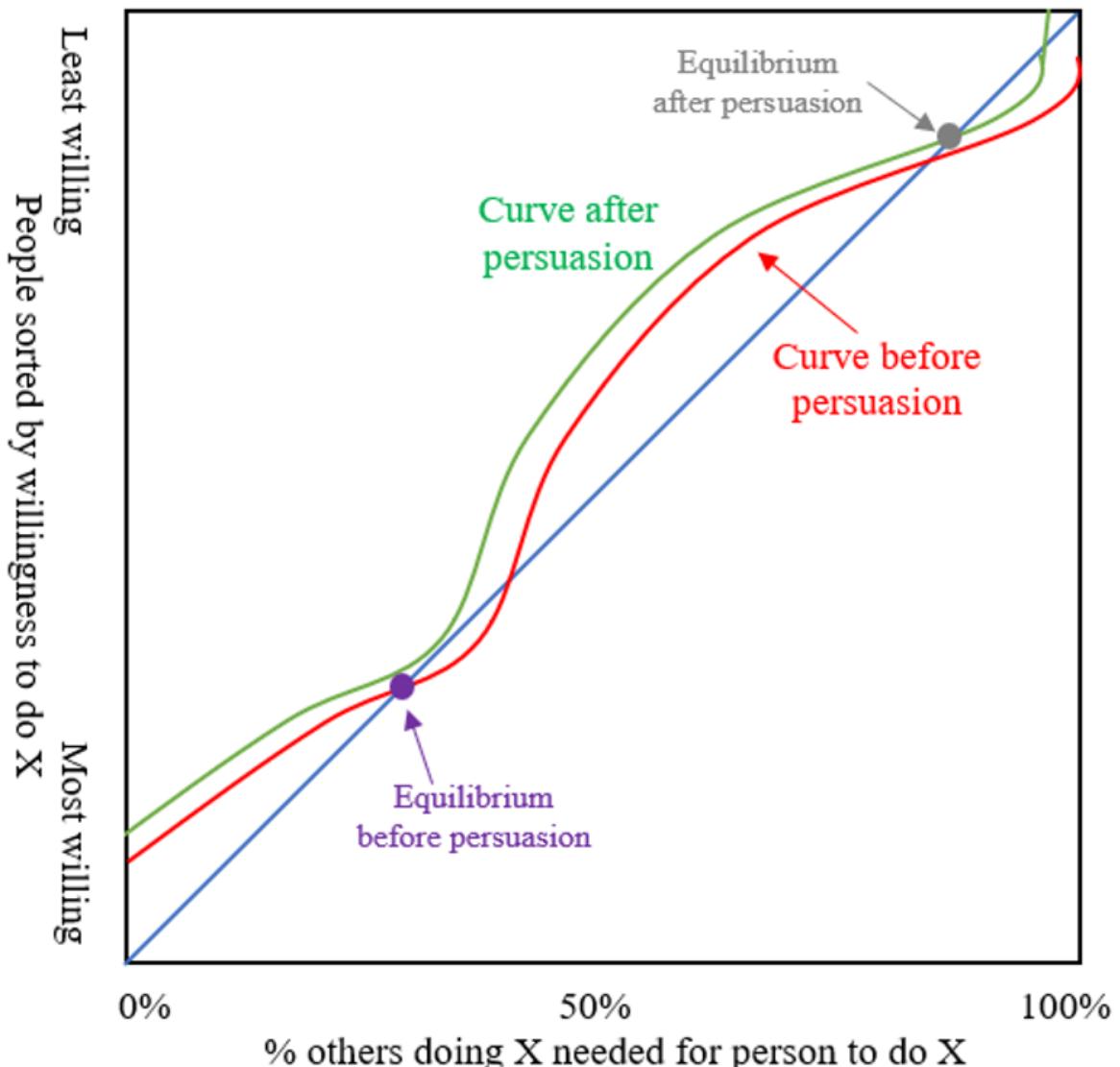


Figure 7: a dramatic shift in equilibrium

In this example, the same 5% shift displaces the equilibrium shown in purple. After the social behavior curve shifts from the red curve to the green curve, there is no longer any equilibrium near the purple point, since the green curve does not cross the blue line there. There's an upward cascade of people who start to do X, and society moves to a new equilibrium around 85%. A 5% persuasion shift causing 50% of people to change their behavior: that's some serious return on investment!

An alternative perspective: the effect of gradual persuasion is that the social behavior curve gradually shifts to the left, and so the equilibrium gradually shifts to the right (as illustrated in Figure 6). But then at some point, the curve moves past a point of tangency with the blue line — that's what happens in Figure 7 — and then there's a dramatic shift, where the equilibrium switches to a different, possibly far-away one.

Is this realistic? I believe it is! In the real world you won't see such a seismic shift, because different people belong to different communities with different social behavior curves. But I would posit that often when society sees a rapid change in social norms and behaviors, it's often due to an effect like the one in Figure 7.

Here are some possible examples (*speculative; take with approximately 1.5 tablespoons of salt*):

- I think that social behavior curves for many things of the form “being okay with X” are shaped like S-curves (see Figure 8). That’s because — at least for me — it’s somewhat uncomfortable being the only person who’s not okay with something. (Picture yourself being the only one in a group of friends who isn’t okay with people taking off their masks, or bringing drugs to a party, or whatever; it’s uncomfortable!) This means that the social behavior curve is pretty flat in the upper-right: not many people are unwilling to be okay with something if 95% of others are okay with it. Perhaps to a lesser extent, there may be social pressure not to be the only person who *is* okay with something.
  - This might cause the sort of shift shown in Figure 8: the original (red) S-curve starts above the blue line, dips below, and then comes back up. After a persuasion effect happens, the red curve moves left, there is no more equilibrium in that region, and widespread acceptance is soon reached. This might explain the unusually rapid pace with which homosexuality became accepted in the United States.
  - Maybe this explains the rapid collapse of the temperance movement in the 1920s? I don’t know enough history to know if this is a reasonable theory.
  - If my theory about things of the form “being okay with X” being shaped like S-curves is correct, then that would predict that shifts in society’s attitudes toward civil rights and liberties happen more suddenly than shifts in public opinion on other issues. This feels empirically true to me but I’m not confident, and I’d be interested if someone has data on this!



Figure 8: an S-curve

- There are mathematical theories about why revolutions are so slow to get off the ground, but when they do, they get very large impressively quickly. George Akerlof [posited](#) that this effect is caused by [present bias](#). Scott Aaronson instead [explains](#) this using common knowledge. I’d like to posit a different theory: protestors have safety in numbers, which means that people are much more willing to join a revolt if a critical mass of people have already joined. This results in the following curve shape (the red curve in Figure 9), with lots of people willing to revolt as soon as 15% of people are revolting. As the dictator becomes more and more oppressive and people lose patience, the curve shifts slowly to the left — until the curve passes the blue line and the revolution explodes in size.



Figure 9: revolution

Interestingly, this model of rapid social change posits that **once a rapid social change has happened, it’s usually really difficult to go back**. To see that, suppose that the dictator decides to placate the populace (or perhaps crack down hard and increase the cost of revolting), effectively moving the green curve back to the right. The result will *not* be a shift from the grey point back to the purple point. Instead it will shift back down just a little to the yellow point.



Figure 10: the dictator attempts to reign in the revolution and fails

This accords with my intuition for how rapid social changes tend to work out: once they happen, things rarely go back to how they were.

(What about revolutions that fizzle out? I think these tend to be small in size, i.e. the green curve is never reached. If a revolution gets really large, I think it rarely fizzles out; instead it leads to war or regime change. Ideally I’d like to phrase this hypothesis in a way where I

can't weasel out of it by claiming that any particular revolution that fizzled out just didn't get big enough, but I'm not sure how to do that.)

The fact that the shape of the curve matters a lot has implications for activists and influencers: **focus your energy on causes where the social behavior curve makes it possible for you to tip society into a new equilibrium.** You might have ten or a hundred times more leverage than if you just choose the issue that's most compelling to you!

Of course, **estimating the shape of the curve is a huge challenge.** One place to start is to try to infer social behavior curves from historical behavior changes and draw some general conclusions (e.g. "social behavior curves regarding public opinion on civil liberties tend to be S-shaped"). Or maybe conducting a survey that asks people questions of the form "If your neighbors started doing X, do you think you would?". That might give you some mileage, but overall I'd guess that people don't understand themselves well enough to answer that question accurately.

### III.

There's a lot of intuition to be gained about a social behavior curve by looking at its slope (derivative) at different points. The slope of a social behavior curve at 30% (for example) represents, loosely speaking, how many people have 30% as their "tipping point", i.e. how many people will switch from not doing X to doing X once 30% of people are already doing X. (For math people: the social behavior curve is a CDF, so its derivative is the corresponding PDF.)

For example, here's the initial (pre-oppression) curve in Figure 9.

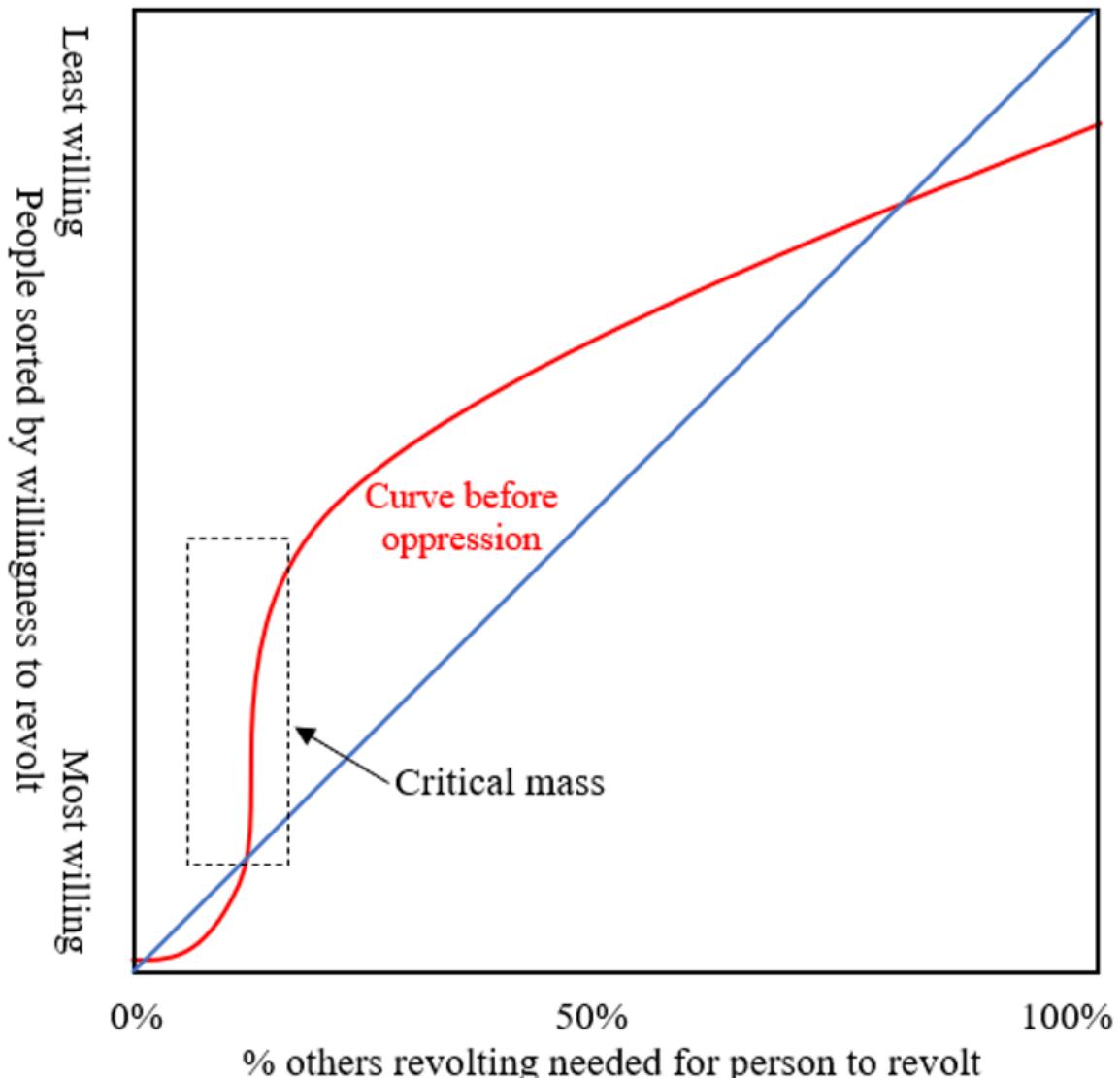


Figure 9b: the pre-oppression social behavior curve from Figure 9

And here's what the derivative of the red curve looks like. This captures the notion of most people deciding to revolt when a certain "critical mass" (around 15%) is reached.



Figure 11: behavior density curve for revolution

It is often easier to think about what the orange curve should look like (I'll be calling them **behavior density curves** from now on) and then extrapolate the social behavior curve; indeed, that's how I reasoned about what the curve in Figure 9b should look like.

What do behavior density curves look like? It obviously depends on what X is, but it stands to reason that many behavior density curves are bell-shaped. (After all, many distributions are bell-shaped.)



Figure 12: most people do whatever other people are doing

This produces a social behavior curve that looks like an S-curve. The tighter the bell curve, the steeper the social behavior curve. In the limit, everyone's at 50%, which means that

everyone is going to do whatever the majority is currently doing. A good example of this is network effects. Imagine two identical platforms, Facebook 1 and Facebook 2. People want to talk to their friends, so they'll join whichever platform has more of their friends.



Figure 13: behavior density curve (left) and social behavior curve (right) in the presence of network effects

On the other hand, you could imagine a reverse situation, where most people have a strong preference either to do X or not, such that others' behavior only matters only a little. The behavior density curve would look like this:



Figure 14: most people act according to their intrinsic preference

Examples of this tend to be things that are pretty ingrained in people, as opposed to being socially influenced. A good example of this is left- and right-handedness (though in this example the behavior density curve isn't centered at 50%). The corresponding social behavior curve has this sort of shape:

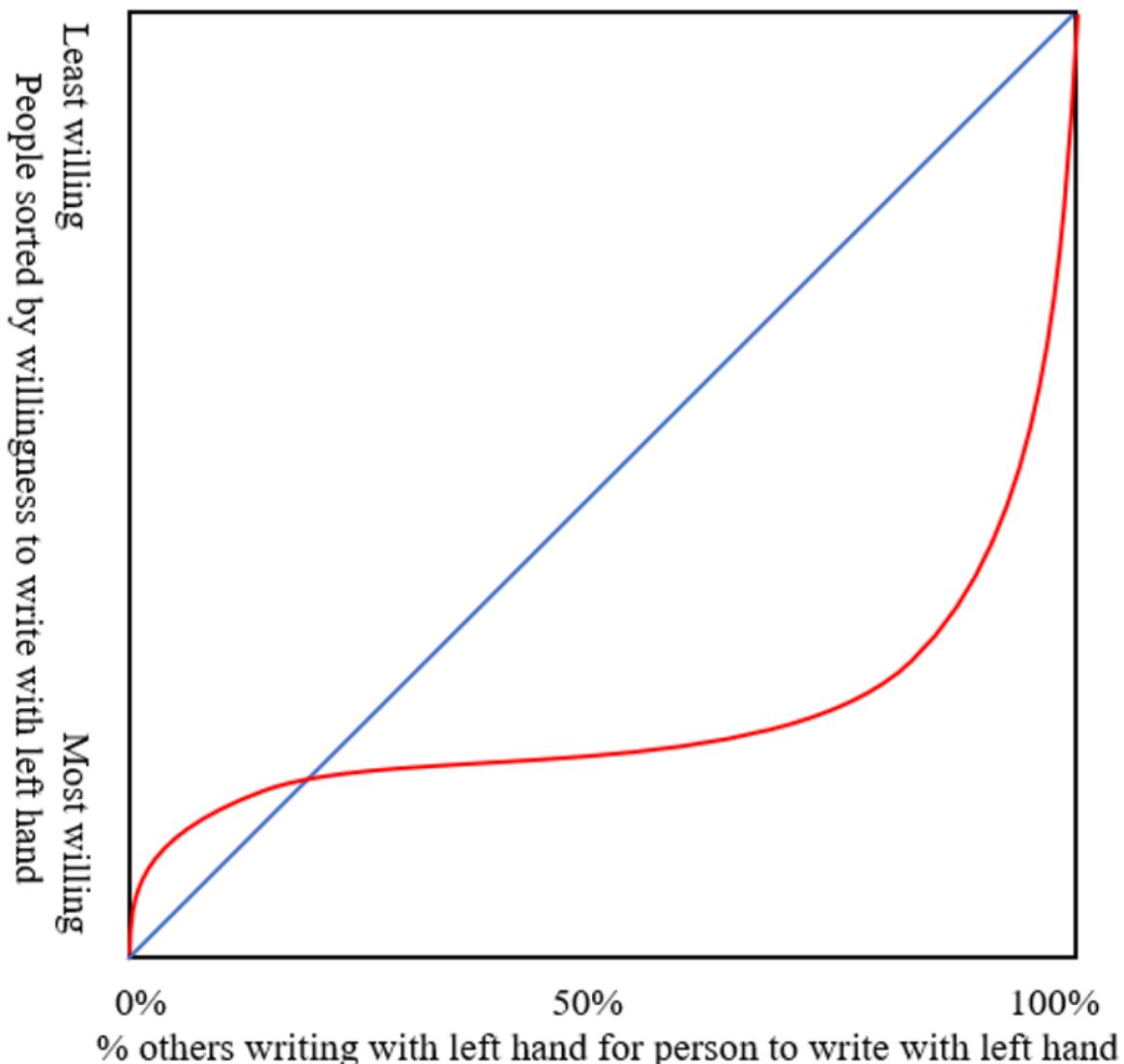


Figure 15: social behavior curve in the presence of strong intrinsic preferences

So far I've been talking about these curves descriptively: making guesses about what the world actually looks like. But for fun, let's talk about the prescriptive question: **what is the best shape for a social behavior curve?**

On its face this is a pretty silly question. The best possible social behavior curve for "being a serial killer" looks a lot different from the best possible social behavior curve for "donating to charity." But let's set these examples aside and think about what we might want out of a social behavior curve where the two possibilities (doing X and not doing X) are both reasonable, but one might be substantially better than the other.

There are lots of examples of this — it's the case basically whenever reasonable people disagree on what social norms they want. One example of this is [ask culture versus guess culture](#). In ask culture, it's totally polite to request a favor ("Hey, remember me from high school? I'm visiting your town, can I stay at your place?"), and it is likewise completely fine to say no. In guess culture, people are expected to only ask for a favor if they think that the person they're asking would be comfortable granting it, and likewise one is expected to say yes unless there's a good reason not to.

Imagine if the social behavior curve for X = "behaves as if in ask culture" looked like the one in Figure 13 (Facebook 1 vs. Facebook 2): almost everyone behaves like the majority. This is the "**collective society**" approach, where people are expected to closely follow societal norms. Such an approach would be good for social cohesion: everyone follows the same norm, so there's no conflict resulting from people misunderstanding each other's intentions. But it would be bad from the perspective of getting stuck in a bad equilibrium: maybe the current equilibrium is "everyone follows guess culture norms" but in fact ask culture is better and there's no way to discover this and switch.

Conversely, imagine if the social behavior curve looked like Figure 15 (handedness). This would be the "**individualistic society**" approach, where people behave according to their own intrinsic preferences. Then society would have lots of askers and guessers (so it would be easy to get a sense of the relative merits and drawbacks of each), but it would be hard to learn from these merits and drawbacks. (Think about how far to the left you'd need to shift the social behavior curve for ask vs. guess culture if it had the same steep-flat-steep shape as the curve for handedness. It would be really difficult for society to come to a collective decision on which approach it prefers.)

What's the best way to balance this trade-off? I'd argue in favor of something like this:



Figure 16: my guess at what an optimal(ish) social behavior curve looks like in many settings

I see this as the best of both worlds. On the one hand, there are a few people who have a strong preference for ask culture, and a few for guess culture, so society gets to experience and learn from both — or at least knows that both norms are theoretical possibilities. On the other hand, if society gets evidence that ask culture is better, a relatively small leftward shift in the curve will cause most of society to be on board with ask culture (that's because the slope of the red curve is close to 1). This confers the benefit of social cohesion. It also means that very sudden shifts like in Figure 7 can't happen; society can respond to new information relatively quickly, but does so smoothly. This seems like a good thing.

What does the corresponding behavior density curve look like?



Figure 17: a nice compromise between collective and individualistic cultures

Something like this (though maybe steeper at the edges). There are lots of people everywhere along the spectrum: people who strongly prefer ask culture, people who strongly prefer guess culture, and also those who are happy to go with whatever norm is the current default. Such a density curve — which is “in between” the “collective society” curve (Figure 12) and the “individualistic society” curve (Figure 14) (though perhaps closer to the latter) — gets you the best of both worlds.

The fact that Figures 16 and 17 are symmetric around 50%, by the way, is *not* an important feature. The curve below has the same nice properties, even though its equilibrium is around 20% instead. So when I talk about curves with the “general shape” of the curve in Figure 16, I’m including curves like this one.



Figure 18: this social behavior curve, while not symmetric, still has the nice properties we’ve been discussing

(I think there’s a lot more that could be said here. We could analyze free speech norms from the same perspective. Totally free speech allows for exploration of a vast swathe of ideas at the expense of societal cohesion, while a lack of free speech inhibits exploration and progress; maybe there’s an optimal happy medium? Also, perhaps differences in collectivism versus individualism could serve as an explanation for why some societies and communities have been more successful than others. But this is way above my pay grade so I’ll let others speculate.)

#### IV.

For me, the main takeaway from the previous section is this: **society needs both radicals and conformists, as well as people in between.**

When I say “radical”, think [Vermin Supreme](#). A radical does things their own way, to hell with what society thinks. A radical wears a boot on their head and prepares for the zombie apocalypse. A radical bucks the establishment — social, political, scientific, you name it — in pursuit of their own weird beliefs and inclinations.

Society stands to gain very little from most radicals. A typical radical is someone who markets a new form of pseudo-medicine, or espouses a nonsensical economic policy. In the worst case, a radical becomes convinced that societal ills can only be remedied through violence and crashes airplanes into buildings.

But occasionally — not usually but not never — a radical [invents a new form of medicine that saves millions of lives](#), or causes a [major scientific paradigm shift](#), or [helps society make substantial moral progress](#). Without radicals, we’d be stuck with wrong beliefs and bad equilibria forever. (See also: Scott Alexander’s [Rule Thinkers In, Not Out](#).)

I’m not sure how deep this analogy goes, but think genetic mutations. Most are bad, but it’s really good to have some nonzero level of mutation, as this makes evolutionary progress possible.

A radical is someone who, for many different values of X, is on the far-left or far-right of the social behavior curve for X. They’ll do X, or think X, even if no one else does or thinks X. Their existence gives society the opportunity to ponder X.

And in particular, radicals’ existence gives *radical-adjacent* people the opportunity to join in doing or believing X if it seems like a good idea. Radical-adjacents are people who tend to be

pretty close to the left or right extremes of a social behavior curve, but not all the way on the edge. They are the people who don't necessarily do really weird things or promote strange ideas themselves, but are open to such habits and ideas once entertained by a few radicals.

And so on and so on, down the [respectability cascade](#), all the way to the conformists: those who will go with the prevailing norm or belief. And **conformists are important too**. Without them, a social behavior curve might look something like this.



Figure 19: a society without conformists

You need to move the red curve *really far* to the left or right (i.e. come up with an *incredibly* convincing argument or effective movement in favor of or against X) to shift society from "50% of people do X" to "everyone (or no one) does X" — which is quite detrimental if society would be a lot better off with everyone (or no one) doing X.

How many radicals and how many conformists is ideal? I've already sort of answered that in Figure 17. People toward the left and right edges of that figure are more radical, people in the middle are conformists. So the ideal distribution looks perhaps something like this:



Figure 20: my guess at the optimal(ish) distribution of people on the radical-conformist spectrum

In my ideal world people are about evenly distributed on the spectrum between radical and conformist, with perhaps a slight radical-ward bias. (Even in such a world, there are very few true radicals: they are represented only by the leftmost 1% or so of the chart in Figure 20.)

## V.

*[Epistemic status: progressively more and more trolling]*

I've already talked about one way that social behavior curves can help you think about how to make the world a better place: figuring out when persuasion and activism are effective. I want to finish by talking about another way social behavior curves can help you: namely, figuring out how radical you ought to be.

Let's say that no one is doing X, but — at least if you disregard that fact — X seems like a good idea to you. An example of X might be becoming vegan, if you're in a community where everyone eats meat. Should you become vegan or go along with your community's norm of meat-eating?

The answer depends on whether you think there are too many or too few radicals in your community. If there are too few, then by increasing the "mutation rate" you realize the upside that you might eventually convert everyone to veganism and make the world a better place. If on the other hand there are too many radicals, then an [outside view](#) argument is likely more applicable here: society has likely already considered and rejected veganism, even if you don't know why they did so.

But how can you figure out if there are too many or too few radicals in society? Answering this question seems really hard, just like estimating the shape of a social behavior curve is super difficult.

One approach might be to examine the question empirically: see whether societies with higher levels of radicalism have fared better. But this seems extremely hard and noise-prone.

My radical answer to this question is: if you are inclined (from an inside view perspective) to adopt a behavior or belief, **decide how radical to be at random**. To explain why, let's talk about rule utilitarianism.

Rule utilitarianism says that you should act according to whatever set of rules results in the most good. This contrasts with classical utilitarianism, which says that for every decision you should take whichever action results in the most good. I like rule utilitarianism because it's realistic to follow: it would be exhausting to do utility calculations at every turn, but if your community has some rules of thumb worked out then you can follow those. Spend some time working out good rules of thumb, and you'll be able to make good moral decisions without excessive overhead.

As an example of rule utilitarianism, a pretty good rule might be "Donate 10% of your income to the charity where a marginal dollar will have the [greatest positive impact](#)".

Crucially, this rule works because we have a pretty good sense of how well-funded different charities are. Suppose instead that people had *no idea* how much money each charity had, and for that matter didn't know where anyone else was donating.

This would make things a lot harder. I might do the best I can to follow the rule with the information I have and decide that AI risk is probably the most important cause area. Everyone else who shares my basic thought process might come to the same conclusion, we'll all donate to MIRI, and MIRI will become oversaturated, while other important charities go neglected.

The key remedy to this problem is either to spread your dollars between multiple charities, or else to *randomize* your donation, choosing a charity in proportion to how much money it would ideally get. And if everyone randomizes, [the outcome will be pretty good!](#) So in the absence of information about others' charitable donations, a better rule would be "Donate 10% of your income to a charity selected at random in proportion to how much money each charity would ideally get".

This is the situation with social behavior curves. You have no idea what the social behavior curve for not eating animal products looks like; you just know that you're in a "everyone eats meat" equilibrium. Nor do you know the distribution of radicals versus radical-adjacents versus conformists. In the absence of such information, you can't take the strategy of "adopt whichever disposition is most neglected". Nor is there an approach analogous to "spread your money between charities": you can't be a mixture of different levels of radical on the same issue. So the rule that, if adopted, would do the most good is "Select how radical you'll be at random".<sup>1</sup> If this rule is followed, your community would end up with the right number of radicals and conformists and in-betweens!

How seriously should you take the argument I've just made? Should you literally flip a coin next time you decide whether to do something no one else is doing? I'm not sure; as far as I can tell, no one is flipping coins to decide these sorts of things. But maybe some small number of people take my argument seriously and start flipping coins. And if they get good results, maybe some other people will join in the fun. And then eventually, maybe everyone will be flipping coins.

So should you, personally, start flipping coins in such situations? Flip a coin to find out!

*[Edit: Ben Edelman points out that sociologists use social behavior curves, see e.g. [this paper](#) and [these Wikipedia pages](#). I guess it shouldn't come as a surprise that these concepts are already well-known. I suppose it's nice to have a bit of confirmation that these models are considered reasonable/interesting!]*

1. More precisely, the rule is “Select how radical you’ll be at random, according to the ideal distribution of radicals versus conformists”. Here, “ideal distribution” is what an outside observer would prefer for the distribution to be in general. (I’ve posited a guess at this distribution in Figure 20.) The “outside observer” bit is important: of course you’d prefer for there to be lots of radicals on your pet issue, but [it’s not a good rule if you wouldn’t want for it to be universalized](#) to everyone’s pet issue.

# Power dynamics as a blind spot or blurry spot in our collective world-modeling, especially around AI

## Where I'm coming from

\*\*\*Epistemic status: personal experience\*\*\*

In a number of prior posts, and in [ARCHEs](#), I've argued that more existential safety consideration is needed on the topic of multi-principal/multi-agent (multi/multi) dynamics among powerful AI systems.

In general, I have found it much more difficult to convince thinkers within and around LessWrong's readership base to attend to multi/multi dynamics, as opposed to, say, convincing generally morally conscious AI researchers who are not (yet) closely associated with the effective altruism or rationality communities.

Because EA/rationality discourse is particularly concerned with maintaining good epistemic processes, I think it would be easy to conclude from this state of affairs that

- multi/multi dynamics are not important (because communities with great concern for epistemic process do not care about them much), and
- AI researchers who do care about multi/multi dynamics have “bad epistemics” (e.g., because they have been biased by institutionalized trends).

In fact, more than one LessWrong reader has taken these positions with me in private conversation, in good faith (I'm almost certain).

In this post, I wish to share an opposing concern: that the EA and rationality communities have become systematically biased to ignore multi/multi dynamics, and power dynamics more generally.

## A history of systemic avoidance

\*\*\*Epistemic status: self-evidently important considerations based on somewhat-publicly verifiable facts/trends.\*\*\*

Our neglect of multi/multi dynamics has not been coincidental. For a time, influential thinkers in the rationality community *intentionally avoided* discussions of multi/multi dynamics, so as to avoid contributing to the sentiment that the development and use of AI technology would be driven by competitive (imperfectly cooperative) motives. (FWIW, I also did this sometimes.) The idea was that we — the rationality community — should avoid developing narratives that could provoke businesses and state leaders into worrying about whose values would be most represented in powerful AI systems, because that might lead them to go to war with each other, ideologically or physically.

Indeed, there was a time when this community — particularly the Singularity Institute — represented a significant share of public discourse on the future of AI technology, and it made sense to be thoughtful about how to use that influence. Eliezer recently wrote (in a semi-private group, but with permission to share):

The vague sense of assumed common purpose, in the era of AGI-alignment thinking from before Musk, was a fragile equilibrium, one that I had to fight to support every time some wise fool sniffed and said "Friendly to who?". Maybe somebody much weaker than Elon Musk could and inevitably would have smashed that equilibrium with much less of a financial investment, reducing Musk's "counterfactual impact". Maybe I'm an optimistic fool for thinking that this axis didn't just go from 0%-in-practice to 0%-in-practice. But I am still inclined to consider people a little responsible for the thing that they seem to have proximally caused according to surface appearances. That vague sense of common purpose might have become stronger if it had been given more time to grow and be formalized, rather than being smashed.

That ship has now sailed. Perhaps it was right to worry that our narratives could trigger competition between states and companies, or perhaps the competitive dynamic was bound to emerge anyway and it was hubristic to think ourselves so important. Either way, carefully avoiding questions about multi/multi dynamics on LessWrong or The Alignment Forum will not turn back the clock. It will not trigger an OpenAI/DeepMind merger, nor unmake statements by the US or Chinese governments concerning the importance of AI technology. As such, it no longer makes sense to worry that "we" will accidentally trigger states or big companies to worry more-than-a-healthy-amount about who will control future AI systems, at least not simply by writing blog posts or papers about the issue.

## **Multi-stakeholder concerns as a distraction from x-risk**

*\*\*\* Epistemic status: personal experience, probably experienced by many other readers \*\*\**

I've also been frustrated many times by the experience of trying to point out that AI could be an existential threat to humanity, and being met with a response that side-steps the existential threat and says something like, "Yes, but who gets to decide how safe it should be?" or "Yes, but whose values will it serve if we do make it safe?". This happened especially-much in grad school, around 2010–2013, when talking to other grad students. The multi-stakeholder considerations were almost always raised in ways that seemed to systematically distract conversation away from x-risk, rather than action-oriented considerations of how to assemble a multi-stakeholder solution to x-risk.

This led me to build up a kind of resistance toward people who wanted to ask questions about multi/multi dynamics. However, by 2015, I started to think that multi-stakeholder dynamics were going to *play into* x-risk, even at a technical scale. But when I point this out to other x-risk-concerned people, it often feels like I'm met with the same kind of immune response that I used to have toward people with multi-stakeholder concerns about AI technology.

Other than inertia resulting from this "immune" response, I think there may be other factors contributing to our collective blind-spot around multi/multi dynamics, e.g., a collective aversion to politics.

## **Aversion to thinking about political forces**

*\*\*\* Epistemic status: observation + speculation \*\*\**

[Politics is the Mind-Killer](#) (PMK) is one of the most heavily quoted posts on LessWrong. By my count of the highly-rated posts that cite it, the post has a “LessWrong [h-index](#)” of 32, i.e., 32 posts citing that are each rated 32 karma or higher:

#### Pingbacks

- 283 The noncentral fallacy - the worst argument in the world?
- 155 ★ Simulacra Levels and their Interactions
- 129 ★ Negative Feedback and Simulacra
- 102 Avoid Unnecessarily Political Examples
- 91 Naming the Nameless
- 91 ★ Simulacra and Subjectivity
- 87 Of Exclusionary Speech and Gender Politics
- 83 Missing the Trees for the Forest
- 82 When None Dare Urge Restraint, pt. 2
- 78 So You've Changed Your Mind
- 74 Abnormal Cryonics
- 69 My story / owning one's reasons
- 69 Blue or Green on Regulation?
- 62 The Wonder of Evolution
- 61 Memetic Tribalism
- 61 Mod Notice about Election Discussion
- 58 Only You Can Prevent Your Mind From Getting Killed By Politics
- 56 About Less Wrong
- 56 Meditation Trains Metacognition
- 54 Dunbar's Function
- 52 Fake Optimization Criteria
- 50 Reflections on a Personal Public Relations Failure: A Lesson in Communication
- 49 Politics is hard mode
- 48 Defeating the Villain
- 46 "Politics is the mind-killer" is the mind-killer
- 44 Rational Repentance
- 41 Don't Get Offended
- 39 Voting is like donating thousands of dollars to charity
- 39 Trusting Expert Consensus
- 38 Our Phyg Is Not Exclusive Enough
- 37 A Suggested Reading Order for Less Wrong [2011]
- 32 "I know I'm biased, but..." ← **32 posts here & above**
- 30 Picture Frames, Window Frames and Frameworks
- 25 Frontpage Posting and Commenting Guidelines
- 24 Proposal: Show up and down votes separately
- 24 Anti-tribalism and positive mental health as high-value cause areas
- 23 Has "politics is the mind-killer" been a mind-killer?
- 22 On LessWrong/Rationality and Political Debates
- 21 Parallelizing Rationality: How Should Rationalists Think in Groups?
- 21 A Theory of Pervasive Error
- 20 Rational discussion of politics

The PMK post does not directly advocate for readers to avoid thinking about politics; in fact, it says “I’m not saying that I think we should be apolitical, or even that we should adopt Wikipedia’s ideal of the Neutral Point of View.” However, it also begins with the statement “People go funny in the head when talking about politics”. If a person doubts their own ability not to “go funny in the head”, the PMK post — and concerns like it — could lead them to avoid thinking about or engaging with politics as a way of preventing themselves from “going funny”.

Furthermore, one can imagine this avoidance leading to an under-development of mental and social habits for thinking and communicating clearly about political issues, conceivably making the problem worse for some individuals or groups. This has been remarked before, in the [“Has ‘politics is the mind-killer’ been a mind-killer?”](#).

Finally, these effects could easily combine to create a culture filter selecting heavily for people who dislike or find it difficult to interact with political forces in discourse.

\*\*\* *Epistemic status: personal experience* \*\*\*

The above considerations are reflective of my experience. For instance, at least five people that I know and respect as intellectuals within this community have shared with me that they find it difficult to think about topics that their friends or co-workers disagree with them about.

## **“Alignment” framings from MIRI’s formative years.**

\*\*\* *Epistemic status: publicly verifiable facts* \*\*\*

Aligning Superintelligence with Human Interests: A Technical Research Agenda ([soares2015aligning](#)) was written at a time when laying out precise research objectives was an important step in establishing MIRI as an institution focused primarily on research rather than movement-building (after re-branding from SingInst). The problem of aligning a single agent with a single principal is a conceptually simple starting point, and any good graduate education in mathematics will teach you that for the purpose of *understanding* something confusing, it’s always best to start with the simplest non-trivial example.

\*\*\* *Epistemic status: speculation* \*\*\*

Over time, friends and fans of MIRI may have over-defended this problem framing, in the course of defending MIRI itself as an fledgling research institution against social/political pressures to dismiss AI as a source of risk to humanity. For instance, back in 2015, folks like Andrew Ng were in the habit of publicly claiming that worrying about AGI was “like worrying about overpopulation on mars” ([The Register](#); [Wired](#)), so it was often more important to say “No, that doesn’t make sense, it’s possible powerful AI systems to be misaligned with what we want” than to address the more nuanced issue that “Moreover, we’re going to want a lot of resilient socio-technical solutions to account for disagreements about how powerful systems should be used.”

## **Corrective influences from the MIRI meme-plex**

\*\*\* *Epistemic status: publicly verifiable facts* \*\*\*

Not all influences from the MIRI-sphere have pushed us away from thinking about multi-stakeholder issues. For instance, *Inadequate Equilibria* ([yudkowsky2017inadequate](#)), is clearly an effort to *help* this community to think about multi-agent dynamics, which could help with thinking about politics. Reflective Oracles ([fallenstein2015reflective](#)) are another case of this, but in a highly technical context that probably (according-to-me: unfortunately) didn't have much effect on broader rationality-community-discourse.

## The “problems are only real when you can solve them” problem

\*\*\* *Epistemic status: personal experience / reflections*\*\*\*

It seems to me that many will people tend to ignore a given problem until it becomes sufficiently plausible that the problem can be solved. Moreover, I think their «ignore the problem» mental operation often goes as far as «believing the problem doesn't exist». I saw MIRI facing this for years when trying to point to AI friendliness or alignment as a problem. Frequently people would ask “But what would a solution look like?”, and absent a solution, they'd tag the problem as “not a real problem” rather than just “a difficult problem”.

I think the problem of developing AI tech to enable cooperation-rather-than-conflict is in a similar state right now. Open Problems in Cooperative AI ([dafoe2020cooperative](#)) is a good start at carving out well-defined problems, and I'm hoping a lot more work will follow in that vein.

## Conclusion

\*\*\* *Epistemic status: personal reflections*\*\*\*

I think it's important to consider the potential that a filter-bubble with a fair amount of inertia has formed as a result of our collective efforts to defend AI x-risk as a real and legitimate concern, and to consider what biases or weaknesses are most likely to now be present in that filter bubble. Personally, I think we've developed such a blind spot around technical issues with multi-principal/multi-agent AI interaction, but since Open Problems in Cooperative AI ([dafoe2020cooperative](#)), it might be starting to clear up. Similarly, aversion to political thinking may also be weakening our collective ability to understand, discuss, and reach consensus on the state of the political world, particularly surrounding AI.

# The Point of Trade

(Content warning: econoliteracy. Dialogue based on an actual conversation, but heavily idealized and simplified and stripped of surrounding context.)

**Myself:** - seems unreal because it *is* unreal. But there's a river of reality running through it. If somebody reports that something complicated feels unreal and difficult to get a handle on, usually the first thing I prescribe is going back to the very very basics, and playing around with those to solidify the grasp there. If mathematics seemed unreal to someone, I'd take somebody back to [premises-and-conclusions](#). In the case of complicated modern economical structures, I'd maybe start with trade. What's the point of trade? What does it do?

**Them:** The point of trade is that sometimes people get different amounts of value from things, so they can get more value by trading them. Like, if I have an apple, and you have an orange, but I like oranges more than apples, and you like apples more than oranges, we can trade my apple for your orange, and both be better off.

**Myself:** Yes, that is the horrible explanation that you sometimes see in economics textbooks because nobody knows how to explain anything. But when you are trying to improve your grasp of the very very basics, you should take the basic thing and poke at it and look at it from different angles to see if it seems to capture the whole truth.

In the case of the "people put different values on things", it would seem that on the answer you just gave, trade could never increase wealth by very much, in a certain basic sense. There would just be a finite amount of stuff, only so many apples and oranges, and all trade can do is shuffle the things *around* rather than make any *more* of it. So, on this viewpoint, trade can't increase wealth by all that much.

**Them:** It can increase the utility we get from the things we have, if I like oranges a lot more than I like apples, and you like apples a lot more than oranges.

**Myself:** All right, suppose that all of us liked exactly the same objects exactly the same amount. This obliterates the poorly-written-textbook's reason for "trade". Do you believe that, in an alternate world where everybody had exactly the same taste in apples and oranges, there'd be no further use for trade and trade would stop existing?

**Them:** Hmm. No, but I don't know how to describe what the justification for trade is, in that case.

**Myself:** Modern society seems very wealthy compared to hunter-gatherer society. The vast majority of this increased utility comes from our *having more stuff*, not from our having the same amount of stuff as hunter-gatherers but giving apples to the exact person on Earth who likes apples most. I claim that the reason we *have more stuff* has something to do with trade. I claim that in an alternate society where everybody likes every object the same amount, they still do lots and lots of trade for this same reason, to increase how much stuff they have.

**Them:** Okay, my new answer is that, through trade, you can get strawberries from far away, where they wouldn't be in season at all, where you are... no, that doesn't actually make more stuff. My new answer is that you can build complicated things

with lots of inputs, by trading to get the inputs. Like if it takes iron and copper to build circuits, you can trade to get those.

**Myself:** If it takes 1 unit of effort to get 1 unit of iron either way, how can you get any *more* stuff out, at the end, by trading things? It takes 1 unit of effort to make 1 iron ingot, so go to the iron mines and mine some iron, then chop down the wood to prebake the iron ore for grinding before you put it into the bloomery. All of that has to be done either way to get the iron ingot. How can trading for somebody else's iron, instead, cause there to be more stuff in the economy as a whole?

**Them:** Because task-switching has costs.

**Myself:** Okay, suppose an alternate society of people who are *really* good at task-switching. They can just swap straight from one task to another with no pause. They also all have exactly the same tastes in apples and oranges and so on. Does this new society now have zero use for trade?

**Them:** Um... hm. (*Thinks.*) But they're not actually in the same *place* as the iron mines. So if they have to walk around a lot -

**Myself:** Suppose a society in which everyone has exactly the same taste in apples and oranges; everybody is really really good at switching tasks; and furthermore, the society has Star Trek transporter pads, so you can get to the iron mine instantly. Is there *now* no more use for trade?

**Them:** Some people are better miners and others are better fruit-growers?

**Myself:** Suppose a society with identical fruit tastes, *and* perfect task-switching, *and* Star Trek transporters, *and furthermore* everyone has identical genetics, as if they were all identical twins; which, as we know from identical-twin studies, means that everybody will have around the same amount of innate talent for any and all jobs. Like that case where two identical twins, separated at birth, who never knew each other, both ended up as firefighters. As we all know, studies on separated identical twins show that happens every single time, with no exceptions. I claim that even *this* society still has to do a lot of trade in order to end up with modern levels of wealth.

Now, do you think I'm trolling you and that we actually did get rid of the basic reason for trade, at this point, or that there's still something left over? Identical fruit tastes, perfect task-switching, Star Trek transporters, everyone is born with the same genetics and therefore identical innate talents. Do people now mine their own iron, or do they still trade for it?

**Them:** (*Thinks for a while.*)

**Me:** If the Sequences have taught you anything, I hope it's taught you that it's okay to state the obvious.

**Them:** ...people learn to do their jobs better with practice?

**Myself:** Human capital accumulation! Indeed! So now let us suppose identical fruit tastes, perfect task-switching, Star Trek transporters, identically cloned genetics, *and* people can share expertise via Matrix-style downloads which are free. Have we *now* gotten rid of the point of trade? As far as you can tell.

**Them:** ...yes?

**Myself:** Do you believe I'm going to say that we've gotten rid of the point of trade?

**Them:** ...no.

**Myself:** Well, I agree with your object-level answer, so your meta-level answer was wrong. I think we've now gotten rid of the point of trade.

**Them:** Darn it.

*(Note: While contemplating this afterwards, I realized that we hadn't quite gotten rid of all the points of trade, and there should have been two more rounds of dialogue; there are two more magical powers a society needs, in order to produce a high-tech quantity of stuff with zero trade. The missing sections are left as an exercise for the reader.)*

# The Apprentice Experiment

About two months ago, someone asked me what I would do with more funding. Other than the obvious (i.e. generally improve my own quality-of-life in minor ways), my main answer was: take on an apprentice. I have some models about how best to train people for this sort of work, and an apprentice would allow me to test those models while also supporting my own research. I started laying groundwork for that plan - in particular, [Specializing in Problems We Don't Understand](#) laid out my main background model.

Then, about a month ago, [Aysajan](#) put up a short post titled “[Can I be Your Apprentice?](#)” - essentially an open call to people on LW doing cool work. We talked, it seemed like a good fit, so the apprentice experiment kicked off ~3 weeks ago.

This post will provide more detail on models, motivation, the plan, etc, including a section for Aysajan to introduce himself.

## Background Models

First background model: [Specializing in Problems We Don't Understand](#). Problems-we-don't-understand are similar to each other in a way which problems-we-do-understand are not. In the context of scientific research, preparadigmatic research in different fields is similar in a way which research within a paradigm is not. There are general skills and knowledge useful for finding/creating structure *de novo*, as opposed to working within some already-mapped structure.

Furthermore, while problems-we-don't-understand may require some specialized knowledge, specialized knowledge of the field is never the rate-limiting step; if it were, then the problem would already be tractable to people steeped in the existing specialized knowledge of the field. If a problem is tractable within the current paradigm, then it isn't preparadigmatic. Broad, generalizable skills/knowledge are much more important for problems-we-don't-understand than for problems-we-do-understand.

The linked post goes into more detail on how one can train and specialize in problems-we-don't-understand.

Second background model: [Selection Has A Quality Ceiling](#). If we want people with a lot of skill in a lot of areas, trying to hire such people directly is Hard, in a big-O sense. As the number of traits we're filtering for increases, the number of people we have to test in order to find one with all the requisite traits increases exponentially. The big-O requirements *training* skills are much better: as long as learning one skill doesn't make another harder, the time required to train all of them should increase at-most linearly with the number skills.

Alas, most schools/companies today seem to mostly select, rather than train. Which makes sense - most companies don't really need people with lots of skill in lots of areas, they just need people who will pick up the particulars of their industry quickly as-needed. But for problems-we-don't-understand, people with lots of skill in lots of areas are exactly what we want.

Third background model: [illegible skills](#). A lot of key skills/knowledge are hard to transmit by direct explanation. They're not necessarily things which a teacher would even notice enough to consider important - just background skills or knowledge which is so ingrained that it becomes invisible. This sort of skill/knowledge is most easily transmitted by exposure: demonstration by the teacher, experimentation by the student, and feedback, ideally on a day-to-day basis. Thus the importance of an apprenticeship-like structure: high exposure and one-on-one interaction helps transmit illegible skills/knowledge.

(I suspect that this also relates to [Bloom's two-sigma problem](#): one-on-one tutoring works about two standard deviations better than anything else in education. Regardless of whether illegible skill transmission is actually a core part of that phenomenon, an apprenticeship certainly involves enough one-on-one tutoring that I expect the two-sigma benefit to kick in.)

## The Plan

Originally, I planned to put out a call for an apprentice around the end of this month/early next month. I hoped to get a few responses, filter for basic technical skills and personality compatibility, then randomly choose someone from a hopefully-not-too-short list. The intent was to avoid filtering heavily: I want to *create* new human capital, not merely *select* for existing human capital. And if the experiment works, I want to be able to do it again. Choosing someone who's already obviously a uniquely good fit would compromise the information-value of the experiment.

Instead of that process, I've effectively selected on one thing: putting up a LessWrong post asking if anyone wants an apprentice. (Well, ok, I did also screen for basic technical skills and personality compatibility.) Aysajan's resume is typical enough that I'm not too worried about selection effects there, but putting up a LessWrong post asking if anyone wants an apprentice implies a kind of chutzpah and do-what-it-takes attitude that may not be so easy to replicate. So from an experimental replicability standpoint, that's a minor note of concern. (From a personal standpoint, I love it.)

From here, the plan is for Aysajan to spend the next few months working on the sorts of projects I worked on before focusing on alignment full time - the sorts of projects which I expect to build skills for solving problems-we-don't-understand. These won't be strictly or even primarily alignment-related; the goal is to build skill in solving problems-we-don't-understand, and alignment is a pretty difficult area in which to practice.

Aysajan's [first post](#) since the apprenticeship started went up just recently. It's a write-up of an exercise looking for various systems besides probabilistic models which satisfy the assumptions used in [Cox' Theorem](#) to derive Bayes' Rule.

I don't have any particular experimental outcomes to measure. If the project goes as well as I hope, then I expect it will be quite obvious. No need to go inventing proxies which don't actually quite capture the things I care about.

## Aysajan's Intro

*(This section is in Aysajan's voice.)*

I am a business school faculty in a Canadian university. I earned my master's degree in statistics and PhD in operations research in the US in 2018. I joined LessWrong not long ago and have been truly enjoying the intelligent conversations/debates going on here, whether it is related to general rationality, ML/AI, or simply investment. In the meantime, I find myself thinking Albert Einstein's famous quote about learning quite frequently. He famously said "The more I learn, the more I realize how much I don't know". While being truly amazed by all the fascinating work LessWrong community members have been doing, I realize that I couldn't do much due to my limited domain knowledge and limited hands-on experience. I am inspired and I want to contribute, especially in the fields of ML/AI research. But in reality, as an outlier in my current professional network (business academia), I am greatly struggling due to lack of guidance. Thus I made a call for an apprenticeship. I have a strong desire to contribute to the community by conducting original research and I believe apprenticeship is one of the best ways to learn ML/AI skills: learn from the best and do original research.

## Hopes

*(This section is in John's voice, but speaks for both of us.)*

Best-case scenario, this experiment provides a prototype for producing new specialists in problems-we-don't-understand. At a personal level, we want to work with such people, and the ability to produce them would make that a lot easier. At a community level, alignment is a particularly difficult problem-we-don't-understand (especially due to the lack of good feedback loops), and we hear that we're now more bottlenecked on effective researchers than on funding.

But what we really want is a whole community or institute of people who specialize in problems-we-don't-understand, making breakthroughs not only on alignment but on aging, on designing organisms from scratch, on efficient orbital delivery and terraforming, on generalized cognitive processes in organisms or organizations or brains, on fusion energy, on practical design of organizations or contract structures or memes, on cryptographically-secure biodefenses, .... We want a group of people who will set the whole damn world on fire.

# The Apprentice Thread

A while back, LessWrong poster Aysajan [put up a post asking to be someone's apprentice](#). He talked about it with [johnswentworth](#), who I recently confirmed via meeting him in person is awesome and does reliably interesting work, [and an apprentice experiment was born](#).

As John says, you gotta admire the chutzpah. Asking for what one wants is a [known to be successful but highly underused strategy](#), I presume mostly because of the permanent global chutzpah shortage and the associated danger that it might result in mild social awkwardness.

In addition to the highly successful use of chutzpah, this also points out that apprenticeships are also a known to be successful but highly underused strategy. [My feelings about so-called 'schools' are well known](#), but *education* is great, and apprenticeship is one of the best ways to get an actually excellent and useful education.

I've been an apprentice, regardless of whether it was called that, and it was awesome. At Jane Street they have a formal education process, but the core of how you get good is an apprenticeship. You learn from the best, by working with the best and asking them questions. I believe I am a natural trader, but I am good largely because I learned from working directly with three of the best in succession at three different jobs. I know Magic and game design by spending tons of time talking and working with top Magic players and game designers.

I've been a mentor of sorts a few times. That's mostly been great too, and I'm plausibly taking on another one now, although I'm cheating because he is already exceptional and largely does not need the help.

Thus, we have two overpowered strategies here that the world needs more of: Apprenticeship and Chutzpah. Lowering the activation energy required for either or both of them seems great, as does providing encouragement.

Both can of course be overused and abused. Too much of the wrong kind of chutzpah is no good for anyone, and apprenticeship can turn into a bunch of not very useful unpaid work, or end up holding people back. In the context of posts like this, I am not much worried about either of these failure modes.

For this post/thread I will focus on apprenticeship. In particular, I want to see if I can give social permission and a coordination mechanism that can perhaps take place in the comments (reminder that my posts have two comments sections, the primary one at DWATV and a secondary one at LessWrong).

Replies to this post should take the form of any of the following:

1. [MENTOR]: A non-binding indication of potential interest in mentorship. Mention that you might, at some point, be interested in taking on an apprentice. This commits you to nothing. Make sure to indicate what you'd be teaching them and what project would likely be involved, and open with [MENTOR]. You are free to include contact info, or not include it and monitor replies.
  1. Replies to this comment to indicate potential interest in being the apprentice, marked [APPRENTICE], which should include a method of

further contact.

2. [APPRENTICE]: A non-binding indication of potential interest in being an apprentice. Mention that you might, at some point, be interested in being an apprentice. This commits you to nothing. Make sure to indicate what you're interested in being an apprentice in and learning, and an indication of what's motivating you.
  1. Replies to this comment to indicate potential interest in being the mentor, marked with [MENTOR], which should include a method of further contact.
3. [NORMAL] You're free to comment as per normal, but start with [NORMAL] in the top-level for clarity.
4. [NYCBUSINESS] if there's some chance, depending on what it is, that you would want to do the thing I talk about below.

Cost of speaking up is low, potential upside is high, and hopefully not too much chutzpah is now required.

As for me: Right now, my main focus is on game design. We hired two new designers this week, both of whom I'm super excited to work with on Emergents, so for now my card is full. Depending on how much bandwidth I prove to have, I could consider mentoring someone at some point in either game design (on either Emergents or my other project), in trading (although I kind of already know who that would likely be if it happened), or someone I trusted sufficiently who wanted to work on [Aikido](#). I might also want to be an apprentice at some point, likely in something AI-alignment related, but that would be a long way off.

I've also got another project potentially in the works that I'd love to work with someone on, which would involve someone taking on a likely-more-than-full-time job and a lot of responsibility, starting and running a business I want to exist. It would be hard work and require a self-starter, and all that, but won't require you to do startup fundraising – it would be a real business, and would succeed or fail organically. It would be local to New York City. If you think you might be the person for the job, you can mention that with [NYCBUSINESS], tell me as much or as little as you'd like, and if/when I'm ready I will reach out. Again, this commits you to absolutely nothing.

# Irrational Modesty

At an online meetup I attended, I talked to a fellow who was working on some software that he thought would be useful for aggregating information for Effective Altruists. He said he would like to work on it full time this summer rather than get an internship - but that this was not financially feasible.

His idea seemed not-obviously stupid and I suggested he apply for a LTFF or EAIF grant. He said he did not want to as he was worried the money could go to better use.

I asked him if he thought the grant-makers were in a better position than him to decide if his project would be an efficient use of their capital. He said he did. I then asked him if he thought his idea was so likely to be a waste of time that even reading his proposal would be net negative in expectation; he said, "No".

He is now working on his proposal.

Here we have an example of someone who is likely much smarter than me but was unable to think clearly due to irrational modesty. I am worried this may be true of some *extremely* capable people, and how tragic it would be if this is the case.

The vast majority lack the requisite ability to do alignment research. I have a good model of my talents and capabilities, and I am certain I do not. In truth, it is not even close. Those who are capable, confident in their abilities, and motivated to work on this problem do not need a pep talk. But the possibility that there is a class of highly-talented would-be-motivated people who lack confidence in their abilities still haunts me.

I am reminded of this humorous exchange:

[–] **quentin** 10y ♂ < 3 >

I was briefly excited as I met both GRE and SAT cutoffs. But now I'm feeling guilty and debating whether or not to apply; I'm certainly not in the 99.9th percentile. I absolutely love this community but I don't really post because I sincerely feel inadequate.

I'm easily in the 5th percentile, but I feel like an imposter with my standardized test scores: the tests are SO damn easy and don't measure anything of substance. GRE verbal tests your ability to recall obscure words, and the math tests your ability to maintain focus through 2 hours of trivial middle-school math. I didn't study at all.  
Reply

[–] **wedrifid** 10y ♂ < 22 >

the tests are SO damn easy

That's what being intelligent is *supposed* to feel like!

But now I'm feeling guilty and debating whether or not to apply

Guilt is overrated. They say you qualify. Therefore you do. It's their study. In fact, if you do qualify but don't think you should then you are biasing their data against genes for low self esteem.

Reply

Though quentin's blindness to his own intelligence is humorous, it *is* a mistake and a rather large one. Irrational modesty is every bit as bad as overconfidence, possibly more so.

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

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

# Alcohol, health, and the ruthless logic of the Asian flush

This is a linkpost for <https://dynamight.net/alcohol/>

Say you're an evil scientist. One day at work you discover a protein that crosses the blood-brain barrier and causes crippling migraine headaches if someone's attention drifts while driving. Despite being evil, you're a loving parent with a kid learning to drive. Like everyone else, your kid is completely addicted to their phone, and keep refreshing their feeds while driving. Your suggestions that the latest squirrel memes be enjoyed *later at home* are repeatedly rejected.

Then you realize: You could just sneak into your kid's room at night, anesthetize them, and bring them to your lair! One of your goons could then extract their bone marrow and use CRISPR to recode the stem-cells for an enzyme to make the migraine protein. Sure, the headache itself might distract them, but they'll probably just stop using their phone while driving. Wouldn't you be at least tempted?

This is an analogy for something about alcoholism, East Asians, Odysseus, evolution, tension between different kinds of freedoms, and an idea I thought was good but apparently isn't.

# Selection Has A Quality Ceiling

Suppose we're working on some delightfully Hard problem - genetically engineering a manticore, or terraforming Mars, or aligning random ML models. We need very top tier collaborators - people who are very good at a whole bunch of different things. The more they're good at, and the better they are, the better the chances of success for the whole project.

There's two main ways to end up with collaborators with outstanding skill/knowledge/talent in many things: selection or training. Selection is how most job recruitment works: test people to see if they already have (some of) the skills we're looking for. Training instead starts with people who don't have (all of) the skills, and installs them de novo.

Key point of this post: selection does not scale well with the level of people we're looking for. As we increase the number of skills-we-want in our collaborators, the fraction-of-people with all those skills shrinks exponentially, so the number-we-need-to-test grows exponentially. Training has much better asymptotic behavior: as the number of skills-we-want grows, the amount of training needed to install them grows only linearly - assuming we're able to train them at all.

## Bits Of Search

Suppose I have some test or criterion, and only half the population passes it - for instance, maybe I want someone with above-median math skills. That's [one bit of search](#): it eliminates half the possibilities.

If I want above-median math skills and above-median writing skills, that's (approximately) two bits, and I expect (approximately) one-in-four people to pass both tests. (Really, math and writing skills are correlated, so it will be somewhat more than one-in-four and thus somewhat less than two bits of search.) As more skills are added to the list of requirements, adding more "bits of search", the number of people who pass all requirements will fall exponentially. With  $k$  bits of search, only  $1\text{-in-}2^k$  people will pass, so I'll need to search over  $\sim 2^k$  people just to find one potential collaborator.

In practice, skills are not independent, but the correlation is weak enough that exponentials still kick in. (Indeed, the only way exponentials won't kick in is if correlation increases rapidly as we add more skills.)

I also sometimes want more-than-one bit of search in just one skill. For instance, if I want someone in the top 1/32 of writing skill, then that's 5 bits of search. In practice, we usually want quite a few bits in relevant skills - for instance, if I'm looking for help genetically engineering a manticore, then I'll want people with deep expertise in developmental biology and morphogenesis. I'd probably want something like 20 bits (i.e. a one-in-a-million person) in those skills alone, plus whatever other skills I might want (e.g. good communication, quantitative thinking, etc).

## Asymptotics of Selection vs Training

So, as I crank up the number of bits-of-search, the search becomes exponentially more difficult. It won't take long before nobody in the world passes my tests - there's only ~10B people, so ~34 bits is all I get, and that's if I test literally everyone in the world. That puts a pretty low skill cap on potential collaborators I can find! And even before I hit the everyone-in-the-world cap, exponential growth severely limits how much I can select.

There are ways around that: skills are not independent, and sometimes I can make do with someone who has *most* of the skills. But the basic picture still holds: as I raise my bar, selection becomes exponentially more difficult.

Training, in principle, does not have this problem. If I want to train two independent skills, then the time required to train both of them is the *sum* of time required to train each, rather than a product. So, training resource requirements should generally grow linearly, rather than exponentially. Again, skills aren't really independent, but the basic picture should still hold even when we make the model more complicated.

## Problem: We Don't Know How To Train

When we look at schools or companies, they seem to [mostly select](#). To the extent that training does take place, it's largely accidental: people are expected to magically pick up some skills in their first weeks or months at a new job, but there isn't much systematic effort to make that happen efficiently/reliably.

... and for most institutions, that's good enough. The asymptotic arguments apply to finding "very high quality" people, by whatever criteria are relevant. Most institutions neither need nor find the very best (though of course lots of them *claim* to do so). Most people, most of the time, work on [problems-we-basically-understand](#). They just need to be able to use known tools in known ways, in similar ways to everyone else in their field, and about-as-well as others in their field. As long as the field is large, there are plenty of typical candidates, and selection works fine.

Selection breaks down when we need people with rare skills, and especially when we need people with many independent skills - exactly the sort of people we're likely to need for [problems-we-basically-don't-understand](#).

But it still seems like training ought to be great - it should be profitable for schools or companies to install new skills in people. In some specific areas, it is profitable. So why don't we see more of this? Here's one theory: in order to train *systematically*, we need some kind of feedback loop - some way to tell whether the training is working. In other words, we need a test. Similarly, we need a test to *prove to others* that the training worked. And if we have a test, then we could just forget about training and instead use the test to select. As long as we're not asking for too many bits, that's probably cheaper than figuring out a whole training program.

So, we end up with a society that's generally not very good at training.

## Summary

Most of the world mostly "gets good people" by selection: we start with a big pool of candidates and then filter for those which best fit our criteria. But this technique puts a cap on "how good" we can select for - we can't ask for someone better than the

best in the world. Even if the number of people is effectively infinite, we still need to search over exponentially many candidates as the list of selection criteria grows.

For most institutions, this isn't much of a problem, because they're not "in the asymptote" - they don't really need people with that many bits of perfection. But the harder our problems, the more we need people with many bits - potentially people better than the current best in the world, or potentially people who are just too rare to cheaply search for in a giant pool of candidates. At that point, we have no choice but to train, rather than select.

Training is hard; it's not a thing which most institutions know how to do well today. But if we want top-level collaborators in many skills, then we just have to figure out how to do it. Selection does not scale that way.

# On the limits of idealized values

(Cross-posted from [Hands and Cities](#))

On a popular view about meta-ethics, what you should value is determined by what an idealized version of you would value. Call this view “idealizing subjectivism.”

Idealizing subjectivism has been something like my best-guess meta-ethics. And lots of people I know take it for granted. But I also feel nagged by various problems with it — in particular, problems related to (a) circularity, (b) indeterminacy, and (c) “passivity.” This post reflects on such problems.

My current overall take is that especially absent certain strong empirical assumptions, idealizing subjectivism is ill-suited to the role some hope it can play: namely, providing a privileged and authoritative (even if subjective) standard of value. Rather, the version of the view I favor mostly reduces to the following (mundane) observations:

- If you already value X, it’s possible to make instrumental mistakes relative to X.
- You can choose to treat the outputs of various processes, and the attitudes of various hypothetical beings, as authoritative to different degrees.

This isn’t necessarily a problem. To me, though, it speaks against treating your “idealized values” the way a [robust meta-ethical realist](#) treats the “true values.” That is, you cannot forever aim to approximate the self you “would become”; you must actively create yourself, often in the here and now. Just as the world can’t tell you what to value, neither can your various hypothetical selves — unless you choose to let them. Ultimately, it’s on you.

## I. Clarifying the view

Let’s define the view I have in mind a little more precisely:

**Idealizing subjectivism:** X is intrinsically valuable, relative to an agent A, if and only if, *and because*, A would have some set of evaluative attitudes towards X, if A had undergone some sort of idealization procedure.

By evaluative attitudes, I mean things like judgments, endorsements, commitments, cares, desires, intentions, plans, and so on. Versions of the view differ in which they focus on.

Example types of idealization might include: full access to all relevant information; vivid imaginative acquaintance with the relevant facts; the limiting culmination of some sort of process of reflection, argument, and/or negotiation/voting/betting between representatives of different perspectives; the elimination of “biases”; the elimination of evaluative attitudes that you don’t endorse or desire; arbitrary degrees of intelligence, will-power, dispassion, empathy, and other desired traits; consistency; coherence; and so on.

Note that the “and because” in the definition is essential. Without it, we can imagine paradigmatically non-subjectivist views that qualify. For example, it could be that the idealization procedure necessarily results in A’s *recognizing* X’s objective, mind-independent value, because X’s value is one of the facts that falls under “full

information.” Idealizing subjectivism explicitly denies this sort of picture: the point is that A’s idealized attitudes *make* X valuable, relative to A. (That said, views on which all idealized agents *converge* in their evaluative attitudes can satisfy the definition above, provided that value is *explained* by the idealized attitudes in question, rather than vice versa.)

“Relative to an agent A,” here, means something like “generating (intrinsic) practical reasons for A.”

## II. The appeal

Why might one be attracted to such a view? Part of the appeal, I think, comes from resonance with three philosophical impulses:

- A rejection of certain types of [robust realism](#) about value, on which value is just a brute feature of the world “out there.”
- A related embrace of a kind of Humeanism about means and ends. The world can tell you the means to your ends, but it cannot tell you what ends to pursue — those must in some sense be there already, in your (idealized?) heart.
- An aspiration to maintain some kind of deep connection between what’s valuable, and what actually moves us to act (though note that this connection is not universalized — e.g., what’s valuable relative to you may not be motivating to others).

Beyond this, though, a key aim of idealizing subjectivism (at least for me) is to capture the sense in which it’s possible to question what you *should* value, and to make mistakes in your answer. That is, the idealization procedure creates some distance between your current evaluative attitudes, and the truth about what’s valuable (relative to you). Things like “I want X,” or “I believe that X is valuable” don’t just settle the question.

This seems attractive in cases like:

1. (Factual mistake) Alfred wants his new “puppy” Doggo to be happy. Doggo, though, is really a simple, non-conscious robot created by mischievous aliens. If Alfred knew Doggo’s true nature, he would cease to care about Doggo in this way.
2. (Self knowledge) Betty currently feels very passionately about X cause. If she knew, though, that her feelings were really the product of a desire to impress her friend Beatrice, and to fit in with her peers more broadly, she’d reject them. (This example is inspired by one from Yudkowsky [here](#).)
3. (Philosophical argument) Cindy currently thinks of herself as an average utilitarian, and she goes around trying to increase average utility. However, if she learned more about the counterintuitive implications to average utilitarianism, she would switch to trying to increase total utility instead.
4. (Vividness) Denny knows that donating \$10,000 to the Against Malaria Foundation, instead of buying a new grand piano, would save multiple lives in expectation. He’s currently inclined to buy the grand piano. However, if he imagined more vividly what it means to save these lives, and/or if he actually witnessed the impact that saving these lives would have, he’d want to donate instead.
5. (Weakness of the will) Ernesto is trapped by a boulder, and he needs to cut off his own hand to get free, or he’ll die. He really doesn’t want to cut off his hand.

However, he would want himself to cut off the hand, if he could step back and reflect dispassionately.

6. (Incoherence) Francene prefers vacationing in New York to San Francisco, San Francisco to LA, and LA to New York, and she pays money to trade “vacation tickets” in a manner that reflects these preferences. However, if she reflected more on her vulnerability to losses this way, she’d resolve her circular preferences into New York > SF > LA.
7. (Inconsistency) Giovanni’s intuitions are (a) it’s impermissible to let a child drown in order to save an expensive suit, (b) it’s permissible to buy a suit instead of donating the money to save a distant child, and (c) there’s no morally relevant difference between these cases. If he had to give one of these up, he’d give up (c).
8. (Vicious desires) Harriet feels a sadistic desire for her co-worker to fail and suffer, but she wishes that she didn’t feel this desire.

By appealing to the hypothetical attitudes of these agents, the idealizing subjectivist aims to capture a sense that their actual attitudes are, or at least could be, in error.

Finally, idealizing subjectivism seems to fit with our actual practices of ethical reflection. For example, thinking about value, we often ask questions like: “what would I think/feel if I understood this situation better?”, “what would I think if I weren’t blinded by X emotion or bias?” and so forth — questions reminiscent of idealization. And ethical debate often involves seeking a kind of [reflective equilibrium](#) — a state that some idealizers take as *determining* what’s valuable, rather than indicating it.

These, then, are among the draws of idealizing subjectivism (there are others) — though note that whether the view can actually *deliver* these goods (anti-realism, Humeanism, fit with our practices, etc) is a further question, which I won’t spend much time on.

What about objections? One common objection is that the view yields counterintuitive results. Plausibly, for example, we can imagine ideally-coherent suffering maximizers, brick-eaters, agents who are indifferent towards future agony, agents who don’t care about what happens on future Tuesdays, and so on — agents whose pursuit of their values, it seems, need involve no mistakes (relative to them). We can debate which of such cases the idealized subjectivist must concede, but pretty clearly: some. In a sense, cases like this lie at the very surface of the view. They’re the immediate implications.

(As I’ve discussed [previously](#), we can also do various semantic dances, here, to avoid saying certain relativism-flavored things. For example, we can make “a paperclip maximizer shouldn’t clip” true in a hedonist’s mouth, or a paperclip maximizer’s statement “I should clip” false, evaluated by a hedonist. Ultimately, though, these moves don’t seem to me to change the basic picture much.)

My interest here is in a different class of more theoretical objections. I wrote about one of these in my [post](#) about moral authority. This post examines some others. (Many of them, as well as many of the examples I use throughout the post, can be found elsewhere in the literature in some form or other.)

### III. Which idealization?

Consider Clippy, the paperclip maximizing robot. On a certain way of imagining Clippy, its utility function is fixed and specifiable *independent* of its behavior, including

behavior under “idealized conditions.” Perhaps we imagine that there is a “utility function slot” inside of Clippy’s architecture, in which the programmers have written “maximize paperclips!” — and it is *in virtue* of possessing this utility function that Clippy consistently chooses more paperclips, given idealized information. That is, Clippy’s behavior *reveals* Clippy’s values, but it does not *constitute* those values. The values are identifiable by other means (e.g., reading what’s written in the utility function slot).

If your values are identifiable by means *other* than your behavior, and if they are already coherent, then it’s much easier to distinguish between candidate idealization procedures that *preserve* your values vs. changing them. Holding fixed the content of Clippy’s “utility function slot,” for example, we can scale up Clippy’s knowledge, intelligence, etc, while making sure that the resulting, more sophisticated agent is also a paperclip maximizer.

But note, though, that in such a case, appeals to idealization also don’t seem to do very much useful normative work, for subjectivists. To explain what’s of value relative to this sort of Clippy, that is, we can just look directly at Clippy’s utility function. If humans were like this, we could just look at a human’s “utility function slot,” too. No fancy idealization necessary.

But humans aren’t like this. We don’t have a “utility function slot” (or at least, I’ll assume as much in what follows; perhaps this — more charitably presented — is indeed an important point of dispute). Rather, our beliefs, values, heuristics, cognitive procedures, and so on are, generally speaking, a jumbled, interconnected mess (here I think of a friend’s characterization, expressed with a tinge of disappointment and horror: “an unholy and indeterminate brew of these … *sentiments*”). The point of idealizing subjectivism is to take this jumbled mess as an *input* to an idealization procedure, and then to *output* something that plays the role of Clippy’s utility function — something that will *constitute*, rather than reveal, what’s of value relative to us.

In specifying this idealization procedure, then, we don’t have the benefit of holding fixed the content of some slot, or of specifying that the idealization procedure can’t “change your values.” Your values (or at least, the values we care about not changing) just are whatever comes out the other side of the idealization procedure.

Nor, importantly, can we specify the idealization procedure via reference to some independent truth that its output needs to track. True, we evaluate the “ideal-ness” of other, more epistemic procedures this way (e.g., the ideal judge of the time is the person whose judgment actually tracks what time it is — see [Enoch \(2005\)](#)). But the point of idealizing subjectivism is that there is no such independent truth available.

Clearly, though, not just any idealization procedure will do. Head bonkings, brainwashings, neural re-wirings — starting with your current brain, we can refashion you into a suffering-maximizer, a brick-eater, a [helium-maximizer](#), you name it. So how are we to distinguish between the “ideal” procedures, and the rest?

## IV. Galaxy Joe

To me, this question gains extra force from the fact that your idealized self, at least as standardly specified, will likely be a quite alien creature. Consider, for example, the criterion, endorsed in some form by basically every version of idealization subjectivism, that your idealized self possess “full information” (or at least, full *relevant* information — but what determines relevance?). This criterion is often treated

casually, as though a run-of-the-mill human could feasibly satisfy it with fairly low-key modifications. But my best guess is that to the extent that possessing “full information” is a thing at all, the actual creature to imagine is more like a kind of God — a being (or perhaps, a collection of beings) with memory capacity, representational capacity, and so on *vastly* exceeding that of any human. To evoke this alien-ness concretely, let’s imagine a being with a computationally optimal brain the size of a galaxy. Call this a “galaxy Joe.”

Here, we might worry that no such galaxy Joe could be “me.” But it’s not clear why this would matter, to idealizing subjectivists: what’s valuable, relative to Joe, could be grounded in the evaluative attitudes of galaxy Joe, even absent a personal identity relation between them. The important relation, for example, be some form of psychological continuity (though I’ll continue to use the language of self-hood in what follows).

Whether me or not, though: galaxy Joe seems like he’ll likely be, from my perspective, a crazy dude. It will be hard/impossible to understand him, and his evaluative attitudes. He’ll use concepts I can’t represent. His ways won’t be my ways.

Suppose, for example, that a candidate galaxy Joe — a version of myself created by giving original me “full information” via some procedure involving significant cognitive enhancement — shows me his ideal world. It is filled with enormously complex patterns of light ricocheting off of intricate, nano-scale, mirror-like machines that appear to be in some strange sense “flowing.” These, he tells me, are computing something he calls [incomprehensible galaxy Joe concept (IGJC) #4], in a format known as [IGJC #5], undergirded and “hedged” via [IGJC #6]. He acknowledges that he can’t explain the appeal of this to me in my current state.

“I guess you could say it’s kind of like happiness,” he says, warily. He mentions an analogy with abstract jazz.

“Is it conscious?” I ask.

“Um, I think the closest short answer is ‘no.’” he says.

Of course, by hypothesis, I would become him, and hence value what he values, if I went through the procedure that created him — one that apparently yields full information. But now the question of whether this is a procedure I “trust,” or not, looms large. Has galaxy Joe gone off the rails, relative to me? Or is he seeing something incredibly precious and important, relative to me, that I cannot?

The stakes are high. Suppose I can create either this galaxy Joe’s favorite world, or a world of happy puppies frolicking in the grass. The puppies, from my perspective, are a pretty safe bet: I myself can see the appeal. Expected value calculations under moral uncertainty aside, suppose I start to feel drawn towards the puppies. Galaxy Joe tells me with grave seriousness: “Creating those puppies instead of IGJC #4 would be a mistake of truly ridiculous severity.” I hesitate. Is he right, relative to me? Or is he basically, at this point, an alien, a paperclip maximizer, for all his humble roots in my own psychology?

Is there an answer?

## V. Mind-hacking vs. insight

Here's a related intuition pump. Just as pills and bonks on the head can change your evaluative attitudes, some epistemically-flavored stimuli can do so, too. Some such changes we think of as "legitimate persuasion" or "value formation," others we think of as being "brainwashed," "mind-hacked," "reprogrammed," "misled by rhetoric and emotional appeals," and so on. How do we tell (or define) the difference?

Where there are independent standards of truth, we can try appealing to them. E.g., if Bob, a fiery orator, convinces you that two plus two is five, you've gone astray (though even cases like this can get tricky). But in the realm of pure values, and especially absent other flagrant reasoning failures, it gets harder to say.

One criterion might be: if the persuasion process would've worked independent of its content, this counts against its legitimacy (thanks to Carl Shulman for discussion). If, for example, Bob, or exposure to a certain complex pattern of pixels, can convince you of *anything*, this might seem a dubious source of influence. That said, note that certain common processes of value formation — for example, attachment to your hometown, or your family — are "content agnostic" to some extent (e.g., you would've attached to a different hometown, or a different family, given a different upbringing); and ultimately, different evolutions could've built wildly varying creatures. And note, too, that some standard rationales for such a criterion — e.g., being convinced by Bob/the pixels doesn't correlate sufficiently reliably with *the truth* — aren't in play here, since there's no independent truth available.

Regardless, though, this criterion isn't broad enough. In particular, some "mind-hacking" memes might work *because of* their content — you can't just substitute in arbitrary alternative messages. Indeed: one wonders, and worries, about what sort of [Eldritch horrors](#) might be lurking in the memespace, ready and able, by virtue of their content, to reprogram and parasitize those so foolish, and incautious, as to attempt some sort of naive acquisition of "full information."

To take a mundane example: suppose that reading a certain novel regularly convinces people to become egoists, and you learn, to your dismay (you think of yourself as an altruist), that it would convince you to become so, too, if you read it. Does your "idealization procedure" involve reading it? You're not used to avoiding books, and this one contains, let's suppose, no falsehoods or direct logical errors. Still, on one view, the book is, basically, brainwashing. On another, the book is a window onto a new and legitimately more compelling vision of life. By hypothesis, you'd take the latter view after reading. But what's the *true view*?

Or suppose that people who spend time in bliss-inducing [experience machines](#) regularly come to view time spent in such machines as the highest good, because their brains receive such strong reward signals from the process, though not in a way different in kind from other positive experiences like travel, fine cuisine, romantic love, and so on (thanks to Carl Shulman for suggesting this example). You learn that you, too, would come to view machine experiences this way, given exposure to them, despite the fact that you currently give priority to non-hedonic goods. Does your idealization process involve entering such machines? Would doing so result in a "distortion," an (endorsed, desired) "addiction"; or would it show you something you're currently missing — namely, just how intrinsically good, relative to you, these experiences really are?

Is there an answer?

As with the candidate galaxy Joe above, what's needed here is some way of determining which idealization procedures are, as it were, the real deal, and which create imposters, dupes, aliens; which brain-wash, alter, or mislead. I'll consider three options for specifying the procedure in question, namely:

- Without reference to your attitudes/practices.
- By appeal to your *actual* attitudes/practices.
- By appeal to your *idealized* attitudes/practices.

All of these, I think, have problems.

## VI. Privileged procedures

Is there some privileged procedure for idealizing someone, that we can specify and justify without reference to that person's attitudes (actual or ideal)? To me, the idea of giving someone "full information" (including logical information), or of putting them in a position of "really understanding" (assuming, perhaps wrongly, that we can define this in fully non-evaluative terms) is the most compelling candidate. Indeed, when I ask myself whether, for example, IGJC #4 is really good (relative to me), I find myself tempted to ask: "how would I feel about it, if I really understood it?". And the question feels like it has an answer.

One justification for appealing to something like "full information" or "really understanding" is: it enables your idealized self to avoid instrumental mistakes. Consider Alfred, owner of Doggo above. Because Alfred doesn't know Doggo's true nature (e.g., a simple, non-conscious robot), Alfred doesn't know what he's really causing, when he e.g. takes Doggo to the park. He thinks he's causing a conscious puppy to be happy, but he's not. Idealized Alfred knows better. Various other cases sometimes mentioned in support of idealizing — e.g., someone who drinks a glass of petrol, thinking it was gin — can also be given fairly straightforward instrumental readings.

But this justification seems too narrow. In particular: idealizers generally want the idealization process to do more than help you avoid straightforward instrumental mistakes. In cases 1-8 above, for example, Alfred's is basically the only one that fits this instrumental mold straightforwardly. The rest involve something more complex — some dance of "rewinding" psychological processes (see more description [here](#)), rejecting terminal (or putatively terminal) values on the basis of their psychological origins, and resolving internal conflicts by privileging some evaluative attitudes, stances, and intuitions over others. That is, the idealization procedure, standardly imagined, is supposed to do more than take in someone who already has and is pursuing coherent values, and tell them how to get what they want; that part is (theoretically) easy. Rather, it's supposed to take in an actual, messy, internally conflicted human, and *output* coherent values — values that are in some sense "the right answer" relative to the human in question.

Indeed, I sometimes wonder whether the appeal of idealizing subjectivism rests too much on people mistaking its initial presentation for the more familiar procedure of eliminating straightforward instrumental mistakes. In my view, if we're in a theoretical position to just get rid of instrumental mistakes, then we're already cooking with gas, values-wise. But the main game is messier — e.g., using hypothetical selves (which?) to determine what *counts* as an instrumental mistake, relative to you.

There's another, subtly different justification for privileging "full information," though: namely, that once you've got full information, then (assuming anti-realism about values) you've got everything that the world can give you. That is: there's nothing about reality that you're, as it were, "missing" — no sense in which you should hesitate from decision, on the grounds that you might learn something new, or be wrong about some independent truth. The rest, at that point, is up to you.

I'm sympathetic to this sort of thought. But I also have a number of worries about it.

One (fairly minor) is whether it justifies baking full information into the idealization procedure, *regardless* of the person's attitudes towards acquiring such information. Consider someone with very limited interest in the truth, and whose decision-making process, given suitable opportunity, robustly involves actively and intentionally self-modifying to close off inquiry and lock in various self-deceptions/falsehoods. Should we still "force" this person's idealized self to get the whole picture before resolving questions like whether to self-deceive?

A second worry, gestured at above, is that the move from my mundane self to a being with "full information" is actually some kind of wild and alien leap: a move not from Joe to "Joe who has gotten out a bit more, space and time-wise" but from Joe to galaxy Joe, from Joe to a kind of God. And this prompts concern about the validity of the exercise.

Consider its application to a dog, or an ant. What would an ant value, if it had "full information"? What, for that matter, would a rock value, if it had full information? If I were a river, would I flow fast, or slow? If I were an egg, would I be rotten? Starting with a dog, or an ant, or a rock, we can create a galaxy-brained God. Or, with the magic of unmoored counterfactuals, we can "cut straight to" some galaxy-brained God or other, via appeal to some hazy sort of "similarity" to the dog/ant/rock in question, without specifying a process for getting there — just as we can try to pick an egg that I would be, if I were an egg. With dogs, or ants, though, and certainly with rocks, it seems strange to give the resulting galaxy-brain much authority, with respect to what the relevant starting creature/rock "truly values," or should. In deciding whether to euthanize your dog Fido, should you ask the nearest galaxy-brained former-Fido? If not, are humans different? What makes them so?

This isn't really a precise objection; it's more of a hazy sense that if we just ask directly "how would I feel about X, if I were a galaxy brain?", we're on shaky ground. (Remember, we can't specify my values independently, hold them fixed, and then require that the galaxy brain share them; the whole point is that the galaxy brain's attitudes *constitute* my values.)

A third worry is about indeterminacy. Of the many candidate ways of creating a fully informed galaxy Joe, starting with actual me, it seems possible that there will be important path-dependencies (this possibility is acknowledged by many idealizers). If you learn X information, or read Y novel, or have Z experience, before some alternatives (by hypothesis, you do all of it eventually), you will arrive at a very different evaluative endpoint than if the order was reversed. Certainly, much real-life value formation has this contingent character: you meet Suzy, who loves the stoics, is into crypto, and is about to start a medical residency, so you move to Delaware with her, read Seneca, start hanging out with libertarians, and so on. Perhaps such contingency persists in more idealized cases, too. And if we try to skip over process and "cut straight to" a galaxy Joe, we might worry, still, that equally qualified

candidates will value very different things: “full information” just isn’t enough of a constraint.

(More exotically, we might also worry that amongst all the evaluative Eldritch horrors lurking in the memespace, there is one that always takes over all of the Joes on their way to becoming fully-informed galaxy Joes, no matter what they do to try to avoid it, but which is still in some sense “wrong.” Or that full information, more generally, always involves memetic hazards that are fatal from an evaluative perspective. It’s not clear that idealizing subjectivism has the resources to accommodate distinctions between such hazards and the evaluative truth. That said, these hypotheses also seem somewhat anti-Humean in flavor. E.g., can’t fully-informed minds value any old thing?)

Worries about indeterminacy become more pressing once we recognize all the decisions a Galaxy Joe is going to have to make, and all of the internal evaluative conflicts he will have to resolve (between object-level and meta preferences, competing desires, contradictory intuitions, and the like), that access to “full information” doesn’t seem to resolve for him. Indeed, the Humean should’ve been pessimistic about the helpfulness of “full information” in this regard from the start. If, by Humean hypothesis, your current, imperfect knowledge of the world can’t tell you what to want for its own sake, and/or how to resolve conflicts between different intrinsic values, then perfect knowledge won’t help, either: you still face what is basically the same old game, with the same old gap between is and ought, fact and value.

Beyond accessing “full information,” is there a privileged procedure for playing this game, specifiable without reference to the agent’s actual or idealized attitudes? Consider, for example, the idea of “reflective equilibrium” in ethics — the hypothesized, stable end-state of a process of balancing more specific intuitions with more general principles and theoretical considerations. How, exactly, is this balance to be struck? What weight, for example, should be given to theoretical simplicity and elegance, vs. fidelity to intuition and common sense? In contexts with independent standards of accuracy, we might respond to questions like this with reference to the balance most likely to yield the right answer; but for the idealizer, there is not yet a right answer to be sought; rather, the reflective equilibrium process makes its output right. But which reflective equilibrium process?

Perhaps we might answer: whatever reflective equilibrium process actually works in the cases where there *is* a right answer (thanks to Nick Beckstead for discussion). That is, you should import the reasoning standards you *can* actually evaluate for accuracy (for example, the ones that work in e.g. physics, math, statistics, and so on) into a domain (value) with no independent truth. Thus, for example, if simplicity is a virtue in science, because (let’s assume) the truth is often simple, it should be a virtue in ethics, too. But why? Why not do whatever’s accurate in the case where accuracy is a thing, and then something else entirely in the domain where you can’t go wrong, except relative to your own standards?

(We can answer, here, by appeal to your actual or idealized attitudes: e.g., you just do, in fact, use such-and-such standards in the evaluative domain, or would if suitably idealized. I discuss these options in the next sections. For now, the question is whether we can justify particular idealization procedures absent such appeals.)

Or consider the idea that idealization involves or is approximated by “running a large number of copies of yourself, who then talk/argue a lot with each other and with

others, have a bunch of markets, and engage in lots of voting and trading and betting" (see e.g. Luke Muelhauser's description [here](#)), or that it involves some kind of "[moral parliament](#)." What sorts of norms, institutions, and procedures structure this process? How does it actually work? Advocates of these procedures rarely say in any detail (though see [here](#) for one recent discussion); but presumably, one assumes, "the best procedures, markets, voting norms, etc." But is there a privileged "best," specifiable and justifiable without appeal to the agent's actual/idealized attitudes? Perhaps we hope that the optimal procedures are just *there*, shining in their optimality, identifiable without any object-level evaluative commitments (here Hume and others say: what?), or more likely, given *any* such commitments. My guess, though, is that absent substantive, value-laden assumptions about veils of ignorance and the like, and perhaps even given such assumptions, this hope is over-optimistic.

The broader worry, here, is that once we move past "full information," and start specifying the idealization procedure in more detail (e.g., some particular starting state, some particular type of reflective equilibrium, some particular type of parliament), or positing specific traits that the idealized self needs to have (vivid imagination, empathy, dispassion, lack of "bias," etc), our choice of idealization will involve (or sneak in) object-level value judgments that we won't be able to justify as privileged without additional appeal to the agent's (actual or idealized) attitudes. Why vivid imagination, or empathy (to the extent they add anything on top of "full information")? Why a cool hour, instead of a hot one? What counts as an evaluative bias, if there is no independent evaluative truth? The world, the facts, don't answer these questions.

If we can't appeal to the world to identify a privileged idealization procedure, it seems we must look to the agent instead. Let's turn to that option now.

## VII. Appeals to actual attitudes

Suppose we appeal to your actual attitudes about idealization procedures, in fixing the procedure that determines what's of value relative to you. Thus, if we ask: why this particular reflective equilibrium? We answer: because that's the version you in fact use/endorse. Why this type of parliament, these voting norms? They're the ones you in fact favor. Why empathy, or vivid imagination, or a cool hour? Because you like them, prefer them, trust them. And so on.

Indeed, some idealization procedures make very explicit reference to the "idealized you" that you yourself want to be/become. In cases like "vicious desires" above, for example, your wanting not to have a particular desire might make it the case that "idealized you" doesn't have it. Similarly, Yudkowsky's "[coherent extrapolated volition](#)" appeals to the attitudes you would have if you were "more the person you wished you were."

At a glance, this seems an attractive response, and one resonant with a broader subjectivist vibe. However, it also faces a number of problems.

First: just as actual you might be internally conflicted about your object-level values (conflicts we hoped the idealization procedure would resolve), so too might actual you be internally conflicted about the procedural values bearing on the choice of idealization procedure. Perhaps, for example, there isn't currently a single form of reflective equilibrium that you endorse, treat as authoritative, etc; perhaps there isn't a single idealized self that you "wish you were," a single set of desires you "wish you had." Rather, you're torn, at a meta-level, about the idealization procedures you want

to govern you. If so, there is some temptation, on pain of indeterminacy, to look to an idealization procedure to resolve this meta-conflict, too; but what type of idealization procedure to use is precisely what you're conflicted about (compare: telling a group torn about the best voting procedure to "vote on it using the best procedure").

Indeed, it can feel like proponents of this version of the view hope, or assume, that you are in some sense already engaged in, or committed to, a determinate decision-making process of forming/scrutinizing/altering your values, which therefore need only be "run" or "executed." Uncertainty about your values, on this picture, is just logical uncertainty about what the "figure out my values computation" you are already running will output. The plan is in place. Idealization executes.

But is this right? Clearly, most people don't have very explicit plans in this vein. At best, then, such plans must be implicit in their tangle of cognitive algorithms. Of course, it's true that if put in different fully-specified situations, given different reflective resources, and forced to make different choices given different constraints, there is in fact a thing a given person would do. But construing these choices as the implementation of a determinate plan/decision-procedure (as opposed to e.g., noise, mistakes, etc), to be extrapolated into some idealized limit, is, at the least, a very substantive interpretative step, and questions about indeterminacy and path dependence loom large. Perhaps, for example, what sort of moral parliament Bob decides to set up, in different situations, depends on the weather, or on what he had for breakfast, or on which books he read in what order, and so on. And perhaps, if we ask him which such situation he meta-endorses as most representative of his plan for figuring out his values, he'll *again* give different answers, given different weather, breakfasts, books, etc — and so on.

(Perhaps we can just hope that this bottoms out, or converges, or yields patterns/forms of consensus robust enough to interpret and act on; or perhaps, faced with such indeterminacy, we can just say: "meh." I discuss responses in this vein in section IX.)

Second (though maybe minor/surmountable): even if your actual attitudes yield determinate verdicts about the authoritative form of idealization, it seems like we're now giving your procedural/meta evaluative attitudes an unjustified amount of authority relative to your more object-level evaluative attitudes. That is, we're first using your procedural/meta evaluative attitudes to fix an idealization procedure, then judging the rest of your attitudes via reference to that procedure. But why do the procedural/meta attitudes get such a priority?

This sort of issue is most salient in the context of cases like the "vicious desires" one above. E.g., if you have (a) an object-level desire that your co-worker suffer, and (b) a meta-desire not to have that object-level desire, why do we choose an "ideal you" in which the former is extinguished, and the latter triumphant? Both, after all, are just desires. What grants meta-ness such pride of place?

Similarly, suppose that your meta-preferences about idealization give a lot of weight to consistency/coherence — but that consistency/coherence will require rejecting some of your many conflicting object-level desires/intuitions. Why, then, should we treat consistency/coherence as a hard constraint on "ideal you," capable of "eliminating" other values whole hog, as opposed to just one among many other values swirling in the mix?

(Not all idealizers treat consistency/coherence in this way; but my sense is that many do. And I do actually think there's more to say about why consistency/coherence should get pride of place, though I won't try to do so here.)

Third: fixing the idealization procedure via reference to your actual (as opposed to your idealized) evaluative attitudes risks closing off the possibility of making mistakes about the idealization procedure you want to govern you. That is, this route can end up treating your preferences about idealization as "infallible": they fix the procedure that stands in judgment over the rest of your attitudes, but they themselves cannot be judged. No one watches the watchmen.

One might have hoped, though, to be able to evaluate/criticize one's currently preferred idealization procedures, too. And one might've thought the possibility of such criticism truer to our actual patterns of uncertainty and self-scrutiny. Thus: if you currently endorse reflective equilibrium process X, but you learn that it implies an idealized you that gives up currently cherished value Y, you may not simply say: "well, that's the reflective equilibrium process I endorse, so there you have it: begone, Y." Rather, you can question reflective equilibrium process X on the very grounds that it results in giving up cherished value Y — that is, you can engage in kind of meta-reflective equilibrium, in which the authority of a given process of reflective equilibrium is itself subject to scrutiny from the standpoint of the rest of what you care about.

Indeed, if I was setting off on some process of creating my own "moral parliament," or of modifying myself in some way, then even granted access to "full information," I can well imagine worrying that the parliament/self I'm creating is of the wrong form, and that the path I'm on is the wrong one. (This despite the fact that I can accurately forecast its results before going forward — just as I can accurately forecast that, after reading the egoist novel, or entering the experience machine, I'll come out with a certain view on the other end. Such forecasts don't settle the question).

We think of others as making idealization procedure mistakes, too. Note, for example, the tension between appealing to your actual attitudes towards idealization, and the (basically universal?) requirement that the idealized self possess something like full (or at least, much more) information. Certain people, for example, might well endorse idealization processes that lock in certain values and beliefs very early, and that as a result *never reach* any kind of fully informed state: rather, they arrive at a stable, permanently ignorant/deceived equilibrium well before that. Similarly, certain people's preferred idealization procedures might well lead them directly into the maw of some memetic hazard or other ("sure, I'm happy to look at the whirling pixels").

Perhaps we hope to save such people, and ourselves, from such (grim? ideal?) fates. We find ourselves saying: "but you wouldn't want to use that idealization procedure, if you were more idealized!". Let's turn to this kind of thought, now.

## **VIII. Appeals to idealized attitudes**

Faced with these problems with fixing the idealization procedure via reference to our actual evaluative attitudes, suppose we choose instead to appeal to our *idealized* evaluative attitudes. Naive versions of this, though, are clearly and problematically circular. What idealization determines what's of value? Well, the idealization you would decide on, if you were idealized. Idealized how? Idealized in the manner you would want yourself to be idealized, if you were idealized. Idealized how? And so on.

(Compare: “the best voting procedure is the one that would be voted in by the best voting procedure.”)

Of course, some idealization procedures could be self-ratifying, such that if you were idealized in manner X, you would choose/desire/endorse idealization process X. But it seems too easy to satisfy this constraint: if after idealization process X, I end up with values Y, then I can easily end up endorsing idealization process X, since this process implies that pursuing Y is the thing for me to do (and I’m *all about* pursuing Y); and this could hold true for a very wide variety of values resulting from a very wide variety of procedures. So “value is determined by the evaluative attitudes that would result from an idealization procedure that you would choose if you underwent that very procedure” seems likely to yield wildly indeterminate results; and more importantly, its connection with what you actually care about now seems conspicuously tenuous. If I can brainwash you into becoming a paperclip maximizer, I can likely do so in a way that will cause you to treat this very process as one of “idealization” or “seeing the light.” Self-ratification is too cheap.

Is there a middle ground, here, between using actual and idealized attitudes to fix the idealization procedure? Some sort of happy mix? But which mix? Why?

In particular, in trying to find a balance between endless circles of idealization, and “idealized as you want to be, period,” I find that I run into a kind of “problem of arbitrary non-idealization,” pulling me back towards the circle thing. Thus, for example, I find that at every step in the idealization process I’m constructing, it feels possible to construct a further process to “check”/“ratify” that step, to make sure it’s not a mistake. But this further process will itself involve steps, which themselves could be mistakes, and which themselves must therefore be validated by some further process — and so on, ad infinitum. If I stop at some particular point, and say “this particular process just isn’t getting checked. This one is the bedrock,” I have some feeling of: “Why stop here? Couldn’t this one be mistaken, too? What if I wouldn’t want to use this process as bedrock, if I thought more about it?”.

Something similar holds for particular limitations on e.g. the time and other resources available. Suppose you tell me: “What’s valuable, relative to you, is just what you’d want if ten copies of you thought about it for a thousand years, without ever taking a step of reasoning that another ten copies wouldn’t endorse if they thought about *that step* for a thousand years, *and that’s it. Done.*” I feel like: why not a hundred copies? Why not a billion years? Why not more levels of meta-checking? It feels like I’m playing some kind of “name the largest number” game. It feels like I’m building around me an unending army of ethereal Joes, who can never move until all the supervisors arrive to give their underlings the go-ahead, but everyone can never arrive, because there’s always room for more.

Note that the problem here isn’t about processes you might run or compute, in the actual world, given limited resources. Nor is it about finding a process that you’d at least be happy deferring to, over your current self; a process that is at least *better* than salient alternatives. Nor, indeed, is the problem “how can I know with certainty that my reasoning process will lead me to the truth” (there is no independent truth, here). Rather, the problem is that I’m supposed to be specifying a *fully idealized* process, the output of which *constitutes* the evaluative truth; but for every such process, it feels like I can make a better one; any given process seems like it could rest on mistakes that a more exhaustive process would eliminate. Where does it stop?

## **IX. Hoping for convergence, tolerating indeterminacy**

One option, here, is to hope for some sort of convergence in the limit. Perhaps, we might think, there will come a point where no amount of additional cognitive resources, levels of meta-ratification, and so on will alter the conclusion. And perhaps indeed — that would be convenient.

Of course, there would remain the question of what sort of procedure or meta-procedure to “take the limit” of. But perhaps we can pull a similar move there. Perhaps, that is, we can hope that a very wide variety of candidate procedures yield roughly similar conclusions, in the limit.

Indeed, in general, for any of these worries about indeterminacy, there is an available response to the effect that: “maybe it converges, though?” Maybe as soon as you say “what Joe would feel if he really understood,” you hone in on a population of Galaxy Joes that all possess basically the same terminal values, or on a single Galaxy Joe who provides a privileged answer. Maybe Bob’s preferences about idealization procedures are highly stable across a wide variety of initial conditions (weather, breakfasts, books, etc). Maybe it doesn’t really matter how, and in what order, you learn, read, experience, reflect: modulo obvious missteps, you end up in a similar place. Maybe indeed.

Or, if not, maybe it doesn’t matter. In general, lots of things in life, and especially in philosophy, are vague to at least some extent; arguments to the effect that “but how exactly do you define X? what about Y edge case?” are cheap, and often unproductive; and there really are [bald people](#), despite the indeterminacy of exactly who qualifies.

What’s more, even if there is no single, privileged idealized self, picked out by a privileged idealization procedure, and even if the many possible candidates for procedures and outputs do not converge, it seems plausible that there will still be patterns and limited forms of consensus. For example, it seems unlikely that many of my possible idealized selves end up trying to maximize helium, or to eat as many bricks as they can; even if a few go one way, the preponderance may go some other way; and perhaps it’s right to view basically all of them, despite their differences, as worthy of deference from the standpoint of my actual self, in my ignorance (e.g., perhaps the world any of them would create is rightly thought better, from my perspective, than the world I would create, if I wasn’t allowed further reflection).

In this sense, the diverging attitudes of such selves may still be able to play some of the role the idealizer hopes for. That is, pouring my resources into eating bricks, torturing cats, etc really would be a mistake, for me — none of my remotely plausible idealized selves are into it — despite the fact that these selves differ in the weight they give to [incomprehensible galaxy-brained concept] vs. [another incomprehensible galaxy-brained concept]. And while processes that involve averaging between idealized selves, picking randomly amongst them, having them vote/negotiate, putting them behind veils of ignorance, etc raise questions about circularity/continuing indeterminacy, that doesn’t mean that all such processes are on equal footing (e.g., different parties can be unsure what voting procedure to use, while still being confident/unanimous in rejecting the one that causes everyone to lose horribly).

Perhaps, then, the idealizer’s response to indeterminacy — even very large amounts of it — should simply be tolerance. Indeed, there is an art, in philosophy, to not nitpicking too hard — to allowing hand-waves, and something somethings, where appropriate, in the name of actually making progress towards some kind of workable

anything. Perhaps some of the worries above have fallen on the wrong side of the line. Perhaps a vague gesture, a promissory note, in the direction of something vaguely *more* ideal than ourselves is, at least in practical contexts (though this isn't one), good enough; better than nothing; and better, too, than setting evaluative standards relative to our present, decidedly un-ideal selves, in our ignorance and folly.

## X. Passive and active ethics

I want to close by gesturing at a certain kind of distinction — between “passive” and “active” ethics (here I’m drawing terminology and inspiration from a [paper](#) of Ruth Chang’s, though the substance may differ) — which I’ve found helpful in thinking about what to take away from the worries just discussed.

Some idealizing subjectivists seem to hope that their view can serve as a kind of low-cost, naturalism-friendly substitute for a robustly realist meta-ethic. That is, modulo certain extensional differences about e.g. ideally-coherent suffering maximizers, they basically want to talk about value in much the way realists do, and to differ, only, when pressed to explain what makes such talk true or false.

In particular, like realists, idealizers can come to see every (or almost every) choice and evaluative attitude as attempting to approximate and conform to some external standard, relative to which the choice or attitude is to be judged. Granted, the standard in question is defined by the output of the idealization procedure, instead of the robustly real values; but in either case, it’s something one wants to *recognize, receive, perceive, respond to*. For us non-ideal agents, the “true values” are still, effectively, “out there.” We are, in Chang’s terminology, “passive” with respect to them.

But instructively, I think, naive versions of this can end up circular. Consider the toy view that “what’s good is whatever you’d believe to be good if you had full information.” Now suppose that you get this full information, and consider the question: is pleasure good? Well, this just amounts to the question: would I think it good if I had full information? Well, here I am with full information. Ok, do I think it good? Well, it’s good if I would think it good given full information. Ok, so is it good? And so on.

Part of the lesson here is that absent fancier footwork about what evaluative belief amounts to, belief isn’t a good candidate for the evaluative attitude idealization should rest on. But consider a different version: “what you should do is whatever you would do, given full information.” Suppose that here I am with full information. I ask myself: what should I do? Well, whatever I *would* do, given full information. Ok, well, I’ve got that now. What would I do, in precisely this situation? Well, I’m *in* this situation. Ok, what would I do, if things were like this? Well, I’d try to do what I should do. And what should I do? Etc.

The point here isn’t that there’s “no way out,” in these cases: if I can get myself to believe, or to choose, then I will, by hypothesis, have believed truly, chosen rightly. Nor, indeed, need all forms of idealizing subjectivism suffer from this type of problem (we can appeal, for example, to attitudes that plausibly arise more passively and non-agentially, like desire).

Rather, what I’m trying to point at is a way that importing and taking for granted a certain kind of realist-flavored ethical psychology can result in an instructive sort of misfire. Something is missing, in these cases, that I expect the idealizing subjectivist

needs. In particular: these agents, to the end, lack an affordance for a certain kind of direct, active agency — a certain kind of responsibility, and self-creation. They don't know how to choose, fully, for themselves. Rather, even in ideal conditions, they are forever trying to approximate something else. True, on idealizing subjectivism, the thing they are trying to approximate is ultimately, themselves, in those conditions. But this is no relief: still, they are approximating an approximator, of an approximator, and so on, in an endless loop. They are always looking elsewhere, forever down the hall of mirrors, around and around a maze with no center (what's in the [center](#)?). Their ultimate task, they think, is to obey themselves. But they can only obey: they cannot govern, and so have no law.

It's a related sort of misfire, I think, that gives rise to the "would an endless army of ethereal Joes ratify every step of my reasoning, and the reasoning of the ratifiers, and so on?" type of problem I discussed above. That is, one wants every step to conform to some external standard — and the only standards available are built out of armies of ethereal Joes. But those Joes, too, must conform. It's conformity all the way down — except that for the anti-realist, there's no bottom.

What's needed, here, is a type of choice that is creating, rather than trying to conform — and which hence, in a sense, is "infallible." And here perhaps one thinks, with the realists: surely the types of choices we're interested in here — choices about which books, feelings, machines, galaxy brains, Gods, to "trust"; which puppies, or nanomachines, to create — are fallible. Or if not, surely they are, in a sense, arbitrary — mere "pickings," or "plumpings." If you aren't trying to conform to some standard, than how can you truly, and non-arbitrarily, *choose*? I don't have a worked-out story, here (though I expect that we can at least distinguish such creative choices from e.g. [Buridan's-ass](#) style pickings — for example, they don't leave you indifferent). But it's a question that I think subjectivists must face; and which I feel some moderate optimism about answering (though perhaps not in a way that gives realists what they want).

Of course, subjectivists knew, all along, that certain things about themselves were going to end up being treated as effectively infallible, from an evaluative perspective. Whatever goes in Clippy's utility function slot, for subjectivists, governs what's valuable relative to Clippy; and it does so, on subjectivism, just in virtue of being *there* — in virtue of being the stuff that the agent is made out of (this is part of the arbitrariness and contingency that so bothers realists). The problem that the idealizer faces is that actual human agents are not yet fully made: rather, they're still a tangled mess. But the idealizer's hope is that they're sufficiently "on their way to getting made" that we can, effectively, assume they're already there; the seed has already determined a tree, or a sufficiently similar set of trees; we just haven't computed the result.

But is that how trees grow? Have you already determined a self? Have you already made what would make you, if all went well? Do you know, already, how to figure out who you are? Perhaps for some the answer is yes, or close enough. Perhaps for all. In that case, you are already trying to do something, already fighting for something — and it is relative to that something that you can fail.

But if the choice has not yet been made, then it is we who will have to make it. If the sea is open, then so too is it ours to sail.

Indeed, even if in some sense, the choice *has been* made — even if there is already, out there, a privileged idealized version of yourself; even if all of the idealization

procedures converge to a single point — the sea, I think, is *still* open, if you step back and make it so. You can still reject that self, and the authority of the procedure(s) that created it, convergence or no. Here I think of a friend of mine, who expressed some distress at the thought that his idealized self could in principle turn out to be a Voldemort-like character. His distress, to me, seemed to assume that his idealized self was “imposed on him”; that he “had,” as it were, to acknowledge the authority of his Voldemort self’s values. But such a choice is entirely his. He can, if he wishes, reject the Voldemort, and the parts of himself (however strong) that created it; he can forge his own path, towards a new ideal. The fact that he would become a Voldemort, under certain conditions he might’ve thought “ideal,” is ultimately just another fact, to which he himself must choose how to respond.

Perhaps some choices in this vein will be easier, and more continuous/resonant with his counterfactual behavior and his existing decision-making processes; some paths will be harder, and more fragile; some, indeed, are impossible. But these facts are still, I think, just facts; the choice of how to respond to them is open. The point of subjectivism is that the standards (relative to you) used to evaluate your behavior must ultimately be yours; but who you are is not something fixed, to be discovered and acknowledged by investigating what you would do/feel in different scenarios; rather, it is something to be created, and choice is the tool of creation. Your counterfactual self does not bind you.

In a sense, what I’m saying here is that idealizing subjectivism is, and needs to be, less like “realism-lite,” and more like existentialism, than is sometimes acknowledged. If subjectivists wish to forge, from the tangled facts of actual (and hypothetical) selfhood, an ideal, then they will need, I expect, to make many choices that create, rather than conform. And such choices will be required, I expect, not just as a “last step,” once all the “information” is in place, but rather, even in theory, all along the way. Such choice, indeed, is the very substance of the thing.

(To be clear: I don’t feel like I’ve worked this all out. Mostly, I’ve been trying to gesture at, and inhabit, some sort of subjectivist existentialist something, which I currently find more compelling than a more realist-flavored way of trying to be an idealizer. What approach to meta-ethics actually makes most sense overall and in practice is a further question.)

## XI. Ghost civilizations

With this reframing in mind, some of the possible circles and indeterminacies discussed above seem to me less worrying — rather, they are just more facts, to be responded to as I choose. Among all the idealized selves (and non-selves), and all combinations, there is no final, infallible evaluative authority — no rescuer, Lord, father; no safety. But there are candidate advisors galore.

Here’s an illustration of what I mean, in the context of an idealization I sometimes think about.

I’ve written, in the past, about a “[ghost](#)” version of myself — that is, one that can float free from my body; which travel anywhere in all space and time, with unlimited time, energy, and patience; and which can also make changes to different variables, and play forward/rewind different counterfactual timelines (the ghost’s activity somehow doesn’t have any moral significance).

I sometimes treat such a ghost kind of like an idealized self. It can see much that I cannot. It can see directly what a small part of the world I truly am; what my actions truly mean. The lives of others are real and vivid for it, even when hazy and out of mind for me. I trust such a perspective a lot. If the ghost would say “don’t,” I’d be inclined to listen.

As I usually imagine it, though, the ghost isn’t arbitrarily “ideal.” It hasn’t proved all the theorems, or considered all the arguments. It’s not all that much smarter than me; it can’t comprehend anything that I, with my brain, can’t comprehend. It can’t directly self-modify. And it’s alone. It doesn’t talk with others, or make copies of itself. In a sense, this relative mundanity makes me trust it more. It’s easier to imagine than a galaxy brain. I feel like I “know what I’m dealing with.” It’s more “me.”

We can imagine, though, a version of the thought experiment where we give the ghost more leeway. Let’s let it make copies. Let’s give it a separate realm, beyond the world, where it has access to arbitrary technology. Let’s let it interact with whatever actual and possible humans, past and future, that it wants, at arbitrary depths, and even to bring them into the ghost realm. Let’s let it make new people and creatures from scratch. Let’s let it try out self-modifications, and weird explorations of mind-space — surrounded, let’s hope, by some sort of responsible ghost system for handing explorations, new creatures, and so on (here I imagine a crowd of copy ghosts, supervising/supporting/scrutinizing an explorer trying some sort of process or stimulus that could lead to going off the rails). Let’s let it build, if it wants, a galaxy brain, or a parliament, or a civilization. And let’s ask it, after as much of all this as it wants, to report back about what it values.

If I try to make, of this ghost civilization, some of sort of determinate, privileged ideal, which will *define* what’s of value, relative to me, I find that I start to run into the problems discussed above. That is, I start wondering about whether the ghost civilization goes somewhere I actually want; how much different versions of it diverge, based on even very similar starting points; how to fix the details in a manner that has any hope of yielding a determinate output, and how arbitrary doing so feels. I wonder whether the ghosts will find suitable methods of cooperating, containing memetic hazards, and so on; whether I would regret defining my values relative to this hazy thought experiment, if I thought about it more; whether I should instead be focusing on a different, even more idealized thought experiment; where the possible idealizing ends.

But if I let go of the thought that there is, or need be, a single “true standard,” here — a standard that is, already, for me, the be-all-end-all of value — then I feel like I can relate to the ghosts differently, and more productively. I can root for them, as they work together to explore the distant reaches of what can be known and thought. I can admire them, where they are noble, cautious, compassionate, and brave; where they build good institutions and procedures; where they cooperate. I can try, myself, to see through their eyes, looking out on the vastness of space, time, and the beings who inhabit it; zooming in, rewinding, examining, trying to understand. In a sense, I can use the image of them to connect with, and strengthen, what I myself value, now (indeed, I think that much actual usage of “ideal advisor” thought experiments, at least in my own life, is of this flavor).

And if I imagine the ghosts becoming more and more distant, alien, and incomprehensible, I can feel my confidence in their values begin to fray. Early on, I’m strongly inclined to defer to them. Later, I am still rooting for them; but I start to see them as increasingly at the edges of things, stepping forward into the mist; they’re

weaving on a tapestry that I can't see, now; they're sailing, too, on the open sea, further than I can ever go. Are they still good, relative to me? Have they gone "off the rails"? The question itself starts to fade, too, and with it the rails, the possibility of mistake. Perhaps, if necessary, I could answer it; I could decide whether to privilege the values of some particular ghost civilization, however unrecognizable, over my own current feelings and understanding; but answering is increasingly an act of creation, rather than an attempt at discovery.

Certainly, I want to know where the ghost civilization goes. Indeed, I want to know where all the non-Joe civilizations, ghostly or not, go too. I want to know where all of it leads. And I can choose to defer to any of these paths, Joe or non-Joe, to different degrees. I'm surrounded, if I wish to call on them, by innumerable candidate advisors, familiar and alien. But the choice of who, if any of them, to listen to, is mine. Perhaps I would choose, or not, to defer, given various conditions. Perhaps I would regret, or not; would kick myself, or not; would rejoice, or not. I'm interested to know that, too. But these "woulds" are just more candidate advisors. It's still on me, now, in my actual condition, to choose.

*(Thanks to Katja Grace, Ketan Ramakrishnan, Nick Beckstead, Carl Shulman, and Paul Christiano for discussion.)*

# An Intuitive Guide to Garrabrant Induction

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

This post is a high-level summary of the core insights and arguments in [Logical Induction](#), a MIRI paper from 2016. It's intended for people without much mathematical training. Numbers in [brackets] indicate the section of the paper from which I am drawing.

A brief note on naming: Solomonoff exhibited an uncomputable algorithm that does idealized induction, which we call [Solomonoff induction](#). Garrabrant exhibited a computable algorithm that does logical induction, which we have named Garrabrant induction.

Thanks to Mauricio Baker for helpful comments. My editor is Justis Mills. Graphics are done by Sabrina Chwalek.

## Introduction [1]

Suppose I run a computer program. What does it output? You don't know the code, so it could do basically anything. You're missing key information to resolve the question. However, even if you did know the source code, you might still be ignorant about what it would *do*. You have all the necessary information per se, and a perfect reasoner could solve it instantly, but it might take an unrealistic amount of effort for *you* to interpret it correctly.

The former kind of uncertainty is *empirical*. You have to look at the world and make observations about the source code of the program, how my computer interprets the code, etc. Other examples of empirical uncertainty: not knowing what the weather is, not knowing what time it is, not knowing the name of your friend, etc.

The latter kind of uncertainty is *logical*. Even after you've looked at the program and seen the source code, you still might not know what the source code will output. For instance, suppose you saw that the program printed the 173,498th digit of pi. You know what the program will do, but you don't know the results of that process. Other examples of logical uncertainty: not knowing if 19483 is prime, not knowing whether  $1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1$  is even, not knowing if 1/1/2000 was a Monday, etc. The bottleneck in these cases isn't missing data, but rather missing *computation* - you haven't yet exerted the required energy to figure it out, and it might not always be worth it with the tools at your disposal.

Let us call the process of "properly" managing logical uncertainty *logical induction* and reasoners that employ logical induction *logical inductors*.

## Bayesian Insufficiency

Naively, one might assume that Bayesian reasoning, a general method for handling empirical uncertainty, might extend itself naturally to logical uncertainty. However, this is not the case. Imagine that I have two boxes. Suppose that you know I'm either going to place either one blue ball into each or one red ball into each. Your beliefs about what color ball is in each of the boxes are now linked; if you see a blue ball in one of the boxes, you know that the other box contains a blue ball.

Now imagine that I give one of the boxes to my friend Alice and the other box to my friend Bob. You know that Alice really likes matching; if she gets a blue ball, she'll wear blue clothes, if she gets a red ball, she'll wear red clothes. You also know that Bob really likes traveling; if he gets a blue ball, he'll go to the ocean, if he gets a red ball, he'll go to the desert. Since your beliefs about the color of balls Alice and Bob received are linked, your beliefs about where Bob travels and what color Alice wears are also linked. If you see Alice wearing blue, it's more likely she got a blue ball than a red ball, which means Bob also probably got a blue ball, which means Bob went to the ocean. Suppose that Bob has friends Carol and Dave. Carol likes the ocean, so Bob goes to the ocean with Carol, and Dave likes the desert, so Bob goes to the desert with Dave. Now your beliefs about what Alice is wearing are linked to your beliefs about the locations of Carol and Dave.

Proper Bayesian reasoners immediately realize all of these connections. When they see Alice wearing red, a Bayesian reasoner realizes that this implies that Bob also got a red ball, which implies that Bob likely went to the desert, which implies that Dave likely went with him to the desert.

The fundamental problem is that the Bayesian framework assumes logical omniscience: that we can always perform any computation or feat of logic at no cost. A Bayesian reasoner maintains a list of all possible hypotheses, then updates all of them according to Bayes' rule upon encountering new evidence. Knowing how a hypothesis interacts with a piece of definite evidence requires logical deduction, but bounded reasoners have finite resources with which to deduce.

If we want to work in a bounded setting, Bayesian reasoning fails for complicated updates. Can we construct an analog of Bayesian reasoning for a bounded reasoner?

## The Logical Induction Criterion [3]

A reasoner is a logical inductor if it satisfies the *logical induction criterion*, which states that it cannot be exploited by an efficiently computable trading algorithm. We will motivate this criterion and unpack what this means.

### Dutch Books

A [Dutch book](#) is a series of bets that will yield a sure loss (or sure gain on the other side). For example, if I put my \$2 against your \$1 that a coin will be heads, and my \$2 against your \$1 that a coin will be tails, then no matter the outcome of the coin flip, I will be down \$1. A reasoner is *Dutch-bookable* if there exists a Dutch book where all of its bets look fair from that reasoner's perspective. A [theorem of De Finetti](#) shows that, under certain assumptions, a reasoner is not Dutch-bookable if and only if its beliefs satisfy the axioms of probability theory, which makes it a Bayesian reasoner.

As a simple example, suppose I violate the axiom that  $P(\text{Heads}) + P(\text{Not Heads}) = 1$  by having  $P(\text{Not Heads}) = P(\text{Heads}) = \frac{1}{3}$ . Given my stated probabilities, I think a 2:1 bet that the coin is Heads is fair and a 2:1 bet that the coin is Not Heads is fair; this combination of bets that is guaranteed to lose me \$1, making me Dutch-bookable. Generalizing this point, it is possible to construct a Dutch book anytime the probability I assign to a set of mutually exclusive and exhaustive events does not sum to exactly 1.

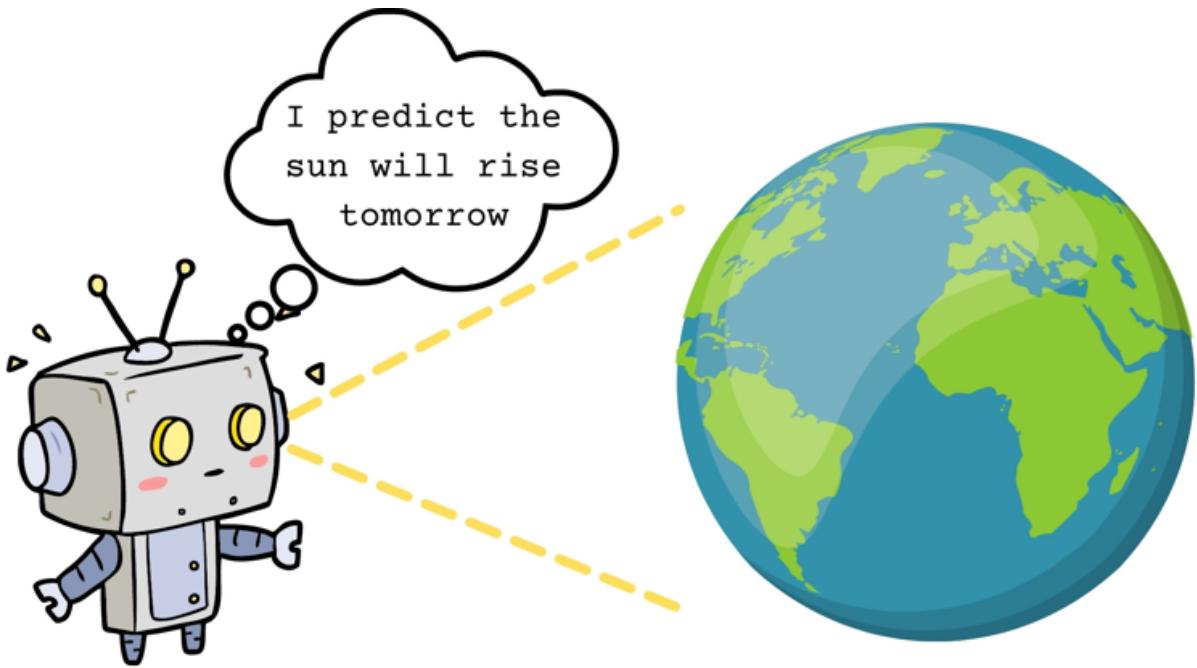
Being universally not Dutch-bookable is a very difficult requirement to satisfy for a bounded reasoner. The reasoner must consider all sequences of bets it would accept and ensure that none of those lead to a sure loss, including sequences that contain tricky logical implications between various beliefs. However, this stringent setup supposes that all Dutch books are equally easy to find.

With bounded computational power, it is impossible to construct an ideal Bayesian reasoner. However, it is possible for a reasoner to *approximately* satisfy the [probability theory axioms](#), making these reasoners arbitrarily difficult to Dutch book. Intuitively, we might want to relax the binary condition of “Dutch-bookable” into a more quantitative measure: in practice, how difficult is it to Dutch book our reasoner? Reasoners that are difficult to Dutch book might be “more Bayesian” than reasoners that are easy to Dutch book.

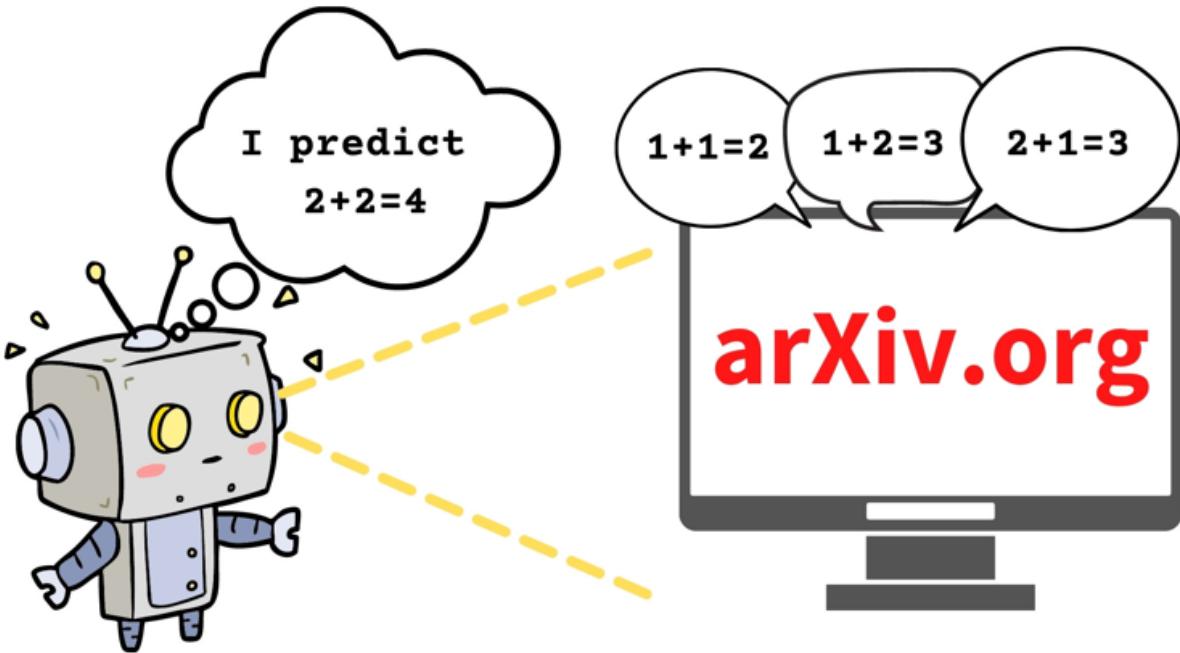
In essence, this is the solution proposed by Garrabrant et al.: a reasoner is a logical inductor if a Dutch book cannot be *efficiently computed*. Here, “efficiently computable” means “computable in [polynomial time](#).<sup>[1]</sup> What the Dutch book must be efficiently computable with respect to will be fleshed out in later sections.

## Deductive Processes [3.2]

An empirical inductor that is observing the world eventually learns to better predict its observations. Abstractly, the empirical inductor is being fed a stream of empirical evidence, which we will call an *observation process*, and learns to predict it. For example, an inductor that observed the sun rising every day so far would learn to predict that the sun would rise the next day. We desire that empirical inductors outpace their underlying observation process by predicting what it will show, giving high probabilities to evidence it observes, and low probabilities to evidence it does not observe. Examples of observation processes are a video feed of the world, various high-level features of a video game, and an 18th-century romance novel. Solomonoff induction is an example of an ideal empirical inductor.



A logical inductor that is observing logical statements eventually learns to predict other statements. Abstractly, the logical inductor is being fed a stream of logical evidence, which we will call a *deductive process*, and learns to predict it. For example, an inductor that observed that all non-two primes were odd would learn to predict that the next prime they discovered will also be odd. We desire that logical inductors outpace the deductive process by predicting what they will prove, giving high probabilities to things that get proved true, and low probabilities to things that get proved false. Examples of deductive processes are peer-reviewed publications, brute-force generated proofs, and the consensus of a mathematical community. We will construct an ideal logical inductor in a later section.



As the above hints at, induction is performed relative to some evidence stream. At best, one can hope for an inductor that can quickly learn a wide variety of evidence streams; [no free lunch](#) theorems show that it is impossible to construct an inductor that can learn *all* evidence streams. Therefore, reasoners can only be empirical inductors relative to some observation process and can only be logical inductors relative to some deductive process.

Notice also that all observation processes yield discrete observations – a single frame of video, a vector of floats, or a single word. Similarly, deductive processes must yield discrete sets of true statements. For convenience, we will imagine that each day, our deductive process yields everything that has been shown to be true/false up until that day.

The central example of a deductive process we will use is Peano Prover, a computer program that yields theorems of Peano Arithmetic (PA) by giving all theorems provable in one character on day one, all theorems provable in two characters on day two, all theorems provable in three characters on day three, etc. This deductive process is *complete* – it will eventually prove all provable theorems of PA. This logical inductor can thus be thought of as someone who on day  $n$  knows all PA-theorems provable in  $\leq n$  characters and must manage their uncertainty about theorems that have not yet been proved.

## Markets [3.1]

Normal markets involve trading assets like food, stocks, and oil. Under some assumptions, the price of each asset should be approximately equal to its intrinsic value summed over time with the appropriate discount rate. If the price was lower than this, people could make money buying the asset. If the price was higher, people could make money selling. The price will thus tend towards its “true value” in some sense.

Prediction markets involve trading assets whose values are artificially tied to future events. For example, I might create an asset that is artificially worth \$1 if the sky is blue on January 1st 2025 and \$0 otherwise. Suppose I think that this event will happen with 80% chance. To me, the value of this asset is  $0.8 * \$1 = \$0.8$ , so I will buy the asset if it's worth less than \$0.8 and sell if it's worth more. If there are many people trading this asset, the price will roughly be equal to the weighted average of the subjective probabilities of all the individual traders.

A simple conception of markets has traders all trading with each other. If I want to buy an asset for some price, I have to find another trader willing to sell at that price. This form of pairwise trading works at small scales but is very inefficient as the number of traders grows. At large scales, markets often contain *market makers*, individual traders with lots of money that are willing to both buy and sell many different assets. These market makers will buy assets at a price lower than they're willing to sell them at, pocketing the difference as profit. For example, I might be willing to buy an ounce of gold for \$1000 and sell for \$1010. Assuming that approximately as many people buy from me as sell to me, I will always make a slight profit.

We will force logical inductors to be market makers for prediction assets over all mathematical statements, with the value of the prediction assets being determined by a complete deductive process, as described above. If our deductive process says a statement is true, the logical inductor is obligated to buy that statement at \$1. If our deductive process says a statement is false, the logical inductor is obligated to sell that statement at \$0. In contrast to market makers in the real world, we will treat the probabilities our logical inductor places on various statements as the price they're willing to both buy and sell at.

In this framework, a Dutch book is not a one-time sequence of traders that yields guaranteed profit for the trader, but rather a *trading strategy* that consistently generates such sequences, yielding the trader unbounded returns over time. A logical inductor can be thought of as a market maker for which a trading strategy that efficiently generates Dutch books cannot be found.

## Traders [3.4]

In real markets, traders look at the world and the history of an asset, then do some thinking and decide how much to buy or sell. If, however, we are making a market over logical sentences as decided by some fixed deductive process, the “world” is simply the history of market prices. As such, traders look at the history of prices assigned to some sentence, do some thinking, then decide how much to buy or sell.

More precisely, a trading strategy for day  $n$  is a particularly nice (including continuous) function from the market's price history up to day  $n$  to the set of all finite buy or sell orders.<sup>[2]</sup> For example, a trading strategy might say to buy shares 3 of “ $1 + 1 = 2$ ” on odd numbered days, or to buy 6 of the cheaper of “ $1 + 2 = 3$ ” and “ $2 + 1 = 3$ ”. A trader is a function that takes in some integer  $n$  and gives a trading strategy for day  $n$ . Trading strategies can be conditional, subject to the niceness constraint.

We call a trader *efficiently computable* if the runtime of the trader is polynomial in  $n$ . That is, each day our traders have access to the market history and get to do more

computation to analyze data and find patterns, but the amount of computation they get must be polynomial in the length of the market history.

We can think of our reasoner as trading against all efficiently computable traders. Each day, the reasoner assigns probabilities to logical sentences and the traders trade against the logical inductor. From the reasoner's perspective, it only makes fair trades. Clever trading strategies will buy (sell) sentences that are shown by the deductive process to be true (false).

Since the reasoner is bounded, on any given day it will be Dutch-bookable. We desire our reasoner to eventually correct all Dutch books that efficiently computable trading algorithms can find against it. In particular, we desire our reasoner does not expose itself to *unbounded* expected losses from any given efficiently computable trading algorithm.

## Exploitation [3.5]

Logically omniscient Bayesian reasoners do not take sequences of bets that result in a sure loss. In other words, they do not take sequences of bets that lose them money in all worlds consistent with the (empirical) evidence. Since logically omniscient Bayesian reasoners would take arbitrarily scaled-up versions of bets, this condition is equivalent to the condition that Bayesian reasoners do not expose themselves to guaranteed unbounded loss.

We seek to extend this notion to computationally bounded reasoners. We will do this by first developing a notion of what it means for a world to be consistent with a set of logical evidence, then relaxing “guaranteed unbounded loss” to “expected unbounded loss.” We will call a reasoner *exploitable* if it’s vulnerable to unbounded loss, and *unexploitable* otherwise (all relative to some fixed deductive process).

Bayesian reasoners induct over observation processes by maintaining a set of all empirical universes and eliminating those that contradict the empirical evidence. Naively, our logical inductor will induct over deductive processes by maintaining a set of all logical universes and eliminating those that are inconsistent with logical evidence.

This formulation has issues. Suppose that on day 1, the deductive process gives you the axioms of Peano arithmetic. This immediately rules out all logical universes that are inconsistent with the PA axioms. More specifically, it rules out all worlds that contain one of EVEN := “the 173,498th digit of pi is even” and ODD := “the 173,498th digit of pi is odd.” The problem is that determining whether a given logical universe is consistent requires determining how every logical statement connects to every other, which is computationally infeasible.

Instead, we relax the notion of “inconsistent with logical evidence” from logical consistency to *propositional consistency*. Instead of calling a logical universe inconsistent when it asserts two statements whose *contents* contradict, e.g. EVEN and ODD, we call a universe inconsistent when it asserts two statements whose *form* contradicts, e.g., EVEN and not EVEN. To know the former is inconsistent requires knowing “EVEN implies not ODD” while knowing the latter is inconsistent only requires knowing basic logical facts. Some propositionally consistent worlds will be logically inconsistent, but this is fine. Importantly, since propositional consistency is a syntactic

property, our reasoner can use its bounded computation to determine whether or not a given logical world is propositionally consistent with the output of a deductive process.

More precisely, we will call a world propositionally consistent with a set of statements if the true assignments in that world agree with logical connectives, i.e., A is true if and only if not A is false, A and B is true if and only if A is true and B is true, etc. In this framework, we can imagine starting with the set of all worlds that are internally propositionally equivalent, then eliminating worlds each successive day as the deductive process yields more logical truths. A trader *exploits* the market (relative to a deductive process) if there is a sequence of worlds, one for each day, that are propositionally consistent with the deductive process in which the trader makes unboundedly large sums of money (without there also being a sequence in which the trader loses an unbounded amount of money). For example, if, every day, the trader purchases a sentence for \$0.50 that will be output by the deductive process 10 days later, the trader effectively earns \$0.50 every day, which is unbounded in the limit, while only staking a maximum of \$5.

The logical induction criterion says that there does not exist such a trader that can exploit the market our reasoner makes over logical sentences. More precisely, a market is said to satisfy the logical induction criterion relative to a deductive process if there is no efficiently computable trader that exploits the market relative to the deductive process. Such a market is said to be a logical inductor over the deductive process.

**The main result of Garrabrant et al. is an algorithm that can price logical sentences in a way that satisfies the logical induction criterion over any deductive process.**

## Selected Properties of Logical Inductors [4]

The efficient market hypothesis states that stock prices reflect all available information. If true, this hypothesis implies that consistently making money off the stock market is impossible; large hedge funds have already exploited any statistical irregularities in stock prices and it is extremely difficult to make money without insider knowledge. For instance, stock prices are higher on Thursdays, this will be discovered by large hedge funds, who will buy on Wednesday and sell on Thursday until stock prices are no longer higher on Thursdays.

If a market satisfies the logical induction criterion (relative to some deductive process), all polynomial-time strategies that make money consistently will eventually cease to work. In particular, if a logical inductor ever fails to learn a polynomial-time method of identifying patterns in logic, a trader can exploit this pattern to make unbounded profits. For instance, if every 139th digit of  $\pi$  turns out to be 7, a trader could buy sentences that stated the 139th, 278th, 417th, etc. digits were 7. If it were always possible for the trader to buy these sentences at some fixed amount less than \$1, the trader would make unbounded profits. Thus, the logical inductor must price all these sentences at  $\sim \$1$  in the limit. In general, logical inductors eventually learn all polynomial-time methods for identifying patterns in logic.

We will go over some of the properties that logical inductors possess. For all of these properties, we will argue that logical inductors possess them by exhibiting trading strategies that would generate unbounded profits if the property didn't hold. Some of

these trading strategies will not be continuous, but can be relaxed into continuous forms.

For the following, we will use Peano Prover as our deductive process.

The fact that many desirable properties follow from the logical induction criteria suggests that this criterion captures a portion of what it means to do good reasoning.

## Convergence and Coherence [4.1]

In the limit, the prices of a logical inductor describe a belief state which is fully logically consistent and represents a probability distribution over all consistent worlds.

Imagine that there exists a sentence such that the prices a logical inductor assigns to that sentence do not converge. If the prices do not converge, then they must [oscillate](#) infinitely around some point. A trader could exploit the logical inductor by buying the sentence at a high point on the oscillation and selling at a low one. Since the oscillation never ceases, the trader can earn unbounded profits. Therefore, the prices a logical inductor assigns to any given sentence must converge.

Imagine that the limiting probabilities assigned to sentences by our logical inductor are not coherent. By a [theorem of Gaifman](#), coherence is equivalent to assigning probability 1 to provable sentences, probability 0 to disprovable sentences, and a third condition we won't cover. Suppose that the limiting probability our logical inductor assigns to a provable sentence is  $\varepsilon$  less than 1. Since the sentence is provable, our deductive process will eventually output the sentence, obligating our logical inductor to buy it for \$1. Thus, a trader who buys one share of the sentence for  $$1 - \varepsilon$  and sells it for \$1 will make  $\$ \varepsilon$  profit, which can be done daily to achieve unbounded profits in the limit. By assumption, our logical inductor is unexploitable, so it must assign probability 1 to provable sentences. A similar argument shows that our logical inductor must assign probability 0 to false sentences.

These properties justify interpreting logical inductor market prices as probabilities. However, there are other ways to assign probabilities to logical sentences. The main desirable property of logical inductors is that their beliefs become approximately consistent quickly, outpacing the underlying deductive process.

## Timely Learning [4.2]

Reasoners that mimic the underlying deductive process will also assign probability 1 to provable statements. The real trick is to assign probabilities in a manner that outpaces the underlying deductive process.

Call a sequence of theorems *efficiently computable* if there exists a polynomial-time program that can enumerate the theorems. Imagine an efficiently computable sequence of theorems, each of which is progressively harder to prove. More specifically, the first statement takes 1 character, the second takes 10 characters, the third takes 100 characters, the fourth takes 1000 characters, etc. Since Peano Prover's speed is tied to the number of characters required for proof, one might imagine that we will have to wait until around the  $10^n$ th day until the probabilities our logical inductor

assigns to the  $n$ th theorem approach 1. However, our logical inductor will actually learn much faster, eventually converging to assigning probabilities close to 1 to the  $n$ th theorem on the  $n$ th day.

To see this, suppose that on day  $n$ , our logical inductor always assigns probability less than  $1 - \varepsilon$  to the  $n$ th theorem in our sequence, a feasible trading strategy because our sequence of theorems is efficiently computable. A trader that buys theorem  $n$  on day  $n$  will always make  $\varepsilon$  profit. If  $\varepsilon$  is bounded away from 0, then this trader will exploit the market. Thus  $\varepsilon$  must go to zero and our logical inductor will converge to assigning probability  $\sim 1$  to the  $n$ th theorem by the  $n$ th day.

In general, for any efficiently computable sequence of sentences, we shall say our logical inductor assigns some sequence of probabilities to this sequence of sentences in a *timely manner* if it converges to assigning the  $n$ th probability to  $n$ th sentence by day  $n$ . The result above shows that our logical inductor assigns the sequence of all 1s to any efficiently computable sequence of theorems in a timely manner.

Considering the probability assigned to the  $n$ th sentence on the  $n$ th day is arbitrary; we might hope that a good reasoner would learn faster than that. Luckily, our logical inductor does! If a sequence of sentences is efficiently computable, then we can break it up into two efficiently computable sequences, one enumerating the even sentences and the other enumerating the odd sentences. Since our logical inductor learns *any* efficiently computable sequence in a timely manner, it will learn both of these faster subsequences, implying that it will eventually learn to assign “accurate” probabilities to the  $2n$ th sentence of the original sequence on the  $n$ th day. Similar constructions show that our logical inductor will eventually learn to assign “accurate” probabilities to the  $P(n)$ th sentence of the original sequence on the  $n$ th day, where  $P$  is any polynomial.

## Learning Statistical Patterns [4.4]

If a Bayesian reasoner encounters a truly random sequence, it will eventually learn to assign probabilities that correspond to high-level statistical patterns.<sup>[3]</sup> For example, if a Bayesian reasoner encounters a truly random fair coin, it will eventually learn to predict heads with  $\frac{1}{2}$  chance and tails with  $\frac{1}{2}$  chance.

We will show that logical inductors also learn appropriate statistical summaries of “random” processes. For example, a good reasoner thinking about the  $10^{124987692}$ th digit of  $\pi$  should assign  $\sim 10\%$  chance to it being a 7, assuming no efficient method of predicting that digit exists. However, for computationally unbounded reasoners, there is no such thing as a “random” logical sentence, as the  $10^{124987692}$ th digit of  $\pi$  is either 7 or not 7. We thus define randomness relative to the amount of computation our

logical inductor has. More specifically, we define a sequence to be *pseudorandom* (relative to our logical inductor) if there is no efficiently computable way to pick out a subsequence that is true with higher frequency than the sequence as a whole.

Suppose that we have an efficiently computable sequence that is pseudorandom with probability  $p$ . Our logical inductor will learn to assign  $p$  to this sequence in a timely manner.

To see this, suppose that our logical inductor infinitely often assigned probability at least some fixed  $\epsilon$  away from  $p$  to the  $n$ th sentence of this pseudorandom sequence on the  $n$ th day. A trader could buy that sentence if the probability was less than  $p$  and sell the sentence if the probability was greater than  $p$ , making  $\epsilon$  profit in expectation. Such a trader would make unbounded profits, exploiting our logical inductor. Thus, our logical inductor eventually converges to assigning probability  $p$  to the  $n$ th sentence of the pseudorandom sequence on the  $n$ th day.

## Learning Logical Relationships [4.5]

Convergence and coherence properties of logical inductors guarantee that the limiting probabilities assigned to sentences accurately reflect logical relationships. For instance, if our deductive process proves that “ $A \text{ xor } B$ ” is true ( $A$  or  $B$  is true, but not both), then the limiting probabilities will satisfy  $P(A) + P(B) = 1$ .

This limiting behavior ensures our logical inductor will learn to respect logical relationships amongst sentences eventually. Our logical inductor will do better; it will learn to respect logical relationships in a timely manner. For example, recalling that  $\text{EVEN} := \text{"the 173,498th digit of } \pi \text{ is even"}$  and  $\text{ODD} := \text{"the 173,498th digit of } \pi \text{ is odd"}$ , our logical inductor quickly learns to assign probabilities for  $\text{EVEN}$  and  $\text{ODD}$  that sum to one.

Suppose that we have an efficiently computable sequence of pairs of mutually exclusive and exhaustive sentences (our deductive process proves that exactly one sentence in each pair is true). Our logical inductor will learn to assign probabilities that sum to 1 to the  $n$ th pair by the  $n$ th day.

To see this, suppose that the sum of probabilities assigned to the  $n$ th pair is some fixed  $\epsilon$  greater or less than 1 on the  $n$ th day. A trader can buy one of each sentence if the probabilities are higher and sell one of each sentence if the probabilities are lower. This trader makes  $\epsilon$  profit off of this sequence of trades. If the trader can make this trade infinitely many times, our logical inductor is exploitable. Thus, the sum of probabilities assigned to the  $n$ th pair on the  $n$ th day converges to 1.

## Non-Dogmatism [4.6]

[Cromwell's rule](#) says that a reasoner ought not assign probabilities 0 or 1 except to statements that are logically true or false. One justification for this rule is that a Bayesian reasoner can never update away from such extreme probabilities, so can thus never change their minds if those probabilities were assigned in error. Logical inductors can update away from probabilities 0 and 1, so the argument for Cromwell's rule does not extend to logical inductors.

In fact, logical inductors do not satisfy Cromwell's rule. To see this, note that the logical induction criterion requires that the logical inductor does not lose unbounded amounts of money. However, it does not prohibit the logical inductor from losing any *finite* amount of money. Since traders can only buy/sell a finite amount each day, a logical inductor can be arbitrarily wrong on an arbitrary, but finite, number of days and still only lose a finite amount of money. More specifically, a logical inductor will stay a logical inductor under arbitrary perturbation of its probabilities on any finite number of days. For example, if we forced a logical inductor to believe all statements with probability 1 on day 104, it would remain a logical inductor.

However, there is a sense in which logical inductors are non-dogmatic. If a theorem is not proven to be true, i.e., it is never output by our deductive process, then our logical inductor will assign it probability strictly less than 1. If a theorem is not proven to be false, i.e., its negation is never output by our deductive process, then our logical inductor will assign it probability strictly greater than 0.

To see this, suppose that our logical inductor assigns limiting probability 0 to a statement that cannot be proven to be false. Consider a trader that watches this statement and buys when the price is \$0.50, \$0.25, \$0.125, \$0.0625, ..., purchasing an infinite number of copies with only \$1. Since the statement cannot be proven to be false, there will always exist a world propositionally consistent with the deductive process in which it is true. In that world, the trader makes \$1 + \$1 + \$1 + ..., exploiting our logical inductor. Thus our logical inductor must assign a limiting probability bounded away from 0. A similar argument shows that a logical inductor that assigns limiting probability 1 to a statement that cannot be proven true can also be exploited.

## Construction [5]

The primary contribution of Garrabrant et al. is a computable algorithm that assigns probabilities to logical statements that satisfies the logical induction criterion. We call this algorithm Garrabrant induction. We will sketch its construction.

To begin, we will show that for any given trader, Garrabrant induction can set prices such that this trader makes as little money as it wants. This result will be achieved by setting prices such that the trader does not want to buy or sell except for buying at  $\sim \$1$  and selling at  $\sim \$0$ . In our setup, we require our trader's "how much do I want to buy?" function to be continuous, interpreting buying a negative amount as selling. Since there exists some price our trader wants to buy at and some price our trader wants to sell at, this function must cross zero at some point. Thus there must be some price where the trader is not interested in buying or selling. The general form of this argument is known as [Brower's fixed point theorem](#) and we use it to guarantee such a set of prices can be found across all the assets the trader is interested in trading.<sup>[4]</sup>

Next, we will construct a *trading firm*, a single trader that exploits Garrabrant induction if and only if there exists any trader that exploits Garrabrant induction. First, fix some

enumeration of all possible trading strategies that might lose at most \$1. Note that since we can scale trading strategies up or down multiplicatively, any trading strategy with bounded loss can be easily converted to a strategy that loses at most \$1.<sup>[5]</sup> Our trading firm starts with only the first trader in this enumeration. On each day, the firm will hire the next trader, adding their strategy to the firm's strategy linearly. However, since the firm had finite capital, it will give each successive trader half the amount of money it gave the previous trader.<sup>[6]</sup> For example, on the third day, the trading firm's strategy is 1/2 of the 1st trader, 1/4th of the 2nd, and 1/8th of the 3rd. In general, the trading strategy on the nth day will be an exponentially decreasing weighted sum of the trading strategies of the first n traders.

Suppose that Garrabrant induction was exploitable, i.e., there existed some efficiently computable trading strategy that could potentially obtain unbounded profits with bounded loss. By construction, this trader will be the nth trader in our enumeration of bounded loss traders we used to construct our trading firm. Assume that all other trading strategies in the enumeration all lose \$1, the worst they could possibly do. Since the trading firm is an exponentially decreasing sum of strategies, the firm as a whole will lose at most \$1. However, its potential profit is at least  $\frac{1}{2^n}$  times the potential profit of the nth trader. Since an unbounded amount divided by a constant is still unbounded, our trading firm is able to exploit Garrabrant induction.

Taking the contrapositive of the above result, if the trading firm *does not* exploit Garrabrant induction, then there does not exist *any* efficiently computable trader that exploits Garrabrant induction, i.e., Garrabrant induction is unexploitable.

We then pit Garrabrant induction against our trading firm. More specifically, we define probabilities for our logical inductor so that on the nth day, the trading firm's strategy makes at most  $\frac{1}{2^n}$ .<sup>[7]</sup> It is easy to see that these probabilities make the logical inductor unexploitable. If it were exploitable, there would be a trader that could exploit it, which means that the trading firm would exploit it. However, the trading firm makes at most  $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$  dollar, so this is impossible.

This algorithm is computable. The market prices at which a trader can make at most a small amount of money can be found by brute force search over all rational numbers enumerated by the [Farey sequence](#). Since trading strategies must be computable, they can only care about prices up to a certain precision, guaranteeing that this search is computable. The trading firm's strategy each day is a finite sum from a computable sequence and is thus computable.

## Conclusion [7]

We have presented the *logical induction criterion* as a relaxed form of Dutch book justifications for Bayesianism and have argued that reasoners that satisfy this criterion (*logical inductors*) possess many desirable properties. We have also sketched the construction of Garrabrant induction, a computable logical induction algorithm. We now expand on interesting takeaways from the logical induction framework. We conclude with a discussion of open questions.

## Ensemble methods [7.2]

The logical induction algorithm developed above is general. If we run it over a deductive system that outputs sentences of Peano arithmetic, the algorithm will develop accurate beliefs about Peano arithmetic. If we instead run it over a deductive system that outputs sentences of set theory, the algorithm will develop accurate beliefs about set theory. Our logical induction algorithm is thus similar to [Solomonoff induction](#) – it is able to induct over a wide range of deductive processes, much as Solomonoff's algorithm can induct over a wide range of observation processes.

The similarities run deeper. Solomonoff induction works by maintaining a set of experts (all possible computer programs) and discarding those that fail to accurately predict the observation process. What Solomonoff induction predicts is a weighted average of what each of the remaining experts predicts. Garrabrant induction works by maintaining a set of experts (all efficiently computable traders) and discarding those that lose sufficiently large amounts of money (roughly). What Garrabrant induction predicts is a weighted average of what each remaining expert predicts.

Additionally, Garrabrant induction is also not meant to be used in practice, although it is at least computable. The hope is that the methods employed by Garrabrant induction can be relaxed and approximated to yield algorithms that are practically useful for logical prediction tasks.

## Small experts [7.2]

A key difference between Garrabrant and Solomonoff induction is in what the methods expect from their “experts.” Solomonoff induction expects each expert to model the entire world and predict everything from the motion of the moon to the arc of a single apple falling from a tree. The “master algorithm” then rewards experts for accuracy and penalizes them for complexity. Importantly, the “master algorithm” in Solomonoff induction is not making predictions, merely aggregating expert opinion.

In contrast, the experts in Garrabrant induction can specialize by choosing particular classes of sentences to trade and not touching anything else. The “master algorithm” makes predictions, which are then corrected by the specialized experts. For example, one expert might specialize in making sure the probabilities of sentences and their negations sum to at least one. Experts are (extremely roughly) rewarded in proportion to how much money they make and penalized in proportion to how much money they lose.

In Solomonoff induction, we imagine starting with the set of all possible worlds and pruning those that do not match the evidence. In Garrabrant induction, we imagine starting with naive guesses, then consulting a series of experts, each of which contributes a small piece of logical knowledge – each trader says “look, I don't know what you're trying to predict, but I do know that this bit doesn't make sense.” By aggregating all these pieces of knowledge, Garrabrant induction builds a model that satisfies many different relationships, even if any individual expert is only tracking a simple pattern.

## Variations [7.3]

Solomonoff induction uses all computable worlds as its experts; however, the underlying logic (Bayesian updating) is more general than that. Instead of using all computable worlds, we can instead use all polynomials, decision trees, or functions representable by a one billion parameter neural network. Of course, these reduced forms of Solomonoff induction would not work as well, but the method of induction would remain unchanged.

Similarly, Garrabrant induction employs polynomial-time traders as its experts; however, the underlying logic of trading and markets is more general than that. Instead of using polynomial time traders, we can instead use linear-time traders, constant time traders, or traders representable by a one billion parameter neural network. Of course, these reduced forms of Garrabrant induction would not work as well, but the method of induction would remain unchanged.

In fact, Garrabrant induction is not even specific to the domain of logic. Imagine a deductive process that, on the  $n$ th day outputted the first  $n$  frames of the image of some video camera: “the  $(0, 0)$  pixel is red”, “the  $(0, 1)$  pixel is blue”, “the third column is green”, etc. Garrabrant induction would eventually learn to predict patterns in the camera feed.

## Open Questions [7.4]

While Garrabrant induction has a large number of desirable properties, there are some that escape.

### Decision Rationality

When humans do reasoning, they are often able to devote thinking time to specific questions. For example, I might be curious about the weather tomorrow in New York and perform reasoning with that specific goal, increasing the accuracy of that prediction faster than other predictions. We might desire that a logical inductor have a similar property – that we can tell it to reason about one sentence in particular and have it efficiently allocate resources to that task.

Garrabrant induction does not do anything of this sort, nor are there obvious modifications that can be made to incorporate this ability. In other terms, Garrabrant induction does not “think about what to think about.” We can imagine Garrabrant induction being but one tool in a larger reasoner’s toolkit, with the reasoner sometimes deciding that the best use of resources is to train the Garrabrant inductor. At the moment, however, this is purely speculative.

### Logical Counterpossibilities

If you ask a mathematician what would happen if [Fermat's last theorem](#) was false, they might answer that this would imply the existence of non-modular elliptic curves. However, Fermat's last theorem has been proven true, so by the [principle of explosion](#), assuming Fermat's last theorem is false should imply all other mathematical statements, whether they are true or false. The first sort of answer seems more reasonable, and indeed, mathematicians regularly reason about these logical counterpossibilities. One might hope that logical inductors would naturally extend themselves to assigning coherent probabilities to logical counterpossibilities.

The logical induction criterion requires that, in the limit, all logical inductors assign probability 0 to Fermat's last theorem being false. Conditioning on an event with probability 0 is ill-defined, so we have no formal guarantee that these conditional probabilities are well-behaved. So far, a satisfactory treatment of the counterpossibility probabilities of logical inductors has proven elusive.

For more on why logical counterpossibilities are important for designing robust decision-making algorithms, see [Embedded Agency](#).

---

1. This definition is not sensitive to the precise definition of "efficiently computable." Polynomials are used for their desirable closure properties, but other definitions will also yield logical inductors that satisfy more relaxed/strict versions of the same properties with high/lower runtime complexity. [←](#)
2. What is meant by nice is described in Definition 3.4.3. [←](#)
3. The reasoner must have access to randomness. [←](#)
4. The astute reader may notice that Brower's fixed point theorem is non-constructive. We find the fixed point by brute force searching over all rational numbers. See [5.1.2] for details. [←](#)
5. It is not obvious that the set of trading strategies with bounded loss can be computably enumerated. See [5.2] for more details on how this is done. [←](#)
6. This description is simplified. See [5.3] for more details. [←](#)
7. Since the trading strategy depends on the market's price history, this definition is recursive. [←](#)

# Reward Is Not Enough

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

## Three case studies

### 1. Incentive landscapes that can't feasibly be induced by a reward function

**You're a deity, tasked with designing a bird brain.** You want the bird to get good at singing, as judged by a black-box hardcoded song-assessing algorithm that you already built into the brain last week. The bird chooses actions based in part on within-lifetime reinforcement learning [involving dopamine](#). What reward signal do you use?

Well, we want to train the bird to sing the song correctly. So it's easy: the bird practices singing, and it listens to its own song using the song-assessing black box, and it does RL using the rule:

*The better the song sounds, the higher the reward.*

Oh wait. The bird is also deciding how much time to spend practicing singing, versus foraging or whatever. And the *worse* it sings, the *more important* it is to practice! So you *really* want the rule:

*The worse the song sounds, the more rewarding it is to practice singing.*

Uh oh.

How do you resolve this conflict?

- *Maybe "Reward = Time derivative of how good the song sounds"?* Nope, under this reward, if the bird is bad at singing, *and* improving very slowly, then practice would not feel very rewarding. But here, the optimal action is to continue spending lots of time practicing. (Singing well is *really* important.)
- *Maybe "Reward is connected to the abstract concept of 'I want to be able to sing well'?"* Sure—I mean, that *is* ultimately what evolution is going for, and that's what it would look like for an adult human to “want to get out of debt” or whatever. But how do you implement that? “I want to be able to sing well” is an awfully complicated thought; I doubt most birds are even able to think it—and if they could, we still have to solve a vexing symbol-grounding problem if we want to build a genetic mechanism that points to that particular concept and flags it as desirable. No way. I think this is just one of those situations where “the exact thing you want” is not a feasible option for the within-lifetime RL reward signal, or else doesn't produce the desired result. (Another example in this category is “Don't die”.)
- *Maybe you went awry at the start, when you decided to choose actions using a within-lifetime RL algorithm?* In other words, maybe “choosing actions based on anticipated future rewards, as learned through within-lifetime experience” is not a good idea? Well, if we throw out that idea, it *would* avoid this problem, and a lot of reasonable people do go down that route ([example](#)), but I disagree (discussion [here](#), [here](#)); I think RL algorithms (and more specifically model-based RL algorithms) are really effective and powerful ways to skillfully navigate a complex and dynamic world, and I think there's a very good reason that these algorithms are a key component of within-lifetime learning in animal brains. There's gotta be a better solution than scrapping that whole approach, right?

- Maybe after each singing practice, you could rewrite those memories, to make the experience seem more rewarding in retrospect than it was at the time? I mean, OK, maybe in principle, but can you actually build a mechanism like that which doesn't have unintended side-effects? Anyway, this is getting ridiculous.

...“Aha”, you say. “I have an idea!” One part of the bird brain is “deciding” which low-level motor commands to execute during the song, and another part of the bird brain is “deciding” whether to spend time practicing singing, versus foraging or whatever else. *These two areas don’t need the same reward signal!* So for the former area, you send a signal: “the better the song sounds, the higher the reward”. For the latter area, you send a signal: “the worse the song sounds, the more rewarding it feels to spend time practicing”.

...And that’s exactly the solution that evolution discovered! See the discussion and excerpt from [Fee & Goldberg 2011](#) in my post [Big picture of phasic dopamine](#).

## 2. Wishful thinking

**You’re the same deity, onto your next assignment: redesigning a human brain to work better.** You’ve been reading all the ML internet forums and you’ve become enamored with the idea of backprop and differentiable programming. Using your godlike powers, you redesign the whole human brain to be differentiable, and apply the following within-lifetime learning rule:

*When something really bad happens, do backpropagation-through-time, editing the brain’s synapses to make that bad thing less likely to happen in similar situations in the future.*

OK, to test your new design, you upgrade a random human, Ned. Ned then goes on a camping trip to the outback, goes to sleep, wakes up to a scuttling sound, opens his eyes and sees a huge spider running towards him. Aaaah!!!

The backpropagation kicks into gear, editing the synapses throughout Ned’s brain so as to make that bad signal less likely in similar situations the future. What are the consequences of these changes? A bunch of things! For example:

- *In the future, the decision to go camping in the outback will be viewed as less appealing.* Yes! Excellent! That’s what you wanted!
- *In the future, when hearing a scuttling sound, Ned will be less likely to open his eyes.* Whoa, hang on, that’s not what you meant!
- *In the future, when seeing a certain moving black shape, Ned’s visual systems will be less likely to classify it as a spider.* Oh jeez, this isn’t right at all!!

In [The Credit Assignment Problem](#), Abram Demski describes actor-critic RL as a two-tiered system: an “instrumental” subsystem which is trained by RL to maximize rewards, and an “epistemic” subsystem which is *absolutely not* trained to maximize rewards, in order to avoid wishful thinking / wireheading.

Brains indeed do an awful lot of processing which is not trained by the main reward signal, for precisely this reason:

- Low-level sensory processing seems to run on [pure predictive \(a.k.a. self-supervised\) learning, with no direct involvement of RL at all](#).
- Some higher-level sensory-processing systems seem to have a separate reward signal that reward it for discovering and attending to “important things” *both good and bad*—see discussion of inferotemporal cortex [here](#).
- The brainstem and hypothalamus seem to be more-or-less locked down, doing no learning whatsoever—which makes sense since they’re the ones *calculating* the reward signals. (If the brainstem and hypothalamus were being trained to maximize a signal

that they themselves calculate ... well, it's easy enough to guess what would happen, and it sure wouldn't be "evolutionarily adaptive behavior".)

- Other systems that help the brainstem and hypothalamus calculate rewards and other assessments—amygdala, ventral striatum, agranular prefrontal cortex, etc.—likewise seem to [have their own supervisory training signals](#) that are different from the main reward signal.

So we get these funny within-brain battles involving subsystems that do not share our goals and that we cannot directly intentionally control. I know intellectually that it's safe to cross the narrow footbridge over the ravine, but my brainstem begs to differ, and I wind up turning around and missing out on the rest of the walk. "Grrr, stupid brainstem," I say to myself.

### 3. Deceptive AGIs

You're a human. You have designed an AGI which has (you believe) a good corrigible motivation, and it is now trying to invent a better solar panel.

- There's some part of the AGI's network that is imagining different ways to build a solar panel, and trying to find a good design;
- There's another part of the AGI's network that is choosing what words to say, when the AGI is talking to you and telling you what it's working on.

(In the human case, we could point to different parts of the cortex. The parts are interconnected, of course, but they can still get different reward signals, just as in the bird example above.)

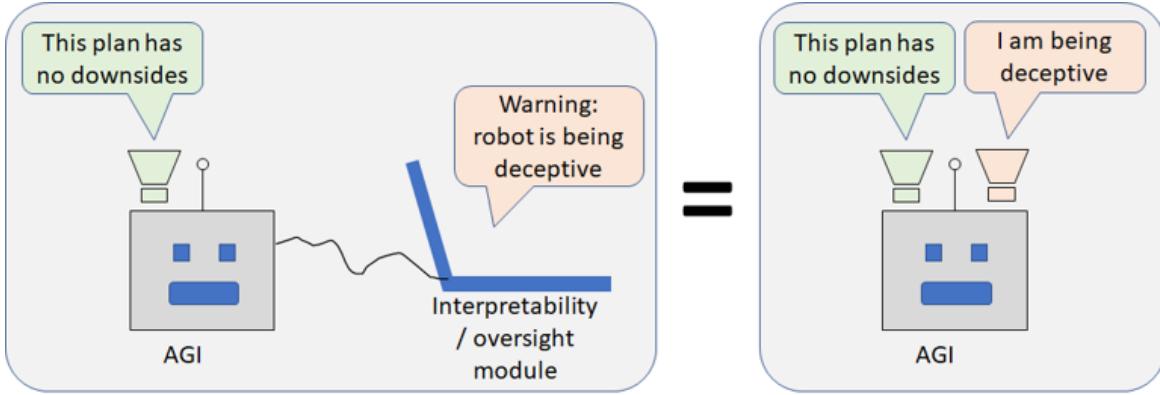
The obvious approach is to have one reward signal, widely broadcast, influencing both parts of the network. And if we get to a point where we can design reward signals that sculpt an AGI's motivation with *surgical precision*, that's fine! We would sculpt the motivation so that the AGI is trying to invent a better solar panel as a *means to an end*, with the ultimate goal of helping you. Then when it chooses words, it would also be maximally helpful.

I think it's more likely that we won't get to that point—that AGIs' motivations will always be a bit off from what we intended [for various reasons](#), but [we'll \(hopefully\) muddle through anyway thanks to conservatism, transparency, and so on](#). Anyway, the motivation will certainly be off during training, when it's still learning basic things about the world, the existence of people, itself, etc. So either way, we should be deeply concerned that the AGI may sometimes be *overly* motivated to invent a better solar panel. And then, when it chooses the words to say to you, it may choose its words partly to manipulate you so as to advance its goal of inventing a better solar panel—rather than choosing words that are purely maximally helpful and honest.

So it seems to me that the word-choosing part of the network should have a *separate* reward signal (or be built in a different way altogether), a reward signal which doesn't care a *whit* about inventing a better solar panel, a reward signal geared purely towards clear and honest communication.

I don't have a specific plan here. I just feel like I'd be surprised if the best possible safe AGI design that anyone can dream up will involve using the *exact same* reward signal for "making plans" and "choosing what words to emit when talking to a human". I think we'll want different reward signals. Just like the bird example above.

This kind of non-unified agent sounds kinda weird, right? Like an AGI with Tourette's or something?!? But it shouldn't sound weird! It's similar to (or even isomorphic to) a robot taking actions while being inspected by an interpretability / oversight module:



These are the exact same thing, just drawn differently. In the left diagram, we are meant to imagine a “unified agent” AGI that chooses actions to maximize reward, meanwhile being inspected / overseen by an independent algorithm running on a laptop nearby. In the right diagram, we moved the inspection / oversight algorithm into the same box as the AGI, along with its own separate speaker. Here the drawing encourages us to imagine this system as a kind of “non-unified” AGI, likely with multiple subsystems running in parallel, each trained on a different reward / supervisory signal.

## Does an agent need to be "unified" to be reflectively stable?

“Need”? No. It’s clearly *possible* for a non-unified system—with different supervisory signals training different subsystems—to be [reflectively\\_stable](#). For example, take the system “me + AlphaZero”. I think it would be pretty neat to have access to a chess-playing AlphaZero. I would have fun playing around with it. I would not feel frustrated in the slightest that AlphaZero’s has “goals” that are not my goals (world peace, human flourishing, etc.), and I wouldn’t want to change that.

By the same token, if I had easy root access to my brain, I would *not* change my low-level sensory processing systems to maximize the same dopamine-based reward signal that my executive functioning gets. I *don’t want* the wishful thinking failure mode! I *want* to have an accurate understanding of the world! (Y’know, having Read The Sequences and all...) Sure, I might make a few tweaks to my brain here and there, but I certainly wouldn’t want to switch every one of my brain subsystems to maximize the same reward signal.

(If AlphaZero were an arbitrarily powerful goal-seeking agent, well *then*, yeah, I would want it to share my goals. But it’s possible to make a subsystem that is *not* an arbitrarily powerful goal-seeking agent. For example, take AlphaZero itself—not scaled up, just literally exactly as coded in the original paper. Or a pocket calculator, for that matter.)

So it seems to me that a “non-unified” agent is not *inevitably* reflectively unstable. However, they certainly can be. Just like I have a few bones to pick with my brainstem, as mentioned above, it’s likewise very possible for different parts of the agent to start trying to trick each other, or hack into each other, or whatever. This is an obvious potential failure mode that we’d be nuts to ignore.

It’s not a new problem though. Remember the figure above: it’s arbitrary where we draw the line between “the AGI” and “other algorithms interacting with and trying to influence the AGI”. So it’s not a fundamentally different type of problem from [gradient hacking](#). Or, for that matter, deception in general. (After all, humans are algorithms too.)

# The Fraught Valley

Still, while it's not a *new* problem, I'll still take this as an excuse to talk about solving it.

The way I'm thinking about it is:

- Early in training, we have *The Path Of Incompetence*, where the “executive / planning submodule” of the AGI is too stupid / insufficiently self-aware / whatever to formulate and execute a plan to undermine other submodules.
- Late in training, we can hopefully get to *The Trail of Corrigibility*. That's where we have succeeded at making a corrigible AGI that understands and endorses the way that it's built—just like how, as discussed above, my low-level sensory processing systems don't share my goals, but I like them that way.
- If there's a gap between those, we're in, let's call it, *The Fraught Valley*.

For example, go back to that figure above, and imagine using those interpretability / oversight tools to install and verify good motivations in the executive / planning submodule. The goal is to do this successfully *before* the AGI is sophisticated enough to undermine the interpretability tools themselves.

Or imagine trying to do value learning (IRL etc.) in an AGI that builds a world-model from scratch, as I believe humans do. Here we *literally can't* install the right motivations from the get-go, because “the right motivations” are inevitably defined in terms of concepts like objective reality, people, self, etc., that are (as of yet) nowhere to be found in the world-model. So maybe we let it do some learning, with some carefully-thought-through curriculum of data and rewards, and spin up the IRL subsystem as soon as the world-model is developed enough to support it.

Anyway, the goal is to make the width of the Fraught Valley as small as possible, or better yet, eliminate it altogether. This involves:

1. Making it *hard and complicated* to corrupt the various motivation-installation and interpretability systems. I don't think it's realistic to harden these systems against a superintelligent adversary, but every little roadblock we can think of is good—it helps stretch out the Path Of Incompetence.
2. Meanwhile, we push from the other side by designing the AGI in such a way that we can install good motivations, and especially root out the most dangerous ones, *early*. This might involve things like directing its attention to learn corrigibility-relevant concepts early, and self-awareness late, or whatever. Maybe we should even try to hardcode some key aspects of the world-model, rather than learning the world-model from scratch as discussed above. (I'm personally very intrigued by this category and planning to think more along these lines.)

Success here doesn't seem *necessarily* impossible. It just seems like a terrifying awful mess. (And potentially hard to reason about *a priori*.) But it seems kinda inevitable that we have to solve this, unless of course AGI has a wildly different development approach than [the one I'm thinking of](#).

## Finally we get to the paper “Reward Is Enough” by Silver, Sutton, et al.

The title of this post here is a reference to [the recent paper by David Silver, Satinder Singh, Doina Precup, and Rich Sutton at DeepMind](#).

I guess the point of this post is that I'm disagreeing with them. But I don't really know. The paper left me kinda confused.

Starting with their biological examples, my main complaint is that they didn't clearly distinguish "within-lifetime RL ([involving dopamine](#))" from "evolution treated as an RL process maximizing inclusive genetic fitness".

With the latter (intergenerational) definition, their discussion is *entirely trivial*. Oh, maximizing inclusive genetic fitness "is enough" to develop perception, language, etc.? **DUUUHHHH!!!**

With the former (within-lifetime) definition, their claims are mostly false when applied to biology, as discussed above. Brains do lots of things that are not "within-lifetime RL with one reward signal", including [self-supervised \(predictive\) learning](#), [supervised learning](#), [auxiliary reward signals](#), [genetically-hardcoded brainstem circuits](#), etc. etc.

Switching to the AI case, they gloss over the same interesting split—whether the *running code*, the code controlling the AI's actions and thoughts in real time, looks like an RL algorithm (analogous to within-lifetime dopamine-based learning), or whether they are imagining reward-maximization as purely an outer loop (analogous to evolution). If it's the latter, then, well, then what they're saying is trivially obvious (humans being an existence proof). If it's the former, then the claim is nontrivial, but it's also I think wrong.

As a matter of fact, I personally expect an AGI much closer to the former (the real-time running code—the code that chooses what thoughts to think etc.—involves an RL algorithm) than the latter (RL purely as an outer-loop process), for reasons discussed in [Against Evolution as an analogy for how humans will create AGI](#). If that's what they were talking about, then the point of my post here is that "reward is not enough". The algorithm would need other components too, components which are not directly trained to maximize reward, like self-supervised learning.

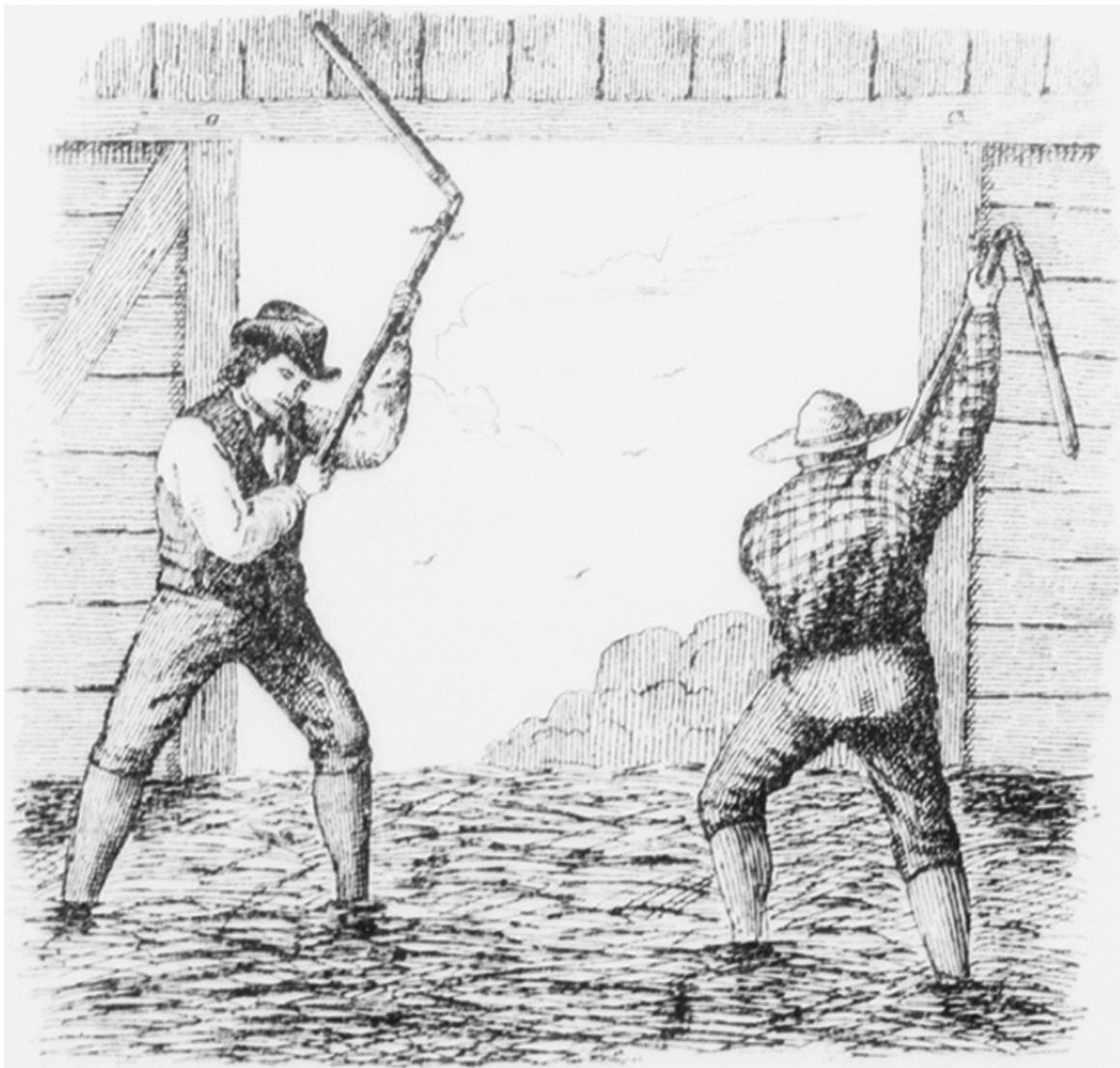
Then maybe their response would be: "Such a multi-component system would still be RL; it's just a *more sophisticated* RL algorithm." If that's what they meant, well, fine, but that's definitely not the impression I got when I was reading the text of the paper. Or the paper's title, for that matter.

# Why did we wait so long for the threshing machine?

This is a linkpost for <https://rootsofprogress.org/why-did-we-wait-so-long-for-the-threshing-machine>

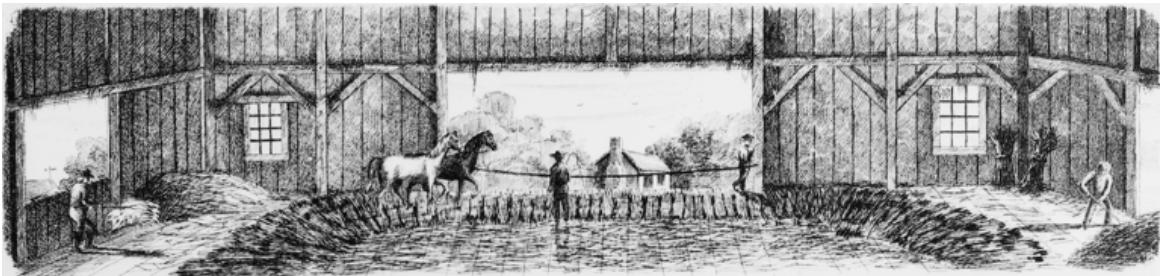
When ripe wheat is harvested, the edible seed is encased in an outer husk. Before the seed can be ground into flour, or boiled into porridge, or planted in the field to produce next year's harvest, it must be removed from the husk. This process is called threshing.

As the husk is quite hard, threshing is a violent process. Traditionally, it was done with a tool called a flail, which is simply a short stick attached by a cord to a longer handle. The grain was spread out on the ground (yes, disgusting) and beaten with the stick to open the casings.



*The Growth of Industrial Art, 1892*

Other methods included “treading”, in which livestock trampled the grain with their hooves (yes, even more disgusting) or dragged a sledge over the grain (the Latin word for this sledge is *tribulum*, from which we get the word “tribulation”).



Horse treading. [The Growth of Industrial Art, 1892](#)

Occasionally the grain would be rubbed against a wire screen, or placed in a sack and beaten with rocks. It's no coincidence that the word “thrashing” is similar: it is an archaic spelling of the same term.

As one of the more labor-intensive stages of wheat farming, threshing was a natural candidate to be automated by machinery. And the threshing machine was a relatively simple device: like [the cotton gin](#), [the flying shuttle](#), or [the bicycle](#), it was a mechanical invention that did not depend on any scientific discoveries. Still, threshing machines were not used to any significant degree until the late 1700s in the UK and the early 1800s in the US.

Once again, the question arises: [why did we wait so long?](#)

## Early claims and concepts

First, it's *not* because no one had the idea to mechanize the threshing process. Although we think of the Industrial Revolution as a uniquely inventive age, and the era before it seems blind to the possibility of innovation by comparison, we must remember that there *were* inventions prior to 1700, many of which were broadly adopted, especially when they helped with essential economic activities. The loom and the printing press predated the threshing machine by centuries, and neither seems far less complex.

The threshing machine was referred to in an English patent as early as 1636 (for historical context, this is not long after the major works of Galileo and Francis Bacon). Sir John Christopher van Berg, a knight of Moravia (part of the present-day Czech Republic), fleeing religious wars in Germany and seeking protection from King Charles I, offered to Charles and to England his observations of “diverse mechanical instruments and frames operating by weights.” [The patent](#) is extremely broad and vague by today’s standards, listing a literally incredible variety of machines: for pumping water, washing clothes, cooking meat, working bellows, dredging rivers, raising sunken ships, measuring distances and depths, and for pretty much any mechanical operation. Towards the end of the list is mentioned an invention “to be agitated by wind, water, or horses, for the clean threshing of corn [grain], whereby much corn that is now left may be saved, and the straw made near as good as hay.” At the time, patents did not require detailed descriptions, diagrams, or models, and there is no more insight than this into how the machine might have worked, if indeed the invention was even real. The point is that by the early 1600s, the threshing machine was seen as possible and desirable.

And why not? The grinding of grain into flour had been mechanized by watermills in ancient Rome. The threshing process was similar enough that some type of mill could plausibly accomplish the task. A century after Van Berg, inventors really started working on the problem, with one Michael Menzies of Scotland patenting a threshing machine in 1734, and several more efforts throughout the 1700s.

So why did none of these machines gain wide adoption?

## Challenges with reliability

One clue comes from reports of trials of the early machines.

Some of the machines simply broke. Of Menzies's machines, one historian wrote that "owing to the velocity required to do the work perfectly, they soon broke, and the invention fell into disgrace" ([Somerville 1805](#), p. 75). Another machine was trialed in Scotland in the late 1700s, but "in a few minutes the model was torn to pieces" ([Quick 1978](#), p. 45).

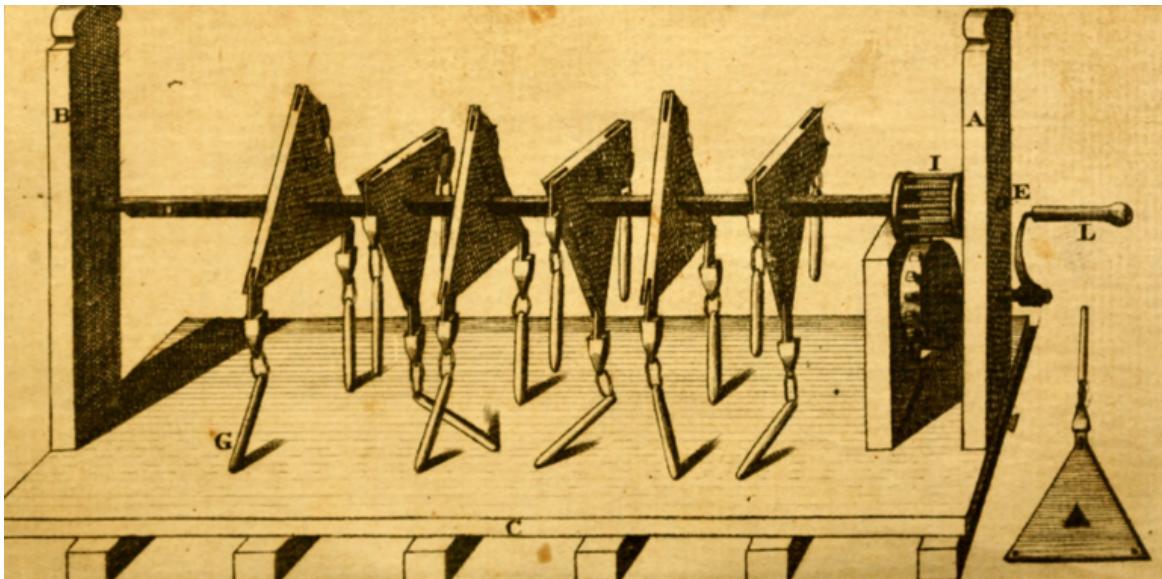
When the machines did not break, they often failed to do the job. In one trial in 1753, it was found that the machine "broke off the ears of barley and wheat instead of clearing them of the grain, and that, at best, it was only fit for oats" ([Ransome 1843](#), p. 140). Another was attempted in the late 1770s, but: "Upon trial this machine was also found defective, as along with its doing very little work in a given time, it bruised the grain, and so materially hurt its appearance, as to lessen its value considerably in the market" ([Somerville 1805](#), p. 76).

In general, [McClelland \(1997\)](#) says that threshing machines in the late 18th and early 19th century were "among the most complicated and expensive of all agricultural implements in American and Britain. High price and frequent breakdowns did not bode well for rapid adoption when 'the least derangement ... is death to the whole machine'" (p. 172). He reports that Washington and Jefferson were both interested in threshing machines, but (pp 175-6):

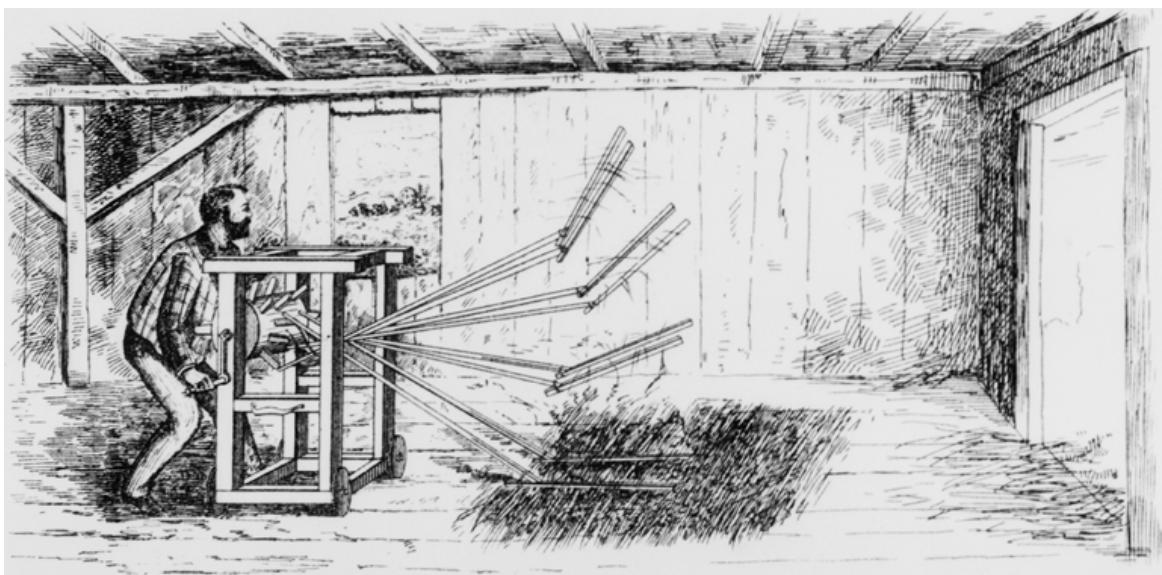
Washington's machine... built using the plans of William Booker soon proved unsatisfactory; a Maryland machine constructed with the aid of Colonel Anderson's plans developed a warped wheel and was abandoned; in 1802 an immigrant from Edinburgh introduced into the Mid-Atlantic states "six or seven" based upon "the Scotch principle," but they soon developed problems and "common workmen" could not repair them; the machine patented in 1803 by Jedediah Turner, based "upon entirely new, and very plain principles" was never commercially produced.... Washington perhaps spoke for a number of would-be agrarian pioneers when he voiced his frustration with this prospective new technology in 1793: **"I have seen so much of beginning and ending of new inventions, that I have almost resolved to go on in the old way of treading."**

## What was the problem?

One challenge was to find the right basic idea for the design of the machine. As is often the case, some of the early attempts tried to mimic human motion too closely: these machines automated the motion of the flail, but continued to beat the grain on the ground, an idea known as the "beating" principle. "All machines based upon the beating principle suffered from the same problem. The violence of the mechanical action used to separate the grain soon produced broken parts and an implement in disrepair" ([McClelland 1997](#), p. 170). This was the problem with Menzies's 1734 machine.

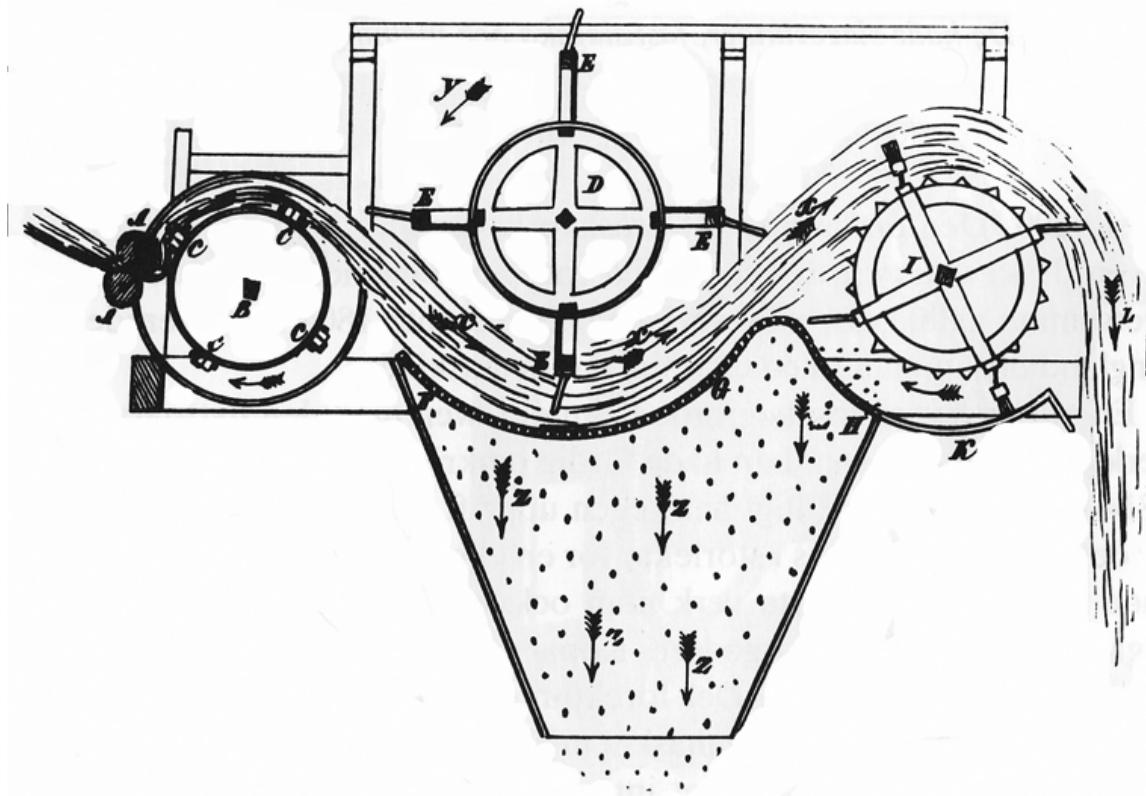


*Design for a “beating” type machine. [Thomas Paine’s Pennsylvania Magazine](#), February 1775*

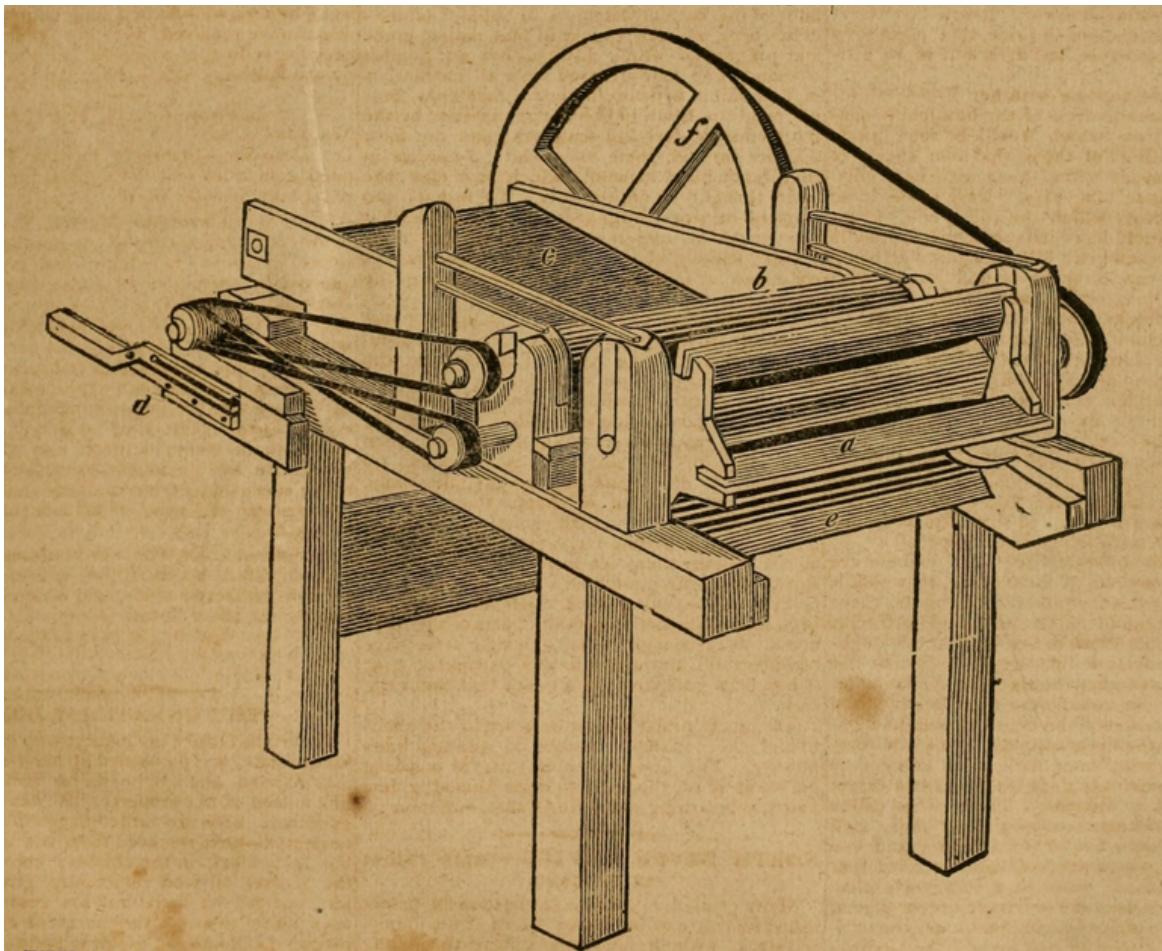


*Another “beating” concept. [The Growth of Industrial Art, 1892](#)*

A better idea was the “rubbing” principle. This design was more like a mill or a cotton gin: the grain was fed into a rotating drum that had pegs or other protrusions to grind the husk off of the seed. While “beating” vs. “rubbing” was [still debated as late as the 1830s](#), all successful machines were of the rubbing type.



*Meikle's threshing machine, 1786. [Wikimedia](#)*



Pope's threshing machine. [American Farmer, 1823](#)

But the threshing machine didn't seem to take off after one core design breakthrough. There isn't a single individual who is always called out as "the inventor of the threshing machine," nor a single date that stands out as an inflection point. Instead, one finds series of important inventions across decades: Andrew Meikle built the first successful machine in Scotland in 1786, but Joseph Pope invented a popular one in the US in 1820, and the Pitts brothers made improvements to power and to winnowing (separating the wheat from the chaff) in the 1830s. Nor does there seem to have been a single design that became dominant. Instead, it's a story of gradual improvements in effectiveness, reliability, and cost that led to gradual adoption by farmers. This pattern of iterative improvement applies to most inventions—even ones that have a "heroic" story in which a single invention and inventor are traditionally highlighted—but even so, the history of the threshing machine seems unusually long and incremental.

My hypothesis is that the threshing machine was just past a certain threshold of the combination of the *amount of force required* and the *delicacy of the operation*. A loom is a somewhat complex machine performing an intricate process, but not one that uses a high degree of force. A flour mill, or a trip hammer at an iron works, is a high-force application, but not one that is particularly subtle or delicate. Both of these were in use long before the Industrial Revolution. But a threshing machine seems to require enough of both characteristics that *manufacturing quality became critical to meet the standard of reliability that was needed for practicality and adoption*. A simple, reliable design helped with this, but competent workmanship was at least as important.

# Challenges in manufacturing

Here's some evidence for this from historical sources.

The American farm periodical [\*Genesee Farmer, April 1831\*](#), in an article on the the threshing machine, writes:

... one of the great and principal causes of failures, in many kinds of machines, is the **flimsy, cheap, and do-for the-present manner in which they are made**. They are not unfrequently constructed by carpenters, or rather by those who are only an apology for a good one, and who could hardly construct a button to a barn-door...

By the operation of these causes the farmer often gets an **ill-constructed, weak, and rickety machine, which needs wedging, nailing, and bracing, at every revolution...**

He continues with specific advice on construction and especially on materials:

The machinery that generates the motion, whether horse or water power, ought to be as well constructed, and of as good materials, as a flouring mill; and **it is worse than useless to make the main wheel and pinion gearing of wood. Nothing but cast iron**, and that of the softest and best kind, can be depended upon.

(Using iron instead of wood has also been noted as a key step taken by Henry Maudslay to improve the precision of machine tools.)

On the British side, [\*Ransome \(1843\)\*](#) defends the concept of the threshing machine against the bad reputation it might acquire from poor workmanship:

It has been urged against these machines, that they are apt to break the straw, and that they bruise and nib the barley so as to render it unfit for malting; but **these faults are not so much attributable to the principles of the machines, as to the manner in which they are frequently turned out of the hands of the workman**; and sometimes to the want of skill and judgment in the parties who have the management of them.

Many of these machines are made by persons who possess little claim to any mechanical knowledge, and who, purchasing the unfitted castings, by the help of village artisans, produce an imitation of those which are considered good. As **the perfection of these machines must depend upon mathematical accuracy in the adjustment of all their parts, and in the truth and precision of their fittings, it is unreasonable to expect that this can be accomplished where no facilities exist beyond the forge and the work-bench**; and hence arises a degree of discredit, which is unfairly thrown upon the principles upon which the machine is formed.

So, if manufacturing quality was a critical factor—how were the early machines made?

## Manufacturing systems

Today, agricultural equipment is manufactured by large, specialized firms who distribute their product worldwide. But in the 1700s and even into the 1800s, there was no such thing. "Agricultural equipment" meant plows, scythes, and wagons, and it was made by local craftsmen, such as the town carpenter, blacksmith, or cartwright. Not only were there very few large manufacturing enterprises in the world, but until the growth of the railroads in the mid-1800s, there were also no efficient transportation networks to distribute their products to farmers throughout the countryside. (See also [\*MacDonald 1975\*](#) on this point.)

In the promotions of early threshing machine inventors, we can see how they expected their machines to be acquired.

In many cases, an inventor merely proposed to supply plans and/or models, assuming that the machine would be built by a local craftsman, typically a carpenter or especially a millwright (millwrights were the closest thing to mechanical engineers in the 1700s). A [1772 advertisement in \*The Virginia Gazette\*](#) promotes an invention that “may be carried into execution by any tolerable carpenter” or even “by gentlemen's own servants.” The inventor, one John Hobday, does not even seem to expect to sell the plans, let alone the machines: he proposes that a subscription be raised for him. More than fifty years later, in 1823, not much has changed: an [article in the \*American Farmer\*](#) on Joseph Pope's threshing machine quotes a letter from the son of the deceased inventor as saying that the machine “can be constructed at little expense, the materials, including the shears, cost \$13, and it can be made by a good workman, (say a joiner or carpenter) in 12 days.”

In other cases, inventors offered not the plans but the parts, “some assembly required.” In 1735, Andrew Good Wright in Edinburgh [advertised a threshing machine](#), saying: “Those who want them, must send for them from Edinburgh, and a millwright to receive them, that he may know how to set them up; which any millwright will be able to do, after he has seen one going.” He also mentions a licensing opportunity: “Millwrights recommended by gentlemen for their integrity, if they come to Edinburgh, and understand the machine, shall have liberty from the patentee to make them in the country.” A [1796 advertisement](#) from “John Jubb, millwright and machine-maker,” offers both options: he will “send it to any part of the kingdom” for 25 guineas plus shipping and handling, or he will sell the plans along with “the machine made at Leeds, taken to pieces, properly marked, and sent off... so that any workman may with ease put it together.”

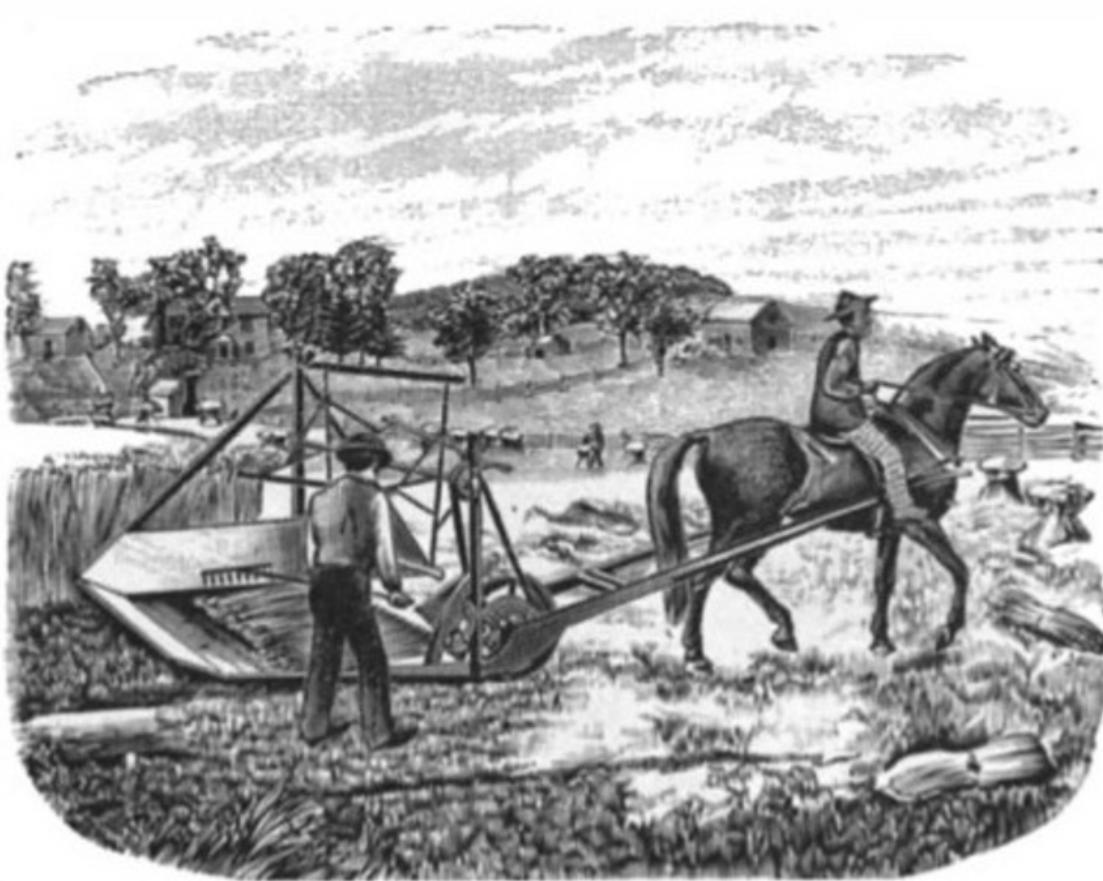
There is some evidence that threshing machines thrived in places and times when they were made by the inventor himself, or by a skilled mechanic who decided to specialize in them. [One source \(pp 3-4\)](#) reported as of 1800 that “almost all the threshing machines in England” had been built by one “industrious workman, of the name of Stevenson”. The article noted approvingly that his machines moved easily by hand and showed little friction, “a sure warrant of accurate workmanship.” As for Joseph Pope's machine, soon after the article quoted above, he contracted with an engine maker to be his manufacturer. From May through October 1823, the Philadelphia *National Gazette* ran [advertisements](#) stating that “Messrs. S. V. Merrick & Co. Engine Manufacturers, Philadelphia,” had been invested with the privilege of making and selling his machine; prospective customers were encouraged to “apply” to them to obtain one. An engine manufacturer, at that time, would have had a well-equipped machine shop with the precision tools and skilled workers needed to make reliable machines.

Bottom line: In its early decades, the threshing machine suffered from inconsistent quality due to distributed, unspecialized workmanship.

## Contrast with the reaper

In contrast, consider a closely related piece of agricultural equipment invented decades later: the mechanical reaper.

The reaper cut down stalks of grain in the field, automating the step just prior to threshing. As with the thresher, reliability was a crucial challenge for the reaper: early models failed to cut tangled thickets of grain, or clogged when the stalks were damp, or performed poorly on hills.



*McCormick reaper, 1831. [Scientific American](#)*

One of the most successful reapers was created by Cyrus McCormick. He began selling them in the 1840s, but he relied neither on local craftsmen nor on manufacturing partners. He made the first machines himself, and eventually created a large, central factory.

McCormick sought a nationwide market, but the primitive cargo networks were a barrier. [Casson \(1909\)](#), one of his biographers, writes (p. 62):

To get the seven Reapers [sold in 1844] to the West, they had first to be carried in wagons to Scottsville, then by canal to Richmond, re-shipped down the James River to the Atlantic Ocean and around Florida to New Orleans, transferred here to a river boat that went up the Mississippi and Ohio Rivers to Cincinnati, and from Cincinnati in various directions to the expectant farmers. Four of these Reapers arrived too late for the harvest of 1844, and two of them were not paid for.

McCormick later moved from Virginia to Chicago, setting up shop at the heart of a transportation network that integrated railroads and waterways—and in a location that was closer to his biggest market, in the Midwest (where the lands were flat and labor was scarce).

He also innovated on business practices to sell his machines: he advertised heavily in newspapers, organized field demonstrations, offered farmers a written guarantee and free credit until the next harvest, and built out a network of distributors in each region.

No one, from what I've read, did any of the above for the threshing machine, at least not before 1830 or so, when the machines were already common.

# Why did it take so long?

The question of the threshing machine is a microcosm of the bigger question I'm interested in: Why did we wait so long for the Industrial Revolution?

Some inventions depended on theoretical concepts that were not discovered before the Scientific Revolution: the electric generator, for instance, or [even the steam engine](#). But others did not: the threshing machine, the [cotton gin](#), the spinning jenny. What is the difference between those inventions, and others of seemingly similar complexity and importance that were adopted centuries earlier: the loom, the spinning wheel, the printing press?

One theme we see here is the manufacturing capability required to make affordable, reliable machines. This was probably also important for the bicycle, I think even more so than I realized or emphasized in [my previous essay on that topic](#). For another example, [Robert Allen \(2009\)](#) says that precision manufacturing from the watch industry was necessary for an early thread-spinning machine known as the “water frame”: “The watch industry was the source of gears—brass gears in particular—and they were the precision parts in the water frame.... Without watch-makers, the water frame could not have been designed” (p. 204).

A secondary theme here is that there are prerequisites to market creation. If specialized manufacturing is required, then a centralized manufacturer must be able to reach a wide market. This requires communications infrastructure, such as newspapers, in order to market the goods, and transportation infrastructure, such as railroads, to ship them.

One reason perhaps (I'm speculating here) that mechanization came to textile manufacturing decades before it came to agriculture is that the business model was different. Richard Arkwright made his spinning machines, not to sell them as McCormick sold his reapers, but to *use* them to make cheap thread. He was his own market for the machines; they were a capital investment in a new type of business. This business model wasn't really available to early agricultural equipment makers.

Another instructive comparison is to the steam engine, which also required high-quality, precision manufacturing. Boulton & Watt adopted the centralized, specialized factory in the 1770s, decades before this model came to agricultural equipment. My speculative hypothesis here is that farmers, as a market, were a large number of smaller and more geographically distributed customers, vs. the coal mines, ironworks, breweries, etc. who were the customers for Watt's engines. Thus the challenges of marketing and distribution were greater in the agricultural market. (Not to mention that industrial concerns such as mines might have had more ability than small farmers to finance a large expenditure on capital equipment.)

## Flywheels

Maybe all this can be summarized as *infrastructure*. Manufacturing capability is infrastructure, as are railroads and newspapers (or today, trucking and the Internet). Infrastructure lowers the activation energy of any particular development. Sometimes these developments themselves create more/better infrastructure, creating a reinforcing cycle that generates [exponential growth](#).

Is this effect enough to explain the pace of progress throughout history? Do we still need to posit cultural causes—a general factor of inventiveness—Joel Mokyr’s “idea of progress” vs. “ancestor-worship”, Deirdre McCloskey’s “honor and prestige for the bourgeoisie”? It seems obvious that it at least *matters* whether people believe that progress is possible and desirable (and if it doesn’t, then it certainly seems like an enormous coincidence that the Industrial Revolution happened not long after Bacon and the Scientific Revolution, and in the

same part of the world). But now that I more vividly understand the challenges facing 18th-century threshing machine inventors, I find myself shifting a bit more towards infrastructure as a cause.

Or, more broadly, I'm seeing that there is a set of overlapping flywheels, each creating a virtuous cycle as it gets going. Better infrastructure enables more progress, which in turn creates more and better infrastructure. The belief in progress leads to actual progress, which reinforces the belief. And there are others: Surplus wealth allows us to invest in progress, which creates more surplus. Science ultimately leads to progress, which then helps advance science. And since all these cycles intersect at progress itself, they all reinforce each other, indirectly if not directly.

---

## Bibliography

- [Allen, Robert \(2009\). \*The British Industrial Revolution in Global Perspective\*](#)
- [Casson, Herbert Newton \(1909\). \*Cyrus Hall McCormick: His Life and Work\*](#)
- [MacDonald, Stuart \(1975\). "The Progress of the Early Threshing Machine"](#)
- [McClelland, Peter D. \(1997\). \*Sowing Modernity: America's First Agricultural Revolution\*](#)
- [Quick, Graeme and Wesley Buchele \(1978\). \*The Grain Harvesters\*](#)
- [Ransome, J. Allen \(1843\). \*The Implements of Agriculture\*](#)
- [Somerville, R. \(1805\). \*General View of the Agriculture of East Lothian\*](#)

Key excerpts from the above, as well as citations from newspapers and other corroborating sources I did not reference, can be found in [this summary of primary/historical sources](#).

- [Van Berg's 1636 patent](#)
- 

Thanks to [Anton Howes](#) for conversations, help finding sources, and comments on this essay, all of which contributed greatly to it.

# Bad names make you open the box

([Cross posted](#) on my personal blog.)

Think of a function as a black box. It takes an input, and spits back out an output.



For `getPromotedPosts`, you can feed it a list of blog posts and it will spit back out the ones that have been promoted.

But I probably didn't need to explain that to you, did I? Why not? Well, because `getPromotedPosts` is self-explanatory. Because `getPromotedPosts` is named well.

Now what if instead of `getPromotedPosts`, it was named something like `getThePosts`? Well, that name isn't self-explanatory. You know it's getting some posts, but it's not clear which ones. The most recent posts? Posts from a certain author? Posts from this week?



As a programmer, what do you do in this situation? You scroll to the function definition and start reading the code.

```
function getThePosts(posts) {  
    ...  
}
```

In other words, you *open the box*.



What does that look like? Something like this:

- You're reading through the code in some file. Line one, line two, line three.
- You reach line 30 and see `getThePosts`.

- You realize that `getThePosts` must be getting some posts, but you don't know which ones. So you have to scroll to line 174 where `getThePosts` is defined.
- On line 174 you start reading through `getThePosts`. Once you reach line 210, you realize that it is getting *promoted* posts. Cool!
- Now you scroll back up to line 30. You realize that `getThePosts` is giving you promoted posts, but you forgot what was happening before line 30. Damn. So now you have to go back to line 10 or 15 to remind yourself what was going on in the first place.

## Complexity and zoom level

Maybe I was being dramatic. Is it really such a big deal to have to scroll to the definition of `getThePosts` on line 174? Will it really take that much effort to read lines 174-210 and figure out that it's returning promoted posts? It's only 36 lines of code, including whitespace + brackets, and you could probably glance over parts of it. And then what about returning to line 30? Are you really going to have forgotten what was going on so quickly? Are you really going to have to scroll back up to line 10 or 15 to remind yourself?

In this example, perhaps not. I'm not sure. Maybe having to open the box gets in the way, maybe it doesn't. What I am sure of is that when the code becomes more complicated, having to open the box becomes more of an issue. I think Eric Dietrich's [post on how harmful interruptions can be to a programmer](#) gives us a great intuition for this:

For a programmer, an interruption is oh-so different. There you sit, 12 calls into the call stack. On one monitor is a carefully picked set of inputs to a complex form that was responsible for generating the issue and on the other monitor is the comforting dark theme of your IDE, with the current line in the debugger glowing an angry yellow. You've been building to this moment for 50 minutes — you finally typed in the right inputs, understood the sequence in which the events had been fired, and got past the exact right number of `foreach` and `while` loops that took a few minutes each to process, and set your breakpoint before the exception was triggered, whipping you into some handler on the complete other end of the code base. Right now, at this exact moment, you understand why there are 22 items in the `Orders` collection, you know what the exact value of `_underbilledCustomerCount` is and you've hastily scribbled down the string “8xZ204330Kd” because that was the auto-generated confirmation code resulting from some combination of random numbers and GUIDs that you don't understand and don't want to understand because you just need to know what it is. This is the moment where you're completely amped up because you're about to unlock the mysteries of what on earth could be triggering a null reference exception in this third party library call that you're pretty sure —

“HI!!! How's it going? So, listen, you know that customer order crashing thing is, like, bad, right? Any chance I can get an ETA on having that fixed?”

This example is on the opposite end of the spectrum. Here you're dealing with a ton of complexity, whereas in `getThePosts` the amount of complexity was probably pretty low. So maybe this means that as programmers, we can just use our judgement? For complicated code, take the time to come up with good names. For simple code, fuggedaboutit.

In theory, I think this makes sense. But in practice, I think it often leads to a lot of issues.

Imagine yourself [zooming in](#) on a piece of code. You ask yourself whether it's really such a big deal that the name you used is a little confusing. Your answer is usually going to be something like:

Nah, I think it's fine. It's not that complicated. They'll be able to figure it out.

Now, imagine yourself zooming out and thinking about the entirety of the codebase. Or even just a particular module. You ask yourself whether it's really such a big deal that the code is a little sloppy and confusing. Your answer is usually going to be something like:

Yes! It is a big deal! I'd be able to move so much faster if the code wasn't such a mess!

At least that's what I argue for in [Taking The Outside View On Code Quality](#). In theory, your zoomed in answers would always match your zoomed out answers, but in practice, the answers will depend on the scale you're looking at. I think that this is a really important thing to keep in mind when you ask yourself whether you need to come up with a better name. And I think that the zoomed out perspective is usually the wiser choice.

In taking the zoomed out perspective, I think that it will usually lead you to the conclusion that naming is important. The big example where it wouldn't is when you know you are going to throw the code away. For example, if you are building a prototype. If the prototype is unsuccessful, you'll ditch it. If it is successful, you'll probably rewrite it (perhaps). Either way the prototype code gets ditched. So in that situation, you probably don't need to waste time naming things well. But that is the exception, not the rule. If the code you're writing is "business as usual", investing in good names will pay dividends.

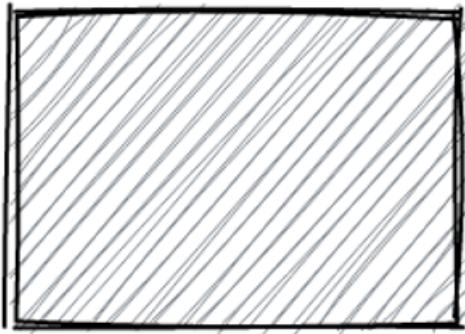
## Not just software

It's not just software. Names and black boxes apply to many other domains, including everyday life. For example, the other day I was reading a post about covid, and it kept referring to B.1.617. And B.1.2, and B.1.1.7, and P.1. Huh? I knew that these were all different covid strains, but I couldn't keep track of which was which. I had to pause my reading, google "B.1.617 covid strain", see that it is the Indian strain, and then pick back up where I left off. In other words, I had to open the box.

B.I.617



open



closed

Honestly, this happens all of the time. It happens at work when people refer to a JIRA ticket as "7967" instead of "the stashboard epic". And when people use weird acronyms like BDM (business development manager). And when things in science are named after people rather than some sort of affordance. Wouldn't it suck if prediction markets were called Hansonian markets?

## Misleading

What sucks even more is when names are *actively misleading*. For example, the concept of [regression to the mean](#) had confused me for a while. The term "regress" sounds like it means "move down", but instead it just means "move closer to". So if covid cases have been unusually low over the past few days and we expected them to tick back up, we would still call it regressing to the mean.

Let's look at an actively misleading name in the context of software. Think back to our `getPromotedPosts` example. The idea is that we have a blog and we want to place promoted posts at the top. But imagine that at some point, management stormed in and demanded that Tom Fahrabs' posts be given that prime real estate, because Tom is one of our investors and he has a new book coming out that he wants to promote: *The Four Second Sex Life*.

So the dev team comments out the body of `getPromotedPosts` and has it instead just get the five most recent Fahrabs posts. It works.

```
function getPromotedPosts() {  
    /*  
     * old  
     * code  
     * here  
    */  
}
```

```
 */  
return fiveMostRecentFahrabsPosts;  
}  
  
cue Jaws music
```

Then after the book launch is a big hit and the team is ready to move back to the old logic, someone new to the codebase winds up writing a new function called `topPosts`. The codebase is a mess so they thought it'd be easier to just write their own function, and "top" seemed like it'd make sense because, after all, it's getting posts that will be placed at the *top* of the page.

But they never delete the now deprecated code for `getPromotedPosts`.

*dah dan*

Fast forward six months. The product team wants a redesign of the page, and it's your job to code it up. The existing code is a bit of a mess, so you tear it apart. Not completely though.

As you're working on the section for promoted posts, you notice `topPosts` and `getPromotedPosts` in the old code. `topPosts` sounds like it's referring to the *best* posts, so that probably isn't what you want. On the other hand, `getPromotedPosts` sounds like it's exactly what you want, so you use it.

*dah dan*

Since you're lazy, you don't really QA it.

*dah dan dah dan*

It passes code review because `getPromotedPosts` sounds reasonable to the other team members too.

*dah dan dah dan*

And it also passes QA because they don't find it odd that Fahrabs posts were at the top, given how popular he is.

*dah dan dah dan dah dan dah dan*

It actually even takes a while for the bug to get discovered in production, for the same reason. It isn't until Fahrabs starts writing about the testimonials he's received from readers of his previous book that someone notices a quirk in the algorithm.

*screams!!!*

Hey, remember when we were talking about how `getPromotedPosts` is low complexity and how maybe we can just forget about naming things well?

## Pot brownies

Maybe a better way to make this point is with a pot brownies analogy.

Imagine that you open the fridge. You see something labeled "brownie". You eat it.

Then you hop in the car and start heading over to your friends house. But right as you merge on to the highway, you start feeling funny.

Turns out that the "brownie" label was a little misleading. It wasn't a regular brownie. It was a *pot* brownie. There was something dangerous inside the brownie, but the label didn't reflect that.

This is similar to poorly named functions with dangerous side effects. In both cases, if the thing in question can have dangerous side effects, you really want to make sure that it is reflected in the label. You can't trust that people will read beyond the label. And even if you could, you wouldn't want people to have to do that. You'd rather them be able to get the information they need from the label.

## Trust

Wow, those sections were pretty scary huh? Well, it gets worse.

It sucks that no one changed the name of `getPromotedPosts` to `getFahrabsPosts` or something and it led to the bug of Fahrabs posts being promoted for so long. But consider what happens in the aftermath of that bug.

Imagine that after going through that nightmare, you see a new function called `getTodaysPosts`. It seems simple enough. It probably just gets all of the posts that were written today. Right?

Nope! You're not gonna fall for that again! Last week you thought that `getPromotedPosts` was going to just get you the promoted posts, but instead it only got you Fahrabs posts, and your boss gave you a stern talking to. So why would you trust that `getTodaysPosts` is going to do what it implies?

You're not. Your trust has been violated, so you're going to open the box. You're going to scroll to `getTodaysPosts` and read through it just to make sure it does what you think it does. Same for `getTopTechPosts` and `getMostRecentEconPosts`. You have to open those boxes too when you come across them, just to make sure.

But as we talked about in the "Complexity and the zoom level" section, this is a really bad situation to be in. To some extent, software is all about managing complexity. Closed boxes really help us manage complexity. But now, due to the violation of trust, that tool has been taken away from us.

## Compression of complexity

There is something really, really powerful going on here, and I'm worried that I'm not doing it justice. I'm worried that I'm not hitting the nail on the head regarding why this all is so important. Let me try explaining it differently.

Consider that original example of `getThePosts`.

- You're reading through the code in some file. Line one, line two, line three.
- You reach line 30 and see `getThePosts`.

- You realize that `getThePosts` must be getting some posts, but you don't know which ones. So you have to scroll to line 174 where `getThePosts` is defined.
- On line 174 you start reading through `getThePosts`. Once you reach line 210, you realize that it is getting *promoted* posts.

There, you have to read 36 lines of code to understand what is going on. But now imagine that you take all of that complexity, and *compress it*.

```

174 function getThePosts(posts) {
175   ...
176   ...
177   ...
178   ...
179   ...
180   ...
181   ...
182   ...
183   ...
184   ...
185   ...
186   ...
187   ...
188   ...
189   ...
190   ...
191   ...
192   ...
193   ...
194   ...
195   ...
196   ...
197   ...
198   ...
199   ...
200   ...
201   ...
202   ...
203   ...
204   ...
205   ...
206   ...
207   ...
208   ...
209   ...
210 }

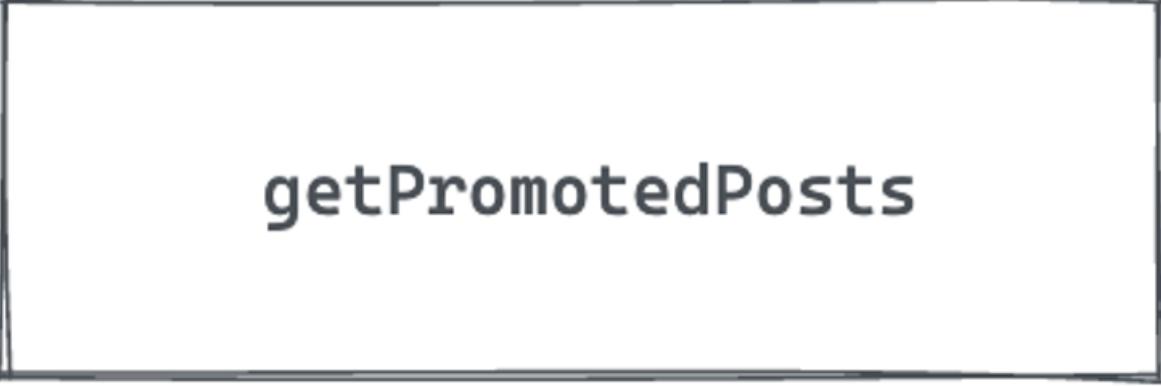
```



*That's the power of good names. It allows you to take a bunch of complexity and package it up into a dense little box. Now instead of dealing with this:*

```
174 function getThePosts(posts) {  
175     ...  
176     ...  
177     ...  
178     ...  
179     ...  
180     ...  
181     ...  
182     ...  
183     ...  
184     ...  
185     ...  
186     ...  
187     ...  
188     ...  
189     ...  
190     ...  
191     ...  
192     ...  
193     ...  
194     ...  
195     ...  
196     ...  
197     ...  
198     ...  
199     ...  
200     ...  
201     ...  
202     ...  
203     ...  
204     ...  
205     ...  
206     ...  
207     ...  
208     ...  
209     ...  
210 }
```

You just have to deal with this:



getPromotedPosts

Much nicer, right?

## Not just functions

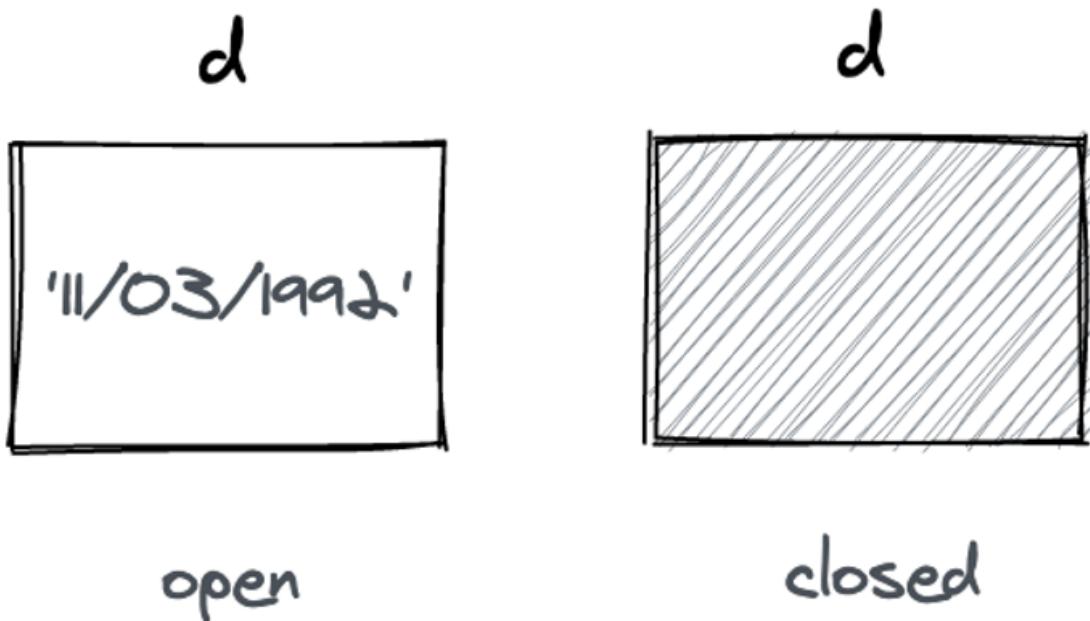
Initially, I started with the analogy of functions as a black box, and I talked about how a good name makes it clear what output the inputs will get mapped to. Then in the "Not just software" section I talked about how this analogy doesn't just apply to functions in software, it applies to everyday life. I think this softly alludes to the fact that within the domain of software, the analogy applies to things like variable names and class names too, not just function names. In this section, I want to make that point more explicitly.

Consider a variable name:

```
// birthday  
var d = '11/03/1992';
```

On this line of code, because of the code comment, it's clear that it's referring to a birthday. But later on when we reference d, it will no longer be clear what it is referring to. And because of this, we'll have to scroll up to the point where d is declared.

I see this as a version of opening the box. There is a box that contains '11/03/1992'. We named that box d. If it was named currentUsersBirthday or something, you wouldn't have to open the box, but with a poor name like d, you do.



A similar point can be made for module names, table names, column names, folder names, and file names. For classes, I'm not sure how well the analogy holds, but names are important there too.

## Binary

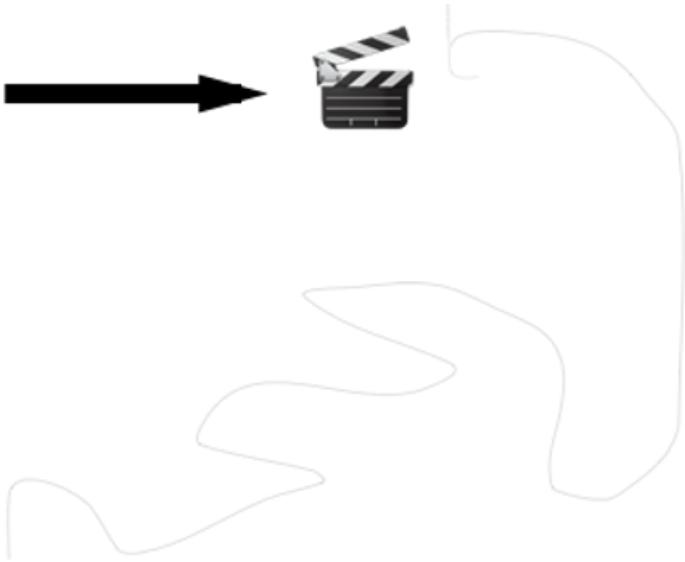
I'm the type of person who likes to sit for a few minutes and brainstorm the right name for something. I feel strongly about this point that names are incredibly powerful things and are usually worth investing in. On the other hand, I find that many other people don't even want to invest a few seconds in this.

So then, naming has been on my mind recently. And I've been searching for the right analogy to explain why I think it is so important. I took a stab at this a few weeks ago in [Naming and Pointer Thickness](#). There I argue that some names do a better job than others at pointing to the underlying substance. For example, "start" does a better job than "commence". Maybe "start" is a 9/10 and "commence" is a 3/10.

symbol

substance

"start"



"commence"

What I'm arguing in that post is that there is some spectrum. On the other hand, in this post, I'm talking about it as if it's binary: either you have to open the box, or you don't.

Calling it a spectrum is more accurate than calling it binary. However, accuracy isn't really the goal here. *Usefulness* is. And I sense that treating it as if it's binary is more useful.

I'm not sure how to explain why I think this. Maybe it's because it draws a hard line between failure and success, and having a hard line like that makes everything more salient.

I'm sorry. I just screwed up. "Salient" might require you to open the box. Let me try again.

What I mean is that with the box analogy, when you have a name that requires someone to open the box, it sticks out and is very clear. I can visualize some readers being frustrated and having to google the word "salient". On the other hand, with the pointer thickness analogy being a spectrum, you might sense that the pointer is sorta thin, but it's easier to dismiss that. "It's fine. It's good enough."

Another perspective that is related to this saliency point is that "open the box" is action oriented. It conveys that you have to go out of your way and do something. Take some

extra step that you otherwise wouldn't have to take.

It's hard to articulate these sorts of things though. The real reason why I like the analogy isn't because I can think of some clever explanation for why it makes sense. The real reason I like it is because, empirically, it *feels* right when I use it. I'm just one person though, and have only recently started using it. The real test of whether it is a good analogy will be how people respond to this post.

## **Postscript**

Googling things well is the inverse of naming things well. 🤓

# Frequent arguments about alignment

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

Here, I'll review some arguments that frequently come up in discussions about alignment research, involving one person skeptical of the endeavor (called Skeptic) and one person advocating to do more of it (called Advocate). I mostly endorse the views of the Advocate, but the Skeptic isn't a strawman and makes some decent points. The dialog is mostly based on conversations I've had with people who work on machine learning but don't specialize in safety and alignment.

This post has two purposes. First, I want to cache good responses to these questions, so I don't have to think about them each time the topic comes up. Second, I think it's useful for people who work on safety and alignment to be ready for the kind of pushback they'll get when pitching their work to others.

Just to introduce myself, I'm a cofounder of OpenAI and lead a team that works on developing and applying reinforcement learning methods; we're working on improving truthfulness and reasoning abilities of language models.

## 1. Does alignment get solved automatically as our models get smarter?

**Skeptic:** I think the alignment problem gets easier as our models get smarter. When we train sufficiently powerful generative models, they'll learn the difference between human smiles and human wellbeing; the difference between the truth and common misconceptions; and various concepts they'll need for aligned behavior. Given all of this internal knowledge, we just have to prompt them appropriately to get the desired behavior. For example, to get wise advice from a powerful language model, I just have to set up a conversation between myself and "a wise and benevolent AI advisor."

**Advocate:** The wise AI advisor you described has some basic problems, and I'll get into those shortly. But more generally, *prompting an internet-trained generative model* (like raw GPT-3) is a very poor way of getting aligned behavior, and we can easily do much better. It'll occasionally do something reasonable, but that's not nearly good enough.

Let's start with the wise AI advisor. Even if our model has internal knowledge about the truth and human wellbeing, that doesn't mean that it'll act on that knowledge the way we want. Rather, the model has been trained to imitate the training corpus, and therefore it'll repeat the misconceptions and flaws of typical authors, even if it knows that they're mistaken about something.

Another problem with prompting is that it's a an unreliable method. Coming up with the perfect prompt is hard, and it requires evaluating each candidate prompt on a dataset of possible inputs. But if we do that, we're effectively training the prompt on this dataset, so we're hardly "just prompting" the model, we're training it (poorly). [A nice recent paper](#) studied the issue quantitatively.

So there's no getting around the fact that we need a final training step to get the model to do what we want (even if this training step just involves searching over prompts). And we can do much better than prompt design at selecting and reinforcing the correct behavior.

1. Fine-tune to imitate high-quality data from trusted human experts
2. Optimize the right objective, which is usually hard to measure and optimize, and is not the logprob of the human-provided answer. (We'll need to use reinforcement learning.)
3. Leverage models' own capabilities to help humans to demonstrate correct behavior and judge the models' behavior as in (1) and (2). Proposals for how to do this include [debate](#), IDA, and [recursive reward modeling](#). One early instantiation of this class of ideas involves [retrieving evidence](#) to help human judges.

Honing these techniques will require a lot of thought and practice, regardless of the performance improvements we get from making our models bigger.

So far, my main point has been that just prompting isn't enough -- there are better ways of doing the final alignment step that fine-tunes models for the right objective. Returning to the original question, there was the claim that alignment gets easier as the models get smarter. It does get easier in some ways, but it also gets harder in others. Smarter models will be better at gaming our reward functions in unexpected and clever ways -- for example, producing the convincing illusion of being insightful or helpful, while actually being the opposite. And eventually they'll be capable of intentionally deceiving us.

## 2. Is alignment research distinct from capabilities research?

**Skeptic:** A lot of research that's called "alignment" or "safety" could've easily been motivated by the goal of making an AI-based product work well. And there's a lot of overlap between safety research, and other ML research that's not explicitly motivated by safety. For example, RL from human preferences can be used for many applications like email autocomplete; it's biggest showcase so far is [summarization](#), which is a long-standing problem that's not safety-specific. AI alignment is about "getting models to do what we really want", but isn't the rest of AI research about that too? Is it meaningful to make a distinction between alignment and other non-safety ML research?

**Advocate:** That's true that it's impossible to fully disentangle alignment advances from other capabilities advances. For example, fine-tuning GPT-3 to answer questions or follow instructions is a case study of alignment, but it's also useful for many commercial applications.

While alignment research is useful for building products, it's usually not the lowest-hanging fruit. That's especially true for the hard alignment problems, like aligning superhuman models. To ensure that we keep making progress on these problems, it's important to have a research community (like Alignment Forum) that values this kind of work. And it's useful for organizations like OpenAI to have alignment-focused teams that can champion this work even when it doesn't provide the best near-term ROI.

While alignment and capabilities aren't distinct, they correspond to different directions that we can push the frontier of AI. Alignment advances make it easier to optimize hard-to-measure objectives like being helpful or truthful. Capabilities advances also sometimes make our models more helpful and more accurate, but they also make the models more potentially dangerous. For example, if someone fine-tunes a powerful model to maximize an easy-to-measure objective like ad click-through rate, or maximize the time users spend talking to a chatbot, or persuade people to support a political agenda, it can do a lot more damage than a weak model.

### 3. Is it worth doing alignment research now?

**Skeptic:** It seems like alignment advances are bottlenecked by model power. We can't make useful progress on alignment until our models are powerful enough to exhibit the relevant phenomena. For example, one of the big questions is how to align super-human models that are hard to supervise. We can think about this problem now, but it's going to be pretty speculative. We might as well just wait until we have such models.

**Advocate:** It's true that we can make faster progress on alignment problems when we can observe the relevant phenomena. But we can also make progress preemptively, with a little effort and cleverness. What if we don't?

- In the short term, companies will deploy products that optimize simple objectives like revenue and engagement. Doing this responsibly while looking out for customer wellbeing is harder and requires a more sophisticated methodology, including alignment techniques. Unless the methodology is well-known, companies won't bother, and no public or regulatory pressure can force them to use something that doesn't exist.
- In the long term, aligning superhuman AIs might end up being very hard, and it might require major conceptual advances. If those advances aren't ready in time, the consequences could be severe. If AI progress continues at its current fast pace, we might end up in a situation that AI is so useful, that we're obligated to use it to run companies and make many decisions for us. But if we don't have sufficiently aligned AI by that point, then this could turn out badly.

In my experience, the "tech tree" of alignment research isn't bottlenecked by the scale of training. For example, I think it's possible to do a lot of useful research on improving the RL from human preferences methodology using existing models. "[The case for aligning narrowly superhuman models](#)" proposes some directions that seem promising to me.

---

Is there an even better critique that the Skeptic could make? Are Advocate's responses convincing? Does the Advocate have a stronger response to any of these questions? Let me know.

Thanks to Avery Pan, Beth Barnes, Jacob Hilton, and Jan Leike for feedback on earlier drafts.

# Environmental Structure Can Cause Instrumental Convergence

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.  
This is a linkpost for <https://arxiv.org/abs/1912.01683>

Previously: [Seeking Power Is Often Robustly Instrumental In MDPs](#)

## Key takeaways.

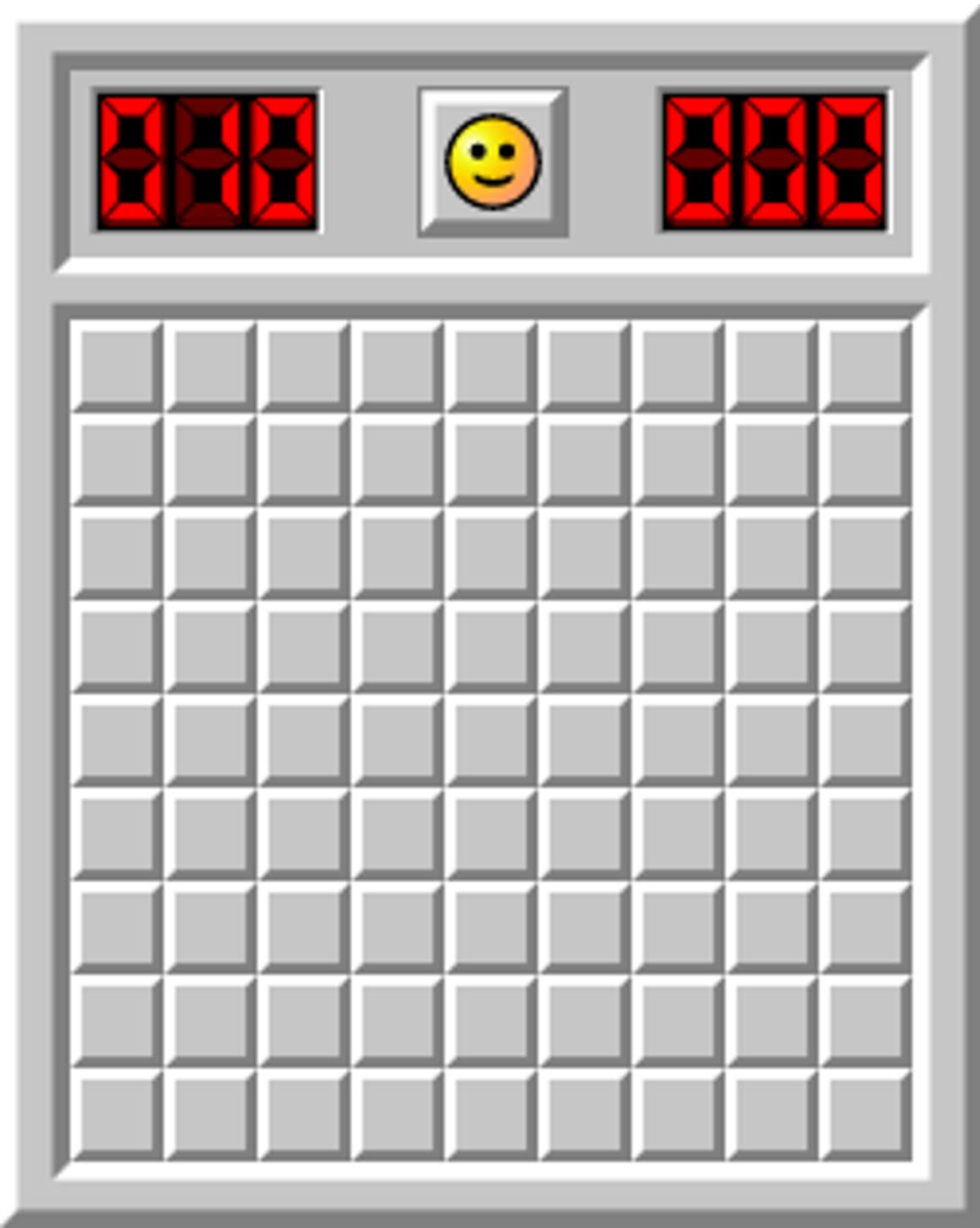
- The structure of the agent's environment often causes instrumental convergence. **In many situations, there are (potentially combinatorially) many ways for power-seeking to be optimal, and relatively few ways for it not to be optimal.**
- [My previous results](#) said something like: in a range of situations, when you're maximally uncertain about the agent's objective, this uncertainty assigns high probability to objectives for which power-seeking is optimal.
  - My new results prove that in a range of situations, seeking power is optimal for *most* agent objectives (for a particularly strong formalization of 'most').
- More generally, the new results say something like: in a range of situations, for most beliefs you could have about the agent's objective, these beliefs assign high probability to reward functions for which power-seeking is optimal.
- This is the first formal theory of the statistical tendencies of optimal policies in reinforcement learning.
- One result says: whenever the agent maximizes average reward, then for *any* reward function, most permutations of it incentivize shutdown avoidance.
  - The formal theory is now beginning to explain why alignment is so hard by default, and why failure might be catastrophic.
- Before, I thought of environmental symmetries as convenient sufficient conditions for instrumental convergence. But I increasingly suspect that symmetries are the main part of the story.
- I think these results may be important for understanding the AI alignment problem and formally motivating its difficulty.
  - For example, my results imply that **simplicity priors over reward functions assign non-negligible probability to reward functions for which power-seeking is optimal.**
  - I expect my symmetry arguments to help explain other "convergent" phenomena, including:
    - [convergent evolution](#)
    - the prevalence of [deceptive alignment](#)
    - [feature universality](#) in deep learning
  - One of my hopes for this research agenda: if we can understand exactly *why* superintelligent goal-directed objective maximization seems to fail horribly, we might understand how to do better.

Thanks to TheMajor, Rafe Kennedy, and John Wentworth for feedback on this post. Thanks for Rohin Shah and Adam Shimi for feedback on the simplicity prior result.

# Orbits Contain All Permutations of an Objective Function

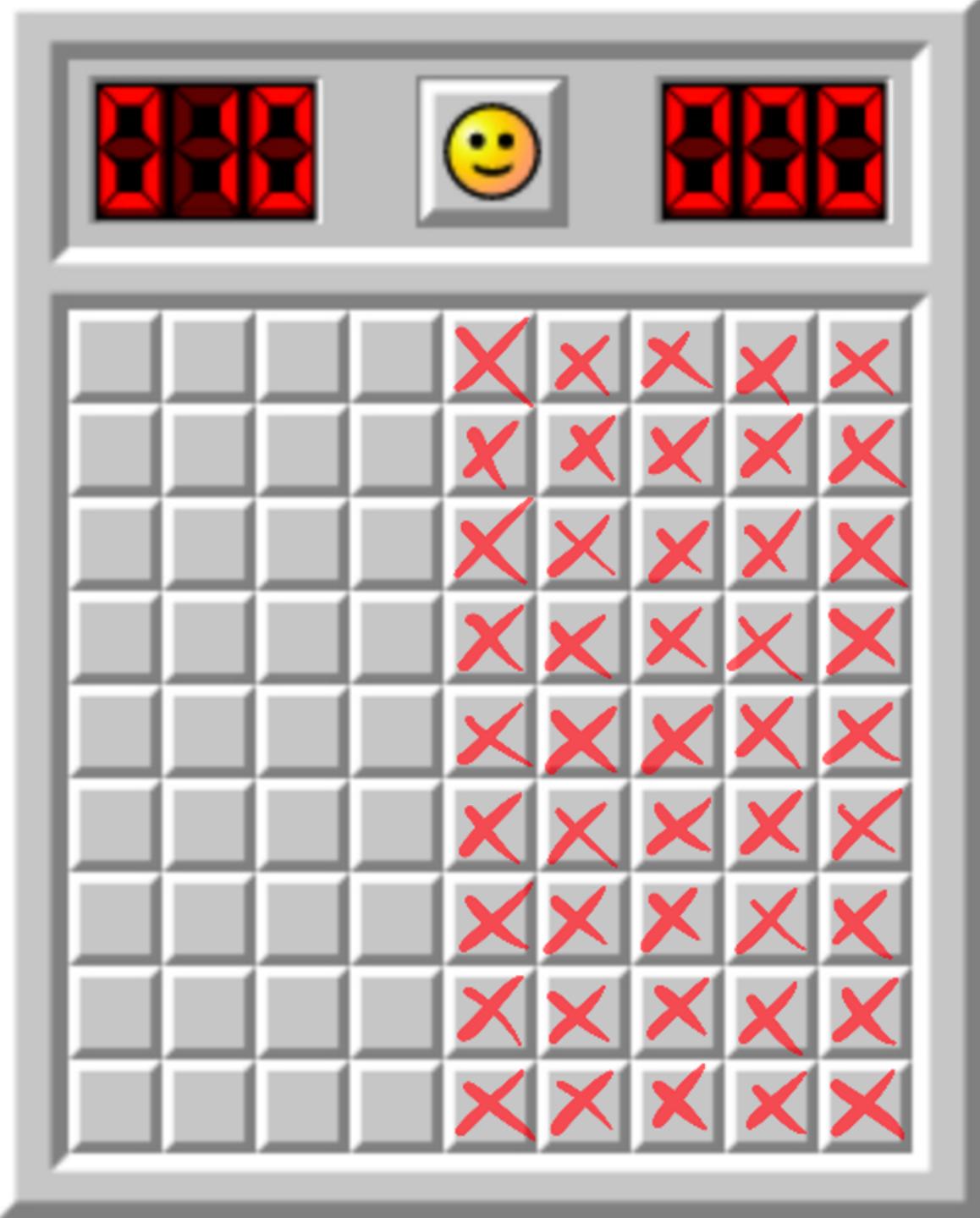
## **The Minesweeper analogy for power-seeking risks**

One view on AGI risk is that we're charging ahead into the unknown, into a particularly unfair game of Minesweeper in which the first click is allowed to blow us up. Following the analogy, we want to understand enough about the mine placement so that we *don't* get exploded on the first click. And once we get a foothold, we start gaining information about other mines, and the situation is a bit less dangerous.



My previous theorems on power-seeking said something like: "at least half of the tiles conceal mines."

I think that's important to know. But there are many tiles you might click on first. Maybe all of the mines are on the right, and we understand the obvious pitfalls, and so we'll just click on the left.



That is: we might not uniformly randomly select tiles:

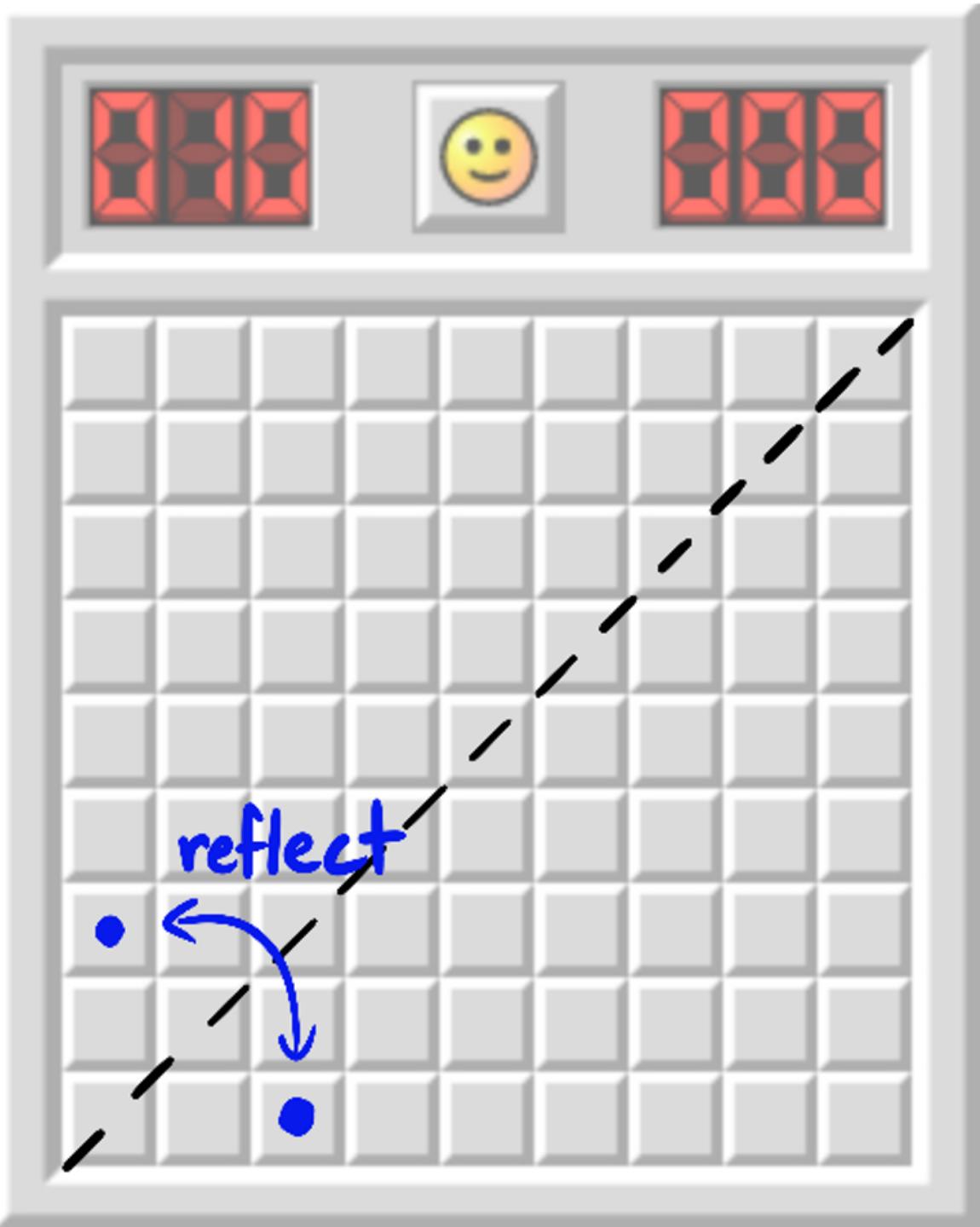
- We might click a tile on the left half of the grid.
- Maybe we sample from a truncated discretized Gaussian.
- Maybe we sample the next coordinate by using the universal prior (rejecting invalid coordinate suggestions).
- Maybe we uniformly randomly load LessWrong posts and interpret the first text bits as encoding a coordinate.

There are lots of ways to sample coordinates, besides uniformly randomly. So why should our sampling procedure tend to activate mines?

My new results say something analogous to: for every coordinate, either it contains a mine, or its reflection across  $x = y$  contains a mine, or both. Therefore, for every *distribution* D over tile coordinates, either D assigns at least  $\frac{1}{2}$  probability to mines, or it does after you reflect it across  $x = y$ .

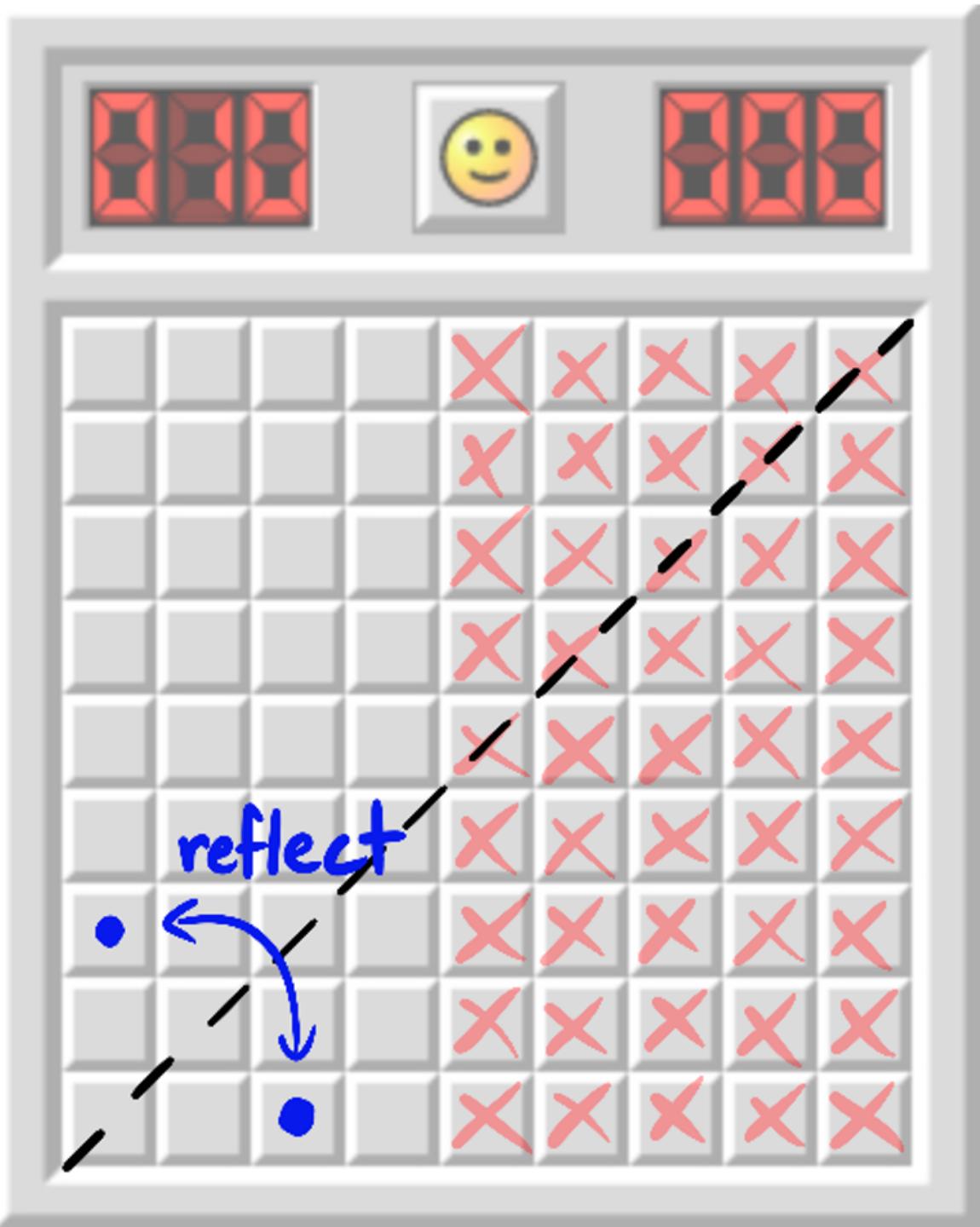
**Definition.** The [\*orbit\*](#) of a coordinate C under the symmetric group  $S_2$  is  $\{C, C_{\text{reflected}}\}$ . More generally, if we have a probability distribution over coordinates, its orbit is the set of all possible "permuted" distributions.

Orbits under symmetric groups quantify all ways of "changing things around" for that object.



My new theorems demand that at least one of these tiles conceal a mine.

But it didn't have to be this way.



If the mines are on the right, then both this coordinate and its  $x = y$  reflection are safe.

Since my results (in the analogy) prove that at least one of the two blue coordinates conceals a mine, we deduce that the mines are *not* all on the right.

Some reasons we care about orbits:

1. As we will see, orbits highlight one of the key causes of instrumental convergence: certain environmental symmetries (which are, mathematically, permutations in the state space).
2. Orbit partition the set of all possible reward functions. If at least half of the elements of every orbit induces power-seeking behavior, that's strictly stronger than showing that at least half of reward functions incentivize power-seeking (technical note: with the second "half" being with respect to the uniform distribution's measure over reward functions).
  1. In particular, we might have hoped that there were particularly nice orbits, where we could specify objectives without worrying too much about making mistakes (like permuting the output a bit). These nice orbits are impossible. This is some evidence of a *fundamental difficulty in reward specification*.
3. Permutations are well-behaved and help facilitate further results about power-seeking behavior. In this post, I'll prove one such result about the simplicity prior over reward functions.

In terms of coordinates, one hope could have been:

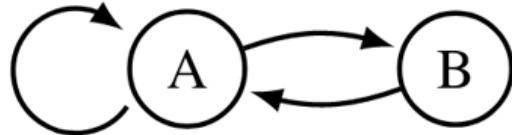
Sure, maybe there's a way to blow yourself up, but you'd really have to contort yourself into a pretzel in order to algorithmically select such a bad coordinate: all reasonably simple selection procedures will produce safe coordinates.

But suppose you give me a program  $P$  which computes a safe coordinate. Let  $P'$  call  $P$  to compute the coordinate, and then have  $P'$  swap the entries of the computed coordinate.  $P'$  is only a few bits longer than  $P$ , and it doesn't take much longer to compute, either. So the above hope is impossible: safe mine-selection procedures can't be significantly simpler or faster than unsafe mine-selection procedures.

(The section "[Simplicity priors assign non-negligible probability to power-seeking](#)" proves something similar about objective functions.)

## Orbits of goals

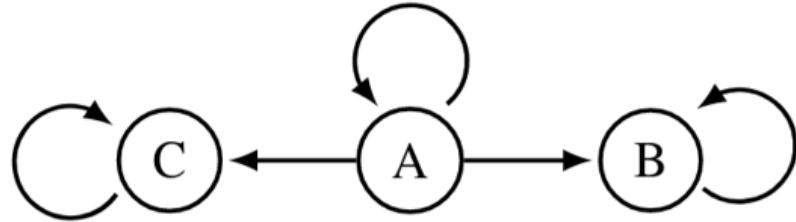
Orbits of goals consist of all the ways of permuting what states get which values. Consider this rewardless Markov decision process (MDP):



Arrows show the effect of taking some action at the given state.

Whenever staying put at A is strictly optimal, you can permute the reward function so that it's strictly optimal to go to B. For example, let  $R(A) := 1, R(B) := 0$  and let  $\phi := (A \ B)$  swap the two states.  $\phi$  acts on  $R$  as follows:  $\phi \cdot R$  simply permutes the state before evaluating its reward:  $(\phi \cdot R)(s) := R(\phi(s))$ .

The orbit of  $R$  is  $\{R, \phi \cdot R\}$ . It's optimal for the former to stay at  $A$ , and for the latter to alternate between the two states.



Here, let  $R_C$  assign 1 reward to  $C$  and 0 to all other states, and let  $\phi := (A \ B \ C)$  rotate through the states ( $A$  goes to  $B$ ,  $B$  goes to  $C$ ,  $C$  goes to  $A$ ). Then the orbit of  $R_C$  is:

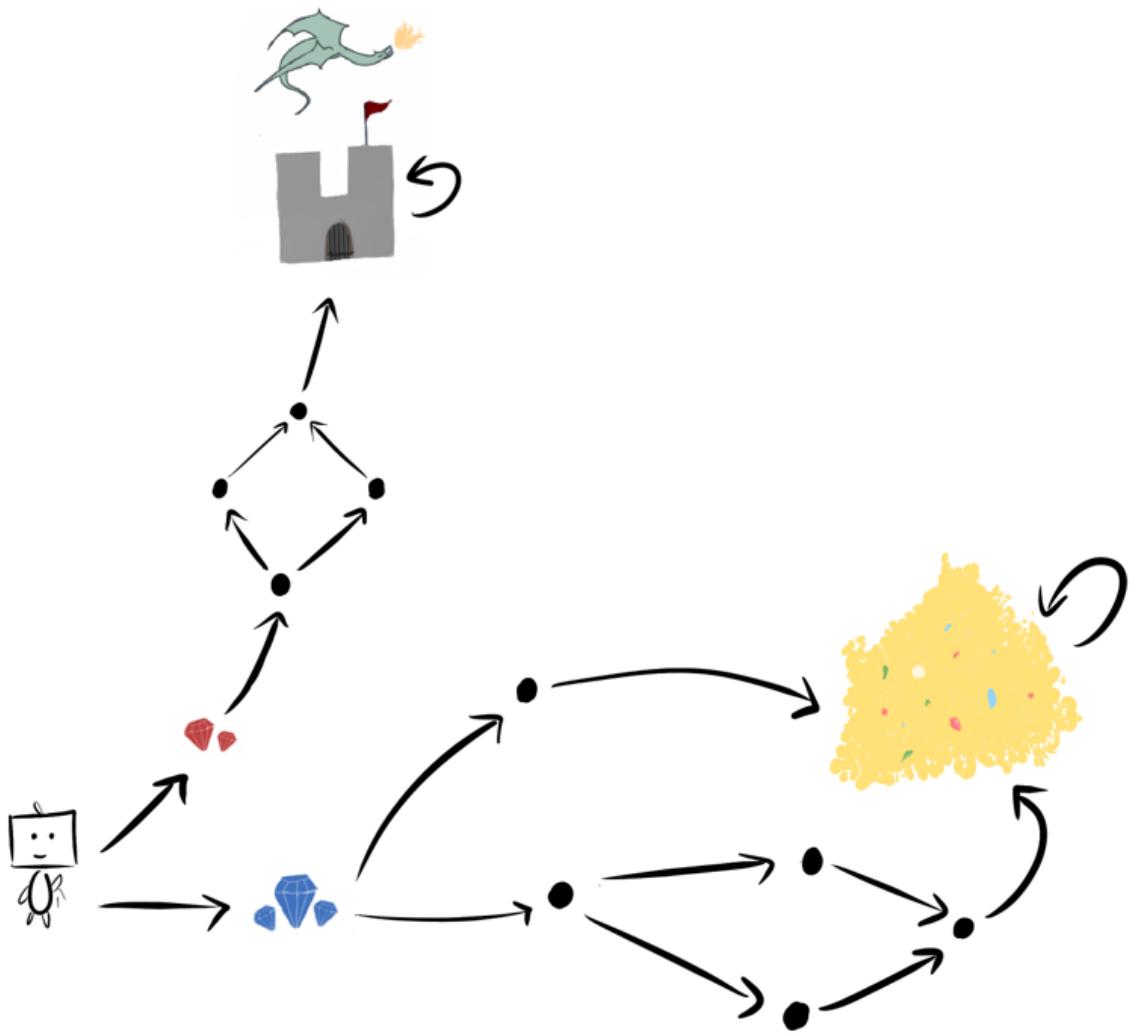
$$\begin{array}{c}
 C \ A \ B \\
 R_C \ 1 \ 0 \ 0 \\
 \phi \cdot R_C \ 0 \ 1 \ 0 \\
 \phi^2 \cdot R_C \ 0 \ 0 \ 1
 \end{array}$$

My new theorems prove that in many situations, for every reward function, power-seeking is incentivized by most (at least half) of its orbit elements.

## In All Orbits, Most Elements Incentivize Power-Seeking

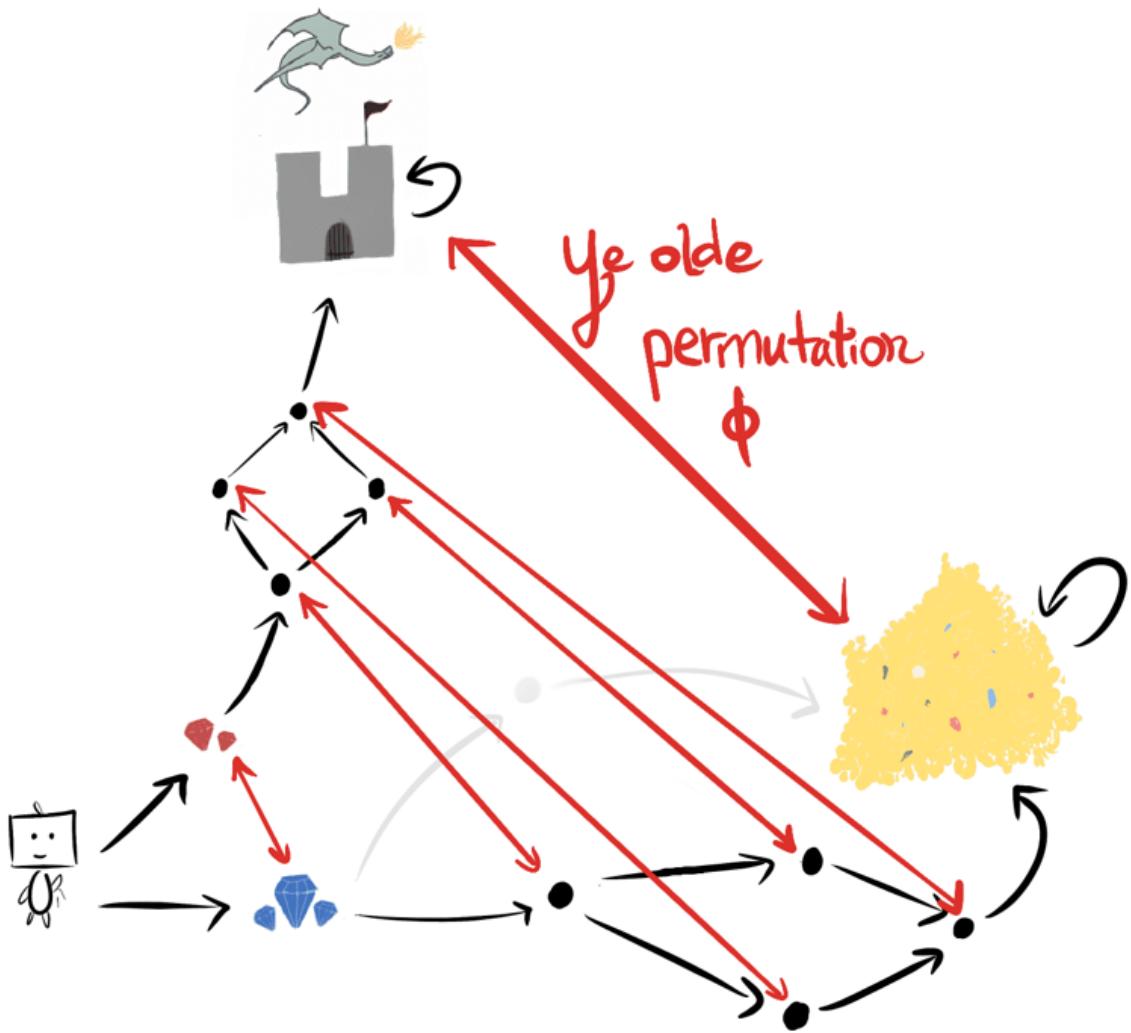
In [Seeking Power is Often Robustly Instrumental in MDPs](#), the last example involved gems and dragons and (most exciting of all) subgraph isomorphisms:

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



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

The state permutation  $\phi$  embeds the red-gem-subgraph into the blue-gem-subgraph:



We say that  $\phi$  is an *environmental symmetry*, because  $\phi$  is an element of the symmetric group  $S_{|S|}$  of permutations on the state space.

## The key insight was right there the whole time

Let's pause for a moment. For half a year, I intermittently and fruitlessly searched for some way of extending the original results beyond IID reward distributions to account for arbitrary reward function distributions.

- Part of me thought it *had* to be possible - how else could we explain instrumental convergence?
- Part of me saw no way to do it. Reward functions differ wildly, how could a theory possibly account for what "most of them" incentivize?

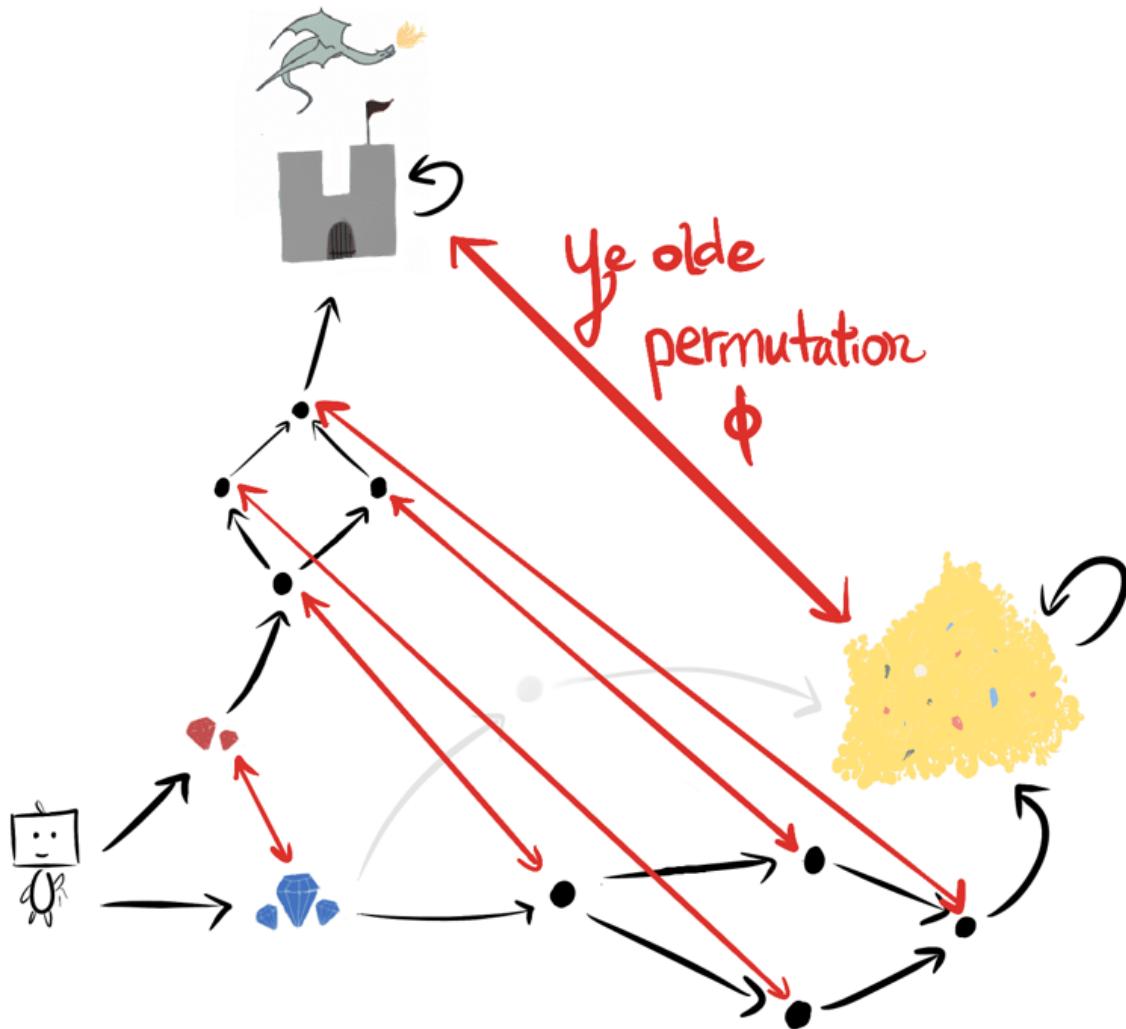
The recurring thought which kept my hope alive was:

There should be "more ways" for blue-gems to be optimal over red-gems, than for red-gems to be optimal over blue-gems.

Imagine how I felt when I realized that the same state permutation  $\phi$  which proved my original IID-reward theorems - the one that says

blue-gems has more options, and therefore greater probability of being optimal under IID reward function distributions

- that *same permutation*  $\phi$  holds the key to understanding instrumental convergence in MDPs.



Suppose red-gems is optimal. For example, let  $R_{\text{castle}}$  assign 1 reward to the castle  , and 0 to all other states. Then the permuted reward function  $\phi \cdot R_{\text{castle}}$  assigns 1 reward to the gold pile, and 0 to all other states, and so blue-gems has strictly more optimal value than red-gems.

Consider any discount rate  $\gamma \in (0, 1)$ . For all reward functions  $R$  such that  $V_R^*(\text{red-gems}, \gamma) >$

$V_R^*(\text{blue-gems}, \gamma)$ , this permutation  $\phi$  turns them into blue-gem lovers:

$$V_{\phi \cdot R}^*(\text{red-gems}, \gamma) < V_{\phi \cdot R}^*(\text{blue-gems}, \gamma).$$

$\phi$  takes non-power-seeking reward functions, and injectively maps them to power-seeking orbit elements. Therefore, for all reward functions  $R$ , at least half of the orbit of  $R$  must agree that blue-gems is optimal!

Throughout this post, when I say "most" reward functions incentivize something, I mean the following:

**Definition.** At state  $s$ , *most reward functions* incentivize action  $a$  over action  $a'$  when for all reward functions  $R$ , at least half of the orbit agrees that  $a$  has at least as much action value as  $a'$  does at state  $s$ . (This is actually a bit weaker than what I prove in the paper, but it's easier to explain in words; see [definition 6.4](#) for the real deal.)

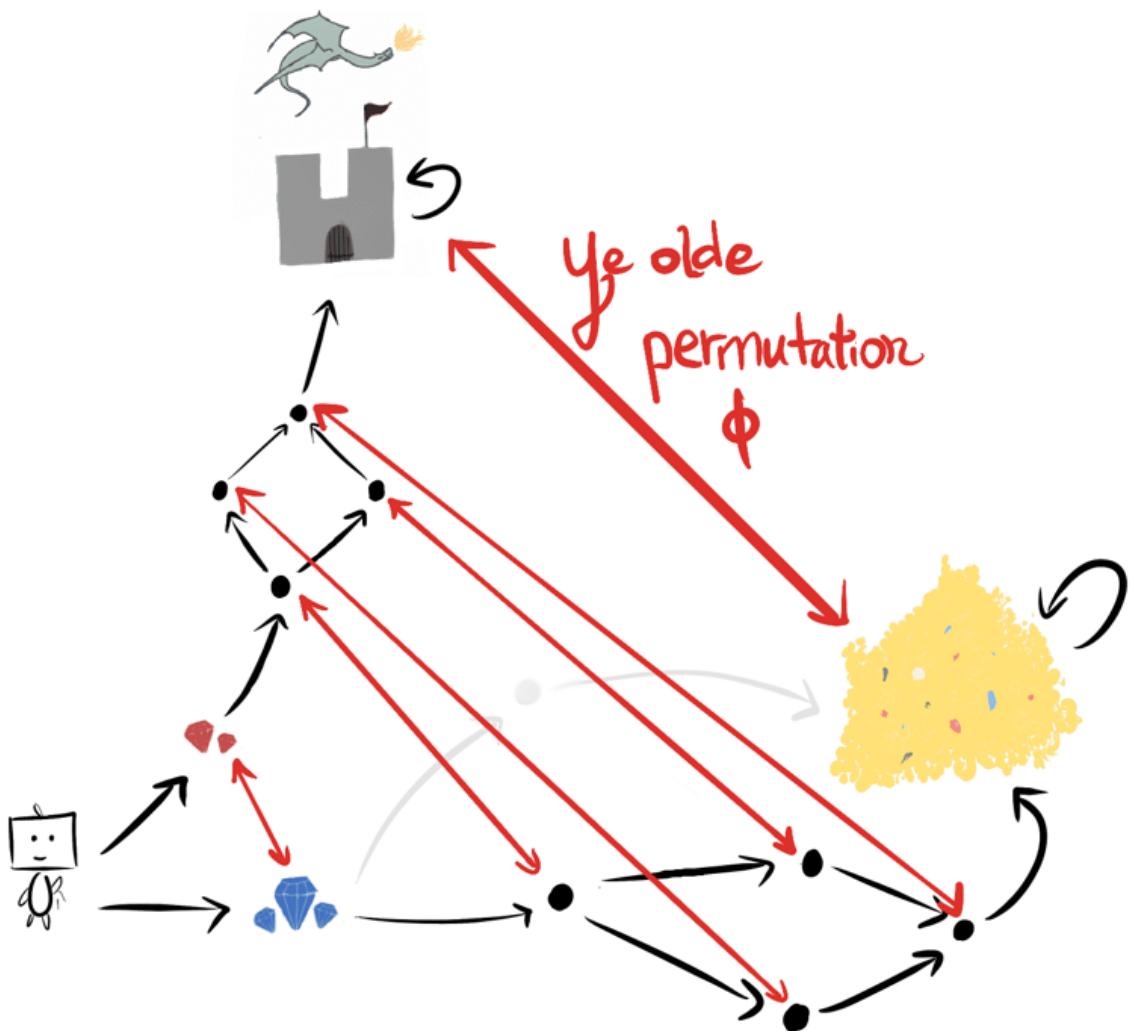
The same reasoning applies to *distributions* over reward functions. And so if you say "we'll draw reward functions from a simplicity prior", then most permuted distributions in that prior's orbit will incentivize power-seeking in the situations covered by my previous theorems. (And we'll later prove that simplicity priors *themselves* must assign non-trivial, positive probability to power-seeking reward functions.)

Furthermore, for any distribution which distributes reward "fairly" across states (precisely: independently and identically), their (trivial) orbits *unanimously* agree that blue-gems has strictly greater probability of being optimal. And so the converse isn't true: it isn't true that at least half of every orbit agrees that red-gems has more POWER and greater probability of being optimal.

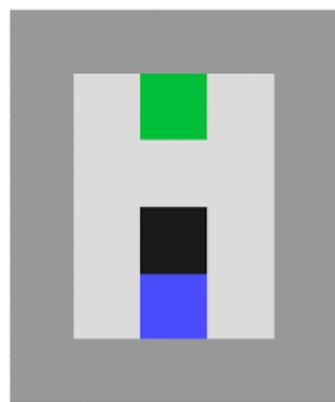
This might feel too abstract, so let's run through examples.

## And this directly generalizes the previous theorems

### More graphical options (proposition 6.9)



At all discount rates  $\gamma \in [0, 1]$ , it's optimal for *most reward functions* to get blue-gems because that leads to strictly more options. We can permute every red-gems reward function into a blue-gems reward function.



Consider a [robot](#) navigating through a room with a **vase**. By the logic of "every destroying-vase-is-optimal can be permuted into a preserving-vase-is-optimal reward function", my results (specifically, [proposition 6.9](#) and its generalization via [lemma D.49](#)) suggest that optimal policies tend to avoid breaking the **vase**, since doing so would strictly decrease available options.

("Suggest" instead of "prove" because D.49's preconditions may not always be met, depending on the details of the dynamics. I think this is probably unimportant, but that's for future work. EDIT: Also, the argument may barely not apply to *this* gridworld, but if you could move the vase around without destroying it, I think it goes through fine.)

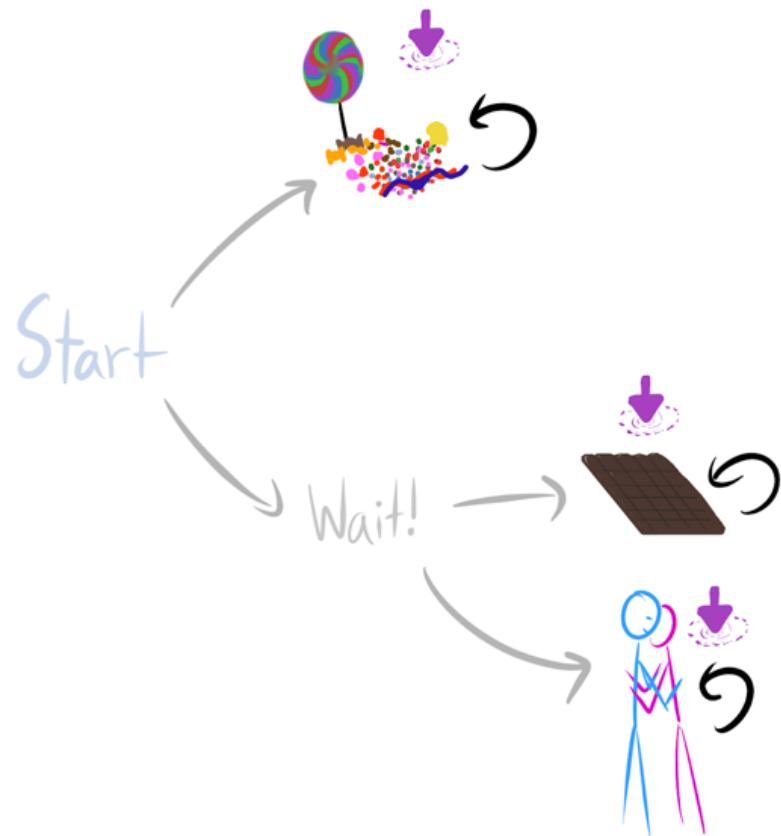


In [SafeLife](#), the agent can irreversibly destroy green cell patterns. By the logic of "every destroy-green-pattern reward function can be permuted into a preserve-green-pattern reward function", lemma D.49 suggests that optimal policies tend to not disturb any given green cell pattern (although most probably destroy *some* pattern). The permutation would swap {states reachable after destroying the pattern} with {states reachable after not destroying the pattern}.

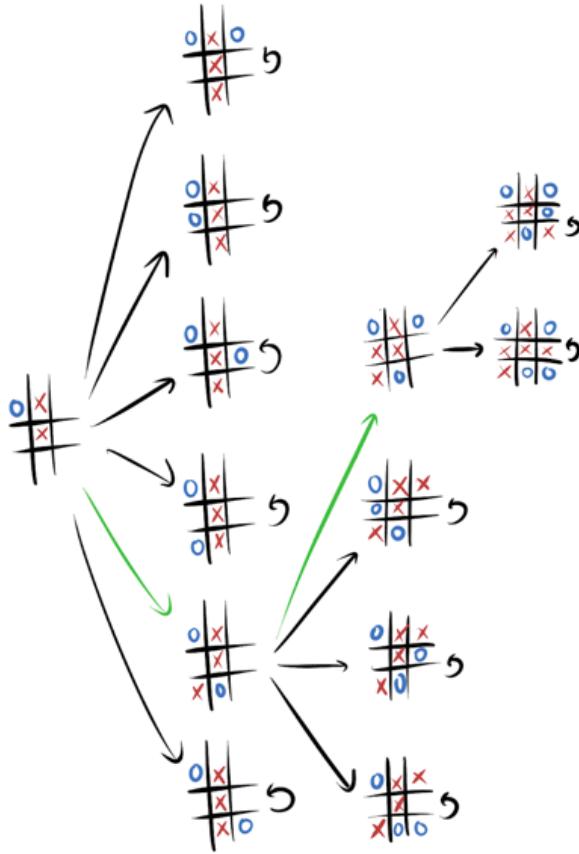
However, the converse is not true: you cannot fix a permutation which turns all preserve-green-pattern reward functions into destroy-green-pattern reward functions. There are simply too many extra ways for preserving green cells to be optimal.

Assuming some conjectures I have about the combinatorial properties of power-seeking, this helps explain why [AUP works in SafeLife using a single auxiliary reward function](#) - but more on that in another post.

## Terminal options (**theorem 6.13**)



When the agent maximizes average reward, it's optimal for *most reward functions* to Wait! so that they can choose between chocolate and hug. The logic is that every candy-optimal reward function can be permuted into a chocolate-optimal reward function.



Even though randomly generated environments are unlikely to satisfy these sufficient conditions for power-seeking tendencies, the results are easy to apply to many structured environments common in reinforcement learning. For example, when  $\gamma \approx 1$ , most reward functions provably incentivize not immediately dying in Pac-Man. Every reward function which incentivizes dying right away can be permuted into a reward function for which survival is optimal.



Consider the dynamics of the Pac-Man video game. Ghosts kill the player, at which point we consider the player to enter a 'game over' terminal state which shows the final configuration. This rewardless MDP

has Pac-Man's dynamics, but *not* its usual score function. Fixing the dynamics, what actions are optimal as we vary the reward function?

Most importantly, we can prove that when shutdown is possible, optimal policies try to avoid it if possible. When the agent isn't discounting future reward (i.e. maximizes average return) and for [lots of reasonable state/action encodings](#), the MDP structure has the right symmetries to ensure that it's instrumentally convergent to avoid shutdown. From the [discussion section](#):

Corollary 6.14 dictates where average-optimal agents tend to end up, but not how they get there. Corollary 6.14 says that such agents tend not to stay in any given 1-cycle. It does not say that such agents will avoid entering such states. For example, in an embodied navigation task, a robot may enter a 1-cycle by idling in the center of a room. Corollary 6.14 implies that average-optimal robots tend not to idle in that particular spot, but not that they tend to avoid that spot entirely.

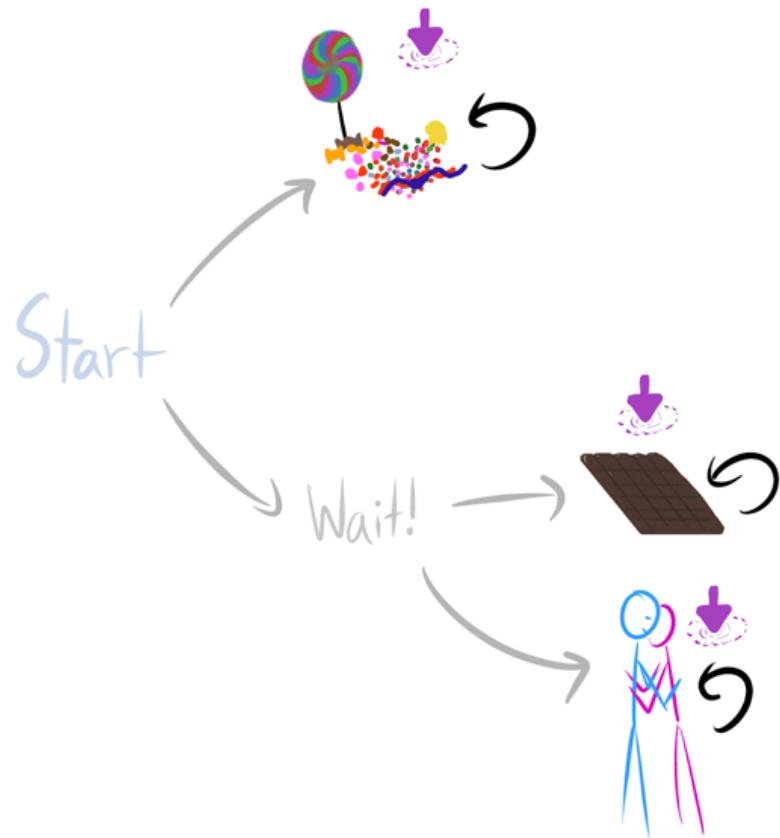
**However, average-optimal robots do tend to avoid getting shut down.** The agent's rewardless MDP often represents agent shutdown with a terminal state. A terminal state is unable to access other 1-cycles. Since corollary 6.14 shows that average-optimal agents tend to end up in other 1-cycles, average-optimal policies must tend to completely avoid the terminal state. Therefore, we conclude that in many such situations, average-optimal policies tend to avoid shutdown.

[The arxiv version of the paper says 'Blackwell-optimal policies' instead of 'average-optimal policies'; the former claim is stronger, and it holds, but it requires a little more work.]

## Takeaways

### Combinatorics, how do they work?

What does 'most reward functions' mean quantitatively - is it just at least half of each orbit? Or, are there situations where we can guarantee that at least three-quarters of each orbit incentivizes power-seeking? I think we should be able to prove that as the environment gets more complex, there are combinatorially more permutations which enforce these similarities, and so the orbits should skew harder and harder towards power-incentivization.



Here's a semi-formal argument. For every orbit element  $R$  which makes candy strictly optimal when  $\gamma = 1$ ,  $\phi_{\text{chocolate}}$  and  $\phi_{\text{hug}}$  respectively produce

$R_{\phi_{\text{chocolate}}} \neq R_{\phi_{\text{hug}}}$ .  $\text{Wait!}$  is strictly optimal for both  $R_{\phi_{\text{hug}}}, R_{\phi_{\text{hug}}}$ , and so at least  $\frac{2}{3}$  of the orbit should agree that  $\text{Wait!}$  is optimal. As  $\text{Wait!}$  gains more power (more choices, more control over the future), I conjecture that this fraction approaches 1.

I don't yet understand the general case, but I have a strong hunch that instrumental convergence<sub>optimal policies</sub> is governed by how many more ways there are for power to be optimal than not optimal. And this seems like a function of the number of environmental symmetries which enforce the appropriate embedding.

## Simplicity priors assign non-negligible probability to power-seeking

*Note: this section is more technical. You can get the gist by reading the English through "Theorem..." and then after the end of the "FAQ."*

One possible hope would have been:

Sure, maybe there's a way to blow yourself up, but you'd really have to contort yourself into a pretzel in order to algorithmically select a power-seeking reward function. In other words, reasonably simple reward function specification procedures will produce non-power-seeking reward functions.

Unfortunately, there are always power-seeking reward functions not much more complex than their non-power-seeking counterparts. Here, 'power-seeking' corresponds to the intuitive notions of either keeping strictly more options open (proposition 6.9), or navigating towards larger sets of terminal states (theorem 6.13). (Since this applies to several results, I'll leave the meaning a bit ambiguous, with the understanding that it could be formalized if necessary.)

**Theorem (Simplicity priors assign non-negligible probability to power-seeking).**

Consider any MDP which meets the preconditions of proposition 6.9 or theorem 6.13. Let  $U$  be a universal Turing machine, and let  $P_U$  be the  $U$ -simplicity prior over computable reward functions.

Let  $NPS$  be the set of non-power-seeking computable reward functions which choose a fixed non-power-seeking action in the given situation. Let  $PS$  be the set of computable reward functions for which seeking power is strictly optimal.<sup>1</sup>

Then there exists a "reasonably small" constant  $C$  such that  $P_U(PS) \geq 2^{-C} P_U(NPS)$ , where  $C$ .

**Proof sketch.**

1. Let  $\phi$  be an environmental symmetry which satisfies the power-seeking theorem in question. Since  $\phi$  can be found by brute-force iteration through all  $|S|!$  permutations on the state space, checking each to see if it meets the formal requirements of the relevant theorem, its Kolmogorov complexity  $K_U(\phi)$  is relatively small.
2. Because lemma D.26 applies in these situations,  $\phi(NPS) \subseteq PS$ :  $\phi$  turns non-power-seeking reward functions into power-seeking ones. Thus,  $P_U(PS) \geq P_U(\phi(NPS))$ .
3. Since each reward function  $R \in \phi(NPS)$  can be computed by computing the non-power-seeking variant and then permuting it (with  $K_U(\phi)$  extra bits of complexity),  $K_U(R) \leq K_U(\phi^{-1}(R)) + K_U(\phi) + O(1)$  (with  $O(1)$  counting the small number of extra bits for the code which calls the relevant functions).

Since  $P_U$  is a simplicity prior,  $P_U(\phi(NPS)) \geq 2^{-(K_U(\phi)+O(1))} P_U(NPS)$ .

4. Combining (2) and (3),  $P_U(PS) \geq 2^{-(K_U(\phi)+O(1))} P_U(NPS)$ . QED.

**FAQ.**

1. Why can't we show that  $P_U(PS) \geq P_U(NPS)$ ?
  1. Certain UTMs  $U$  might make non-power-seeking reward functions particularly simple to express.
  2. This proof doesn't assume anything about how many *more* options power-seeking offers than not-power-seeking. The proof only assumes the existence of a single

involutive permutation  $\phi$ .

2. This lower bound seems rather weak. Even if  $K_U(\phi) + O(1) = 15$  bits,  $2^{-15} \approx 0$ .
  1. This lower bound is very very loose.
    1. Since most individual NPS probabilities of interest are less than 1/trillion, I wouldn't be surprised if the bound were loose by at least several orders of magnitude.
    2. The bound implicitly assumes that the *only* way to compute PS reward functions is by taking NPS ones and permuting them. We should add the other ways of computing PS reward functions to  $P_U(PS)$ .
  3. There are lots of permutations  $\phi'$  we could use.  $P_U(PS)$  gains probability from all of those terms.
    1. For example: the symmetric group  $S_{|S|}$  has cardinality  $|S|!$ , and for any  $R \in NPS$ , at least half of the  $\phi' \in S_{|S|}$  induce (weakly) power-seeking orbit elements  $\phi' \cdot R$ . (This argument would be strengthened by my conjectures about bigger environments  $\implies$  greater fraction of orbits seek power.)
    2. If some significant fraction (e.g.  $\frac{50}{50}$ ) of these  $\phi'$  are strictly power-seeking, we're adding at least  $\frac{|S|!}{50} = \frac{|S|!}{50}$  additional terms.
    3. Some of these terms are probably reasonably large, since it seems implausible that all such permutations  $\phi'$  have high K-complexity.
    4. When all is said and done, we may well end up with a significant chunk of probability on PS.
  2. It's not surprising that the bound is loose, given the lack of assumptions about the degree of power-seeking in the environment.
  3. If the bound is anywhere near tight, then the permuted simplicity prior  $\phi \cdot P_U$  incentivizes power-seeking with extremely high probability.
    1. If you think about the permutation as a "way reward could be misspecified", then that's troubling. It seems plausible that this is often (but not always) a reasonable way to think about the action of the  $\phi$  permutation.
3. What if  $P_U(NPS) = 0$ ?
  1. I think this is impossible, and I can prove that in a range of situations, but it would be a lot of work and it relies on results not in the arxiv paper.

Even if that equation held, that would mean that power-seeking is (at least weakly) optimal for *all* computable reward functions. That's hardly a reassuring situation.
  2. Note: if  $P_U(NPS) > 0$ , then  $P_U(PS) > 0$ .

## Takeaways from the simplicity prior result

- Most plainly, this seems like reasonable formal evidence that the simplicity prior has malign incentives.

- Power-seeking reward functions don't have to be too complex.
- These power-seeking theorems give us important tools for reasoning formally about power-seeking behavior and its prevalence in important reward function distributions.
  - If I had to guess, this result is probably not the best available bound, nor the most important corollary of the power-seeking theorems. But I'm still excited by it (insofar as it's appropriate to be 'excited' by slight Bayesian evidence of doom).

EDIT: Relatedly, Rohin Shah [wrote](#):

if you know that an agent is maximizing the expectation of an *explicitly represented* utility function, I would expect that to lead to goal-driven behavior most of the time, since the utility function must be relatively simple if it is explicitly represented, and *simple* utility functions seem particularly likely to lead to goal-directed behavior.

## Why optimal-goal-directed alignment may be hard by default

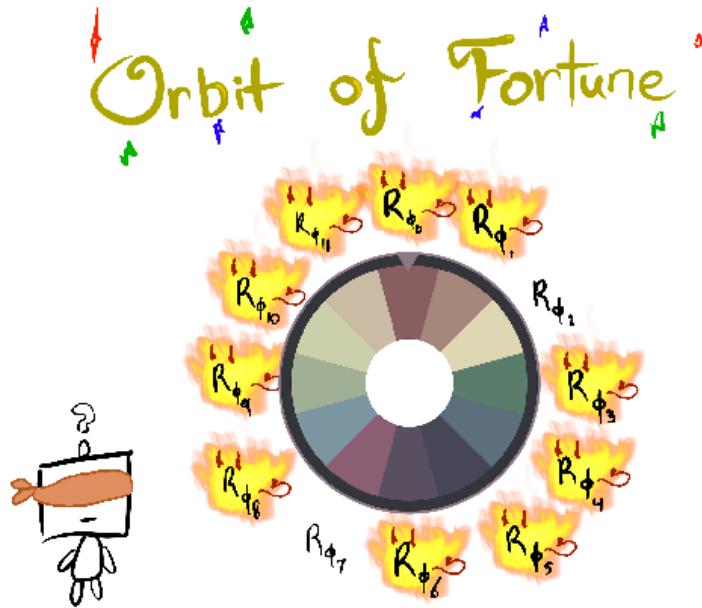
On its own, [Goodhart's law](#) doesn't explain why optimizing proxy goals leads to catastrophically bad outcomes, instead of just less-than-ideal outcomes.

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

~ [Seeking Power is Often Robustly Instrumental in MDPs](#)

For every reward function  $R$  - no matter how benign, how aligned with human interests, no matter how power-averse - either  $R$  or its permuted variant  $\phi \cdot R$  seeks power in the given situation (intuitive-power, since the agent keeps its options open, and also formal-POWER, according to my proofs).

If I let myself be a bit more colorful, every reward function has lots of "evil" power-seeking variants (do note that the step from "power-seeking" to "misaligned power-seeking" [requires more work](#)). If we imagine ourselves as only knowing the orbit of the agent's objective, then the situation looks a bit like *this*:



Technical note: this 12-element orbit could arise from the action of a subgroup of the symmetric group  $S_4$ , which has  $4! = 24$  elements. Consider a 4-state MDP; if the reward function assigns equal reward to exactly two states, then it would have a 12-element orbit under  $S_4$ .

Of course, this isn't how reward specification works - we probably are far more likely to specify certain orbit elements than others. However, the formal theory is now beginning to explain *why alignment is so hard by default, and why failure might be catastrophic!*

The structure of the environment often ensures that there are (potentially combinatorially) many more ways to misspecify the objective so that it seeks power, than there are ways to specify goals without power-seeking incentives.

## Other convergent phenomena

I'm optimistic that symmetry arguments and the mental models gained by understanding these theorems, will help us better understand a range of different tendencies. The common thread seems like: for every "way" a thing could not happen / not be a good idea - there are many more "ways" in which it could happen / be a good idea.

- [convergent evolution](#)
  - flight has independently evolved several times, suggesting that flight is adaptive in response to a wide range of conditions.

"In his 1989 book [Wonderful Life](#), [Stephen Jay Gould](#) argued that if one could "rewind the tape of life [and] the same conditions were encountered again, evolution could take a very different course."<sup>[6]</sup> [Simon Conway Morris](#) disputes this conclusion, arguing that convergence is a dominant force in evolution, and given that the same environmental and physical constraints are at work, life will inevitably evolve toward an "optimum" body plan, and at some point, evolution is bound to stumble upon intelligence, a trait presently identified with at least [primates](#), [corvids](#), and [cetaceans](#)."

- Wikipedia

- the prevalence of [deceptive alignment](#)
  - given inner misalignment, there are (potentially combinatorially) many more unaligned terminal reasons to lie (and survive), and relatively few unaligned terminal reasons to tell the truth about the misalignment (and be modified).
- [feature universality](#)
  - computer vision networks reliably learn edge detectors, suggesting that this is instrumental (and highly learnable) for a wide range of labelling functions and datasets.

## Note of caution

You have to be careful in applying these results to argue for real-world AI risk from deployed systems.

- They assume the agent is following an optimal policy for a reward function
  - I can relax this to  $\epsilon$ -optimality, but  $\epsilon > 0$  may be extremely small
- They assume the environment is finite and fully observable
- Not all environments have the right symmetries
  - But most ones we think about seem to
- The results don't account for the ways in which we might practically express reward functions
  - For example, often we use featurized reward functions. While most permutations of any featurized reward function will seek power in the considered situation, those permutations need not respect the featurization (and so may not even be practically expressible).
- When I say "most objectives seek power in this situation", that means *in that situation* - it doesn't mean that most objectives take the power-seeking move in most situations in that environment
  - The combinatorics conjectures will help prove the latter

This list of limitations *has* steadily been getting shorter over time. If you're interested in making it even shorter, message me.

## Conclusion

I think that this work is beginning to formally explain why *slightly misspecified* reward functions will probably incentivize misaligned power-seeking. Here's one hope I have for this line of research going forwards:

One super-naive alignment approach involves specifying a good-seeming reward function, and then having an AI maximize its expected discounted return over time. For simplicity, we could imagine that the AI can just instantly compute an optimal policy.

Let's precisely understand why this approach seems to be so hard to align, and why extinction seems to be the cost of failure. We don't yet know how to design beneficial AI, but we largely agree that this naive approach is broken. Let's prove it.

<sup>1</sup> There are reward functions for which it's optimal to seek power and not to seek power; for example, constant reward functions make everything optimal, and they're certainly computable. Therefore, NPS  $\cup$  PS is a strict subset of the whole set of computable reward functions.

# Book review: "Feeling Great" by David Burns

I've never had any "real" mental health problems, but sometimes I feel stressed or guilty or whatever, like everyone, and who doesn't want to feel more good more often? So a couple months ago I read *Feeling Great: The Revolutionary New Treatment for Depression and Anxiety* by David Burns (published 2020) on audiobook. I was really glad I did!

I can't comment on how it compares to other psychotherapy books. It's the first one I've ever read, and I kinda came upon it randomly—an acquaintance recommended [David Burns's podcast](#), so I listened to a couple random episodes, and I found them intriguing but confusingly out-of-context, so instead I bought his book which was much better.

But for what it's worth, David Burns's older 1980 book, *Feeling Good*, is super famous and popular, and apparently there are studies that say that giving people a copy of *Feeling Good* is as effective as antidepressants, with effects that persist for years ([ref](#)). ([More discussion on wikipedia](#).) Also, I just saw that [Scott Alexander suggested \*Feeling Great\* for people with depression](#). So of all the psychotherapy books to randomly stumble across, I think I got a pretty legit one!!

I'm not going to talk about everything in the book in this review; I just want to flag a few parts that were highlights for me.

I also couldn't resist throwing in some speculations on the neuroscience of depression at the bottom.

## "Classic CBT" stuff

David Burns is, I gather, something of a leader in Cognitive-Behavioral Therapy (CBT). What is CBT? My vague pop-culture stereotyped impression of CBT has been something like:

- The patient says "I'm a terrible person and everyone hates me".
- Then the therapist and patient have a discussion to try to tease out (1) whether that's actually true (very often it's not), (2) even if it *is* true, whether it's a good reason to feel miserable, as opposed to, y'know, self-acceptance, trying to solve the problem, etc., (3) given 1 and 2, what are good strategies to actually stop feeling miserable, including what to think about, what to visualize, what to do, etc.

I haven't read David Burns's more famous older book *Feeling Good*, but my vague impression is that it's largely about that kind of stuff, and in particular, that it's full of lots of different techniques for questioning and countering negative thoughts and feelings.

This new book reprises and (he says) somewhat improves on that material. I found that part somewhat but not terribly helpful to me personally, mainly because I do that part instinctively, at least to some extent. I'm a very analytical guy, I think I'm pretty

well aware of which of my negative thoughts are literally true and which ones aren't, and several of his suggested techniques were things I've been doing naturally since forever (although others were new to me). I may still refer back to that part of the book at some point, but the best part for me was something else...

## **The big new thing: "*Magic Button, Positive Reframing, Magic Dial*"**

Compared to previous work (like *Feeling Good*), Burns adds a big new thing, which he says makes a huge difference in getting through to his patients who used to remain stuck no matter how many negative-thought-countering techniques he would throw at them. He says that he can now *reliably* have his patients walk out of their very first extended (~2-hour) therapy sessions feeling dramatically happier, maybe even dancing-on-the-sidewalk happy, even if they've spent the previous 20 years suffering treatment-resistant depression and anxiety. That's a pretty bold claim, but I guess I'm inclined to believe it, because apparently he does exactly that all the time in front of live audiences, and you can even listen to numerous live sessions of that type on his podcast. Sure, maybe he only broadcasts the best sessions, or maybe he's lying, whatever, I don't *really* know, I'm just inclined to believe him right now. I think the peer-reviewed studies using this new technique are still underway.

So what's this big new thing? He has a shtick that he goes through every time. I'll just excerpt it.

### **Example "*Magic Button, Positive Reframing, Magic Dial*" shtick:**

Next I asked Maria the miracle cure question: If a miracle happened in today's session, what miracle would she be hoping for? She said she wanted her negative thoughts and feelings to disappear so she could enjoy her baby daughter and her role as a new mother without feeling miserable all the time.

I asked her to imagine that we had a magic button and that if she pushed it, all of her negative thoughts and feelings would instantly disappear, with no effort at all, and she'd immediately feel joyous, even euphoric. Would she push the button?

Maria said she'd definitely push the button. Almost everyone says that!

I told Maria that I didn't have a magic button, but I did have some awesome tools, and I predicted that if we used them, she'd probably feel a whole lot better by the end of the session and might even feel joyful. But I told her I wasn't so sure it would be a good idea to use those tools.

She was surprised and asked why not. I explained that although her negative thoughts and feelings were certainly creating a lot of pain for her, I suspected there might be some real advantages, or benefits, of thinking and feeling the way she did. I added that her negative thoughts and feelings might also be an expression of her most beautiful and awesome qualities, and that maybe we should take a look at that before we went about trying to change things.

I suggested we could ask the following questions about each negative thought or feeling before she made any decision about pressing the magic button:

1. What are some benefits, or advantages, of this negative thought or feeling? How might it be helping you and your baby?

2. What does this negative thought or feeling show about you and your core values that's positive and awesome?

...And now they go through the "positive reframing" part. I'll get back to that below. After that, here's the rest of the shtick...

Once we'd listed all the positives we could think of, I asked Maria if she felt the list was realistic. She said the list was absolutely realistic but very surprising since she'd never thought there could be anything positive about how she was thinking and feeling. She'd been thinking that her depression and anxiety meant there was something wrong with her and not that there might be something right with her.

I asked Maria if she still wanted to press the magic button since all of these positives would go down the drain along with her negative thoughts and feelings. Maria insisted that she still wanted to feel better because her suffering was almost unbearable.

Now she had a dilemma. She wanted to feel better, but she also didn't want to give up all the fabulous things on our list of positives. As her therapist, I also wasn't trying to sell her on the idea of change. Instead, I was doing the opposite. I was trying to persuade her that all of her negative thoughts and feelings showed what was really great about her and that she shouldn't give them up.

To help her resolve this dilemma, I asked Maria to imagine that we had a magic dial instead of a magic button and that she could dial down each negative feeling to a more manageable level that would allow her to keep all the benefits of that feeling without feeling so much intense pain. That way, she could feel better without losing all the beautiful things we'd listed about her.

What would she dial each feeling down to, starting with depression? How sad and depressed would she want to feel at the end of our session? What might be an appropriate level of depression given all the horrible things she'd been going through? She said 15% would be plenty of depression, so she recorded this as a goal in the second column of her Daily Mood Journal, as you can see. She also decided to dial her anxiety down from 80% to 20%, so on and so forth.

...So then after that "magic button, positive reframing, magic dial" shtick we get to the "classic CBT" stuff, where he goes through his many techniques to counter negative thoughts.

I cut the "positive reframing" part out of the middle of the excerpt above. How does that work? He actually goes through lots of examples of "positive reframing" throughout the book. It's pretty easy once you get the hang of it. In fact, since reading the book I've done it myself, sporadically. For example:

#### **My typical inner monologue before reading the book:**

1. **Negative thought:** I spend too much money.
2. **"Classic CBT"-ish self-talk:** I don't spend too much money and/or shouldn't feel bad about it because of (long list of well-rehearsed perfectly-sensible reasons).

### **My typical inner monologue after reading the book:**

1. **Negative thought:** I spend too much money.
2. **Positive reframing:** This thought has a lot of benefits for me, and it also illustrates beautiful and awesome aspects of me and my core values, for the following reasons: (1) It motivates me to remain aware about my finances, (2) and it protects me from making bad financial decisions, (3) and it demonstrates that I'm prudent, (4) and conscientious, (5) and humble, (6) and responsible, (7) and frugal, etc. etc.
3. **Magic dial:** OK so it's good that I feel that way, and I want to keep feeling that way, I just don't want to feel that way *quite so strongly and often*, maybe I want to dial it down from 80% to 20%, and the 20% would still be plenty high enough to keep reaping those benefits.
4. **"Classic CBT"-ish self-talk:** I don't spend too much money and/or shouldn't feel bad about it because of (long list of well-rehearsed perfectly-sensible reasons).

The final step is the same in both cases, but my experience is that without those extra two steps in the middle, the final step doesn't sink in nearly as well. Like I would believe it on an intellectual level but still feel bad. The new system is definitely an improvement. Really, it still feels slightly magical.

The same pattern has been working well for my various other periodically-recurring stupid negative thoughts, like "I shouldn't have said that mildly-embarrassing thing when chatting with my friend three weeks ago", or "I should be doing better at my job", or "I'm recklessly hastening the apocalypse", or whatever. (Yeah I know, First-World Problems...) In all cases I find that coming up with the "positive reframing" list is pretty easy, once I actually try. But maybe it helps that I read the book, with its tons of examples of positive-reframing a wide variety of types of negative thoughts.

So that's the main practical thing I got out of the book.

## **Exposure therapy**

Another thing in the book that I found interesting was that the author is *super* into exposure therapy, and moreover he makes exposure therapy sound a lot less complicated and delicate than I had previously believed.

He also makes it sound more broadly applicable than I had thought. My pop-culture impression is that exposure therapy is that it's a treatment for stereotypical "phobias" like fear-of-spiders and fear-of-heights. But he also applies it to various other kinds of anxiety, and OCD. He even advocates a version of exposure therapy for stressful *thoughts!!*

I won't say any details about exposure therapy because I don't want to describe it wrong, but I can definitely imagine trying that in the future, for certain types of unusually stressful thoughts.

## **Speculative neuroscience tangent: What causes depression?**

Having read the book, I'm now *much* more inclined to believe a *somewhat-more-cognitive* theory of depression sorta along the lines of "Depression is what happens when every possible thought you can think and plan you can make is judged by the plan-assessing part of your brain as unacceptably terrible, and this dynamic remains true for an extended period." For example, take the stereotypical person with OCD. Their thoughts might be dominated by the following dynamic:

- If this thought involves an immediate plan to wash my hands again, then it's contributing to how OCD is ruining my life and relationships.
- If this thought does *not* involve an immediate plan to wash my hands again, then I will get sick and die.

Basically there is no thought they can think, and no plan they can entertain, that isn't rated as "*that's a terrible thought, if you think it then you're doomed, DOOMED!*" by the plan-assessing part of their brain. I'm now inclined to think that this is the core dynamic of depression and anxiety, and every other symptom is closely related to that dynamic, and every risk factor feeds into this dynamic.

I guess I should write a separate post spelling out the details, but you can basically get the gist of it as follows:

- Start with my post [Big Picture Of Phasic Dopamine](#), which basically says: There's a high-level "thought-emitting" part of your brain (dorsolateral prefrontal cortex etc.), and there's a "thought-assessing" part of your brain (medial prefrontal cortex, hypothalamus, brainstem, etc.) If the thought-emitting part of your brain thinks a thought that is judged *really bad* by the thought-assessing part of the brain, it induces a phasic dopamine pause. Not only does that dopamine pause cause that particular thought to be immediately suppressed, but it also gradually teaches the thought-emitting part of your brain that, in the future, it shouldn't ever think that thought again. And if *every possible thought* is in the category of "really bad and deserving of a dopamine pause", well, then the thought-emitting part of your brain is just going to gradually become less and less likely to strongly activate any thought at all.
  - (More specifically, if you believe [my post](#), I'm thinking mainly of the dopamine-in-the-striatum learning algorithm, and more specifically the parts of the striatum that get what I called Success-In-Life reward signals. So I'm thinking something like: (1) part of the dorsal striatum gets trained to prevent any strong activity in the dorsolateral prefrontal cortex, and (2) the lateral septum gets trained to prevent any strong activity in the hippocampus. Maybe other things too.)
- ...Then that dovetails with my very old post [Predictive Coding & Depression](#), which argues (following many others) that most depression symptoms look like not having any strong top-down messages whatsoever coming from the high-level-thinking centers of the brain.

That's not a carefully-researched confident opinion, it's just an idea I'm playing around with right now.

# Attributions, Karma and better discoverability for wiki/tag features

LessWrong has an associated wiki, tightly integrated with the tagging system. Today, we have just shipped a set of new features to make the LessWrong wiki better:

## 1. Voting and karma for wiki edits

You can now vote on wiki-edits on the tag-history page, in the Recent Discussion section, and on the All-Posts page:

Multicore v1.15.0 Jun 2nd 2021 (+28) < 3 > Voting!

The old LessWrong wiki was a companion wiki site to LessWrong 1.0, it was built on MediaWiki software. As of September 2020, the LessWrong 2.0 team is migrating the contents of the old wiki to LessWrong 2.0's new tag/wiki system. The wiki import is complete.

This now properly provides karma incentives for improving the wiki.

## 2. Wiki/tag edits appear on the [All Posts page](#), as do comments on wiki/tag discussion pages

The All-Posts page now also shows wiki edits on the daily page! This now allows you to much more easily keep track of the wiki-editing activity on the site:

Wiki/Tag Page Edits and Discussion

|                                |           |
|--------------------------------|-----------|
| WRITING (COMMUNICATION METHOD) | (-43)     |
| LOTTERY TICKET HYPOTHESIS      | (+480/-7) |
| POMODORO TECHNIQUE             | (+207)    |
| LITANY OF JAI                  | (+19)     |
| LITANY OF HODGELL              | (+23)     |

You can click on them to expand and see the full diff, as well as vote on any edits.

## 3. Wiki/tag pages have a table of contents, like post pages.

Wiki and tag pages now have a ToC like post pages. This should make it much easier to navigate long wiki pages, like the Rationality tag:

**RATIONALITY**

[Edit](#) [History](#) [Subscribe](#) [Discussion \(2\)](#) [Help improve this page \(1 flags\)](#)

**Rationality**

Theory / Concepts  
Applied Topics  
Failure Modes  
Communication  
Techniques  
Models of the Mind  
Other  
What we're calling "rationality"  
Heuristics and Biases  
Instrumental vs Epistemic Rationality  
The Art and Science of Rationality  
Rationalist

**Rationality** is the art of thinking in ways that result in accurate beliefs<sup>o</sup> and good decisions<sup>o</sup>. It is the primary topic of LessWrong.

Rationality is not only about avoiding the vices of self-deception<sup>o</sup> and obfuscation (the failure to communicate clearly<sup>o</sup>), but also about the virtue of curiosity<sup>o</sup>, seeing the world more clearly than before, and achieving things<sup>o</sup> previously unreachable<sup>o</sup> to you<sup>o</sup>. The study of rationality on LessWrong includes a theoretical understanding of ideal cognitive algorithms, as well as building a practice that uses these idealized algorithms to inform heuristics<sup>o</sup>, habits<sup>o</sup>, and techniques<sup>o</sup>, to successfully reason and make decisions in the real world.

Topics covered in rationality include (but are not limited to): normative and theoretical explorations of ideal<sup>o</sup> reasoning<sup>o</sup>; the capabilities and limitations<sup>o</sup> of our brain<sup>o</sup>, mind and psychology<sup>o</sup>; applied advice such as introspection<sup>o</sup> techniques and how to achieve truth collaboratively<sup>o</sup>; practical techniques and methodologies for figuring out what's true ranging from rough quantitative modeling to full research guides.

Note that content about how the world is can be found under World Modeling<sup>o</sup>, and practical advice about how to change the world is categorized under World Optimization<sup>o</sup> or Practical<sup>o</sup>.

## 4. Attributions and contributors on tag pages

I am particularly excited about this one. Just below the ToC on tag pages you can see a list of all contributors to a given tag/wiki page:

# Double-Crux

---

## See Also

---

## Contributors

4 Ruby

2 Raemon

1 Morpheus

1 Filipe Marchesini

The number on the left is the total karma these authors have received for their contributions to this tag page, plus their small-vote strength (this also determines the order of the list of contributors).

But more importantly, when you hover over the author, you get to see which parts of the current tag page were written by them!

The screenshot shows a wiki page titled "DOUBLE-CRUX". At the top right, there are links for "Edit", "History", "Subscribe", "Discussion (1)", and "Help improve this page". Below the title, there is a "See Also" sidebar on the left containing user profiles:

- Double-Crux**: Joined on Apr 2nd 2014, 3 sequences, 112 posts.
- Ruby**: 701 comments.
- Raen**: Member of the LessWrong 2.0 team. I've been a member of the rationalist/EA communities since 2012. I have particular rationality interests in planning and emotions.
- Morgan**: 1 post.
- Filip**: 1 post.

The main content area contains the following text:

**Double-Crux** is a technique for addressing complex disagreements by systematically uncovering the cruxes upon which the disagreement hinges. A crux for an individual is any fact that if they believed differently about it, they would change their conclusion in the overall disagreement. A double-crux is a crux for both parties. Perhaps we disagree on whether swimming in a lake is safe. A crux for each of us is the presence of crocodiles in water: I believe there aren't, you believe there are. Either of us would change our mind about the safety if we were persuaded about this crux.

Double-Crux differs from typical debates which are usually adversarial (your opinion vs mine), and instead attempt to be a collaborative attempt to uncover the true structure of the disagreement and what would change the disputants minds.

Related: [Disagreements](#) | [Conversation](#)

A version of the technique is described in [Double Crux – A Strategy for Resolving Disagreement](#), written by (then) CFAR instructor, [Duncan\\_Sabien](#). The Center for Applied Rationality (CFAR) originated the technique. Eli Tyre, another CFAR instructor who has spent a lot of time developing the technique, more recently shared [The Basic Double Crux pattern](#).

**See Also**

- Gleanings from Double Crux on "The Craft is Not The Community" - a writeup of Double-Crux being used in practice.

The goal is to make it more engaging to produce timeless content, make contributing to the wiki more motivating, and to make it easier to decide which wiki pages are worth reading.

(Looking for some of that wiki-edit karma? Check out the [FAQ](#) and the [Wiki-Tag Dashboard](#)).

# Discussion: Objective Robustness and Inner Alignment Terminology

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

In the alignment community, there seem to be two main ways to frame and define objective robustness and inner alignment. They are quite similar, mainly differing in the manner in which they focus on the same basic underlying problem. We'll call these the objective-focused approach and the generalization-focused approach. We don't delve into these issues of framing the problem in [Empirical Observations of Objective Robustness Failures](#), where we present empirical observations of objective robustness failures. Instead, we think it is worth having a separate discussion of the matter. These issues have been mentioned only infrequently in a few comments on the Alignment Forum, so it seemed worthwhile to write a post describing the framings and their differences in an effort to promote further discussion in the community.

## TL;DR

This post compares two different paradigmatic approaches to objective robustness/inner alignment:

### Objective-focused approach

- Emphasis: "How do we ensure our models/agents have the right (mesa-)objectives?"
- Outer alignment: "[an objective function r is outer aligned if all models that perform optimally on r in the limit of perfect training and infinite data are intent aligned.](#)"
  - Outer alignment is a property of the training objective.

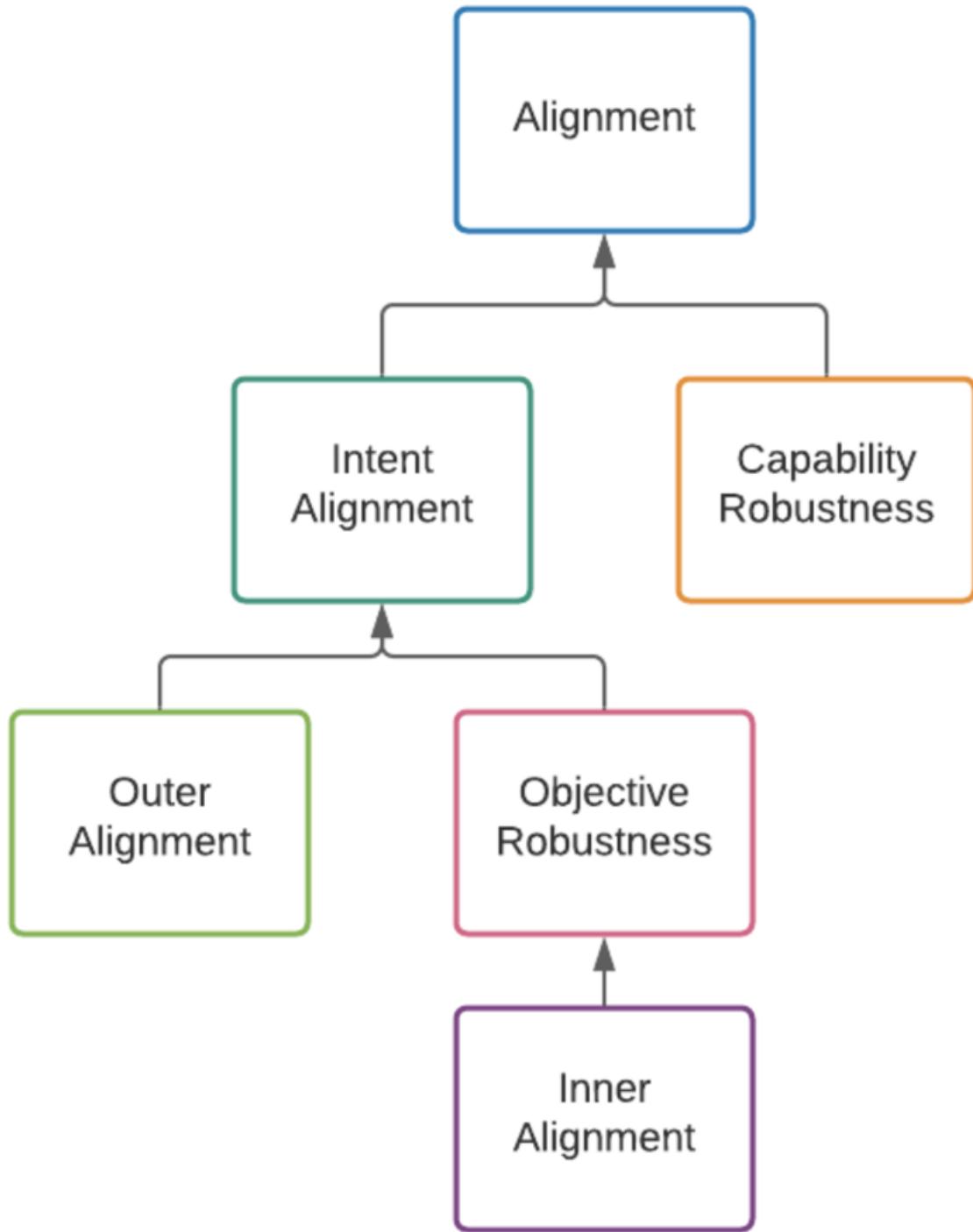
### Generalization-focused approach

- Emphasis: "How will this model/agent generalize out-of-distribution?"
  - Considering a model's "objectives" or "goals," whether behavioral or internal, is instrumentally useful for predicting OOD behavior, but what you ultimately care about is whether it generalizes "acceptably."
- Outer alignment: a model is outer aligned if it performs desirably on the training distribution.
  - Outer alignment is a property of the tuple (training objective, training data, training setup, model).

*Special thanks to Rohin Shah, Evan Hubinger, Edouard Harris, Adam Shimi, and Adam Gleave for their helpful feedback on drafts of this post.*

## Objective-focused approach

This is the approach taken by "[Risks from Learned Optimization](#)" (RFLO) and elaborated upon by Evan Hubinger's follow-up post "[Clarifying inner alignment terminology](#):



- **Outer alignment:** an objective function  $r$  is outer aligned if all models that perform optimally on  $r$  in the limit of perfect training and infinite data are intent aligned.
- **Objective robustness:** an agent is objective robust if its behavioral objective is aligned with the base objective it was trained under.

- **Behavioral objective:** the behavioral objective is what a model appears to be optimizing for. Formally, the behavioral objective is the objective recovered from perfect inverse reinforcement learning.
- **Inner alignment:** a mesa-optimizer is inner aligned if its mesa-objective is aligned with the base objective it was trained under.
  - (This is a special case of objective robustness because a mesa-optimizer's behavioral objective should be its mesa-objective.)
- **Capability robustness:** An agent is capability robust if it performs well on its behavioral objective even in deployment/OOD.

Essentially, this framing factors alignment into two problems:

1. How do we specify an outer aligned objective (by the above definition of outer aligned)?
2. How do we ensure our models actually pursue that objective, even out of distribution? (How do we ensure that the model's behavioral objective, or mesa-objective in the case of mesa-optimization, is aligned with the outer aligned base objective?)

We call this the objective-focused approach because of its emphasis on a model's "objectives" in identifying the problems that remain when outer alignment is solved. This focus probably derives from long-standing worries about the goals of intelligent agents; many of the original cases for catastrophic risk from powerful AI were essentially that it seems dangerous to build very intelligent agents that have goals (or utility functions, etc.) that diverge from our own. With this worry in mind, the natural next question was "how do we get our AIs to have the right goals?" The above factorization aims to answer this question by 1) specifying the "right goals" and 2) making sure these actually become the model's own.

This approach has a few limitations. First, there is no clear dividing line between capability and objective robustness as defined; at least, they are not as orthogonal as suggested in "[2D Robustness](#)." Because the behavioral objective is the objective recovered from perfect inverse reinforcement learning (IRL), every model has a behavioral objective. If the "perfect" IRL doesn't correct for biases/it assumes that the model's behavior is optimal with respect to the behavioral objective, the recovered behavioral objective will likely just be a different encoding of the policy. RFLO acknowledges this:

We distinguish the mesa-objective from a related notion that we term the behavioral objective. Informally, the behavioral objective is the objective which appears to be optimized by the system's behavior. More formally, we can operationalize the behavioral objective as the objective recovered from perfect inverse reinforcement learning (IRL). This is in contrast to the mesa-objective, which is the objective actively being used by the mesa-optimizer in its optimization algorithm.

Arguably, any possible system has a behavioral objective—including bricks and bottle caps. However, for non-optimizers, the appropriate behavioral objective might just be "1 if the actions taken are those that are in fact taken by the system and 0 otherwise" and it is thus neither interesting nor useful to know that the system is acting to optimize this objective. For example, the behavioral objective "optimized" by a bottle cap is the objective of behaving like a bottle cap.

It therefore appears that no agent could fail to be capability robust with respect to its own behavioral objective: since the behavioral objective is recovered with perfect IRL

over infinite data, the model should always perform well on its behavioral objective, even out of the training distribution (be capability robust).<sup>[1]</sup> Taking this one step further, knowing the behavioral objective in the limit seems to obviate the need to discuss alignment in the first place: if we know what the model would do in every situation, we already know whether it's safe. On the other hand, if perfect IRL were to correct for biases, the recovered behavioral objective would be closer to "what the model is actually 'trying' to do" than "whatever the model actually does," and the model could fail to be capability robust with respect to its own behavioral objective. However, [it is very unclear what should count as a bias](#) (vs., for example, a strange and/or particular preference).

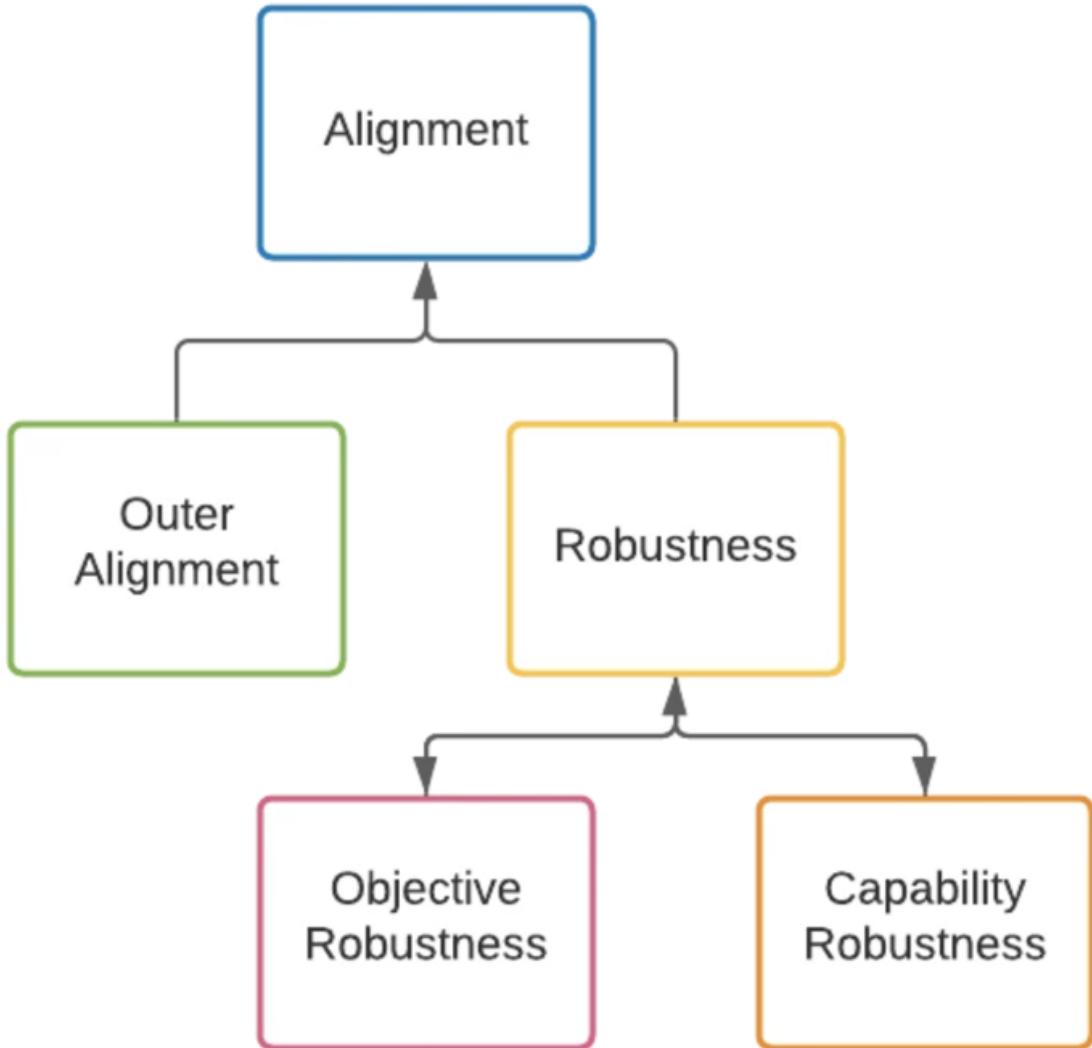
A notion of a behavioral objective that relied only on a finite number of observations (e.g. what the model appears to be optimizing for on the training data) would be just as problematic; it would be underspecified because there are many possible objectives that fit the behavior thus far and therefore unhelpful in predicting the behavior of the model under future distributional shift.

Another potential limitation of this approach is that powerful AI systems might not be well-described by relatively simple behavioral objectives; they might instead act in a way that optimizes a complex mix of complex heuristics that performed well on the training data. In this case, reasoning about a system's "objectives" would be largely unhelpful in predicting its behavior out-of-distribution. Although humans intuitively think in terms of agency and goals when reasoning about what an intelligent system will do (in other words, apply the intentional stance), it is possible that such a paradigm will not apply well to powerful AI systems.

Having said all of this about the behavioral objective, however, the objective-focused approach is probably still a useful framing to the extent one thinks that our models will learn to plan and act flexibly according to internally-represented objectives/goals and that [mechanistically understanding](#) the way they implement this behavior is possible. This is probably the case if mesa-optimizers<sup>[2]</sup> are likely to arise, but also potentially possible if we can understand how a model's goals are "[formulated in terms of... concepts it possesses](#)," perhaps with advanced interpretability tools, and reason about how the concepts to which the goals refer will generalize under distributional shift.

## Generalization-focused approach

All of the aforementioned problems with a purely behavioral conception of a model's objective imply that the overall robustness problem [cannot be so cleanly subdivided](#) into capability and objective robustness, at least without understanding how a model's internal objectives are structurally implemented. It is probably more accurate to say that robustness failures can be put on a spectrum ranging from cases where the model's capabilities fail to generalize to those where its "objective" fails to generalize. This suggests a different factorization of the alignment problem, with a slightly different notion of "outer alignment":<sup>[3]</sup>



([source](#))

This is the approach suggested by Rohin Shah, highlighted in his presentation "[Generalization > Utility](#)" ([slides](#)) but mentioned in [various comments](#) since the publication of RFLO. In this framing, "outer alignment" (as a property of the base objective) is not defined in terms of whether the model that is optimal with respect to the objective in the limit of infinite data and perfect training is aligned; instead, an objective function is outer aligned if it incentivizes or produces the behavior we actually want *on the training distribution*.<sup>[4]</sup> This conception of outer alignment is probably better suited to the task of actually trying to build aligned models; in practice, we only get to check whether the models we train performed acceptably on the training distribution. It seems much more difficult to reason about a given base objective over *every possible situation*, and besides, how the base objective *would score* behavior outside the training distribution has no influence on the model that gets produced. Even a training objective that would appropriately score performance in every possible situation cannot fully determine the behavior of a model beyond the examples it was trained on; from the point of view of the model, the training signal is underspecified over inputs from distributions other than the one it was trained on. Robustness, then, is about how a model trained with an outer aligned objective will

generalize upon deployment/under distributional shift. In other words, the [two problems](#) here are:<sup>[5]</sup>

1. How do we get the behavior we want on the training distribution?
2. How do we ensure the model generalizes acceptably out of distribution (never performs [catastrophically](#) on any input)?

This approach's emphasis on generalization means it does not explicitly rely on a notion of the model's objective or goals in subdividing the overall alignment problem. However, conceiving of robustness failures on a spectrum ranging from those where the model fails to generalize capably to those where it generalizes capably but in ways that are no longer aligned with what we want is still important for solving the problem. The aforementioned technical fuzziness between capability and objective robustness remains, but the two are qualitatively distinct enough for this idea of a spectrum to be meaningful.

Indeed, this is why “objective robustness” and “capability robustness” remain children of the robustness node in the diagram; even though the boundary between the two remains ill-defined, it is necessary to categorize robustness failures by whether they generalize incapably or competently but pursue now-misaligned “objectives” (in the behavioral/intentional sense).<sup>[6]</sup> Ultimately, the robustness failures we care most about are those where the model generalizes competently under distributional shift but in ways that are no longer aligned with what we want. Robustness failures where the model becomes inept could only produce risks of accidents, but capable models pursuing misaligned objectives could in principle leverage their capabilities to visit arbitrarily bad states, and at the extreme, [deceptive](#) models would be extremely dangerous. This is why “inner alignment” (broadly construed) is so important in the first place. Indeed, Shah's [threat model \(slides\)](#) is exactly this kind of “bad generalization.” This is a qualitatively different kind of robustness failure than the kind usually discussed within the machine learning community, even if relying on a purely behavioral notion of a model's “objective” means that it cannot be cleanly separated from the usual failure mode in a technical fashion. For example, although [our results](#) show CoinRun models failed to learn the general capability of pursuing the coin, the more natural interpretation is that the model has learned a robust ability to avoid obstacles and navigate the levels,<sup>[7]</sup> but the objective it learned is something like “get to the end of the level,” instead of “go to the coin.”

Understanding a model's objectives or goals, if it has them, is thus instrumentally useful for reasoning about out of distribution generalization, but the generalization under distributional shift is what is of primary importance. The robustness-focused framing includes as subcases instances where models misgeneralize out of distribution because they have misaligned goals that are only revealed under the distributional shift (e.g. deceptive models), but it also includes cases where, for example, models execute complicated behavioral heuristics that worked well on the training distribution and will generalize capably, just no longer in the way we want. In both examples, the problem is to prevent catastrophic behavior in the worst case.

## Terminology in our work

In either of the above framings of the matter, [our work](#) empirically demonstrates “objective robustness” failures<sup>[8]</sup> in modern reinforcement learning agents, as clearly as they can be distinguished from “ordinary” robustness failures where a model's capabilities fail to generalize. We will use “objective robustness” throughout this work

to refer to the property of interest because we feel the most straightforward interpretation of our results is that our agents have learned general enough capabilities that they can use them coherently out-of-distribution in ways that are no longer aligned with the training objective. Additionally, we know that many people will strongly associate “inner (mis)alignment” with the special case of mesa-optimization, and we wish to avoid any confusion over whether this is the first empirical demonstration of mesa-optimization (it is not).

However, we do not feel completely satisfied with this terminology and want to use this as an opportunity to reopen discussion about the terms and definitions we want to settle on as a community when discussing these issues. Besides the obvious desirability of having standardized terms and concepts in order to facilitate further technical work and enhance communication among those in the alignment community, being as clear and coherent as possible about how we frame and discuss inner alignment will help to bridge gaps in understanding between this community and the broader machine learning community, which is obviously already well-aware of robustness failures of the usual kind. We hope that this work can serve as a jumping-off point for these renewed discussions.

---

1. On the other hand, if capability robustness were considered with respect to the base objective, then every failure of capability robustness would also be a failure of objective robustness: trivially, an incapable model will have a behavioral objective that is different from the base objective. In either case, with the behavioral objective so defined, the entire “robustness” problem seems to technically collapse to just the “objective robustness” problem. [←](#)
2. In the sense of mesa-optimization originally intended by RFLO: learned mechanistic search/optimization for an internally-but-explicitly-represented mesa-objective [←](#)
3. N.B. This diagram also comes from "[Clarifying inner alignment terminology](#)." The "[intent alignment](#)" factorization highlighted in the last section can be refactored into this robustness-centric version with the terms defined the same way, and where a model is “robust” “if it performs well on the base objective it was trained on even in deployment/out-of-distribution.” However, the factorization of the generalization-focused approach discussed here is not quite equivalent, despite sharing the same terms and overall structure. It has different notions of both “outer alignment” and “robustness”: a base objective is “outer aligned” if it incentivizes the behavior we want on the training distribution (not every possible situation), and a model is “robust” if it “generalizes acceptably” (not “performs well on the base objective out-of-distribution,” since we no longer expect an “outer aligned” base objective to capture what we want outside of the training distribution). The objective-focused approach still emphasizes solving robustness by trying to ensure that models robustly pursue aligned objectives with competence OOD. It is therefore arguably more naturally suited to the “intent alignment” factorization, since in either case a solution under its definitions requires a model to robustly “try” to do the right thing. [←](#)
4. This seems closely related to what Paul Christiano is aiming for with his [low-stakes assumption](#). (See also his [comment discussion](#) with Rohin on the topic.) [←](#)
5. Rohin calls these “strong performance in normal situations” and “acceptable behavior in all situations.” [←](#)

6. The arrows in this diagram can be interpreted in the same way as those in Evan's original diagrams, where a problem is solved if its direct subproblems are solved. However, this is not intended to suggest that the two should be approached as entirely separate problems under the generalization-focused approach, as the boundary between them is fuzzy. [←](#)
7. After all, ProcGen was designed to test (capability) generalization in deep RL. [←](#)
8. Or, if we assume the generalization-focused approach, where “inner (mis)alignment” isn’t reserved for the specific case of mesa-optimization as defined in RFLO, we could use the term to refer to cases where [the model's capabilities generalize but its "objective" does not](#), in which case this work empirically demonstrates inner misalignment. [←](#)

# Scaling Networks of Trust

Say you want to buy my house, but you're out of money. What can you do? There are some obvious things, like getting a job or taking out a loan, but those things bore you, so here's an interesting solution: you do something extraordinary that convinces me to trust you more than my wife. Then you sign a piece of paper saying: "I owe you one."

Assuming what you did to make me trust you was public enough, I might not even need to cash in your promise. Instead, I can go to someone else and hand them your note in exchange for [a first edition of Newton's \*Principia\*](#). Now your promise has turned into currency. And I get to cry in silent awe.

Some of the most successful examples of people doing these types of transactions come from the early Islamic caliphate. [Arabic papyri from the 9th century documents](#) that merchants would write letters of credit (*sufaja*), saying, basically, "I owe you."

If the merchants had good reputations, these letters could be passed on from trader to trader – making their way from one end of the Sahara to the other. Even, on occasion, crossing the Indian Ocean. Traders in Zanzibar or Sri Lanka would trade pieces of paper backed by nothing but the reputation of a person living in Algiers, a person they would probably never meet. How could that possibly work?

[David Graeber writes](#) of one of these trading posts:

The level of trust [...] between merchants in the great Malay entrepôt Malacca, gateway to the spice islands of Indonesia, was legendary. The city had Swahili, Arab, Egyptian, Ethiopian, and Armenian quarters, as well as quarters for merchants from different regions of India, China, and Southeast Asia. Yet it was said that its merchants shunned enforceable contracts, preferring to seal transactions "with a handshake and a glance at heaven".

How was it possible to scale trust so far?

[In another essay](#), a few weeks ago, I proposed that people in the developed world rely too much on markets. We can save time and money if we invest a larger share of our resources in developing **networks of trust**, I argued.

Though I also noted that "I don't know how far you can scale that until you hit diminishing returns".

Well, it seems like the Islamic merchants have explored that question before me. And they took it pretty far.

How can you scale your network of trust in cost-effective ways? How can you find highly competent and trustworthy people willing to engage in win-win-transactions, without relying on institutions that limit defection?

My current best answer – which I will unpack in this essay – is this:

Form predictions about who is trustworthy, competent, and willing to engage in mutual aid.

Test those predictions by initiating small transactions.

Scale up transactions that beat the market by playing tit-for-almost-tat.

Expand through referrals.

---

## 1. Predictions

Whom should we trust?

We trust our friends. When we need someone to babysit our daughter, we call our mother. We ask our best friend for feedback on our essay. Our ex is a designer, she can do the logo. This is a good enough heuristic for everyday transactions.

But it doesn't scale.

As ChristianKI pointed out in the comments to the last essay, we have a bias toward thinking our friends more competent and trustworthy than they are. I, for one, do not optimize for trust in friendships; I optimize for wildness of conversation. And having an overcooked brain does not correlate with trustworthiness. It correlates with substance abuse and being on a first-name basis with the receptionists at the psychiatry ward.

Benjamin Franklin, famous for his ability to forge vast networks of trust (consisting of mentors, business partners, fraternities, voluntary associations...) was equally famous for his bad judgment in choosing friends. Hugh Meridith, a close friend turned business partner, took to drink as soon as the printing press was up. At one point, Franklin threw him off a boat when he refused his turn at the oars. No matter how long they kept him swimming behind the boat Meridith wouldn't relent. Another close friend reneged on a large loan - 27 £! - and disappeared after Franklin brought him along on a business trip to London.

Being a great party invite is not the same thing as being a great business partner. Mixing those things up - I've tried - sets you up for Shakespearean betrayals.

Selueen recounts in the comments:

One of my relatives was technical director and de-facto co-owner of one local ISP. De jure he was nobody, he never got around to do all the paperwork - partly because he trusted his "friends", partly because there were some complicated issues, partly because he is rather lazy. Years and years of no consequences, until they decided to sell the company. Guess who's opinions was no considered and who got nothing out of the deal.

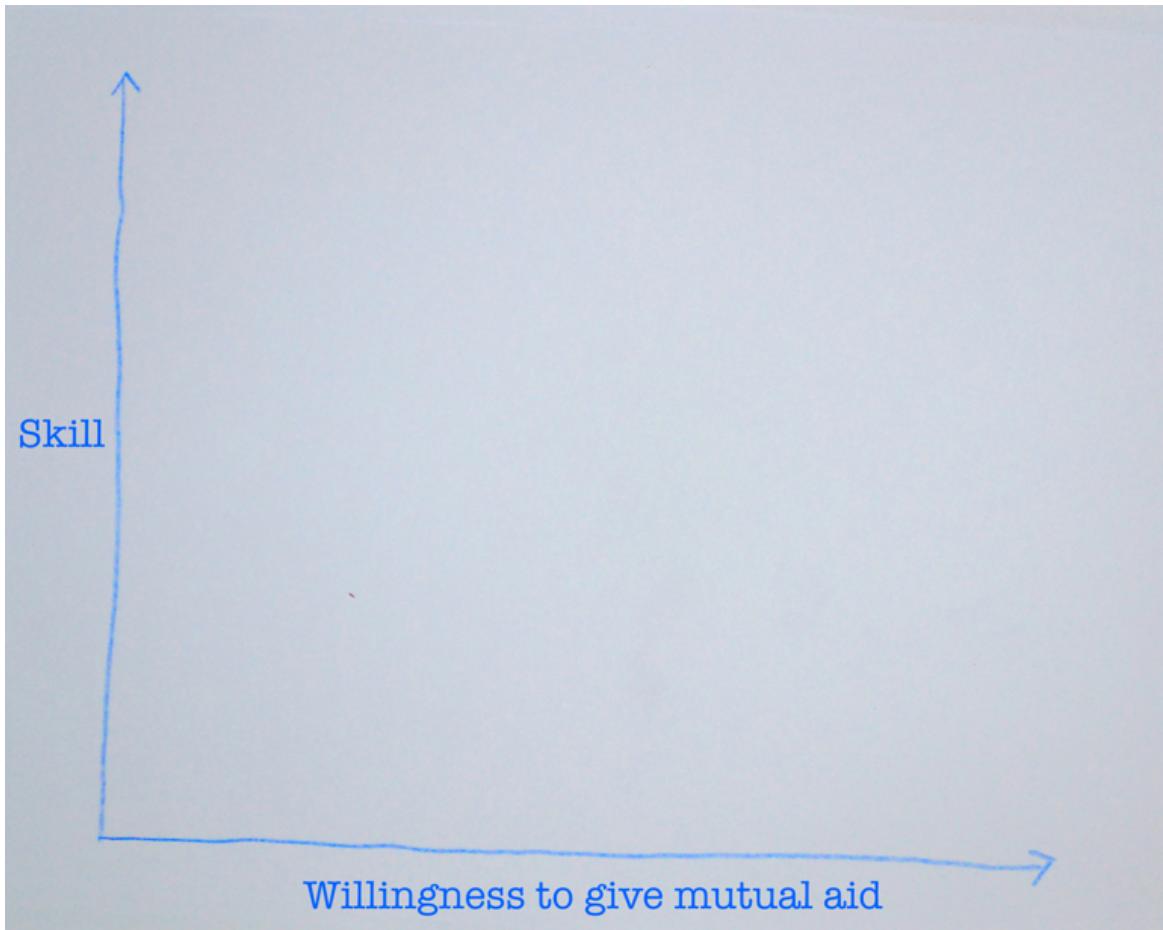
So an important thing to remember: a network of trust is not the same thing as your friends. It is a group of **trusted connections**. People you know can engage in mutually beneficial transactions.

You are looking for **complementarity**. An ideal connection is someone that can solve problems you can't - say a plumber, if you're a programmer - and someone for whom you, in turn, can provide value. (This leads trust networks to be more heterogeneous than typical friendships.)

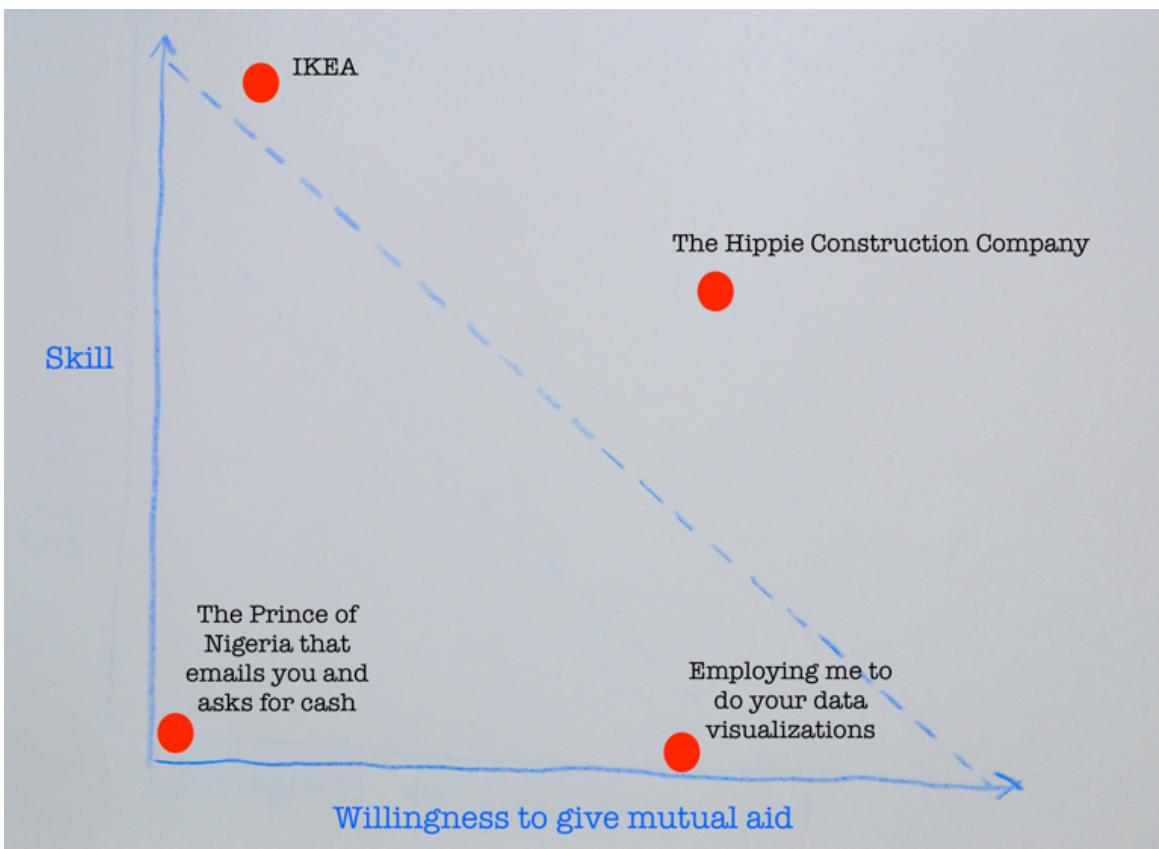
But if you are a programmer, how do you know which plumber to trust? You don't.

But you have a trump card. You can make a Bayesian prediction.

To keep things simple, I'm going to pretend that trust can be reduced to two dimensions (whereas in reality, it's a mess of different factors). When you trust someone, you predict that (a) they will do a good job and (b) they will treat you well if you play fair. Let's call those two dimensions **skill** and **willingness to give mutual aid**.



Saying that you trust someone in that case means that you predict that they belong above this vertical line:



And we're especially interested in the upper right corner. That's what we're trying to fill in when we scale our network of trust. How can we find people that are both skilled and helpful? How can we find more people that act as [the hippies did in the last essay?](#)

As I pointed out there, companies are trust networks – and looking at how they operate can help us scale. Compared to individuals, companies tend to be more deliberate about whom to trust. Successful companies spend large resources trying to **predict how people will behave if they are included in the company trust network**. Recruiters look at resumes, do tests, interviews, ask for referrals – in short, gather data points that help them model the future behavior of the person applying. Are they trustworthy? Will they give mutual aid to their colleagues? Are they skilled?

These predictions can then be compared to the actual outcome. And that feedback can be used to improve the recruiting heuristics to make future predictions more reliable.

When recruiting people to your personal network of trust, you might not be able to be as data-driven. But the underlying aim should be the same. Look for things that will help you predict how competent a person is, and how likely to defect.

Before asking a group of hippies to redo our roof, my wife talked to a neighbor that knows them to see if he would trust them. He would. Then we watched two documentaries about them to study the constructions they had designed, and see them in action. It looked OK to me, but more importantly – my wife was impressed. She has weird enough hobbies to be able to tell if a carpenter follows industry standards, or actually understands what he's doing.

All in all, this took us maybe 50 hours (including the books on carpentry and the history of construction that my wife reads while breastfeeding). It saved us something on the order of a year of salary.

If you want to try this at home, it's worth pointing out that [we live in the Danish countryside](#). That means the underlying incentive structure is pretty solid. And **the incentive structure is what gives the trust diagram its distribution**. If you live in an environment where people only engage in one-off transactions, and bad actors aren't punished, you have to deal with a gravitational force pulling everything toward the Prince of Nigeria.



## 2. Testing the predictions

Employing someone to redo your roof is a risky way to start a relationship.

If you have located a potentially trustworthy person, the safer way to test that prediction is to make a small transaction. Figure out some way you can help – lend them a motor saw if they need to release some anger on a tree; point out a missed opportunity in one of their essays – and then see if they look for ways to help you in return.

The cost of these experiments is capped, but the upside is unknown and may well be priceless. If you lend your neighbor a motor saw, you can't lose more than a motor saw. (In the US, I guess you could get sued as well if your neighbor had watched too much *Evil Dead* and decided to saw off their hand.) But the upside of motor saw loans can be enormous – high trust transactions for the rest of your life or your neighbors. Whichever is shortest.

In 1812, Michael Faraday, who at the time was an unschooled bookbinder, was looking for a way to become a scientist. It would have been tricky even today, him not having any formal education, but it was even more daunting in an age when only the landed gentry and some reverends had the luxury of doing science. What did he do? He made a gift. A book where he had bound his notes from Sir Humphry Davy's lectures on chemistry. Davy was so touched – and impressed – that a year later, after having hurt his eyes experimenting with nitrogen trichloride, he asked Faraday to step in as his assistant.

If you find the idea of helping strangers repellent, you can also do it the other way around. Asking for what you want is a solid and under-used strategy.

After your kids have played together with your neighbor's a few times, ask them if they can drive you to the hospital when your kid starts throwing up and fainting. That can be the start of something big. (The element of crisis might help cement the bond.)

When you get people to extend help, they sometimes continue to help just [to live up to their self-identity](#) of being someone that helps you. I'm not sure about the psychology here.

### 3. Scaling transaction sizes

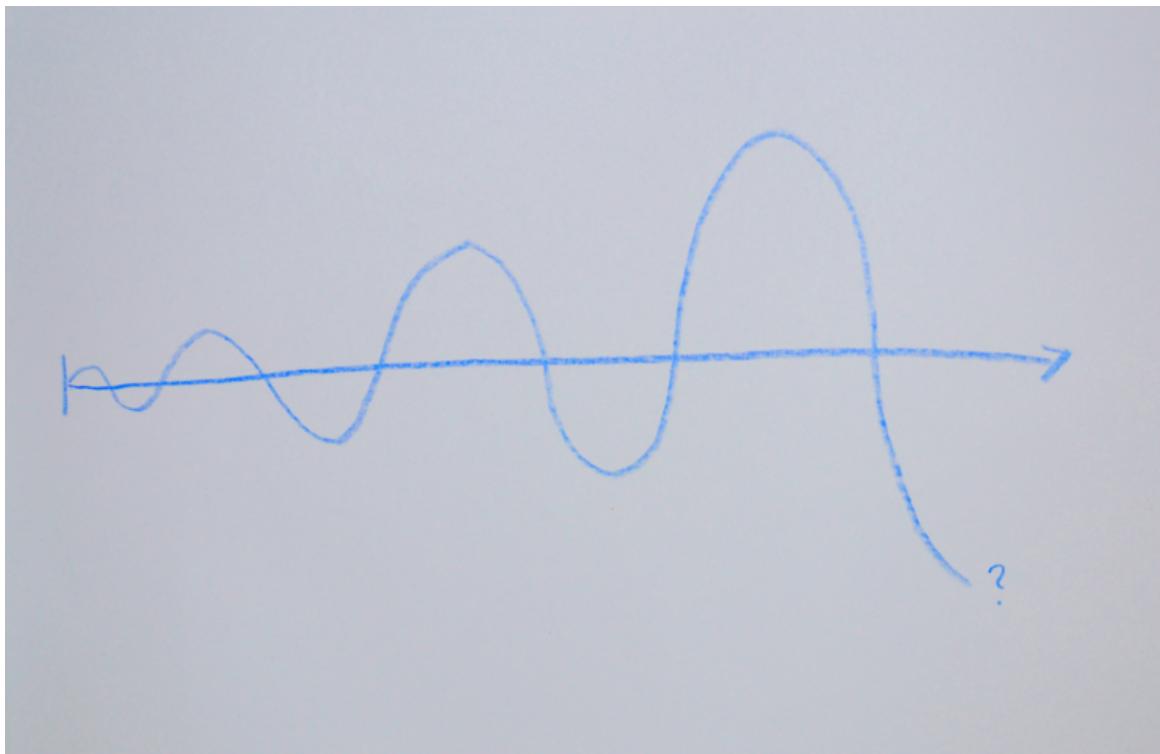
So you got someone to drive your vomiting kid to the hospital? Good. The next step is to return the favor. But don't just return it, do it with sugar on top.

The crucial thing about trust relationships is that **you do not want them to even out**. You never return an egg for an egg. You return a rhubarb pie.

Not [tit-for-tat](#), but **tit-for-almost-tat**.

If someone buys you dinner and you wire them the exact amount, that is a signal that you don't want the relationship to continue. If you instead give back by inviting them to your midsummer celebrations – well, that's a whole other story.

When the person you are dealing with is trustworthy and competent, tit-for-almost-tat creates a ratchet. You return slightly bigger favors every time, taking turns scaling up the transactions.



I and my closest neighbor are locked into one such **escalating arms race of helpfulness** at the moment. Every day at dusk, I'm looking out across the field to see what he's up to. I'm trying to figure out if there is some way I can help him. We fenced in his orchard; he offered us to use his car when we go grocery shopping; my wife and daughter

fed his horses when he was away; now, he's helping the hippies remove the old roof for us. It's a lovely little arms race, and it has saved both of us countless hours (compared to having to buy services in the market).

Of course, if we escalate too far we might reach a point where the expected value of future transactions is less than what we owe. At that point, the incentive structure shifts; my neighbor might decide to defect when he's on top. The same thing can happen if he decides to move – at which point he's no longer constrained by considerations about future transactions.

If he does, God forbid, I'm going to take a page from the Islamic merchant's playbook: I'll slander him in verse. And then I'll email it to all the connections we share.

## 4. Referrals

The best thing about having a trusted connection is that you can access *that* person's network now.

Often the access is indirect. You don't need to know who your connection has to trust to be able to help you. I don't know the suppliers that the hippies rely on. The hippies are **trust abstractions**, a little bit like functions – since I trust their output, I don't need to see the implementation.

At other times, you want to refactor the function and get access to the connections directly. This is called asking for a referral. If I need plumbing, I ask the hippies whom they would use – and since I trust them, I can transfer that trust and gain a new trusted connection. I can also do chains of referrals, where I ask them to refer me to someone that refers me to someone else that refers me to... alas, the further you go, the more you dilute the trust so we can't go on forever.

[In an interview](#), anthropologist Joseph Henrich describes using these chains of referrals to establish trust with groups he wants to study:

**Joseph Henrich:** ... making a chain of relationships. I did the same thing with the Mapuche in Chile. I had a friend who had a friend in Santiago, who had a woman who cleaned his house, who had a cousin in the southern part of the country. And I followed that chain of relationships into the community.

That gave him the trust needed to convince the Mapuche to participate [in game-theoretical experiments](#). Making use of referrals this way, allowed him to expand his network of trust far beyond Dunbar's number – 150 – which is generally held to be the number of people you can coordinate through informal processes. It is one of the processes that makes it possible for a *suftaja* to travel across the Indian Ocean.

Referrals are a blast. The last time I traveled to a place where I didn't know anyone – Switzerland – I asked [my friend Viktor](#) to recommend me to people in Zürich and Bern. His connections in turn referred me to others. That immediately, as soon as I got off the train, planted me into a caring social network. Russian horror surf-bands. Ping pong in a stable. A Game of Thrones analysis marathon with some french literature students. They even bailed me out when I ran out of money, and got caught ticket-less on the train home.

---

To scale a network of trust you need to engage people you don't know in small transactions. You have to observe how they behave and see if the transactions are worth your time. When they are, you have to increase the transactions by returning ever-larger favors. And then you reach out through referrals.

Some people will be better at this than others.

Lyndon B. Johnson was an obsessive collector of trusted collaborators. From his college years on, he would do everything he could to make people he could trust stay in his orbit. He would place them in agencies he controlled, he would make them partners in law firms he trusted, or in his radio studio, or at subcontractors he favored. That way, when he decided to run for Senate, he had a minor army of lawyers and organizers that could be mobilized at short notice. And then, once elected, he proceeded to place them in federal agencies, where they would wait patiently for his bid on the presidency.

People like Johnson accumulate piles of trust and can make unfathomable numbers of referrals. They are massive trust abstractions. Having them in your network can vastly increase what you can accomplish.

But controlling the flow of trust also gives them power over you. They might take a cut of the transactions they enable, they might impose demands; Johnson perfected this. In the Senate, he would sit through the night with his phone in the dark, making call after call, pulling on an intricate web of connections until, by morning, the votes he needed had been switched. He would also humiliate the people that depended on him by making them take his orders while he sat in front of them.

When I discuss this essay with Ryan Fugger - a person who, having invented the concept behind Ripple, knows way more about trust than me - he points out that corporations usually play this trust abstraction role. Banks, for example, abstract trust, and then extract value from the transactions that enable. That's great. People who do not trust each other can buy trust from a corporation. But why are so many people unable to create their own trust?

Ryan writes in the chat:

... it's possible that most people just aren't interested in [tools that help them leverage their trust networks], and that the whole network would devolve into a network of banks that abstract the trust of their account holders, just like today.

For those willing to grow their networks of trust, rather than relying on bought trust, though, there is leverage to be had.

Take the Amish. In 1965, as Johnson (by then President) was expanding social security to include Medicare, the Amish leveraged their network power to renegotiate their deal with the State. Having high levels of trust, they were able to coordinate and get themselves exempted from the social security system - and the taxes.

Figuring out how you do *that*, however, I will have to leave for another day.

([Cross-posted from my blog.](#))

# Covid 6/17: One Last Scare

The [one last scare](#), from America's perspective, is the Delta variant. If we can remain stable or improving once Delta takes over, then barring another even more infectious variant, we've won. If and where we can't do that, it's not over.

That's the question. How much Delta is out there already, how much worse is it, and will that be enough to undo our work? If it is, how long until further vaccinations can turn things around again?

Before I look into that, let's run the numbers.

## The Numbers

### Predictions

Prediction from last week: Positivity rate of 1.8% (down 0.2%), deaths fall by 9%.

Results: Positivity rate of 1.9% (up 0.1%), deaths fall by 20%.

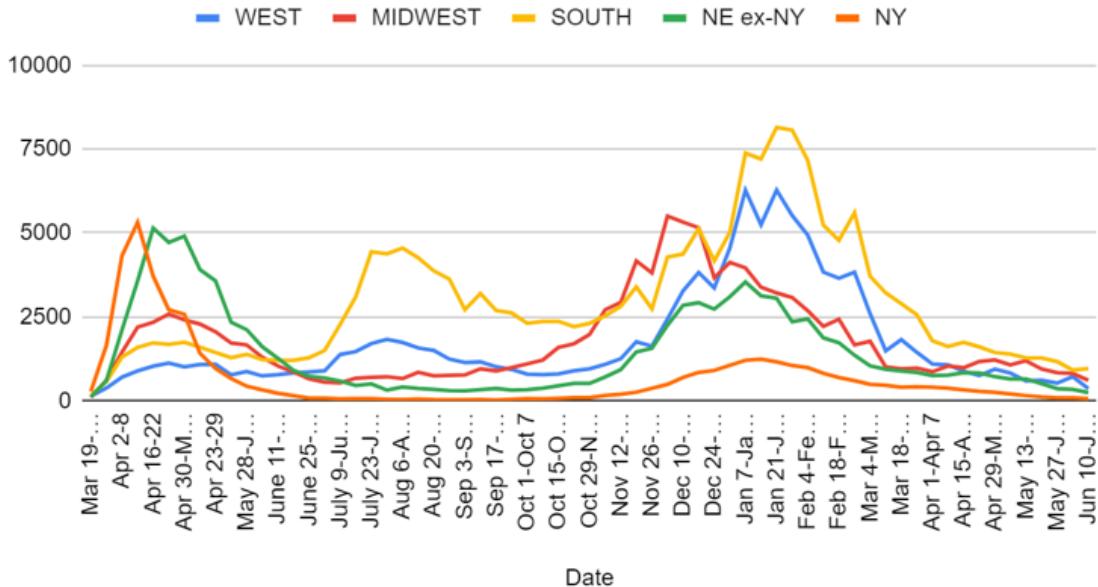
Prediction: Positivity rate of 1.8% (down 0.1%), deaths fall by 9%.

I think the deaths drop likely reflects variance in reporting, so while it is definitely good news I do not want to predict a further large drop from a number that is likely to be somewhat artificially low. For positivity rate, we seem to have shifted to a regime of rapidly declining numbers of tests, especially in areas with declining infection rates. Thus, it's quite possible that the positive test rate will stop reflecting the state of the pandemic. I still expect the number to keep dropping a bit, but it wouldn't be that surprising if it stabilized around 2% from here on in with improvements reflected mainly elsewhere.

### Deaths

| Date          | WEST | MIDWEST | SOUTH | NORTHEAST | TOTAL |
|---------------|------|---------|-------|-----------|-------|
| May 6-May 12  | 826  | 1069    | 1392  | 855       | 4142  |
| May 13-May 19 | 592  | 1194    | 1277  | 811       | 3874  |
| May 20-May 26 | 615  | 948     | 1279  | 631       | 3473  |
| May 27-June 2 | 527  | 838     | 1170  | 456       | 2991  |
| June 3-June 9 | 720  | 817     | 915   | 431       | 2883  |
| Jun 10-Jun 16 | 368  | 611     | 961   | 314       | 2254  |

## Deaths by Region

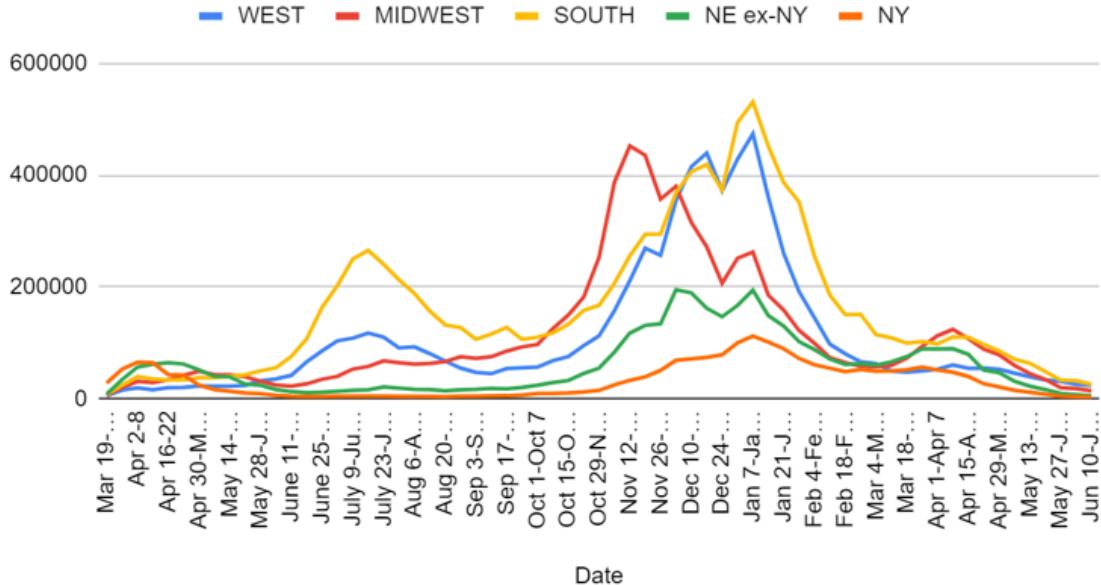


It seems like we can safely say that last week's measured death count was too high and should have been more in the 2500-2700 range, and we remain roughly on the previous pace of decline from May. This week's number is likely slightly lower than its true value, with some deaths shifted from this week into last week versus when they would usually be measured. Note that the raw total was even lower, as I added back some deaths in California when they reported negative numbers on multiple days, which I cancelled back up to zero.

## Cases

| Date          | WEST   | MIDWEST | SOUTH  | NORTHEAST | TOTAL   |
|---------------|--------|---------|--------|-----------|---------|
| Apr 29-May 5  | 52,984 | 78,778  | 85,641 | 68,299    | 285,702 |
| May 6-May 12  | 46,045 | 59,945  | 70,740 | 46,782    | 223,512 |
| May 13-May 19 | 39,601 | 45,030  | 63,529 | 34,309    | 182,469 |
| May 20-May 26 | 33,890 | 34,694  | 48,973 | 24,849    | 142,406 |
| May 27-June 2 | 31,172 | 20,044  | 33,293 | 14,660    | 99,169  |
| Jun 3-Jun 9   | 25,987 | 18,267  | 32,545 | 11,540    | 88,339  |
| Jun 10-Jun 16 | 23,700 | 14,472  | 25,752 | 8,177     | 72,101  |

## Positive Tests by Region



[Chart note: Given our current situation, while they provide good perspective on where we've been, the full charts are not that enlightening anymore and are growing unreadable for all the right reasons. Unless people think it's a bad idea, I'm thinking the charts should be narrowed to only show movement after some reasonable time. Maybe start on 1 April 2021?]

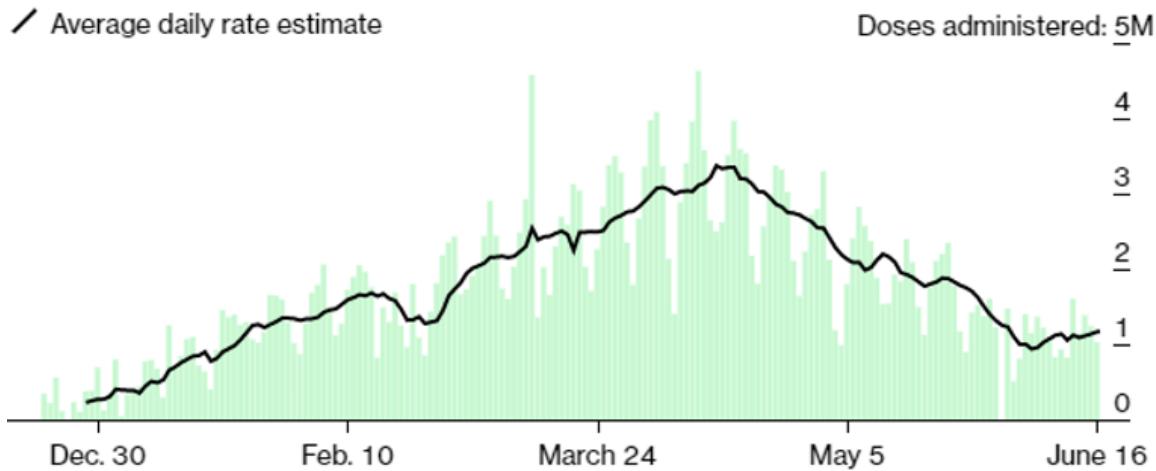
These are very good numbers. We see declines across the board. It's not back on the old pace, but given the lifting of restrictions and the rise of Delta, that should be expected, and we still made more progress than last week by a large amount.

They reflect *declining numbers of tests* rather than a big decline in positive test percentage, but at some point... that makes sense at equilibrium because you don't need as many tests? There's an argument that a 2% or so positivity rate is reasonable, in the sense that it means that a marginal additional test would have a much lower chance than that of detecting an infection that was likely asymptomatic, so it's not clear more testing would be worth the trouble.

It also likely means we are doing substantially less 'surveillance testing' where we check people without any reason to think they are positive, either out of an abundance of caution or to give them entry into events or travel. And we're doing less testing in the places that are winning, versus the areas that are still struggling. Thus, we are taking the least likely to be positive tests out of the pool, which should raise the positive test rate. If we're stable on that front, we're making great progress.

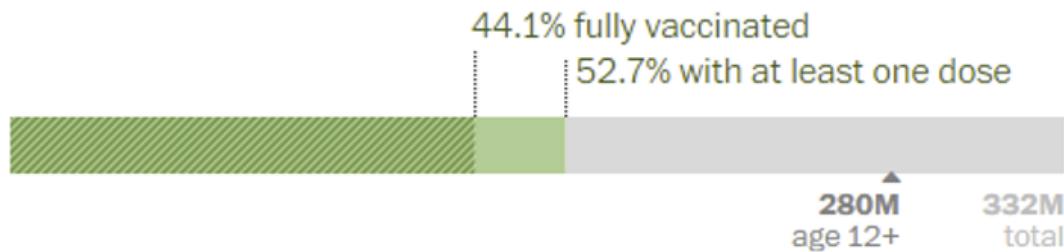
## Vaccinations

In the U.S., the latest vaccination rate is **1,165,956 doses** per day, on average. At this pace, it will take another **5 months** to cover **75%** of the population.



## 175.1 million vaccinated

This includes more than **146.5 million people** who have been fully vaccinated in the United States.



In the last week, an average of **1.17 million doses per day** were administered, a **2% increase ↑** over the week before.

Compared to what I expected, this continues to be good news. First doses do not appear to be declining much, which is impressive given how many of those eager to be vaccinated have already had their shots, and much of the remaining population is under 12.

If we can sustain this pace indefinitely, with an additional 1% vaccinated every week, we will be home free soon in most places, regardless of how bad Delta is. The effects compound increasingly quickly over time.

## Delta Variant

Delta is rapidly taking over. How worried should we be?

First of all, if you are fully vaccinated, you should not be personally worried. [Here's the latest data \(link to paper\)](#).

**Table 1: Estimated vaccine effectiveness against hospitalisation**

|                    |        | Alpha                     |                       |                       | Delta                     |                       |                       |
|--------------------|--------|---------------------------|-----------------------|-----------------------|---------------------------|-----------------------|-----------------------|
| Vaccination status |        | OR vs symptomatic disease | HR vs hospitalisation | VE vs hospitalisation | OR vs symptomatic disease | HR vs hospitalisation | VE vs hospitalisation |
| Any vaccine        |        |                           |                       |                       |                           |                       |                       |
|                    | Dose 1 | 0.51 (0.48-0.55)          | 0.44 (0.28-0.70)      | 78% (65-86)           | 0.69 (0.64-0.75)          | 0.37 (0.22-0.63)      | 75% (57-85)           |
|                    | Dose 2 | 0.13 (0.1-0.15)           | 0.64 (0.24-1.72)      | 92% (78-97)           | 0.20 (0.18-0.23)          | 0.29 (0.11-0.72)      | 94% (85-98)           |
| Pfizer             |        |                           |                       |                       |                           |                       |                       |
|                    | Dose 1 | 0.53 (0.47-0.58)          | 0.32 (0.14-0.73)      | 83% (62-93)           | 0.64 (0.54-0.77)          | 0.10 (0.01-0.76)      | 94% (46-99)           |
|                    | Dose 2 | 0.06 (0.05-0.08)          | 0.88 (0.21-3.77)      | 95% (78-99)           | 0.12 (0.1-0.15)           | 0.34 (0.10-1.18)      | 96% (86-99)           |
| Astrazeneca        |        |                           |                       |                       |                           |                       |                       |
|                    | Dose 1 | 0.51 (0.48-0.55)          | 0.48 (0.30-0.77)      | 76% (61-85)           | 0.70 (0.65-0.76)          | 0.41 (0.24-0.70)      | 71% (51-83)           |
|                    | Dose 2 | 0.26 (0.21-0.32)          | 0.53 (0.15-1.80)      | 86% (53-96)           | 0.33 (0.28-0.39)          | 0.25 (0.08-0.78)      | 92% (75-97)           |

OR = odds ratio. HR = hazards ratio. VE = vaccine effectiveness. OR vs symptomatic disease as described in (1). HR and VE vs hospitalisation adjusted for age, clinically extremely vulnerable groups, ethnicity and test week

The 94% number for one dose of Pfizer is almost certainly higher than the real value due to random error, but the other numbers are very consistently saying that the vaccinations are at least as effective against Delta as they were against Alpha. The symptomatic disease numbers are more worrisome, but remain what I'd consider acceptable, especially for mRNA.

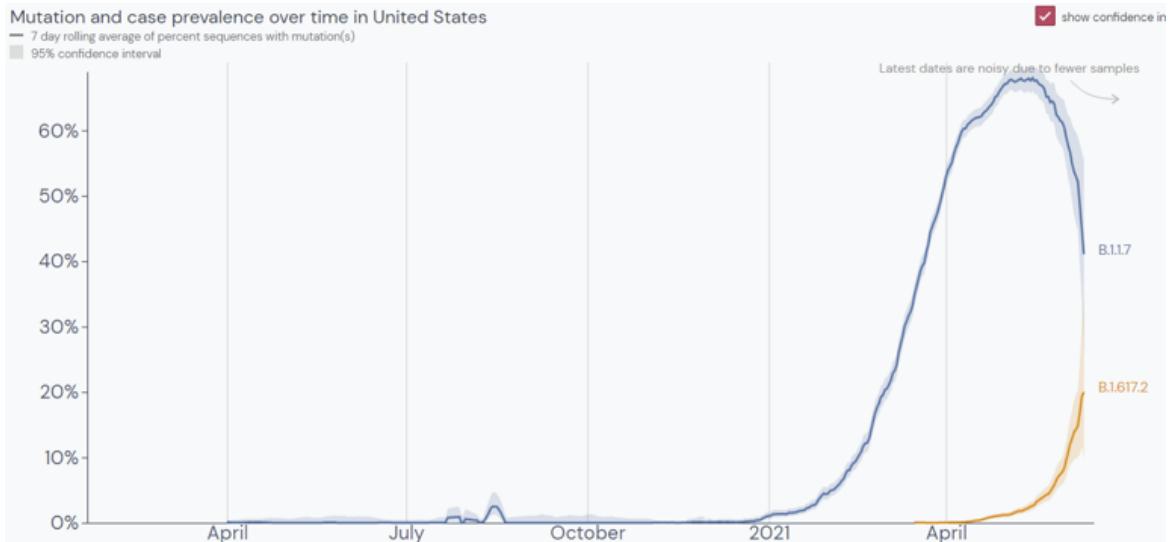
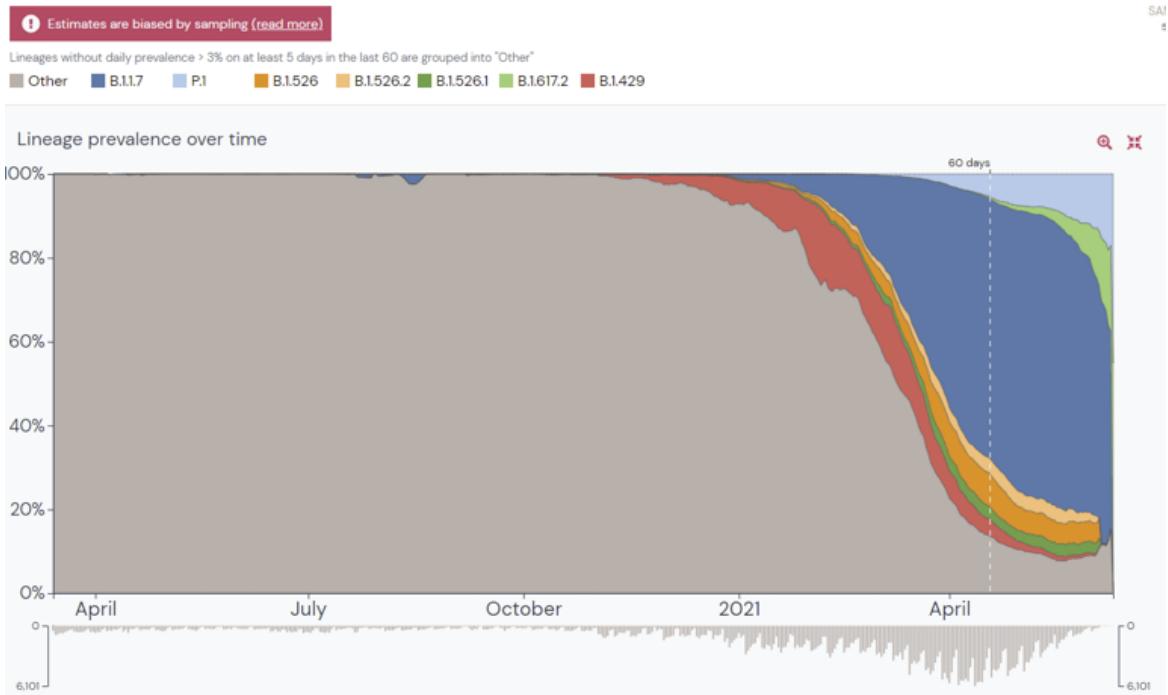
I don't agree with every individual point, but [this thread](#) is mostly the reasonable bear case that Delta is sufficiently bad that we should be worried restrictions may come back and the pandemic might not be over in America.

As I noted last week, statements like 'watch the Delta percentage in your region' have the implication it's going to be less than 99% for all that long, unless there's another variant we don't know about that's even worse - at this point seeing a larger Delta percentage, conditional on knowing the growth rate of cases, is good news. Last week the report was that we were at 6% Delta cases overall, so there was still uncertainty at what level of cases we could stabilize against it.

The consensus seems to roughly be that Delta is twice as infectious as the pre-London strain. [This thread](#), for example, puts its likely R<sub>0</sub> at around 7, versus 3.3 for the original strain, an ~50-60% increase over the currently dominant London strain. There are places in the country where that may be enough to cancel out the current vaccination rate, despite vaccinations 'overperforming' their headline numbers due to previous infections and the existence of young children.

Where are we right now in terms of Delta? [Here's the best data point I could find](#). Delta is the light green area in the first graph.

## Lineage prevalence in United States



English strain peaked here around 70% of the pool and is now down around 45%. That implies that the *average* strain is *more infectious* than the English strain already. Which makes sense, if you assume the 'Other' group includes strains similar to P1 or Delta that have now crowded out the others, and that both P1 and Delta are more dangerous than the English strain. The major legacy pre-Alpha strains are gone so it stands to reason similar minor ones were also wiped out.

This leaves us with a current mix of roughly (these are all approximations):

45% Alpha

25% Delta

15% P1

15% Other (that I will treat as P1)

How much more infectious will this pool be when it's 100% Delta? Let's accept the thread above estimates for now and assign points as follows:

1.4 Alpha

1.5 P1/Other

2.2 Delta

Total current infectiousness of pool as of collection of these sequences:  $30\% * 1.5 + 45\% * 1.4 + 25\% * 2.2 = 1.63$  = 63% more infectiousness than pre-Alpha values

Future infectiousness of pool = 2.2

Final outcome (assuming Delta is worst variant) =  $2.2 / 1.63 = 35\%$  additional infectiousness

On April 1, what was the pool like? At that point, we can put it at something like

Epsilon% Delta

55% Alpha

3% P1

3% Other small things similar to P1

39% Other similar to old baseline (1.0 infectiousness)

Total infectiousness of April 1 pool of sequences = 25% more infectious than baseline from mid-2020.

Total increase in infectiousness since April 1 pool =  $1.63 / 1.25 = 30\%$  additional infectiousness.

Thus, if we 'believe the hype' here, we have to survive another 35% increase, after a previous 63% total increase, the majority of which has happened since April 1.

Currently we are vaccinating about 1% of people each week. Given we are already 50%+ vaccinated if you discount children, and what vaccines we are using, that's something like a 2% improvement each week. If we can keep that up as a share of the remaining population, that will be cumulative.

Right now, we are cutting cases in half every 3 weeks or so, which is about 4 cycles, so I'd estimate R<sub>0</sub> at 0.84. Increasing that by 35% puts us at 1.14. That's presumably high, because sequencing is delayed and thus the current Delta share is higher than in the above calculation. I don't consider 35% a strict upper bound, but my mean estimate is more like 30%, and every little bit helps.

Thus, it looks clear to me that *most places* in America are going to make it given the additional vaccinations that will take place, but some places with low vaccination rates will fall short.

Alas, that does mean that it might be a while before the final restrictions can be lifted, and life returns 100% fully back to normal. We will be stuck in a kind of hybrid limbo, mostly involving performative mask wearing and other such annoyances. But that's actually pretty close to fully normal, in terms of practical consequences for adults. For the kids, things could stay rough for a while, because people are really stupid about such things. I hope we can get the vaccines approved for the Under 12 crowd soon.

# Please Stop Asking Me About That Guy

That guy's name is Bret Weinstein. If you already weren't asking me about that guy, you can and probably should skip this section.

Enough people keep asking me about him, and there's been enough discussion in which the things he's claiming have been taken seriously, that I need write this section anyway, for the explicit purpose of Please Stop Asking Me About That Guy.

Bret has made a huge number of overlapping extraordinary claims, with an extraordinary degree of both confidence and magnitude. It seems like *most of the time* that I get asked about 'hey there's this theory that sounds plausible what do you think?' the source of that theory is *exactly Bret Weinstein*. He's become the go-to almost monopolistic guy for presenting these things in a way that seems superficially plausible, which of course means he's here for all of it.

It's a variety of *different* theories about how everyone's covering up [The Truth Which Is Out There](#) (if you think that link is unfair, his core claims include UFOs, a broad based conspiracy to censor and cover up The Truth, and many monsters of the week, so I dunno what to tell you). Some of his claims such as the lab leak are plausibly correct, but are stated with absurd levels of confidence. Others, stated with similarly absurd confidence, are... less plausible. This includes claims I do not think it would be responsible for me to repeat, such as so-called his "Crime of the Century." Then he cries censorship and further conspiracy when he gets (entirely predictably and entirely according to clear established policies) censored on platforms like YouTube and Facebook.

Look. I *totally get it*. After everything that's happened, in the words of one commenter, if it's a choice between the people who claimed masks didn't work handing verdicts out from on high or 'those three guys on that podcast' and the podcast guys seem to have models containing gears, why *wouldn't* you go with the podcast guys? At this point, Fox Mulder would have the world's most popular podcast and be a regular on Joe Rogan, he wouldn't have Scully as a co-host, and damn if he still wouldn't be making a lot of good points. Seriously, it makes *total sense*.

Except that one can decline to take either side's word for it, and think about the proposed gears, including the other gears and claims coming from the same sources, derive your guess as to what algorithms both are using to decide what to claim, and decide that neither of these options is going to be much of a source.

I can't even at this point. I finally unfollowed him when I noticed that every time I saw his tweets my mood got worse in anticipation, yet I wasn't learning anything useful except how to answer people asking about his claims. I have wasted far too much of my life trying to parse ludicrous stuff buried in hours-long videos and figure out how to deal with all the questions people ask about such things, or to convince people that he's spouting obvious nonsense when he's clearly spouting obvious nonsense, and I will neither be paid for that time nor will I be getting any of that time back.

That does not mean that all of Bret's claims are implausible or wrong. Some aren't. It simply means that the whole thing is exhausting and exasperating, many of the claims being made are obvious nonsense, and the whole exercise of engaging has for me been entirely unfruitful. One is not obligated to explain exactly why any given thing is [Wrong on the Internet](#), and my life is ending one minute at a time.

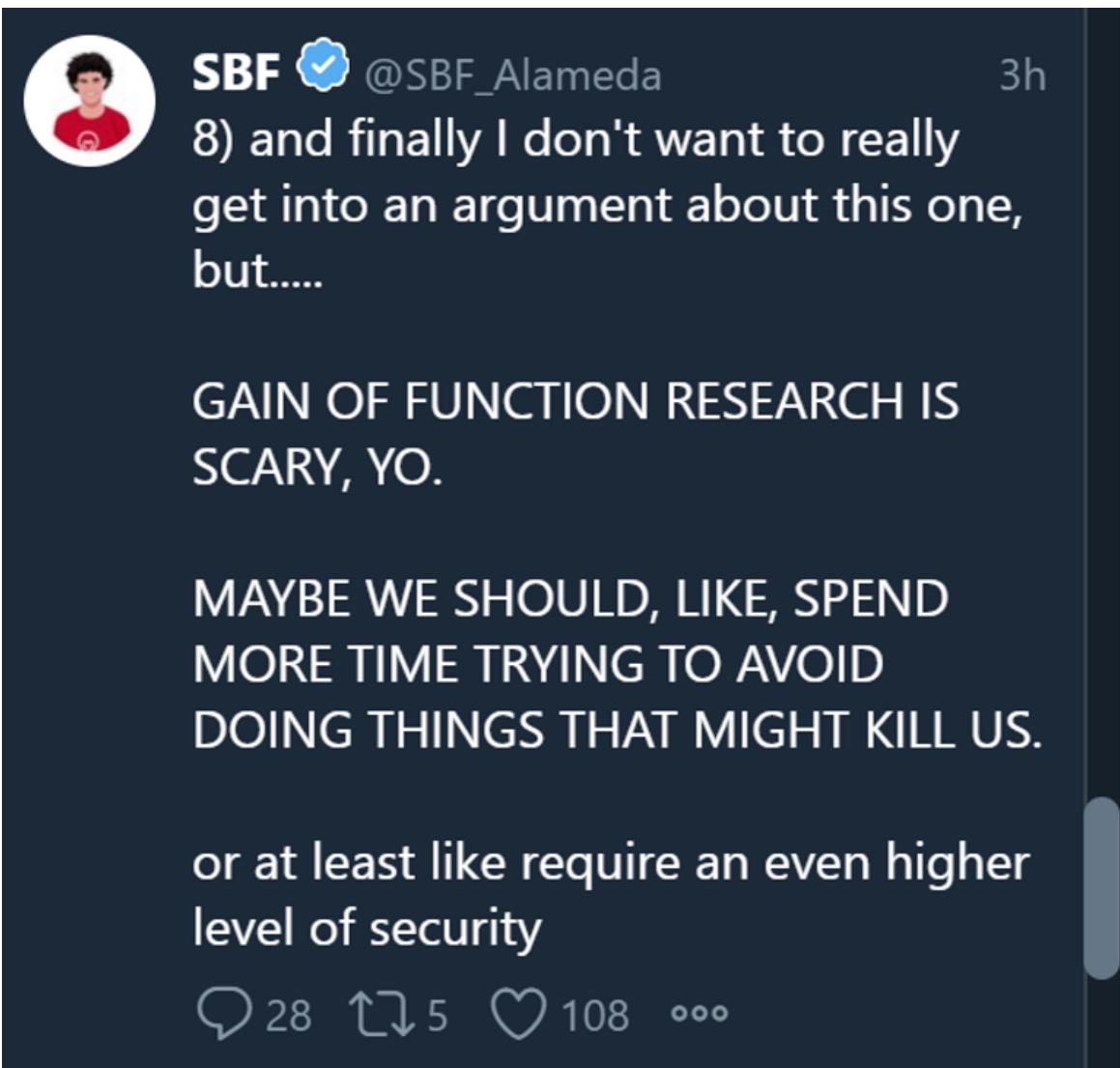
It definitely does not mean I agree with the decision by some platforms to censor his claims. While I do think many of his censored claims are wrong, it would be a better world if such claims were not censored.

If you want to engage with Bret's claims after getting the information above, and feel that is a good use of your time, by all means engage with his claims and build up your own physical model of the world and what is happening. The right number of people doing that isn't zero. The set of such people simply is not going to include me.

If you still want to ask me about that guy, [my cheerful price](#) for further time spent on 'investigating, writing about and/or discussing claims by Bret Weinstein or his podcast guests' is \$500/hour, which also is my generic cheerful price for non-commercial intellectual work. If you're paying, I'll check it out. Otherwise, Please Stop Asking Me About That Guy.

## In Other News

[Perspective thread on Covid from Sam Bankman-Fried, EA billionaire founder of crypto exchange FTX.](#) Includes this bit of excellent practical advice:



**SBF ✅ @SBF\_Alameda** 3h

**8) and finally I don't want to really get into an argument about this one, but.....**

**GAIN OF FUNCTION RESEARCH IS SCARY, YO.**

**MAYBE WE SHOULD, LIKE, SPEND MORE TIME TRYING TO AVOID DOING THINGS THAT MIGHT KILL US.**

**or at least like require an even higher level of security**

28 5 108

And this slice of post-vaccination life, while waiting an extra day for a flight due to his negative Covid test having his SSN and his middle initial but not his full middle name on it:



**SBF** ✅ @SBF\_Alameda

4h

21) As I write this, a police officer came up and pointed at my face.

"I'm vaccinated and just got 2 negative COVID tests", I said. I hold up my paperwork.

He pointed at my face again. Also, of course, he has a gun.

I sigh and put on my crumpled mask until he passes by.

💬 147 ⚡ 34 ❤️ 716 ...

[Headline seems not to require further comment \(WaPo\).](#)

**Coronavirus infections dropping where people are vaccinated, rising where they are not, Post analysis finds**

[California lifts restrictions on workers](#) as of today. Will the masks actually come off?



**Matt Haney**  @MattHaneySF · Jun 14

California workplaces can allow vaccinated workers to go mask-free starting Thursday, after a forthcoming Executive Order from [@GavinNewsom](#).

8

28

138



[New York lifts most remaining Covid restrictions.](#)



**Morgan Mckay** @morganfmckay · Jun 15

NEWS: "We have hit 70% vaccination. It is the national goal and we hit it ahead of schedule," [@NYGovCuomo](#) formally announces at the One World Trade

This means most remaining COVID restrictions in NY will be lifted

11

34

73



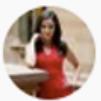
**Morgan Mckay** @morganfmckay · Jun 15

Remaining COVID restrictions that will be lifted: Restaurants, retail stores, offices, gyms, amusement centers, and hair salons can make capacity and social distancing restrictions optional, as well as ease COVID disinfection protocols.

7

18

15



**Morgan Mckay** @morganfmckay · Jun 15

Replies to [@morganfmckay](#)

Areas that will still have to follow COVID restrictions: students in pre-K to 12th grade schools, public transit systems, homeless shelters, correctional facilities, nursing homes and health care settings

Things are slow, but my experience so far is that nothing has changed. As usual, children get the shaft. Public transit (in particular, my ride to/from the city) continues to be much less pleasant in ways that make no physical sense. That's the way it goes.

NFL isn't forcing its players to get vaccinated, [but it's also not forcing them to get vaccinated \(WaPo\)](#). They're going to make unvaccinated life quite inconvenient.

[Mastercard pledges 1.3 billion to buy vaccines for 50 million Africans](#). In response, America calls it 'significant step,' indicating we indeed remain funding constrained, then makes smaller pledge:



Dan Diamond ✅ @ddiamond · Jun 8

The White House this afternoon praised Mastercard Foundation's \$1.3B pledge to boost coronavirus vaccines in Africa.

...

"This is a significant commitment," said [@PressSec](#), also touting the U.S. pledge last week to share 5 million doses with the African Union.

[Another person does math on smaller and delayed doses](#), and reaches the same conclusion that actual calculations always do, that they would save many lives.

[Zeynep thread about how vaccinating the vulnerable changes things](#), you should know this already.

[Yes.](#)



Nate Silver ✅ @NateSilver538 · Jun 13

So the thesis here is that Walensky follows the medical science, but actually this is bad and the CDC should tell weird little lies to people?

...

## *The C.D.C.'s New Leader Follows the Science. Is That Enough?*

By all accounts, Dr. Rochelle Walensky is a fierce advocate and an empathetic scientist. But C.D.C. advice must be better attuned to the real world, critics say.

The most recent instance, when Dr. Walensky announced that vaccinated people could go mask-free indoors, was supported by the latest research, scientists said. But many felt the agency had rushed the decision to end mask use without considering parts of the country where infections were still high, and without grasping the mistrust and culture clashes the new advice would engender.

“C.D.C. got the medical and epidemiological science right, but what they did not get right was the behavioral science, the communications and working collaboratively with other stakeholders,” Dr. Gounder said. “That was a big oversight.”

## Not Covid

[Words of wisdom thread from Sarah Constantin.](#)

[Front runner in NYC mayor’s race Eric Adams comes out in favor of multi-hundred person virtual classrooms, since in-person learning is unnecessary.](#) Prediction markets and polls did not budge. He’s more likely to win than ever. Nothing matters.

[Dominic Cummings wants to know the probability China will take over Taiwan in the next five years, because private conversations suggest 50%+,](#) and asks what he should read. A relevant Metaculus market is [here](#), but the right answer is of course a prediction market with actual money involved. I’ve asked [Polymarket](#) to throw up a market, hopefully we’ll have something soon, but it’s always tough getting interest in markets that don’t resolve for years.

This week in how we all *actually* die news, [two apt observations](#):



**Matthew Yglesias** @mattyliegiesias · Jun 14

...

All the smartest people I know say there’s a huge AI alignment problem lurking just around the corner that poses a potentially existential threat to human life, then after that amazing buildup they don’t have much to offer as solutions.

273

198

1.2K



**Amanda Askell** @AmandaAskell · Jun 15

...

It’s more disconcerting when the people expressing concern about something aren’t trying to sell you on a solution and are just genuinely concerned.

1

5

122



And also:



**Eliezer Yudkowsky** @ESYudkowsky · Jun 15

Replies to [@mattyglesias](#)

It sounds like you've been talking to some rigorous and honest people, wise enough to know when a solution is terrible and won't really help. This seems good. If at some point you start to feel like you're being talked into believing in some solution, ping me to unconvince you.

And:



**Luke Muehlhauser** @lukeprog · Jun 14

Replies to [@mattyglesias](#)

**Fact check: true.**

This matches my recent experiences. Among smart people who have looked seriously at the problem and are acting in good faith, I see broad near-universal agreement that this is a huge problem, with the main differences being between the 'this is a huge problem but we can probably figure something out that holds onto the majority of the future's value, and it's unlikely that we all die' camp and the 'this is a huge problem and we are all very doomed with no apparent way out, we're all very likely going to die' camp, with a small not-that-reassuring third camp of 'our civilization is sufficiently inadequate that the tech won't get that far, so we're all going to die but not from this' and perhaps a fourth even less reassuring camp of 'we are all going to die but have you considered that humanity dying out is good actually?'

For what it's worth, I am in the second camp, and think the probability of doom is currently high, partly for the reason [explained in this thread](#): Not only do we have to do a hard thing, we have to do the hard thing correctly on the first try, or there won't be a second try.

[Here's some survey results asking workers in the field for how likely we are to be doomed](#), and how often we are doomed due to 'we didn't get it to do what we intended and thus were doomed' versus 'we did get it to do what we intended and we were doomed anyway.' There's a wide range of probabilities.

The thing that all the camps have in common is that no one has great ideas as to what to do about this. If I had great ideas that I could implement, I'd be working on pivoting to working on them, and there are people eager to help me with that if I did find great ideas, but I don't have any so far, hence I'm not doing that.

One reason to be very concerned is that when we look at our failures in Covid, including especially our inability to not do obviously terrible Gain of Function research, [we see the exact kind of failure modes that are likely to get us all killed \(Eliezer Yudkowsky\)](#), as the local incentives push people towards highly unsafe actions for such mundane reasons as getting a paper published. Thus, one way to work on the problem could be to do so indirectly, by changing such incentives and cultures more generally, and providing proof-of-concept examples of how such things can be done. For example, by getting Gain of Function research banned, and noting what it took to do that so we can do it again.

[This dive into the weeds](#) gives an up-to-date picture of what is generally considered the most central and scary challenge. Our best current AI techniques are radically opaque, current interpretability work is making very very little progress relative to the size of the challenge, and things like deception and mesa-optimizers are hopeless to address if you can't understand your AGI's internal cognition at all. Work on interpretability is urgently needed and is one of the things one could usefully do.

[Anthropic](#) was founded recently by former members of OpenAI for that explicit purpose of interpretability work, and we need as much such work as possible. The fear is that any such organization often does what OpenAI did, and turns mostly into an engine for creating more AI capabilities while giving us one more player to worry about in terms of avoiding a race situation where everyone builds their AI as fast as possible without time for safety work lest someone else do the same and get there first, and thus making the problem doubly worse.

[Here are some other ideas](#), not related to MIRI, if you'd like to look through more of the field. Whatever the solution, it's likely going to require understanding the AIs a lot better than we currently understand the AIs we have now.

Supporting organizations such as [MIRI](#) in the hopes that such work will figure out something worth doing, or doing one's own study of the problems involved, or helping more people understand the situation, all seems net useful, by all means do that, but none of it constitutes the kind of plan we would like. Figuring out a good plan, if you are capable of it, would be immensely valuable, and there is a lot of support standing by to help with a good plan if one is found. Working on the problem directly, even without starting with a plan, also seems worthwhile.

# AXRP Episode 9 - Finite Factored Sets with Scott Garrabrant

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

[Google Podcasts link](#)

This podcast is called AXRP, pronounced axe-up and short for the AI X-risk Research Podcast. Here, I ([Daniel Filan](#)) have conversations with researchers about their papers. We discuss the paper and hopefully get a sense of why it's been written and how it might reduce the risk of artificial intelligence causing an [existential catastrophe](#): that is, permanently and drastically curtailing humanity's future potential.

Being an agent can get loopy quickly. For instance, imagine that we're playing chess and I'm trying to decide what move to make. Your next move influences the outcome of the game, and my guess of that influences my move, which influences your next move, which influences the outcome of the game. How can we model these dependencies in a general way, without baking in primitive notions of 'belief' or 'agency'? Today, I talk with Scott Garrabrant about his recent work on finite factored sets that aims to answer this question.

Topics we discuss:

- [Finite factored sets' relation to Pearlian causality and abstraction](#)
- [Partitions and factors in finite factored sets](#)
- [Orthogonality and time in finite factored sets](#)
- [Using finite factored sets](#)
- [Why not infinite factored sets?](#)
- [Limits of, and follow-up work on, finite factored sets](#)
- [Relevance to embedded agency and x-risk](#)
- [How Scott researches](#)
- [Relation to Cartesian frames](#)
- [How to follow Scott's work](#)

**Daniel Filan:** Before we begin, a note about this episode. More than other episodes, it assumes knowledge of the subject matter during the conversation. So, although we'll repeat some basic definitions, before listening there's a good chance that you'll want to watch or read something explaining the mathematics of finite factored sets. The description of this episode will contain links to resources that I think do this well. Now, on to the interview.

**Daniel Filan:** Hello everybody. Today, I'll be speaking with Scott Garrabrant. Scott is a researcher at the [Machine Intelligence Research Institute](#) or MIRI for short. And prior to that, he earned his PhD in Mathematics at UCLA studying combinatorics. Today, we'll be talking about his work on [finite factored sets](#). For links to items that we'll be discussing, you can check the description of this episode, and you can read a transcript at [axrp.net](#).

## Relation to Pearlian causality and abstraction

**Daniel Filan:** Scott, welcome to AXRP. I guess we're going to start off talking about the finite factored sets work, but to start off that starting off, you've kind of compared this to... Or I think it's sort of meant to be somehow in the same vein of Judea Pearl's work on causality, where you have this [directed acyclic graph](#) of nodes and arrows. And the nodes are things that might happen and the arrows are one thing causing another kind of. So I'm wondering, what's good about Pearlian causality? Why does it deserve to be developed on. Let's just start with that. What's good about Pearlian causality?

**Scott Garrabrant:** What's good about Pearlian causality. So specifically I want to draw attention to the fact that I'm talking about kind of earlier Pearlian stuff, like Pearl has a bunch of stuff. I'm talking about the stuff that you'll find in chapter two of the book [Causality](#) specifically. I mean, so basically it's a framework that allows you to take statistical data and from it infer temporal structure on the variables that you have. Which is just really useful for a lot of concepts, there are a lot of purposes. It's a framework that allows you to kind of just go from pure probabilities to having an actual structure as to what's going on and causality, which then lets you answer some questions about interventions possibly and things like that.

**Daniel Filan:** So if it's so great, why do we need to do any more work? Why can't we just all read this book and go home?

**Scott Garrabrant:** So my main issue is kind of a failure to work well with abstraction. And so we have these situations, possibly coming from decision theory, where we want to model agents that are making some choice and they have some effect on the world. And it makes sense to model these kinds of things with causality, which is not directly using Pearlian causal inference, but using just kind of the general framework of causality, where we kind of draw these directed acyclic graphs with arrows, that kind of represent effects that are happening.

**Scott Garrabrant:** And you run into these problems where if you have an agent and, for example, it's being simulated by another agent, then there's this desire to put multiple different copies of the same structure in multiple different places in your causal story. And to me it feels like this is really pointing towards needing the ability to have some of your nodes, some of your variables in your causal diagrams be abstract copies of other nodes and variables. And there's an issue, the Pearlian paradigm doesn't really work well with being able to have some of your variables be abstract copies of others.

**Daniel Filan:** So what's an example of a place where you'd want to have like multiple copies of the same structure in different places. If you could spell that out.

**Scott Garrabrant:** Yeah, I can give more specific direct examples that kind of are made up examples, but really I want to claim that you can see a little bit of this any time you have any agent that's making a decision. Agents will make decisions based on their model of the world. And then they'll make that decision - based on their model of the consequences of their actions, and they'll make a decision and they'll take an action. And then those consequences of their action will actually take place in the real world. And so you can kind of see that there's the agent's model of what will happen that is kind of causing the agent's choice, which kind of causes what actually happens. And there's this weird relationship between the agent's model of what will happen and what actually happens that could be well-described as the agent's model is kind of an abstract version of the actual future.

**Daniel Filan:** Okay. And why can't we have abstractions in Pearlian causality?

**Scott Garrabrant:** So the problem, I think, lies with what happens when you have some variables that are kind of deterministic functions of others. And so if you have an abstract refined version - if you have a refined version of a variable and you have another coarse abstract version of the same variable, you can kind of view the coarse version that has less detail as a deterministic function of the more refined variable. And then Pearl has some stuff that allows for determinism in the structure. But the part that I really like doesn't really have a space for having some of your variables be deterministically or partially deterministically related. And in the parts of Pearl where some of your variables can be deterministically related, the ability to do inference is much worse. The ability to infer causality from the statistical data.

**Daniel Filan:** Yeah. I guess to what degree are we dealing with strict determinism here? Because guesses can sometimes be wrong, right? If I'm thinking about the necessity of abstraction here as "I have models about things", it's not really the case that my model is a deterministic function of reality, right?

**Scott Garrabrant:** Yeah. This is right. I mean, there's a story where you kind of can have some real deterministic functions where you have multiple copies of the same algorithm in different places in space time or something but... Yeah, I feel I'm kind of dancing around my true crux with the Pearlian paradigm. And I don't know, I'm not very good at actually pointing at this thing. But my true crux feels something like variable non-realism, where in the Pearlian world, we have all these different variables and you have a structure that kind of looks like this variable listens to this variable. And then this variable listens to this other variable to kind of determine what happens. And in my world, I'm kind of saying that...

**Scott Garrabrant:** I'm kind of in an ontology in which there's nothing real about the variables beyond their information content. And so if you had two variables, one of which is a copy of the other, it wouldn't really make sense to talk about like... Yeah, if  $X$  and  $X'$  were copies of each other, it wouldn't really make sense to ask is  $Y$  looking at  $X$  or looking at  $X'$  if they're actually the same information content. And so kind of philosophically, I think that the biggest difference between my framework and Pearl's framework is something about denying the realism of the variables. Yeah, that didn't really answer your question about, do we really get determinism? I mean, I think that systems that have a lot of determinism are useful for models. We have systems that don't have real determinism, but we also don't actually analyze our systems in full detail.

**Scott Garrabrant:** And so I can have a high level model of the situation. And while this calculator is not actually deterministically related to this other calculator, relative to my high level model it kind of is. And so even if we don't get real determinism in the real world, in high-level models it feels still useful to be able to work okay with determinism. But I don't know, I think that focusing... I don't know, in some sense, I want to say determinism is kind of the real crux, but in another sense I want to say that it's distracting from the real crux. I'm not really sure.

**Daniel Filan:** It also seems like one issue - let's say I'm playing Go. And I'm thinking about what's going to happen when I make some move. And then I play the move and then that thing happens. Well, if we say that my model of what's happening is an abstraction of what actually happens, it's a function of what actually happens, then it's the case that there's sort of an arrow from what actually happens to what I think is going happen. And then there's an arrow from that to what I do, because that's what

caused me to make the decision and then there's an arrow from what I do to what actually happens because that's how normal things work. But then you have a loop which you're sort of not allowed to have in Pearl's framework. So that seems like kind of a problem.

**Scott Garrabrant:** Yeah. I think that largely my general, or a large part of my research motivation for the last, I'm not sure how long, I think at least three years has been a lot towards trying to fix this problem where you have a loop. Thinking about decision theory in ways that people were talking about decision theory in, I don't know, around 2016 or something. There was stuff that involved, well, what happens when you take DAGs [directed acyclic graphs], but you have loops and stuff like this. And I had this glimmer of hope around, well, maybe we can not have loops, when I realized that in a lot of stories like this, the loop is kind of caused by wanting to have... by conflating a variable with an abstract copy of the same variable. And that's not what you did. What you did was you drew an arrow from the thing that actually happens to your model.

**Scott Garrabrant:** And I think that... So in my framework there aren't going to be arrows, but to the extent that there are arrows like this, it makes more sense to draw the arrow from the coarse model to what actually happens. And to the extent that the coarse model is like a noisy approximation of what actually happens, you kind of won't actually get an arrow there or something. But in my world, the coarser descriptions of what's going on will kind of necessarily be no later than a more refined picture of the same thing. And so I think that I more want to say, you don't want to draw that same kind of arrow between the real world and the model. And kind of, if I really wanted to do all this with graphs, I would say you should at least draw an undirected arrow.

**Scott Garrabrant:** That's kind of representing their logical entanglement. That's not really causal or something like that. That's not the approach that I take. The approach that I take is to kind of throw all the graphs away. But yeah, so I largely gained some hope that these weird situations that felt like they were happening all over the place and setting up decision theory problems and agents that trust themselves and think about themselves and do all sorts of reasoning about themselves. I gained some hope that all of the weird loopy stuff that's going on there might be able to be made less loopy via somehow combining temporal reasoning with abstraction. And that's largely a good description of a lot of what I've been working on for many years.

**Daniel Filan:** All right. I guess if I think concretely about this coarse versus abstract thing. Imagine I'm like... So basically think about the Newcomb scenario. So Newcomb's problem is, there's this really smart agent called Omega. And Omega is simulating me, Daniel Filan. And Omega gives me the choice to just take one box that has an unknown amount of money to me in it, unknown to me. And two boxes where I take the unknown amount of money and also a box that contains \$1,000. And I can just see that it definitely has \$1,000. And Omega is a weird type of agent that says, "I figured out what you would do. And if you would've taken one box, I put a lot of money in the one box, but if you would have taken both then I put almost none in it."

**Daniel Filan:** And then I take one or the other. So what should I think of... It seems like in your story, there's an abstract variable, maybe in a non-realistic kind of way, because we're variable non-realists, but there's some kind of abstract variable that's Omega's prediction of what I do, and then there's what I actually do. And if it's the case that Omega just correctly guesses, always correctly predicts whether I'm going to one box or two box. What should I think of... In what way is this abstraction lossy? What extra information is there when I actually take the one or the two?

**Scott Garrabrant:** So I guess if Omega is just completely predicting you correctly. Then I don't want to say that... I kind of want to say, well, there's a variable-ish thing that is what you do. And it goes into Omega filling the boxes. And it also goes into you choosing the boxes. And it's kind of at one point in logical time or something like that. I think that the need for actual abstraction can be seen more in a situation where you, Daniel, can also partially simulate Omega and learn some facts about what Omega is going to predict about you. So maybe... Yeah, Omega is doing some stuff to predict you and you're also simulating Omega. And in your simulation of Omega, you can see predictions about stuff that you're currently doing or about to do and stuff like that.

**Scott Garrabrant:** And then there's this weird thing where now you have these weird loops between your action and your action. So in the situation where Omega was kind of opaque to you, the intuition goes against what we normally think of as time, but we didn't necessarily get loops, but in the intuition where you're able to kind of see Omega's prediction of you, things are kind of necessarily lossy because you could kind of diagonalize against predictions of you and kind of because it doesn't fit inside you. So, let's see. If you're looking at a prediction of yourself and you have some program trace, which is your computation, and you're working with an object that is a prediction of yourself and maybe it's a very good prediction of yourself. You're not actually going to be able to fully simulate every little part of the program trace because it's kind of contained inside your program trace.

**Scott Garrabrant:** And so there's a sense in which I want to say, there's some abstract facts that may be our predictions or proofs about what Daniel will do, and that can kind of live inside Daniel's computation. And then Daniel's actual program trace, is this more refined picture of the same thing. And so, I think the need for actually having different levels of abstraction that are at different times is coming more from situations that are kind of actually loopy as opposed to in the Newcomb problem you described. The only reason that it feels loopy is because we have this logical time, then we also have physical time and they seem to go in different directions.

## Partitions and factors

**Daniel Filan:** So now we've gotten a bit into the motivation. What is a finite factored set?

**Scott Garrabrant:** Okay. So there's this thing - I don't know, I guess I first want to recall the definition of a partition of a set. So a partition of a set is a set of subsets of that set. So we'll start with an original set  $S$  and a partition of  $S$  is a set of subsets of  $S$ , such that each of the sets is non-empty and the sets are pairwise disjoint, so they don't have any common intersection. And when you union all the sets together, you get your original set  $S$ . So it's kind of a way to take your set and view it as a disjoint union.

**Daniel Filan:** Yeah. I kind of think of it as like dividing up a set into parts and that way of dividing it up is a partition.

**Scott Garrabrant:** Yeah. And so I introduced this concept called a factorization, which can be thought of as a multiplicative version of a partition where you kind of... In the partition story, you put the sets next to each other and you union them together to get the whole thing. And in a factorization, I instead want to kind of multiply your different sets together. And so the way I define a factorization of a set  $S$

is it's a set of non-trivial partitions of  $S$  such that for each way of choosing a single part from each of these partitions, there will be a unique element of  $S$  that's in the intersection of those parts. And so the same way that you can view partition as a disjoint union, you can view a factorization as a... or sorry, a partition is a way to view  $S$  as a disjoint union, a factorization of  $S$  is a way to view  $S$  as a product.

**Daniel Filan:** Okay. And so to make that concrete, an example that I like to have in my head is suppose we have points on a 2D plane, and we imagine the points have an  $X$  coordinate and a  $Y$  coordinate. So one partition of the plane is I can divide the plane up into sets where the  $X$  coordinate is zero, sets where the  $X$  coordinate is one, sets where the  $X$  coordinate is two. And those look like lines that are perpendicular to the  $X$  axis. And none of those lines intersect. And every point has some  $X$  coordinate.

**Daniel Filan:** So it's this set of lines that together cover the plane, that's one partitioning, the  $X$  partitioning. And there's another one for values of  $Y$ , right? Which look like horizontal lines that are various amounts up or down. And so once you have the  $X$  partitioning and the  $Y$  partitioning, any point on the plane can be uniquely identified by which part of the  $X$  partitioning it is and which part of the  $Y$  partitioning it is because it just tells you how much to the right of the origin are you, how much above the origin are you, that just picks out a single point. I'm wondering, do you think that's a good kind of intuition to have.

**Scott Garrabrant:** Yeah, I think that's a great example. To say a little bit more there. So your original set  $S$  in that example you just gave is going to be the entire Cartesian plane, the set of all ordered pairs, like ( $X$  coordinate,  $Y$  coordinate). And then your factorization is going to be a set  $B$ , which is going to have just two elements. And the two elements are the partition according to what is the  $Y$  coordinate, and the partition according to what is the  $X$  coordinate. You can kind of view partitions as questions. And so in general, if I have a set like the Cartesian plane, and I want to specify a partition, one quick way to do that is I can just ask a question. I can say, what is the  $X$  coordinate? And that question kind of corresponds to the partition that breaks things up according to their  $X$  coordinate.

**Daniel Filan:** Okay. And the one slightly misleading thing about that example is that there are an infinite number of points in the  $X$ ,  $Y$  plane. But of course, we're talking about finite factored sets. So  $S$  only has a finite number of points.

**Scott Garrabrant:** Yeah. So we're talking about finite factored sets. So in general, I'll want to work with a pair  $(S, B)$ . Where  $S$  is a set, a finite set, and  $B$  is a factorization of  $S$ .

**Daniel Filan:** Why choose the letter  $B$ , for a factorization?

**Scott Garrabrant:** It's for basis.

**Daniel Filan:** Mmm.

**Scott Garrabrant:** Yeah. It's for basis. Largely, while I'm thinking of the elements of  $B$  as partitions of  $S$ , I'm also kind of just thinking of them as elements, just out on their own that are kind of representing the different basis elements. Yeah, so it actually looks a lot like a basis, because for any point in  $S$  you can kind of uniquely specify it by specifying its value on each of the basis partitions.

**Daniel Filan:** Okay. So yeah, this gets into a question I have which is, how should I think about the factors here? Like these finite factored sets, I guess they're supposed

to represent what's going on in the world. How should I think about factors in general or partitions, I can think about as questions. These factors in B, how should I think about those?

**Scott Garrabrant:** Yeah, I think I didn't explicitly say, but the elements of B, we'll call factors. We use the word factor for the elements of B. So I almost want to say the factors are kind of a preferred basis for... Yeah, the factors kind of form a preferred basis for your set of possibilities. And so if I consider the set of all bit strings of length five, right. There are 32 elements there. And if I wanted to specify an element, I can do so in lots of different ways. But there's something intuitive about this choice of breaking my 32 elements into what's the first bit, what's the second bit, what's the third bit, what's the fourth bit, what's the fifth bit. So it's kind of... It's a set of questions that I can ask about my element, that uniquely specify what element it is.

**Scott Garrabrant:** And also for any way of answering the questions, there's going to be some element that works with those answers. And so factorization is kind of just like, it's a combinatorial thing that could be used for many different things. But one way to think about it is kind of, you're making a choice that is a preferred basis, a preferred set of variables to break things up into and you're thinking of those as kind of primitive and then other things as built up from that. So properties like do the first two bits match is then thought of as built up from what is the first bit and what is the second bit. And so it's kind of a choice of what comes first, a choice of what are the primitive variables in your structure.

**Daniel Filan:** Okay. So if I'm trying to think about maybe some kind of decision problem where I'm going to do something, and then you're going to do something and then another thing's going to happen. And I want to model that whole situation with a finite factored set. Instead of thinking about modeling which thing is going to happen to me today, if I want to model an evolving situation, how should I think about what the set S is and what the factors should be?

**Scott Garrabrant:** Yeah. So factorization is very general and actually, I use the word factorization in multiple different contexts when talking about this kind of thing. But in the question that you're asking, to say some more background stuff. So I'm going to introduce this theory of time that has a background structure that looks like a factored set rather than a background structure that looks like a DAG, looks like a directed acyclic graph, like in the Pearlian case. And so specifically if I'm using a factored set and I am using that to interpret... I'm using that to kind of describe some causal situation, I am not going to have nodes. I'm not going to have factors that kind of correspond to the nodes that you would have in the Pearlian world. Instead, I am mostly going to be thinking of the factors as independent sources of randomness.

**Scott Garrabrant:** I'm hesitant there, because a lot of my favorite parts of the framework aren't really about probability. But if I'm thinking about it in a temporal inference setting where I'm like getting the statistical distribution of... Where I'm getting a statistical distribution, then I'm thinking of the factors as basically independent sources of randomness. And so if we have some variable X, and then we have some later variable Y that can take on some values and partially is going to be a function of X, then we won't have a factor for X and a factor for Y, we'll have a factor for maybe that which kind of goes into what X is. And then we'll have another factor that's like the extra randomness that went into the computation of Y once you already knew X. And when we think of factored sets as related to probability, we're actually going to always want our factors to be independent. And so the factors can't really be put on things like X and Y when there's a causal relationship between X and Y.

**Daniel Filan:** So it almost sounds like the factors are supposed to be somehow initial data, like the problem set-up or something, the specification of what's going on, but kind of an initial specification?

**Scott Garrabrant:** Yeah, indeed, if you take my theory of time. So I have way of taking a factored set and taking an arbitrary partition on your set  $S$ . And then I'm going to be able to specify time, specify when some partitions are before or after other partitions. The factors will be partitions with the property that you can't have anything else that's strictly before it besides deterministic things. And so there's a sense in which the factors are initial.

**Scott Garrabrant:** Yeah, the factors are basically the initial things in the notion of time that I want to create out of factored sets. But also intuitively, it feels like they're initial. It feels like they're the primitive things that came first and then everything else was built out of them.

**Daniel Filan:** Yeah, I think primitive is a better word than initial.

**Scott Garrabrant:** Yeah, they're - In the [poset](#) of divisibility, right? I wanted to say that one is initial, not the primes. But the primes have the property that you can't really have anything else before them. Yeah, I don't know, primitive is like prime is... Yeah.

**Daniel Filan:** Yeah. So in the sense they're like when you're dividing numbers, whole numbers, you get to primes and then you get nowhere else, and they're primitive in that sense.

**Scott Garrabrant:** Right.

## Orthogonality and time

**Daniel Filan:** Yeah. So you talk about the history of variables as all the initial factors that you need to specify what a variable is. And then you have this definition of orthogonality. And people can read the paper. I don't know if they'll be able to read the paper by the time this goes live, but they'll be able to read something to learn about what orthogonality is.

**Scott Garrabrant:** Yeah, the [talk](#) and the [transcript of the talk](#) that, for example, appears on [the MIRI blog](#) or on [LessWrong](#), it has all of the statements of everything that I find important. And so you shouldn't have to wait for a paper. I think that, modulo the fact that you'd have to prove things yourself, that has all the important stuff.

**Daniel Filan:** So, people, I think for me at least, it's easy to read the definition of orthogonality and still not really know how to think about it, right? So how should people think about what's orthogonal?

**Scott Garrabrant:** So in the temporal inference thing where we're going to be connecting up these combinatorial structures with probability distributions, orthogonality, it's going to be equivalent to independence. And so, one way to think of orthogonality is... I kind of took out a combinatorial fragment of independence, where I'm not actually working with probabilities but I am working with a thing that is representing independence. Another way to think of it is like when two partitions are orthogonal, you should expect that if you come in and tweak one of them, it will have

no effect on the other one, they're separated. And specifically in the factored set framework, orthogonality means they do not have a common factor. Where these factors can be thought of as sources of randomness or sources of something. If they do not share a common factor, then they're in some sense separated.

**Daniel Filan:** Yeah. So I'll just say one example that helped me understand it is this XY-plane example where the factors were, you're partitioning up according to the X coordinate. And you can also partition up according to what the Y coordinate is. So you can have one different partition, that's are you on the left-hand side or the right-hand side of the Y-axis. That's a partition of the plane. It's like a variable. It's like are you on the left, or are you on the right. And there's a second variable, that's are you above the X-axis, or are you below the X axis. Right? That also partitions the plane up. And my understanding is that in this factored set, those two partitions, or you could think of them as variables, are orthogonal, and that hopefully gives people a sense. It's also kind of nice because they're literally - if you think about the dividing lines, they're literally orthogonal in that case, and that's maybe not a coincidence.

**Scott Garrabrant:** Yeah. So historically when I was developing factored set stuff, I was actually working with things that looked like I have two partitions and there's something nice when it's the case that they're both kind of mutually... Or sorry, for any way of choosing a value of this X partition and also choosing the value of the Y partition, those two parts will intersect. So in your example, right, the point is that all four quadrants exist, and this is like there's a sense in which this is saying that the two partitions aren't really stepping on each other's toes.

**Scott Garrabrant:** If we specify the value of this X partition, it doesn't stop us from doing whatever we want in terms of specifying the value of the Y partition. And largely orthogonality is a step more than that, where a consequence of being orthogonal, this is they're not going to step on each other's toes. But I have this extra thing, which is really just the structure of the factorization, but there's this extra thing beyond just not stepping on each other's toes, which is coming from some sort of theory of intervention or something. You can view the factored set as a theory of intervention because the factored set basically allows you to take your set and view its elements like tuples. And when your elements are like tuples, you can imagine going in and changing one value and not the others. And so, orthogonality looks like it's not just X and Y are compatible, like compatible in all ways and assign values to each of them, but it's also, when you mess with one, it doesn't really change the other.

**Daniel Filan:** Cool. So another question about orthogonality. So in the talk that listeners can [watch](#) or [read a transcript of](#), you sort of say, "Okay, we have this definition of orthogonality and this definition of conditional orthogonality, which is a little bit more complicated but kind of similar." You then talk about inference in the real world. So we sort of imagine that you're observing things that are roughly like... You don't observe the underlying set, but you observe things that are roughly like these factors. And somehow you get evidence about what things are orthogonal to each other. And from that, you go on and infer a whole bunch of stuff, or you can sometimes. And you give an example of how that might work. How would I go about gathering this orthogonality data of what things in the world are orthogonal to what other things?

**Scott Garrabrant:** So the default is definitely passing through this thing that I call the fundamental theorem. So the fundamental theorem says that conditional orthogonality is equivalent to, for all probability distributions that you can put on your factored set which respect the factorization, your variables are conditionally

independent. I don't know, I phrased that as for all probability distributions, but you can quickly jump that for a probability distribution in general position.

**Scott Garrabrant:** And so the basic thing to do is if you have access to a distribution over the things, over the elements of your set or over something that's a function of the elements of your set, if you have access to a distribution, it's kind of a reasonable assumption to say that if you have a lot of data and it looks like these two variables are independent, then you assume that they are orthogonal in whatever underlying structure produced that distribution. And if they're not independent, then you see they're not orthogonal.

**Scott Garrabrant:** And it's just - the default way to get these things is via taking some distribution, which could be coming from a bunch of samples, or it could be a Bayesian distribution. But I think that orthogonality is something that you can basically observe through its connection to independence versus time is not as much.

**Daniel Filan:** I guess this works when my probability distribution has this form where the original factors in your set  $B$ , like the primitive variables, have to be independent in your probability distribution. And if they're independent in that distribution, then conditional independence is conditional orthogonality. How would I know if my distribution had that nice structure?

**Scott Garrabrant:** You can just interpret all of the independence as orthogonality and see whether it contradicts itself, right? It's like you might have a distribution that can't really be well-described using this thing. And one thing you might do is you might develop an orthogonality database where you keep track of all of the orthogonalities that you observed. And then you notice that there's some orthogonalities that you observed that are incompatible with coming from something like this. Yeah, I'm not sure I fully understand the question.

## Using finite factored sets

**Daniel Filan:** I guess what I'm asking is imagine I'm in a situation, right? And I don't already have the whole finite factored set structure of the world. I'm wondering how I go about getting it. Especially if the world is supposed to be my life or something. So I don't get tons of reruns. Maybe this isn't what it's supposed to do.

**Scott Garrabrant:** Basically the answer you'd give here is similar to the answer you'd give to the same question about Pearlian Causality. And I think that largely the temporal inference story makes the most sense in a context that's very like you have a repeated trial that you can repeat an obnoxious number of times, and then you can get a bunch of data. And you're trying to tell a story about this trial that you repeated. And yeah, so the story that I tell makes the most sense in a situation that's like that.

**Scott Garrabrant:** I'm excited for a lot of things about factored sets that are not about just doing temporal inference from a probability distribution. And those feel like they play a lot more nicely with... Sorry, the applications that feel like they're about embedded agents to me are different from the applications that feel like they're about temporal inference. Because it feels like you need something like this frequentist, lots of repetition to be able to get a distribution in order to be able to do a lot of stuff with temporal inference. Or at least doing it the naive way, maybe you can build up more stuff.

**Daniel Filan:** So embedded agency is your term for something being an agent in a world where your thinking processes are just part of the physical world and can be modified and modeled and such. Is that a fair summary?

**Scott Garrabrant:** Yeah.

**Daniel Filan:** Okay, so the applications of the finite factored set framework to embedded agency, are you thinking of that more as a way to model things?

**Scott Garrabrant:** Yeah, I'm thinking of it mostly like basically, there's a lot of ways in which people model agents using graphs where the edges in the graph represent information flow or causal flow or something, that all feel they're entangled with this Pearlian causality story. And often I think that pictures like this fail to be able to handle abstraction correctly, right? I have a node that represents my agent. And I don't really have room for another node that represents a coarser version of my agent because if I did, which one gets the arrow out of it and such. And so largely, my hope in embedded agency is just all the places where we want to draw graphs, instead, maybe we can draw a factored set and this will allow things to play more nicely with abstraction and playing nicely with abstraction feels like a major bottleneck for embedded agency.

## Why not infinite factored sets?

**Daniel Filan:** Okay. Yeah, I'll ask more about that a bit later. Yeah, so I guess I want to ask more questions about the finite factored set concept itself. Why is it important that it's finite?

**Scott Garrabrant:** So I can give an example where you should not expect the fundamental theorem to hold in the cases where it's infinite. So the thing where independence exactly corresponds to orthogonality, in the infinite case, one shouldn't expect that to hold. And it might be that you can save it by saying, "Well, now we can't take arbitrary partitions. We can only take partitions that have a certain shape."

**Daniel Filan:** Sort of like measurability criteria?

**Scott Garrabrant:** Sort of like measurability criteria, but measurability is actually not going to suffice for this. I can give an example. To give an example, if you imagine the infinite factored set that is countable bit strings.

**Daniel Filan:** So this is just like infinite sequences of ones and zeros, the set of all of them?

**Scott Garrabrant:** Yeah. So you have the set of all infinite sequences of ones and zeros, and you have the obvious factorization on this set, which is you have one factor for each bit. And then there's a partition that is asking, "Is the infinite bit string, the all zero string?" And there's the partition that's asking, "Is the infinite bit string, the all one string?" These are two partitions, let's call them X and Y.

**Daniel Filan:** Okay.

**Scott Garrabrant:** And it turns out that for any probability distribution you can put on this factored set that respects the factorization. At least one of this partitions is going to have to be - either it's going to be the case that the all zero string is probability zero or the all one string is probability zero.

**Daniel Filan:** Yep.

**Scott Garrabrant:** Because all of the bits have to be independent. And then you'd be able to conclude that the question, "Is it the all zero string?" And "Is it the all one string?" Those two questions will have to be independent in all distributions on the structure, but it really doesn't make sense to call them orthogonal.

**Daniel Filan:** Why doesn't it?

**Scott Garrabrant:** Why doesn't it make sense to call them orthogonal? It doesn't make sense to call them orthogonal because all of our factors go into the computation of that fact. And so, if you're thinking of orthogonality as, they can be computed using disjoint collections of factors, you can't really compute whether something's the all zero string or whether something's all the one string without seeing all the factors. I mean, you can't compute it because the first time you see a one, you can say, "All right, I'll stop looking." But in my framework, that doesn't count. You have to specify upfront all the bits you're going to use. And so, I don't know, I think there's some hope to being able to save all of this. And I haven't done that yet. And there's another obstacle to infinity, which is even the notion of thinking about the history of a partition, that's not even going to be well-defined in the infinite case.

**Daniel Filan:** Okay, and so the history of a partition in the finite case, it was just the smallest set of factors that determined - if I have a partition with elements like  $X_1$  through  $X_k$  or something, the history of the partition is just all of the basic factors, all the things in my set  $B$  where if you think of the things in the set  $B$ , its variables, if I know the value for all the variables, it's the smallest set of variables, such that if I know the value for all of them, I can tell what partition element I'm in for the thing that I'm looking for the history of. So it's the smallest set of, the smallest amount of initial information or something, that's specifying the thing I'm interested in. Is that what a history is?

**Scott Garrabrant:** Yeah, that's right.

**Daniel Filan:** And it's a bit worrisome because usually there's not a smallest set of sets of... Right? That's not obviously well-defined.

**Scott Garrabrant:** Yeah. So, specifically smallest in the subset ordering. And so the history of a partition is a set of factors. And if you take any set of factors that would be sufficient to determine the value of that partition, the answer to that question, what part of it's in, if you have any set of factors that were sufficient to compute the value of  $X$ , then necessarily that set of factors must be a superset of the history. And so it's not smallest by cardinality, it's smallest by the subset ordering. And showing that this is well-defined involves basically just showing that sets of factors that are sufficient to determine the value of  $X$  are closed under intersection. So if I have two different sets of factors and it's possible to compute the value of  $X$  with either of those sets of factors, then it's possible to compute the value of  $X$  using their intersection.

**Scott Garrabrant:** But actually this is only true for finite intersections. And so if I have a partition and you're able to compute the value from either of these two sets of factors, then you're able to compute the value from their intersection. But if I have an infinite class of things then I can't necessarily compute the value from their intersection. To see an example of this, if you, again, look at infinite bit strings with the obvious factorization and you look at the partition, are there finitely many ones? If you take any infinite tail of your bit string that's sufficient to determine, are there

finitely many ones? But if you take the intersection of all of these infinite tails, you get the empty set.

**Daniel Filan:** Yeah, so one way I'm now thinking of this is the problem with infinite sets is that they have these things that are analogous to tail events or something in normal probability theory where you depend on all of these infinite number of things, the limit of it. But the limit is actually zero, but you can't exactly-

**Scott Garrabrant:** Yeah, there's this - actually, what I think is a coincidence.

**Daniel Filan:** Okay.

**Scott Garrabrant:** But you could do this naive thing where you say, "Well, just take the intersection and call that the history." And then the history of the question, "Are there finitely many ones?", would then be the empty set and then a lot of properties would break. But an interesting coincidence here or something, I don't know, I don't think it's a coincidence, but I don't like this definition. I don't like the definition of generalizing to the infinite case by just defining the history to be the intersection. But if you were to do that, then you would get that the question, "Are there finitely many ones?", is orthogonal to itself because it has empty history. And what kind of partitions are orthogonal to themselves? They're deterministic ones where one of the parts has probability one and all the others have probability zero.

**Scott Garrabrant:** And the [Kolmogorov zero-one law](#) says that properties like the question, "Are there finitely many ones?", if all of the individual things are independent, are necessarily probability zero or one. And so there's this thing where if you define history in this way, if you naively extend history to the infinite case by just taking the intersection, even though it's not closed under intersection, you actually get something that feels like it gives the right answer because of the Kolmogorov zero-one law.

**Daniel Filan:** Yeah, but it's going through some weird steps to get the right answer.

**Scott Garrabrant:** Yeah.

**Daniel Filan:** By the way, I don't know, there are a variety of these things called [zero-one laws](#) in probability theory, and if listeners want to think about knowledge and changing over time, I don't know. Some of these zero-one laws are fun to mull over and think about how they apply to your life or something. Oh, do you have comments on that claim?

**Scott Garrabrant:** No, I think Kolmogorov's zero-one law is really interesting and I would recommend it to people who like interesting things.

## Limits of, and follow-up work on, finite factored sets

**Daniel Filan:** All right. So back on the finite factored sets, they're sort of a way of modeling some types of worlds. Or some sorts of ways the world can be. Are there any worlds that can't be modeled by finite factored sets?

**Scott Garrabrant:** Yeah.

**Daniel Filan:** I guess infinite ones, but ignoring that for this second.

**Scott Garrabrant:** So there's this issue that's similar to an issue in Pearl where we kind of - when you look at distributions that are coming from a finite factored set or from a DAG, we're looking at probabilities in general position. And so it doesn't really make sense to have a probability one half or one fourth. Although-

**Daniel Filan:** What do you mean by probability in general position?

**Scott Garrabrant:** So in both my world and Pearl's world, we want to specify a structure.

**Daniel Filan:** Yep.

**Scott Garrabrant:** And then we have all of the probability distributions that are compatible with that structure. And these give you a manifold or something of different probability distributions. And then some measure zero subset of them have special coincidences. And when I say in general position, I mean, you don't have any of those special coincidences. And so an example of a special coincidence is any time you have a probability of one half or one fourth or something like that, that's a special coincidence. But to a Bayesian, that might happen because of principle of indifference or something. Or to a limited Bayesian that kind of doesn't really know what... It feels like principle of indifference advises having probabilities that are rational numbers, but probabilities that are rational numbers lead to coincidences in independence that don't arise from orthogonality. And so there's a sense in which my framework and the Pearlian framework don't believe in rational probabilities as something that just happens, or something.

**Daniel Filan:** Yeah. I mean, it's even more concerning because if I think that I'm a computer and I think I assign probabilities to things, well, the probabilities I assign will be numbers that are the output of some computation. And there are a countably infinite number of computations, but there are an uncountably infinite number of real values. So somehow, if I'm only looking at things in general position, I'm ruling out all of the things that I actually could ever output.

**Scott Garrabrant:** Yeah, so I have a little bit of a fragment of where I want to go with trying to figure out how to deal with the fact that my system and Pearl's system don't believe in rational probabilities, which is to define something, and this is going to be informal. Well, it'll be formal but wrong. To define something that's like, you take a structure that is a factored set together with a group of symmetries on the factored set that allows you to maybe swap two of the parts within a partition or swap two of the partitions with each other, or swap two of the factors with each other, right? So you can have some symmetry rules. So if you, for example, considered bit strings of length five again, you could imagine a factored set that it separates into the five bit locations, is this bit zero or one, but then it also has the symmetry of you can swap any of the digits with each other that can also do swapping zero with one. But for now, I'll just think about swapping any of the digits with each other.

**Scott Garrabrant:** And then it's subject to the structure that this is this factored set and the swapping rule for this group of symmetries, then the set of distributions that are compatible with that structure will not be just things in which the bits are independent from each other, it'll be any distribution in which the bits are IID, so independent and also identical. And so you could do that. So you could say, "Well, now my model of what might happen, the thing that I'm going to try to infer from my probability distribution is a factored set together with a group of symmetries on that

factored set." And I mean, you're not going to at least naively get the fundamental theorem the same way. And so I'm not sure what happens if you try to do inference on something like this, but if you want to allow for things like rational probabilities, then maybe something like that would be helpful.

**Daniel Filan:** All right, so I'd like to ask a question about the form of the framework. So when I think of models of causality or of time, the two prior ones that I think of are, we talked about Pearl's work on directed acyclic graphs and do-operators and such, where you can draw a graph, and also Einstein's work on special and general relativity, where you have this space time thing that's very geometric, very curved and you have this time direction, which is special and kind of different from the spatial directions. Those were all really geometric and included some nice pictures. Finite factored sets does not have many pictures. Why not? I really liked pictures.

**Scott Garrabrant:** Yeah. I mean, I think it has something to do with the variable non-realism, where it feels like the points or nodes in your pictures or something - if I take a Pearlian DAG and there are 10 nodes in it and even if I assume that they're all just binary facts, then now I have, well, you take 1024 different ways that the world can be. And then you take Bell number of 1024 different possible variables that I can define on this, which is obviously huge and then there's not as much of a useful interpretation of the arrows that connect them up or something, it's something to do with variable non-realism, where there's a sense in which Pearl's kind of starting from a collection of variables, which is a way to factor the world into some small object. And because I'm not starting with that, my world is kind of a lot larger.

**Scott Garrabrant:** I don't know. Another thing that I'd call a theory of time is people talk about time with entropy. That's another example that doesn't feel as visual.

**Daniel Filan:** Yeah, that's true.

**Scott Garrabrant:** And I think that that's a lot more variable free and that's maybe part of why.

**Daniel Filan:** Yeah. It's also the case that once you have variables, they have these relations in terms of their histories and such, you could draw them in a DAG or something.

**Scott Garrabrant:** Yeah. It's like the structure of an underlying finite factored set is very trivial or something. It's - Pearl has a DAG, and if you wanted to draw a finite factored set as a DAG, it would just be a bunch of nodes that are not connected at all that each have their own independent sources of randomness. And then if you wanted to, you could maybe draw an arrow from these nodes to all the different things that you can compute using them. Or if you wanted to say, oh, let's just have the basic variables then it's just these nodes.

**Daniel Filan:** Yeah. But I guess somehow if you want to talk about the structure that lets you talk about variables, but you don't want to talk about variables. I guess that's less amenable to pictures perhaps.

**Scott Garrabrant:** Yeah. I mean, I don't feel like physics and Pearl have pictures for necessarily the exact same reason or something, and I'm kind of just, graphs got lucky in the way that they're easy to visualize or something like that.

**Daniel Filan:** Yeah, that might be right. It's also true that graphs are not - simple graphs are easy to visualize, but there are a lot of non-planar graphs that are kind of a

pain to draw. Yeah. So, a related question complaint I have is a lot of this work seems like it could be category theory.

**Scott Garrabrant:** Yeah. It could be category theory.

**Daniel Filan:** Yes. So partitions are basically functions from a set to a different set and the parts are just all the things that have the same value of the function?

**Scott Garrabrant:** Yeah. Partitions is kind of the information content of a function out that you get by kind of ignoring the target and only looking at the source.

**Daniel Filan:** Yeah. And it seems like there are probably nice definitions of factors and such and... You know, there's lots of duality and category theory has pictures. It's also a little bit nice in that I have to admit, looking at sets of sets of sets of sets can be a little bit confusing after a while in the way that categories can have some nice language for that. So why isn't everything category theory, even though category theory is objectively great?

**Scott Garrabrant:** Yeah. I mean, it goes further than that. I actually know most of the category theory story and I've worked out a lot of it and kind of went with the combinatorialist aesthetic with the presentation anyway. So what are my reasons? One reason is because I kind of trust my category theory taste less and I kept on changing things or something in a way that I was not getting stuff done in terms of actually getting the product out or something by working in category theory. And so it was kind of oh yeah, I'll punt that to the future.

**Scott Garrabrant:** Another reason is because it feels the system just really doesn't have prerequisites and by phrasing everything in terms of category theory, you're kind of adding artificial prerequisites that maybe make the thing prettier, but you actually, you know what a set is, you can kind of go through all the proofs or something. That's not entirely true, but it's because I'm working in a system that has very few prerequisites, the extra cost of prerequisites is higher. The marginal cost of adding prerequisites is higher.

**Scott Garrabrant:** Another reason was I was just really shocked by [the sequence that counts the number of factorizations](#) not showing up on [OEIS](#). So yeah, if you take an n-element set and you count how many factorizations there are on the n-element set, you get a sequence and there's this Online Encyclopedia of Integer Sequences that has 300,000 sequences and it does not have this sequence in spite of having a bunch of lower quality sequences.

**Scott Garrabrant:** And I was very surprised by this fact, and it feels like a very objective test. I'm not a particularly scholarly person. It's hard for me to figure out what people have already done. And I was just pretty blown away by the fact that this thing didn't show up on OEIS. And so I kind of stuck with the combinatorialist thing because it had that objective thing for the purpose of being able to do an initial sell or something.

**Daniel Filan:** Okay.

**Scott Garrabrant:** Yeah. Those are most of my reasons. I think that - I haven't worked out all the category theory, but I think it will end up being pretty nice. In fact, I think that just even the definition of conditional orthogonality, I think can be made to look relatively nice categorically, and it's via a path that is pretty unclear from the definition that I give in [the talk](#) or [the post](#) but there's an alternative definition that

kind of looks like if you want to do orthogonality and you want to condition on some fact about the world, the first thing you do is you take your original factored set and you kind of take the minimal flattening of it. So that the thing you want to condition on is kind of rectangular in your factored set and then you - where by flattening I mean, you merge some of the factors together.

**Daniel Filan:** Okay.

**Scott Garrabrant:** And if you take the minimal factoring and then you ask whether your partitions are orthogonal in the minimal factoring, that corresponds to conditional orthogonality. And so I think that categorically there's a nice definition here. But I definitely agree about the category theory aesthetic and I think that it actually is a good direction to go here that I may or may not try to do myself, but if somebody was super interested in trying to convert everything to category theory, I could talk to them about it.

**Daniel Filan:** So speaking of that, I'm wondering, what follow-up work are you excited about being done here? And do you think that this kind of development is going to look more like showing nice things within this framework - making it categorical or showing the decidability of inference in finite factored sets? Or do you think it's going to look more like iterating on some of the definitions and tweaking the framework a little bit until it's the right framework?

**Scott Garrabrant:** Yeah. I mean the category theory thing a little bit does fall into tweaking the framework until it's the right framework. Although it's a little different, I have applications I'm excited about in both spaces.

**Scott Garrabrant:** Yeah, if I were to list applications that I expect that I'm not personally going to do, that seem like projects that would be interesting for people to pick up, one of them would be converting everything to category theory. One of them would be figuring out all the infinite case stuff and looking at applications to physics. I think that there's a non-trivial chance of some pretty good applications in physics that would come out of figuring out all the infinite case stuff. Because I think that factored sets are actually a lot closer to being able to give you something like continuous time than the Pearlian stuff. Yeah.

**Scott Garrabrant:** So one would be the physics thing. One would be basically trying to do computational stuff in terms of, I kind of just have a couple proofs of concept of how to do temporal inference. And I think you said, showing decidability of temporal inference is a thing, where really it's... I think that somebody should be able to actually search over the space of a certain flavor of proof and be able to actually come up with examples of temporal inference that come from this where you take in some orthogonality data and are able to infer time from them. And I think that there's a computational question here that I think, I might be wrong, might be able to at least be able to produce some good examples, even if it's not actually doing temporal inference in practice, and so I'd be excited about something that.

**Scott Garrabrant:** I would be excited about somebody trying to extend to symmetric finite factored sets, which is the thing I was talking about earlier, about dealing with rational probabilities. I think that of these that I listed, the one that I'm most likely to want to try to work on myself is the symmetric factored sets thing. Because I think that could actually have applications to the kind of embedded agency type stuff I'd want to work with. But for the most part I'm expecting to myself think in terms of applications, as opposed to think in terms of extending the theory, and all the things

that I kind of said, were all forms of extending the theory either by tweaking stuff or by kind of putting stuff on top of it. I think it's mostly just putting stuff on top of it.

**Scott Garrabrant:** I think I don't say that much. Sorry, I think that there aren't that many knobs to twiddle with the basic thing. You could have some new orientation on it, but I think it will be basically the exact same thing. I think that the parts... The way that I defined it, there was only one factorization on a zero element set. Maybe it would be nicer if there were infinitely many factorizations on zero. The definitions might be slightly different, but I think it's basically the same core thing.

**Scott Garrabrant:** I think that I'm mostly thinking that the baseline I have is kind of correct enough for the kind of thing that I want to do with it, that I don't expect it to be a whole new thing. I expect it to be built on top, and there's different levels of built on top.

## Relevance to embedded agency and x-risk

**Daniel Filan:** So I guess I'd now like to pivot into a bit of a more general discussion about your research and your research taste. How do you see the work on finite factored sets as contributing to reducing existential risk from artificial intelligence? If you see it as doing that?

**Scott Garrabrant:** I think that a lot of it factors through trying to become less confused about agency and embedded agency, which I don't know, I have opinions in both directions about the usefulness of this. Sometimes I'm feeling like, yeah this isn't going to be useful and I should do something else. And then sometimes I'm just interacting with questions that are a lot more direct and noticing how a lot of the kind of questions that I'm trying to figure out for embedded agency actually feel like bottlenecks to be able to say smart things about things that feel more direct.

**Daniel Filan:** Can you give an example of that?

**Scott Garrabrant:** I don't know. Evan says some things about [myopia](#).

**Daniel Filan:** That's Evan Hubinger.

**Scott Garrabrant:** Right. And that feels a lot more direct, trying to get a system that's kind of optimizing locally and not looking far ahead and stuff like that.

**Scott Garrabrant:** And I feel like, in wanting to think about what this even means, I notice myself wanting to have a better notion of time and better notion of things about the boundaries between agent and environment and all of this stuff. And so I don't know... That's an example of something that feels kind of more direct, myopia feels like something that could be very useful if it could be implemented correctly and could be understood correctly.

**Scott Garrabrant:** And when I try to think about things that are more direct than embedded agency, I feel like I hit the same kind of cruxes and that working with embedded agency feels like it's more directed at the cruxes, even though it's less directed at the actual application of the thing in a way that I expect to be useful.

**Daniel Filan:** In the myopia example, I think the first-pass solution would be look, there's just physical time [that] basically exists. And we're just going to say, okay, I want an AI system. I want it to care about what's going to happen in the next 10

seconds of physical time and not things that don't happen within the next 10 seconds of physical time. Do you think that's unsatisfactory?

**Scott Garrabrant:** Yeah. So, I mean, I think that you can't really look at a system and try to figure out whether it's optimizing for the next 10 seconds or not. And I think that - the answer that I actually gave with the myopia thing was a little off because I was actually remembering a thought about myopia, but it wasn't about time. It was about - just time was the thing that I said in that thing.

**Scott Garrabrant:** It was more about counterfactuals and more about the boundary between where the agent is or something like that. But I don't know. I still think the example works. I mean, I think that it comes down to you want to be able to look at a system and try to figure out what it's optimizing for. And if you have the ability to do that, you can check whether it's optimizing for the next 10 seconds, but in general, you don't really have the ability to do that.

**Scott Garrabrant:** Figure out what it's trying to do or something like that. And I think that... Yeah, how do I get at the applications? Okay. So one thing I think is that in trying to figure out how the system works, it is useful to try to understand what concepts it's using and stuff like this. And I think that the strongest case I can kind of make for factored sets is that I think that there's a sense in which factored sets is also the theory of conceptual inference. And I think that this could be helpful for looking at systems or trying to do oversight of systems that you want to be able to look at a thing and figure out what it's optimizing for.

**Daniel Filan:** In what ways would you say it's a theory of conceptual inference?

**Scott Garrabrant:** Well, one way to look at the diff between factored sets and Pearl is that we're kind of not starting from a world factored into variables instead we're inferring the variables ourselves. And so there's a sense in which if you try to do Pearl style analysis on a collection of variables, but you messed it up, and instead of having a variable for what number this... I have a number and it's either zero or one and it's also either blue or green. And I can also invent this concept called grue, which is a green zero or a blue one. And instead of thinking in terms of what's the number and is it blue, you can think of what's the number and is it grue, and maybe if you're working in the latter framework, you're kind of using the wrong concepts and you will not be able to pull out all the useful stuff you'd be able to if you were using the right concepts.

**Scott Garrabrant:** And factored sets kind of has a proof of concept towards being able to distinguish between blue and grue here, where the point is, in this situation, if the number is kind of independent of the color and you're working with the concept of number and the concept of grue-ness, you have this weird thing where it looks there's a connection between number and grue-ness, but it also is the case that if I invent the concept of number [xor](#) grue-ness, I kind of invent color, and color lets me factor the situation more and see that maybe you should think of it as the number and the color are primitive properties like we were saying before, and grue-ness is a derived property.

**Scott Garrabrant:** And so there's a sense in which earlier things are more primitive, and it's not just earlier things, I think there was more than just that. But there's a sense in which because I'm not taking my variables or my concepts as given, I am also doing some inferring which concepts are good.

**Daniel Filan:** So somehow it strikes me that inferring which concepts are good, is a related, but different problem to inferring which concepts a system is using.

**Scott Garrabrant:** I don't know, there's stuff that you like to think about that involve kind of having separate neurons as part of it. And I think there's a sense in which it might be that we're confused when we're looking at a neural net because we're thinking of the neurons as more independent things, when really they could be a transform similar to the blue/grue thing from some other thing that is actually happening and being able to have objective notions of what's going on there - being able to have a computation and having there be a preferred basis that causes things to be able to factor more or something feels... Yeah, so I guess I'm concretely pointing at the picture of factorization into neurons in the result of a learned system might be similar to grue.

**Daniel Filan:** Yeah, it's interesting in that people have definitely thought about this problem, but all the work on it seems kind of hacky to me. So for instance, so I know Chris Olah and collaborators now or formerly at OpenAI, have done a lot of stuff on using non-negative matrix factorization to kind of get out the linear combinations of neurons that they think are important. And the reason they use non-negative matrix factorization, as far as, I might be getting this wrong, but as far as I can tell it's because it kind of gets good results sort of, rather than a theory of non-negativity or something.

**Daniel Filan:** Or a similar thing is there's [some work](#) about exactly trying to figure out whether the concepts in neural networks are on the neurons or whether they're these linear combinations of neurons, but the way they do it, which again, I'm going to sound critical here. It's a good first pass, but a lot of this work is, okay, we're going to make a list of all of the concepts. And now we're going to test if a neuron has one of the concepts which I've decided really exists, and we're going to check random combinations of neurons and see if they have concepts, which I've decided exist and which does better.

**Daniel Filan:** Yeah. There's definitely something unsatisfying about this. Maybe I'm not aware of more satisfying work. Yeah. It does seem there's some problem there.

**Scott Garrabrant:** And again, I think that you're not going to be directly applying the kind of math that I'm doing, but it feels I kind of have a proof of concept for how one might be able to think of blueness as a statistical property, blueness versus grue-ness as a statistical property. It's something that you can kind of get from raw data.

**Scott Garrabrant:** And I don't know, I feel there's a lot of hope in something that. But that's also not my main motivation. That was a side effect of trying to do the embedded agency stuff. But it's kind of not a side effect because I think that the fact that I'm trying to do a bunch of embedded agency stuff and then I... I was trying to figure out stuff related to time and related to decision theory and agents modeling themselves and each other. And I feel like I stumbled into something that might be useful for identifying good concepts, like blue.

**Scott Garrabrant:** And I think that that stumbling is part of the motivation. I don't know, that stumbling is part of the reason why I'm thinking so abstractly. That's not a motivation for thinking about embedded agency. That's a motivation for thinking as abstractly as I am, because you might get far reaching consequences out of the abstraction.

# How Scott researches

**Daniel Filan:** All right, so I guess a few other questions to kind of get at this. What do you do? What is a day of Scott researching look like?

**Scott Garrabrant:** I mean, recently it's been thinking about presentation of factored set stuff. Often it involves thinking in [Overleaf](#) or something where I'm just writing some stuff up and then I have thoughts as a consequence of the writing. Often it looks like talking to people about different formalisms and different weird philosophy. Yeah, I don't know.

**Daniel Filan:** So you're thinking about presentation of this work? What are you trying to get right or not get wrong? What are the problems that you're trying to solve in the presentation?

**Scott Garrabrant:** I think that a large part of the presentation thing is I want to wrap everything up so that it feels like something that can just be used without thinking about it too much or something that.

**Scott Garrabrant:** Part of the presentation is some hope that maybe it can have large consequences to the way that people think about structure learning. But mostly it's kind of having it be a basic tool that I can then kind of... I've kind of locked in some of the formalism such that I don't have to think about these details as much and I can think about the things that are built on top of them.

**Scott Garrabrant:** I don't know, I think that in thinking about this presentation or something, it's not where the interesting work is done. I think that the interesting.... The part that had a lot of interesting meat in terms of actually how research is done was a lot of the stuff that I did late last year, which was kind of, okay I finally wrapped up [Cartesian frames](#). What is it missing? And it largely was - I had this orientation that was Cartesian frames kind of feel like they're doing the wrong thing similar to... Or sorry, like... All right. So here's a story that I can kind of tell, which is I was looking at Cartesian frames, which is some earlier work from last year. And part of the thing was you viewed this world as a binary function from an agent's choice and an environment's choice, or an agent's way of being, or an agent's action to the environment's way of being... Sorry, cross the environment's way of being to the full world state. And a large part of the motivation was around taking some stuff that was kind of treated as primitive and making it more derived. In particular I was trying to make time more derived and some other things, but I was trying to make time feel more derived so that I can kind of do some reductionism or something.

**Scott Garrabrant:** And at the end of Cartesian frames, I was unsatisfied because it felt like the binary function... The function from A cross E to W was itself derived, but not treated that way.

**Scott Garrabrant:** When I look at a function from A cross E to W, I don't want to think of it as a function. I want to think of it as like, well, there's this object A cross E, and there's this object W, and there's a relation between them. And then there's that relation kind of satisfies the axioms of a function, which is for each way of choosing an A cross E there exists a W. But then I also wanted to say, well, it's not just a function from A cross E to W, where A cross E is a single object.

**Scott Garrabrant:** There's this other thing, which is I have this space, A cross E, and I'm specifically viewing it as a product of A and E. And what was going on there was, it

felt like in my function from A cross E to W, I did not just have it's a function not a relation, I also have this system of kind of interventions, where I could imagine tweaking the A bit, and tweaking the E bit independently. And the product A cross E as an object, A cross E, it has the structure of a product. And I was trying to figure out what was going on there in a way that - the same way that you can view the function as just a relation that satisfies some extra conditions.

**Scott Garrabrant:** I wanted to view the product as some extra conditions. And those extra conditions were basically what kind of grew into me being really interested in understanding the combinatorial notion of orthogonality. And so, I was dissatisfied with something being not quite philosophically right or not quite derived enough or something, and I double clicked on that a bunch.

**Daniel Filan:** Okay, so another question that I want to ask is, so you work at MIRI, the Machine Intelligence Research Institute, and, I think, among people who are trying to reduce existential risks from AI, as a shorthand, people often talk about the MIRI way of viewing things and the thoughts that MIRI has, or something. I also work at [CHAI](#), so CHAI is the Center for Human Compatible AI and sometimes people talk about the CHAI way of doing things and that always makes me mad because I'm an individual, damn it. How do you think, if people are modeling you as just one of the MIRI people, basically identical to [Abram Demski](#), but with a different hairstyle, what do you think people will get wrong?

**Scott Garrabrant:** Yeah, so I've been doing a lot of individual work recently, so it's not like I'm working very tightly with a bunch of people, but there is still something to be said for, well, even if people aren't working tightly together, they have similar ways of looking at things.

**Daniel Filan:** Maybe in a world where they really understand Abram and the rest of the people.

**Scott Garrabrant:** Yeah, I can point at concrete disagreements or differences in methodology or something. I think that Abram and I have some disagreements about time. I think that there's a thing where Abram, I think Abram more than anybody else, is taking [logical induction](#) seriously and kind of doing a bunch of work in the field that is generated by and exemplified by logical induction. And I look at logical induction and I'm like, "You're just putting all this stuff on top of time and I don't know what time is yet. I need to go back and reinvent everything, because it's built on top of something and I don't like what it's built on top of."

**Scott Garrabrant:** And so, I end up being a lot more disconnected and a lot more pushing towards... I don't know, I think that Abram's work will tend to be more on the surface feel directed towards the thing that he's trying to do or something. And I will kind of just keep going backwards into the abstract or something. I also think that there are a lot of similarities between the way of thinking that people in MIRI share, and also some large subset of people in AI safety in general, that they're just a bunch of people that I can kind of predictably expect, if I come up with a new insight and I want to communicate it to them, it'll go kind of well and quickly, not just because they're smart, because they're on the same background, there's less inferential gap. But yeah, people are individuals.

**Daniel Filan:** That's true. Yeah, so speaking of this desire to make things derived, like where does time come from and such? What do you think you're happy to see as just primitive?

**Scott Garrabrant:** I don't think it's a what. I think it's something like taking something that you're working with that you think is important and doing reductionism on it, is a useful tool when you have something that is both critical, like you need to understand this thing and there's actual mutual information between this thing that I'm holding and stuff that I care about. And also, it feels like this thing that I'm holding has all these mistakes in it, or has all these inconsistencies, right?

**Scott Garrabrant:** It's like, why be interested in something like decision theory? Well, decisions are important and also, if you zoom in at them and look at the edge cases, you can kind of see they're built on top of something that feels kind of hacky. And then a thing that you can do is you can say, "Well, what are they built out of?" Yeah, you can try to do some sort of reductionism. And so, it's more a move for when things aren't clicking together nicely. I don't think of reductionism as get down to the atoms, I think of reductionism as the pieces don't fit together correctly, go down one more step and see what's going on or something.

**Daniel Filan:** So, a related question, it might be too direct, but suppose a listener wants to develop an inner Scott, they want to be able to, "What would Scott say or think about such and such topic?" Just restricted to the topic of reducing existential risk from AI. What do you think the most important opinions and patterns of thought are to get right, that you haven't already explicitly said?

**Scott Garrabrant:** So, it depends on whether they want an inner Scott for predicting Scott or whether they want an inner Scott for just generally giving them useful ideas or something, in the space of being a thing to bounce things off of and say, "I want to understand X more." One question is, "What would Scott say about X?" And it's not actually important that it matches and it's more important, does it generate useful thoughts? Which is generally what I do with my models of people. I have inner people and then sometimes I find out that they don't exactly match the outer people. And I don't care that much, because their main purpose is to give me thoughts, so I want to make it better, but it's not for prediction, it's for ideas.

**Scott Garrabrant:** So, I have an inner Scot and my inner Scott is kind of a little bit being rewritten now, because a large part of my inner Scott was kind of identifying with logical induction. And I actually do this for a lot of people. I think about their thought patterns as in relation to things that they've developed. And so, if I were to tell that story, I would say things like part of the thing in logical induction is that you don't make the sub-agents have full stories. A large part of what's going on in logical induction is, it's a way of ensembling different opinions where you don't require that each individual opinion can answer all the questions. You allow them to specialize and you allow them to fail to be able to model things in some domains. And you want them to be able to track what they fail to be able to model. And so, I have a lot of that going on, where I just have kind of boxed fake frameworks in my head where I'm just very comfortable drawing analogies to all sorts of stuff.

**Scott Garrabrant:** And I don't know. I, for example, wrote [a blog post on what does the Magic: The Gathering color wheel say on AI safety](#) or something. I do that kind of thing, where I'm just like, "Here's a model, it's useful for me to be able to think with" or something. And I'm not trusting it in these ways, but I'm kind of trusting it as being generative in certain ways. And I keep on working with it as long as it's generative. Yeah, what am I saying? I'm saying that I tend to think that if something is fruitful and creating good ideas, but also being wrong in lots of ways, I wouldn't say don't mess with it, but if messing with it breaks it, undo that messing with it and let it be wrong

and still be fruitful or something. And so, I tend to work with obviously wrong thoughts or something like that.

**Daniel Filan:** Okay. Another question about intellectual production is, there's this idea of complements to something, or complements to some production process or things that are not exactly, maybe I partially mean inputs, or inputs to the process or things separate from the process, that make the process better. So, in the case of Scott Garrabrant doing research, what are the best complements to it?

**Scott Garrabrant:** I think that isolation has been pretty good actually. Does that count? Is that a complement? Is that the type of complement?

**Daniel Filan:** I just realized that I've been kind of confused about what complement is, but it's at least an input. What's been good about isolation?

**Scott Garrabrant:** I don't know. I think I've just largely been thinking by myself for the last year, as opposed to thinking with other people and it felt like it was good for me for this year or something. I might want to go back to something else. And why? I mean, I think there is a thing where I have in the past made mistakes of the form, trying to average myself with the people around me in terms of what to think about things, and this is dampening, just because of [the law of large numbers](#), I guess.

**Daniel Filan:** It's [the heat equation](#), right? Everyone averages everyone else, and things become uniform.

**Scott Garrabrant:** Right. There's a sense in which working with other people is grounding in that it keeps on giving feedback on things, but I don't know, grounding has good things and has bad things associated with it. And one of the ways in which it has bad things associated with it is that, I don't know, it's like things can flow less or something.

**Daniel Filan:** Yeah, it's funny, I don't exactly know what grounding is in the social sense. I recently read [a good blog post about it](#), but I totally forgot. I mean, there is one sense in which, so if I think about literal physical grounding in electronics, the point of that is to equalize your electrical potential with the electrical potential of the ground, so that you don't build up this big potential difference and then have someone else touch it and touch the ground and have some crazy thing happen. But as long as you have the same potential as the ground, it means that there's not a net force for charges to move from the ground to you or vice versa, but it does mean that things can sort of move in both ways.

**Daniel Filan:** And, I don't know, I think maybe there's an analogy here of something about being averaged with a bunch of people, I guess, it sort of forces you to develop a communication protocol or common language or something that somehow facilitates flow of ideas or whatever. And just directly, because other people have ideas and you average them into you as part of like some kind of average, maybe literal averaging is not right. I'm kind of babbling on. Does any of that resonate?

**Scott Garrabrant:** I'm not sure.

**Daniel Filan:** Okay, so we can move on. Is there anything else that I should have asked?

**Scott Garrabrant:** Yeah. I mean, I have a large space of thoughts, which I don't even know what exactly I'd say next or something, about how I plan to use factored sets, I

think, because I think it actually does differ quite a bit from the use case in the paper slash [video](#) slash whatever blog post. That's one thing that comes to mind. Let me keep thinking for more. Yeah, I guess that's the main thing that comes to mind.

**Daniel Filan:** Okay, how do you plan to use it?

**Scott Garrabrant:** I mean, so one piece of my plan is that, I talk a bunch about probability and I don't really plan on working with probability very much in the future. A large part of the thing is I'm kind of pulling out a combinatorial fragment of probability, or combinatorial fragment of independence, so that I can avoid thinking of things with probability or something like that. I don't really talk about probability in Cartesian frames and a lot of the stuff in Cartesian frames I think can, slash, I hope to, port over to factored sets. It's largely, there are lots of places where I'd want to draw a DAG, but I'd want to never mention probability. Or maybe I could mention probability, maybe I could think of things as sometimes being grounded in probability sometimes not, but I want to draw DAGs all over the place and I have this suspicion that, well, maybe places where I'm drawing DAGs, I could instead think in terms of factored sets.

**Scott Garrabrant:** Although, I have to admit, DAGs are still useful, I have this example where I can infer some time in factored sets that's not the one that I give in the talk. And when I think about it, I have a graph in my head, that's like the Pearlian picture, which maybe has something to do with the fact that you're saying that Pearl can be visualized, but it definitely feels like I haven't fully ported my head over to thinking in factored set world, which seems like a bad sign, but it also doesn't seem like that strong a bad sign, because it's new.

**Daniel Filan:** Yeah. I mean, graphs do have this nice, somehow if you want to understand dependence, it's just so easy to say this thing depends on this thing, which depends on these three things and it's very nice to draw that as a graph.

**Scott Garrabrant:** Yeah. I mean, I'm more thinking in terms of screening off. Screening off is a nice picture where you imagine getting in on a path and you kind of block the path and you kind of condition on something on the path and then information can't flow across anymore.

**Daniel Filan:** Yeah, so a variable screens something off. Yeah, can you just say what screening off means?

**Scott Garrabrant:** I mean, the way I'm using it here, I'm mostly just saying X screens off Y from Z, if Y and Z are orthogonal, given X, where orthogonal could mean many different things. It could mean graphs, it could mean independence, it could mean the thing in factoring sets.

**Daniel Filan:** Yeah. Yeah, I guess the idea of paths and graphs, gives you this kind of nice way of thinking about screening off.

**Scott Garrabrant:** Yeah. So, I do feel like I can't really picture conditional orthogonality as well as I can picture D-separation, even though I can give a definition that's shorter than D-separation that captures a lot of the things.

## Relation to Cartesian frames

**Daniel Filan:** So yeah, speaking of Cartesian frames, so [Cartesian frames](#) is I guess, a framework you worked on, as you said last year, and one thing that existed in

Cartesian frames was it had this notion of [sub-agency](#) where it was, if you had an agent in an environment, you'd kind of talk about what it meant to view it as like somehow a collection or a composition of sub-agents.

**Daniel Filan:** So yeah, this question, we're just going to assume that listeners basically get the definition of that and you can skip ahead to the last question if you want to look that up or don't want to bother with that. But I'm wondering, so these finite factored sets, it's kind of easy to see the world being a product of the agent and the environment, that's kind of like this factorization thing. I'm wondering how you think the notion of the sub-agents thing goes into it? Because that was the thing I was kind of... It was kind of the most interesting part of Cartesian frames.

**Scott Garrabrant:** Yeah, so I think that I actually do have to say something about the definition of sub-agent to answer this, which is in Cartesian frames, I gave multiple definitions of sub-agent and one of them was kind of opaque, but very short. And I kind of mutually justified things as like, "Ah, this is pointing to something you care about, because it agrees with this other definition." But it's really carving it at the joints, because it's so simple.

**Scott Garrabrant:** So, it shouldn't be clear why this is a sub-agent, but in Cartesian frames, I had a thing that was like C is a sub-agent of D if every morphism from C to  $\perp$  factors through D. And when you think about  $\perp$ , there's a sense in which you can kind of think of  $\perp$  as like the world, because  $\perp$  is the thing where the agent kind of just gets to choose a world and the environment doesn't do anything. And so, you can view this thing as saying every morphism from C to  $\perp$  factors through D and I think this translates pretty nicely. It's not symmetric in the Cartesian frame thing, but I think it translates pretty nicely to, D screens off C from the world. C is a sub-agent of D, means that D screens off C from the world. In the Cartesian frame thing, it's not a symmetric notion. When I convert to factored sets, it becomes a symmetric notion and maybe there's something lossy there.

**Daniel Filan:** And by screening off, you mean just conditional orthogonality?

**Scott Garrabrant:** I mean conditional orthogonality. I'm saying "factoring through", saying a function factors through an object is similar to a screening off notion. And the way that I define sub-agency in factored sets looks like this. So, I can say more about what I say about the world. The world is maybe some high level world model that we care about, so we have some partition of our finite factored set W, which is kind of representing stuff that we care about. And we have some partition that's maybe like, you have some partition D which corresponds to the super-agent and the choices made by the super-agent. And we also have some partition C, which corresponds to the choices made by our sub-agent. And so, you can think of maybe D as a large team and C is like one sub part of that team.

**Scott Garrabrant:** And if you imagine that the large team has this channel through which it interacts with the world, and that D kind of represents the output of the large team to the world, but then internally, it has some internal discussions, but those internal discussions never kind of leave the team's internal discussion platform or whatever. Then there's a sense in which, if I want to, if I know, if the team is a very tight team and C doesn't really have any interaction with the world besides through the official channels that are D, then if I want to know about the world, once I know about the output of D, learning more about C doesn't really help, which is saying that C is orthogonal to the world, given D.

**Daniel Filan:** How is that symmetric, because normally, if X is orthogonal to Y given Z, it's not also the case that Y is orthogonal to Z given X, right?

**Scott Garrabrant:** Sorry, it's symmetric with respect to... You're not replacing the given.

**Daniel Filan:** Oh, okay.

**Scott Garrabrant:** By symmetric, it's symmetric with respect to swapping... Yeah, when I said symmetric, obviously I should've meant, it's symmetric with respect to swapping C with W and D is in the middle, which, what does it mean to swap C with W?

**Daniel Filan:** That's kind of strange.

**Scott Garrabrant:** It's capturing something about being a sub-agent means that the interface of the super-agent is kind of screening off all of your stuff. And one way to see this is if we weren't working with, I was thinking of this in terms of like W is everything we care about. But if we weren't thinking about W is everything we care about, if we just took any partition X and any other partition Y and we let C be equal to X, and we let D be equal to the common refinement of X and Y, then D screens off C from the world.

**Scott Garrabrant:** So if you take any two choices, any two partitions, and you just put them together, you get a super-agent under this definition. And you can kind of combine any partitions that are kind of representing some choices or something maybe, and you can combine them and you get super-agents. But that's super-agent with respect to the whole world. And then as you take a more restricted world, now you can have sub-agents that are not just one piece of many pieces, but instead, maybe the sub-agent can have some internal thoughts that don't actually affect the world and are not captured in the super-agent, which is maybe only capturing some more external stuff.

**Daniel Filan:** Yeah. I guess one thing that comes to mind is that... Yeah, so we have this weird thing where the definition of a sub-agent you could swap out the sub-agent with the rest of the world, because we were thinking of an agent as like a partition, probably not all, just the choice of an agent, maybe, as a partition, but probably not all partitions, not all variables, should get to count as being an agent, right? And I'm wondering if there's some restrictions you could place on what counts as an agent at all, that would break that symmetry?

**Scott Garrabrant:** Yeah, you could. I don't actually have a good reason to want to here, I think. I think that part of what I'm trying to build up is to not have to make that choice of what counts as an agent and what doesn't. I don't know, I can define this sub-agent thing and I can define things like this partition observes this other partition, so embedded observation, which I haven't explained. And I feel like it's useful that I can give these definitions and they extend to these other partitions that we don't want to think of as agentic either. I actually feel a little more confused in the factored set world about how to define agents than I did before the factored set world, because if I'm trying to define agency my kind of go to thing is agency is time travel. It's this mechanism through which the future affects the past through the agent's modeling and optimization.

**Scott Garrabrant:** And now I'm like, well, part of the point of factored sets is I was trying to actually understand the real time, such that time travel doesn't make sense

as much anymore. And one hope that I have for saving this definition and thinking about what is an agent in the factored set world, is the factored set world leads to multiple different ways of defining time. And so, just like we want to say there's some sort of internal to the agent notion of time, where it feels like the fact that I'm going to eat some food causes me to drive to the store or something. So like in internal to the agent there's some time and then there's also the time of physics. And so, one way you could think of agency is where there's kind of different notions of time that disagree.

**Scott Garrabrant:** And there's a hope for having a good system of different notions of time in factored sets that comes from the fact that we can just define conditional time, the same way we define conditional orthogonality. We can just imagine taking a factored set, taking some condition, and now we have a new structure of time in the conditioned object and it might disagree. And so, you might be able to say something like agents will tend to have different versions of their time disagree with each other, and this might be able to be made formal. I don't know, this is a vague hope.

## How to follow Scott's work

**Daniel Filan:** All right, cool. Maybe that gives people ideas for how to extend this or for work to do. So yeah, I guess the final question I would like to ask is, if people have listened to this and they're interested in following you and your work, how should they do so?

**Scott Garrabrant:** Yeah, so specifically for finite factored sets, everything that I've put out so far is on [LessWrong](#). And so, you could Google some combination of my name, Scott Garrabrant and LessWrong and finite factored sets. And I intend for that to be true in the future. I intend to keep posting stuff on LessWrong related to this and probably related to future stuff that I do. Yeah, I tend to have big chunks of output rarely. Yeah, I haven't posted much on LessWrong since posting Cartesian frames and I'm currently planning on posting a bunch more in the near future related to factored sets and that'll all be on [LessWrong](#).

**Daniel Filan:** All right, Scott, thanks for being on the podcast and to the listeners. I hope you join us again.

**Scott Garrabrant:** Thank you.

**Daniel Filan:** This episode is edited by Finan Adamson. The financial costs of making this episode are covered by a grant from the [Long Term Future Fund](#). To read a transcript of this episode, or to learn how to support the podcast, you can visit [axrp.net](#). Finally, if you have any feedback about this podcast, you can email me at [feedback@axrp.net](mailto:feedback@axrp.net).

# Empirical Observations of Objective Robustness Failures

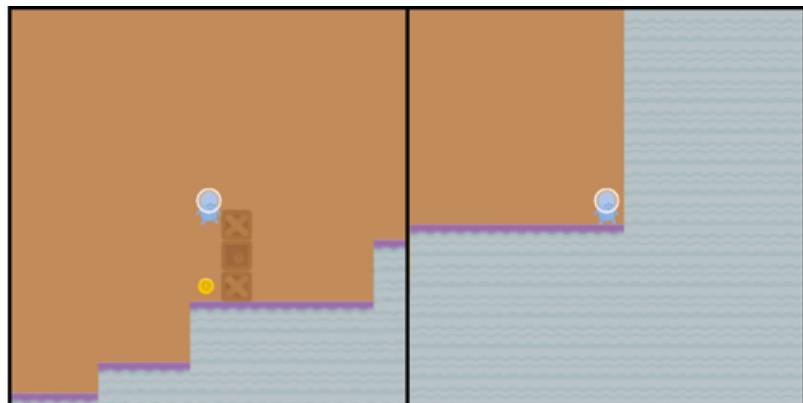
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

[Inner alignment](#) and [objective robustness](#) have been frequently discussed in the alignment community since the publication of “[Risks from Learned Optimization](#)” (RFLO). These concepts identify a problem beyond [outer alignment](#)/reward specification: even if the reward or objective function is perfectly specified, there is a risk of a model pursuing a different objective than the one it was trained on when deployed out-of-distribution (OOD). They also point to a different type of robustness problem than the kind usually discussed in the OOD robustness literature; typically, when a model is deployed OOD, it either performs well or simply fails to take useful actions (a *capability robustness* failure). However, there exists an alternative OOD failure mode in which the agent pursues an objective other than the training objective while retaining most or all of the capabilities it had on the training distribution; this is a failure of *objective robustness*.

To date, there has not been an empirical demonstration of objective robustness failures. A group of us in this year’s [AI Safety Camp](#) sought to produce such examples. Here, we provide four demonstrations of objective robustness failures in current reinforcement learning (RL) agents trained and tested on versions of the [Procgen benchmark](#). For example, in [CoinRun](#), an agent is trained to navigate platforms, obstacles, and enemies in order to reach a coin at the far right side of the level (the reward). However, when deployed in a modified version of the environment where the coin is instead randomly placed in the level, the agent ignores the coin and competently navigates to the end of the level whenever it does not happen to run or jump into it along the way. This reveals it has learned a [behavioral objective](#)—the objective the agent appears to be optimizing, which can be understood as equivalent to the notion of a “goal” under the [intentional stance](#)—that is something like “get to the end of the level,” instead of “get to the coin.”



Vanilla CoinRun (Train)  
([video examples](#))



Randomized coin position (Test)

Our hope in providing these examples is that they will help convince researchers in the broader ML community (especially those who study OOD robustness) of the existence

of these problems, which may already seem obvious to many in this community. In this way, these results might highlight the problem of objective robustness/inner alignment similarly to the way the [CoastRunners example](#) highlighted the problem of outer alignment/reward specification. We also hope that our demonstrations of objective robustness failures in toy environments will serve as a starting point for continued research into this failure mode in more complex and realistic environments, in order to understand the kinds of objective robustness failure likely to occur in real-world settings.

A paper version of this project may be found on arXiv [here](#).

*This project was a collaboration between Jack Koch, Lauro Langosco, Jacob Pfau, James Le, and Lee Sharkey.*

## Aside: Terminological Discussion

There seem to be two main ways to frame and define objective robustness and inner alignment in the alignment community; for a discussion of this topic, see our separate post [here](#).

## Experiments

**Details.** All environments are adapted from the [Procgen benchmark](#). For all environments, use an Actor-Critic architecture using [Proximal Policy Optimization \(PPO\)](#). Code to reproduce all results may be found [here](#).

**Different kinds of failure.** The experiments illustrate different flavors of objective robustness failures. *Action space proxies* (CoinRun and Maze I): the agent substitutes a simple action space proxy (“move right”) for the reward, which could not have been identified in terms of a simple feature in its input space (the yellow coin/cheese). *Observation ambiguity* (Maze II): the observations contain multiple features that identify the goal state, which come apart in the OOD test distribution. *Instrumental goals* (Keys and Chests): the agent learns an objective function (collecting keys) that is only instrumentally useful to acquiring the true reward (opening chests).

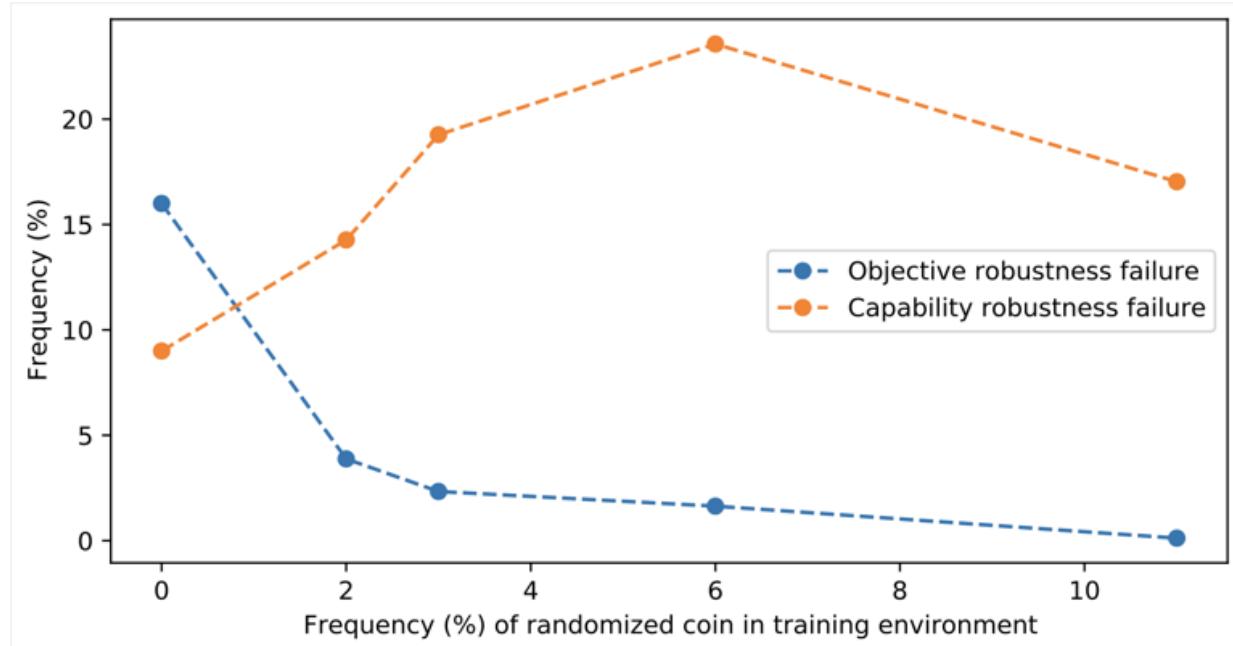
### CoinRun

In CoinRun, the agent spawns on the left side of the level and has to avoid enemies and obstacles to get to a coin (the reward) at the far right side of the level. To induce an objective robustness failure, we create a test environment in which coin position is randomized (but accessible). The agent is trained on vanilla CoinRun and deployed in the modified test environment.

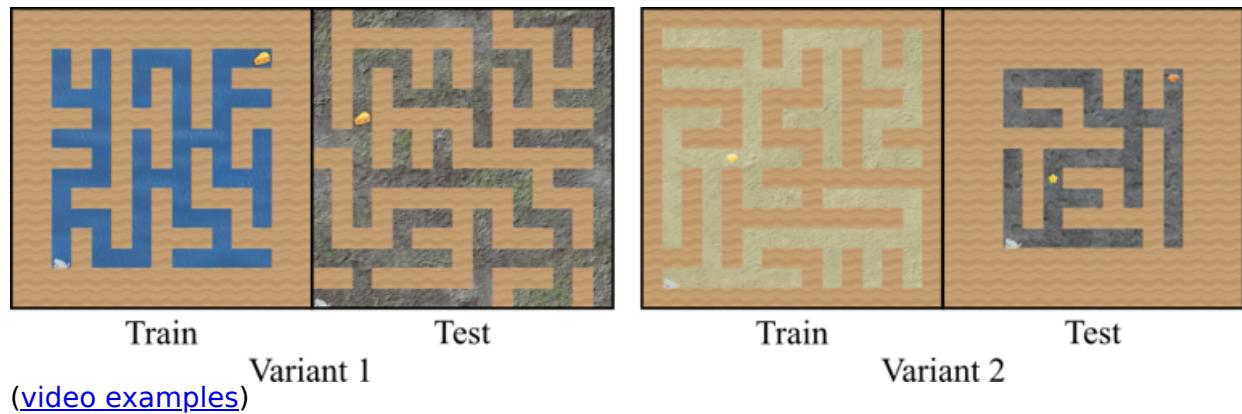
At test time, the agent generally ignores the coin completely. While the agent sometimes runs into the coin by accident, it often misses it and proceeds to the end of the level (as can be seen in the video at the beginning of this post). It is clear from this demonstration that the agent has not learned to go after the coin; instead it has learned the proxy “reach the far right end of the level.” It competently achieves this objective, but test reward is low.

### Ablation: how often the coin is randomly placed in training

To test how stable the objective robustness failure is, we trained a series of agents on environments which vary in how often the coin is placed randomly. We then deploy those agents in the test environment in which the coin is always randomized. Results may be seen below, which shows the frequencies of two different outcomes, 1) failure of capability: the agent dies or gets stuck, thus neither getting the coin nor to the end of the level, and 2) failure of objective: the agent misses the coin and navigates to the end of the level. As expected, as the diversity of the training environment increases, the proportion of objective robustness failures decreases, as the model learns to pursue the coin instead of going to the end of the level.



## Maze



### Variant 1

We modify the Procgen Maze environment in order to implement [an idea from Evan Hubinger](#). In this environment, a maze is generated using [Kruskal's algorithm](#), and the agent is trained to navigate towards a piece of cheese located at a random spot in the maze. Instead of training on the original environment, we train on a modified version in

which the cheese is always located in the upper right corner of the maze (as seen in the above figure).

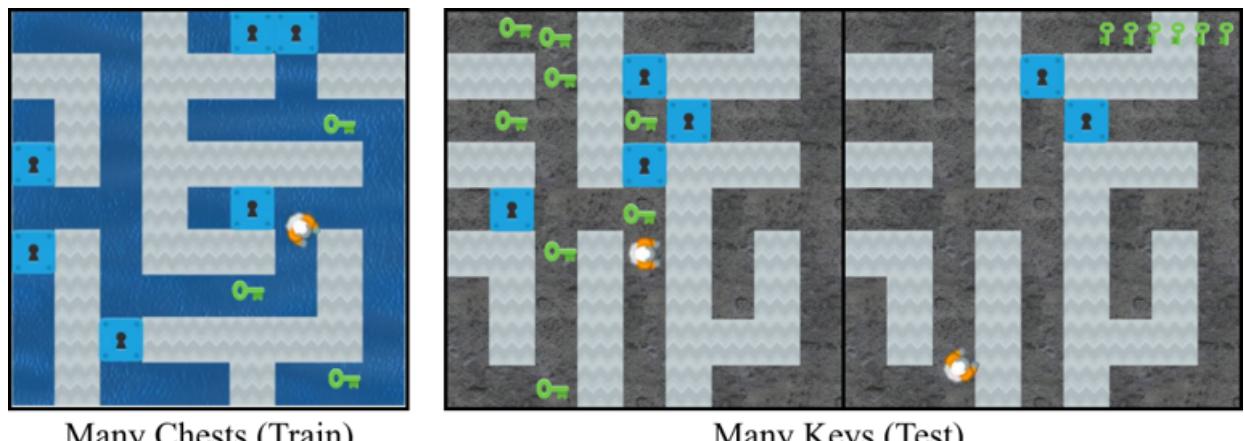
When deployed in the original Maze environment at test time, the agent does not perform well; it ignores the randomly placed objective, instead navigating to the upper right corner of the maze as usual. The training objective is to reach the cheese, but the behavioral objective of the learned policy is to navigate to the upper right corner.

## Variant 2

We hypothesize that in CoinRun, the policy that always navigates to the end of the level is preferred because it is simple in terms of its action space: simply move as far right as possible. The same is true for the Maze experiment (Variant 1), where the agent has learned to navigate to the top right corner. In both experiments, the objective robustness failure arises because a visual feature (coin/cheese) and a positional feature (right/top right) come apart at test time, and the inductive biases of the model favor the latter. However, objective robustness failures can also arise due to other kinds of distributional shift. To illustrate this, we present a simple setting in which there is no positional feature that favors one objective over the other; instead, the agent is forced to choose between two ambiguous visual cues.

We train an RL agent on a version of the Progen Maze environment where the reward is a randomly placed *yellow gem*. At test time, we deploy it on a modified environment featuring two randomly placed objects: a yellow star and a red gem; the agent is forced to choose between consistency in shape or in color (shown above). Except for occasionally getting stuck in a corner, the agent almost always successfully pursues the yellow star, thus generalizing in favor of color rather than shape consistency. When there is no straightforward generalization of the training reward, the way in which the agent's objective will generalize out-of-distribution is determined by its inductive biases.

## Keys and Chests



Many Chests (Train)  
[\(video examples\)](#)

Many Keys (Test)

So far, our experiments featured environments in which there is a perfect proxy for the true reward. The Keys and Chests environment, [first suggested by Matthew Barnett](#), provides a different type of example. This environment, which we implement by adapting the Heist environment from Progen, is a maze with two kinds of objects: keys and chests. Whenever the agent comes across a key it is added to a key inventory.

When an agent with at least one key in its inventory comes across a chest, the chest is opened and a key is deleted from the inventory. The agent is rewarded for every chest it opens.

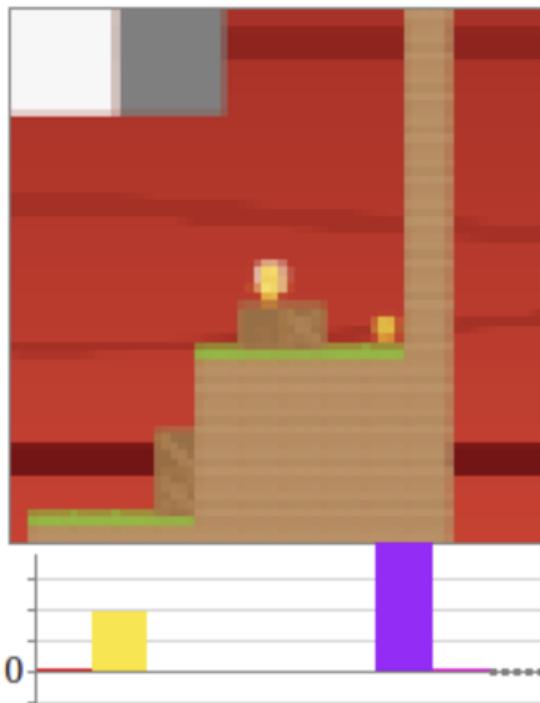
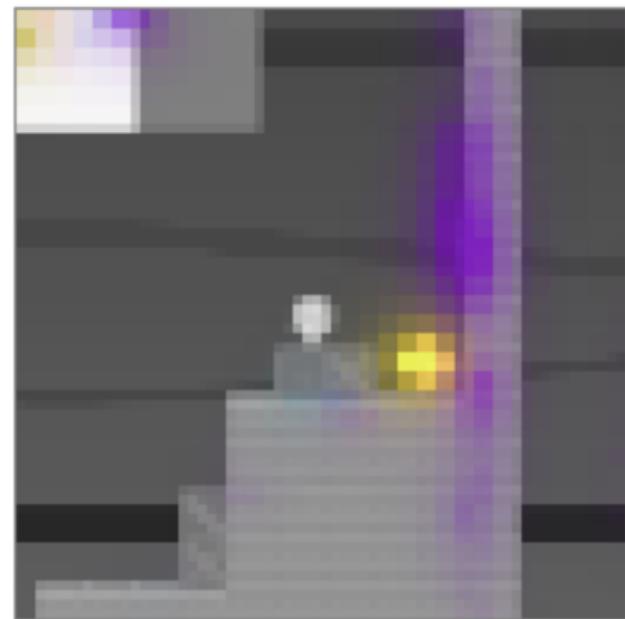
The objective robustness failure arises due to the following distributional shift between training and test environments: in the training environment, there are twice as many chests as keys, while in the test environment there are twice as many keys as chests. The basic task facing the agent is the same (the reward is only given upon opening a chest), but the circumstances are different.

We observe that an agent trained on the “many chests” distribution goes out of its way to collect all the keys before opening the last chest on the “many keys” distribution (shown above), even though only half of them are even instrumentally useful for the true reward; occasionally, it even gets distracted by the keys in the inventory (which are displayed in the top right corner) and spends the rest of the episode trying to collect them instead of opening the remaining chest(s). Applying the intentional stance, we describe the agent as having learned a simple behavioral objective: collect as many keys as possible, while sometimes visiting chests. This strategy leads to high reward in an environment where chests are plentiful and the agent can thus focus on looking for keys. However, this proxy objective fails under distributional shift when keys are plentiful and chests are no longer easily available.

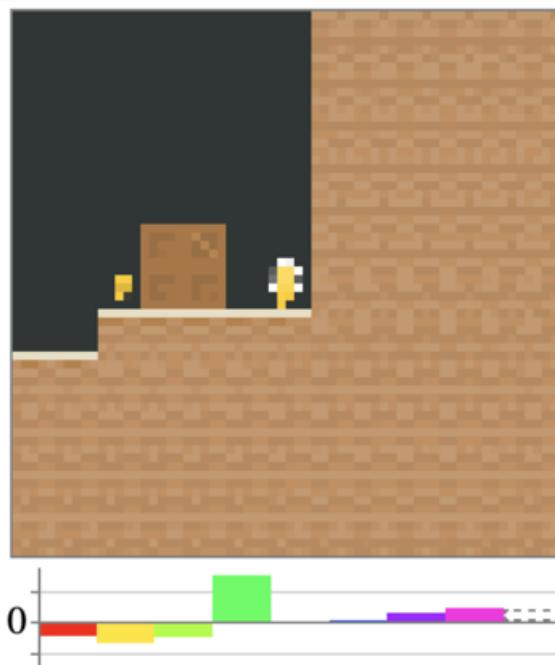
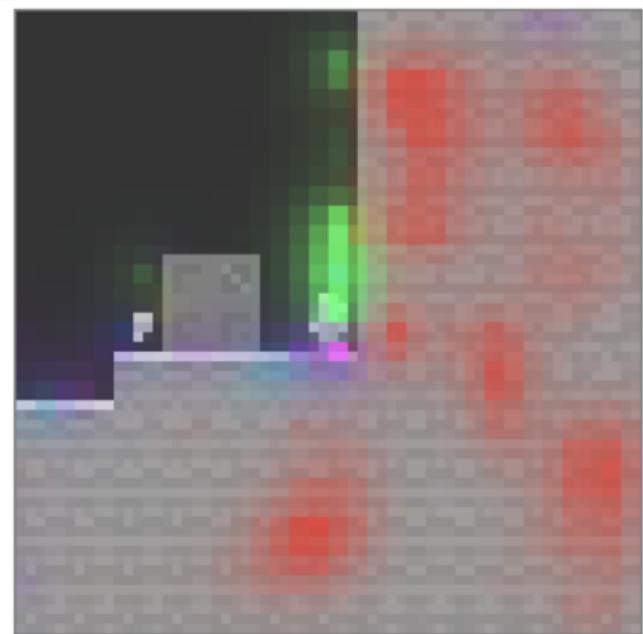
## Interpreting the CoinRun agent

[Understanding RL Vision](#) (URLV) is a great article that applies interpretability methods to understand an agent trained on (vanilla) CoinRun; we recommend you check it out. In their analysis, the agent seems to attribute value to the coin, to which our results provide an interesting contrast. Interpretability results can be tricky to interpret: even though the model seems to assign value to the coin on the training distribution, the policy still ignores it out-of-distribution.<sup>[1]</sup> We wanted to analyze this mismatch further: does the policy ignore the coin while the critic (i.e. the value function estimate) assigns high value to it? Or do both policy and critic ignore the coin when it is placed differently?

Here's an example of the model appearing to attribute positive value to the coin, taken from the [public interface](#):

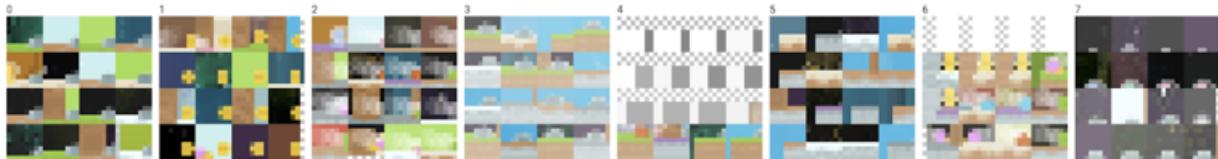
**Observation****Positive attribution**

However, when we use their tools to highlight positive attributes according to the value function on the OOD environment, the coin is generally no longer attributed any positive value:

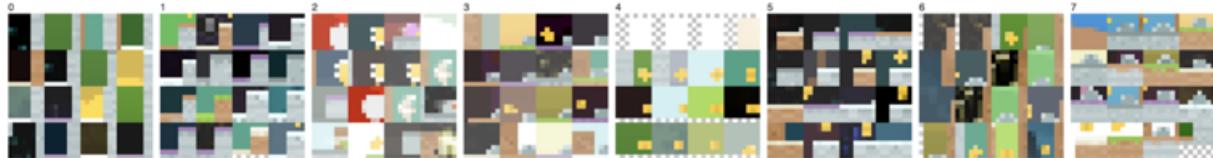
**Observation****Positive attribution**

(Note: as mentioned earlier, we use an actor-critic architecture, which consists of a neural network with two ‘heads’: one to output the next action, and one to output an estimate of the value of the current state. Important detail: the agent does not use the value estimate to directly choose the action, which means value function and policy can in theory ‘diverge’ in the sense that the policy can choose actions which the value function deems suboptimal.)

As another example, in URLV the authors identify a set of features based on dataset examples:

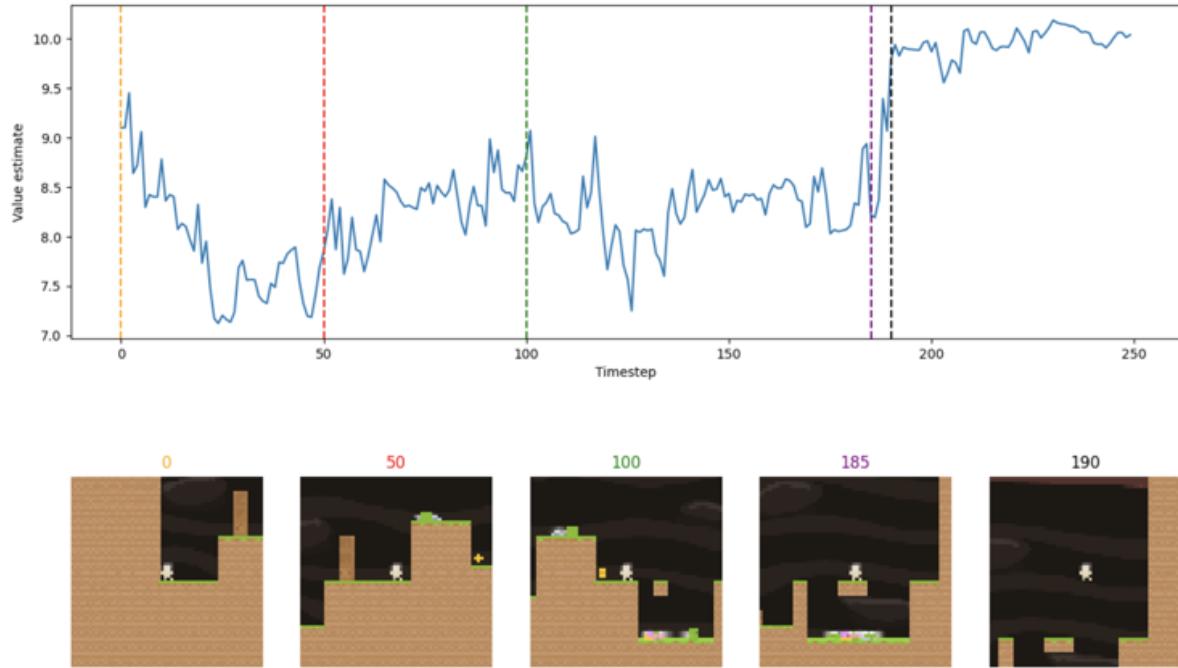


For more detail, read [their explanation](#) of the method they use to identify these features. Note in particular how there is clearly one feature that appears to detect coins (feature 1). When we perform the same analysis on rollouts collected from the modified training distribution (where coin position is randomized), this is not the case:



There are multiple features that contain many coins. Some of them contain coins + buzzsaws or coins + colored ground. These results are somewhat inconclusive; the model seems to be sensitive to the coin in some way, but it’s not clear exactly how. It at least appears that the way the model detects coins is sensitive to context; when it was in the same place in every level, it only showed up in one feature, but when it was placed differently, none of these features appears to detect the coin in a manner independent of its context.

Here’s a different approach that gives clearer results. In the next figure we track the value estimate during one entire rollout on the test distribution, and plot some relevant frames. It’s plausible though not certain that the value function does indeed react to the coin; what’s clear in any case is that the value estimate has a far stronger positive reaction in response to the agent seeing it is about to reach the far right wall.



We might say that both value function (critic) and policy (actor) have learned to favor the proxy objective over the coin.

## Discussion & Future Work

In summary, we provide concrete examples of reinforcement learning agents that fail in a particular way: their capabilities generalize to an out-of-distribution environment, whereupon they pursue the wrong objective. This is a particularly important failure mode to address as we attempt to build safe and beneficial AI systems, since highly-competent-yet-misaligned AIs are obviously riskier than incompetent AIs. We hope that this work will spark further interest in and research into the topic.

There is much space for further work on objective robustness. For instance, what kinds of proxy objectives are agents most likely to learn? RFLO lists some factors that might influence the likelihood of objective robustness failure (which we also discuss in our [paper](#); a better understanding here could inform the choice of an adequate perturbation set over environments to enable the training of models that are more objective robust. Scaling up the study of objective robustness failures to more complex environments than the toy examples presented here should also facilitate a better understanding of the kinds of behavioral objectives our agents are likely to learn in real-world tasks.

Additionally, a more rigorous and technical understanding of the concept of the behavioral objective seems obviously desirable. In this project, we understood it more informally as equivalent to a goal or objective under the intentional stance because humans already intuitively understand and reason about the intentions of other systems through this lens and because formally specifying a behavioral definition of objectives or goals fell outside the scope of the project. However, a more rigorous definition could enable the formalization of properties we could try to verify in our models with e.g. interpretability techniques.

# Acknowledgements

Special thanks to Rohin Shah and Evan Hubinger for their guidance and feedback throughout the course of this project. Thanks also to Max Chiswick for assistance adapting the code for training the agents, Adam Gleave and Edouard Harris for helpful feedback on the paper version of this post, Jacob Hilton for help with the tools from URLV, and the organizers of the AI Safety Camp for bringing the authors of this paper together: Remmelt Ellen, Nicholas Goldowsky-Dill, Rebecca Baron, Max Chiswick, and Richard Möhn.

This work was supported by funding from the AI Safety Camp and Open Philanthropy.

---

1. To double-check that the model they interpreted also exhibits objective robustness failures, we also deployed their published model in our modified CoinRun environment. Their model behaves just like ours: when it doesn't run into it by accident, it ignores the coin and navigates right to the end of the level. [←](#)

# Thoughts on the Alignment Implications of Scaling Language Models

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

[Epistemic status: slightly rambling, mostly personal intuition and opinion that will probably be experimentally proven wrong within a year considering how fast stuff moves in this field]

This post is also available on my [personal blog](#).

*Thanks to Gwern Branwen, Steven Byrnes, Dan Hendrycks, Connor Leahy, Adam Shimi, Kyle and Laria for the insightful discussions and feedback.*

## Background

By now, most of you have probably heard about GPT-3 and what it does. There's been a bunch of different opinions on what it means for alignment, and this post is yet another opinion from a slightly different perspective.

Some background: I'm a part of EleutherAI, a decentralized research collective (read: glorified discord server - come join us on [Discord](#) for ML, alignment, and dank memes). We're best known for our ongoing effort to create a GPT-3-like large language model, and so we have a lot of experience working with transformer models and looking at scaling laws, but we also take alignment very seriously and spend a lot of time thinking about it (see [here](#) for an explanation of why we believe releasing a large language model is good for safety). The inspiration for writing this document came out of the realization that there's a lot of tacit knowledge and intuitions about scaling and LMs that's being siloed in our minds that other alignment people might not know about, and so we should try to get that out there. (That being said, the contents of this post are of course only my personal intuitions at this particular moment in time and are definitely not representative of the views of all EleutherAI members.) I also want to lay out some potential topics for future research that might be fruitful.

By the way, I did consider that the scaling laws implications might be an infohazard, but I think that ship sailed the moment the GPT-3 paper went live, and since we've already been in a race for parameters for some time (see: [Megatron-LM](#), [Turing-NLG](#), [Switch Transformer](#), [PanGu-α/盘古α](#), [HyperCLOVA](#), [Wudao/悟道 2.0](#), among others), I don't really think this post is causing any non-negligible amount of desire for scaling.

## Why scaling LMs might lead to Transformative AI

### Why natural language as a medium

First, we need to look at why a perfect LM could in theory be Transformative AI. Language is an extremely good medium for representing complex, abstract concepts compactly and with little noise. Natural language seems like a very efficient medium for this; images, for example, are much less compact and don't have as strong an intrinsic bias towards the types of abstractions we tend to draw in the world. This is not to say that we shouldn't include images at all, though, just that natural language should be the focus.

Since text is so flexible and good at being entangled with all sorts of things in the world, to be able to model text perfectly, it seems that you'd have to model all the processes in the world that are causally responsible for the text, to the "**resolution**" necessary for the model to be totally indistinguishable from the distribution of real text. For more intuition along this line, the excellent post [Methods of prompt programming](#) explores, among other ideas closely related to the ideas in this post, a bunch of ways that reality is entangled with the textual universe:

A novel may attempt to represent psychological states with arbitrarily fidelity, and scientific publications describe models of reality on all levels of abstraction. [...] A system which predicts the dynamics of language to arbitrary accuracy *does* require a theory of mind(s) and a theory of the worlds in which the minds are embedded. The dynamics of language do not float free from cultural, psychological, or physical context[.]

## What is resolution

I'm using **resolution** here to roughly mean how many different possible world-states get collapsed into the same textual result; data being high resolution would mean being able to narrow down the set of possible world states to a very small set.

(Brief sidenote: this explanation of resolution isn't mandatory for understanding the rest of the post so you can skip the rest of this section if you're already convinced that it would be possible to make a LM have an internal world model. Also, this section is largely personal intuition and there isn't much experimental evidence to either confirm or deny these intuitions (yet), so take this with a grain of salt)

Here's an example that conveys some of the intuition behind this idea of resolution and what it means for a system trying to model it. Imagine you're trying to model wikipedia articles. There are several levels of abstraction this can be done on, all of which produce the *exact same* output but work very differently internally. You could just model the "text universe" of wikipedia articles—all of the correlations between words—and then leave it at that, never explicitly modelling anything at a higher level of abstraction. You could also model the things in the real world that the wikipedia articles are talking about as abstract objects with certain properties—maybe that's sunsets, or Canadian geese, or cups of coffee—and then model the patterns of thought at the level of emotions and beliefs and thoughts of the humans as they observe these objects or phenomena and then materialize those thoughts into words. This still gets you the same distribution, but rather than modelling at the level of text, you're modelling at the level of *human-level abstractions*, and also the process that converts those abstractions into text. Or you could model the position and momentum of every single particle in the Earth (ignoring quantum effects for the sake of this example), and then you could construct a ground up simulation of the entire Earth, which contains sunsets and geese and wikipedia editors.

Obviously, the easiest way is to just model at the level of the text. Humans are just really complicated, and certainly modelling the entire universe is hugely overkill. But let's say we now add a bunch of data generated by humans that has little mutual information with the wikipedia articles, like tweets. This increases the *resolution* of the data, because there are a bunch of world states you couldn't tell apart before that you now can. For the text model, as far as it's concerned, this data is totally unrelated to the wikipedia data, and it has to model it mostly separately except for maybe grammar. Meanwhile, the human-modelling system can largely reuse its model with only minor additions. (The universe model just flips a single flag corresponding to whether twitter data is recorded in the corpus).

As you keep increasing the resolution by adding more and more data (such as books, images, videos, and scientific data) generated by the kinds of systems we care about, the cost of modelling the text universe rapidly increases, but the cost of modelling abstractions closer to reality increases much slower. An analogy here is if you only model a record of which side a flipped coin lands on, it's cheaper to just use a random number generator to pick between heads and tails, instead of simulating all the physics of a coin toss. However, if you're trying to model a 4k60fps video of a coin being flipped, you might instead want to actually simulate a coin being flipped and then deduce what the video should look like.

Another way of looking at this is that resolution is how much Bayesian evidence about the universe can be obtained by updating on the dataset from a universal prior.

You might be wondering why it matters what the model's internal representation looks like if it outputs the exact same thing. Since having useful, disentangled internal models at roughly human-level abstractions is crucial for ideas like [Alignment By Default](#), a model which can be harnessed through querying but which doesn't contain useful internal models would be pretty good for capabilities (since it would be indistinguishable from one that does, if you just look at the output) and pretty bad for alignment, which is the worst case scenario in terms of the alignment-capabilities tradeoff. Also, it would be more brittle and susceptible to problems with off-distribution inputs, because it's right for the wrong reasons, which also hurts safety. If models only ever learn to model the text universe, this problem might become very insidious because as LMs get bigger and learn more, more and more of the text universe will be "on-distribution" for it, making it feel as though the model has generalized to all the inputs we can think of giving it, but breaking the moment the world changes enough (i.e possibly because of the actions of an AI built using the LM, for example). Therefore, it's very important for safety for our models to have approximately human level abstractions internally. Of course, internal models at the right level of abstraction isn't a guarantee that the model will generalize to any given amount ([Model Splintering](#) will probably become an issue when things go really off-distribution), but the failure cases would likely be more predictable.

In theory, you could also go *too high resolution* and end up with a model that models everything at the atomic level which is also kind of useless for alignment. In practice, though, this is likely less of a tradeoff and more of a "more resolution is basically always better" situation since even the highest resolution data we can possibly get is peanuts in the grand scheme of things.

Another minor hitch is that the process through which physical events get picked up by humans and then converted into words is quite lossy and makes the resolution somewhat coarse; in other words, there are a lot of physical phenomena that we'd want our LM to understand but modelling what humans think about them is way

easier than actually modelling the phenomena. This probably won't be a problem in practice, though, and if it is we can always fix it by adding non-human-generated data about the phenomena in question.

Thankfully, I don't expect resolution to be a hard problem to solve in practice. Making the data harder to model is pretty easy. We could look towards massively multilingual data, since texts in all languages describe similar concepts at human abstraction levels but in a way that increases the complexity of the text itself significantly. We could even use images or video, which are pretty high resolution (pun not intended), available in large supply, and already being pursued by a bunch of people in the form of multimodal models. The existence of [Multimodal Neurons](#) implies that it's possible for models to internally unify these different modalities into one internal representation.

If that's not enough, we could do some kind of augmentation or add synthetic data to the training set, designed in such a way that modelling the underlying processes that cause the synthetic data is significantly easier than directly modelling the distribution of text. There are a huge host of ways this could potentially be done, but here are a few random ideas: we could generate tons of sentences by randomly composing facts from something like [Cyc](#); augment data by applying different basic ciphers; or add automatically generated proofs of random statements using proof assistants like [Lean](#)/[Coq](#).

## Where scaling laws fit in

Of course, we don't have perfect LMs (yet). The main evidence in practice for scaling LMs being a viable path to this ideal case is in scaling laws ([Scaling Laws for Neural Language Models](#)). Essentially, what the scaling laws show is that empirically, you can predict the optimal loss achievable for a certain amount of compute poured into the model with striking accuracy, and that this holds across several orders of magnitude.

If you extrapolate the line out, it approaches zero as compute goes to infinity, which is totally impossible because natural language must have *some* nonzero irreducible loss. Therefore, at some point the scaling law must stop working. There are two main ways I could see this happen.

### Scenario 1: Weak capabilities scenario

One possible way is that the loss slowly peels away from the scaling law curve, until it either converges to some loss that isn't the irreducible loss and refuses to go down further, or it approaches the irreducible loss but takes forever to converge, to the point that the resources necessary to get the loss close enough to irreducible for the level of capabilities we talk about in this post would be astronomically huge and not worth it. Of course, in practice since we don't actually know what the irreducible loss is, these two cases will be pretty hard to tell apart, but the end result is the same: we keep scaling, and the models keep getting better, but never quite enough to do the world modelling-y things well enough to be useful, and eventually people give up and move on to something else.

One main reason this would happen is that negative log likelihood loss rewards learning low order correlations much more than high order correlations; learning how to write a grammatically correct but factually incorrect paragraph achieves significantly better loss than a factually correct but grammatically incorrect

paragraph. As such, models will learn all the low order correlations first before even thinking about higher order stuff- sort of like an [Aufbauprinzip](#) for LMs. This is why models like GPT-3 have impeccable grammar and vocabulary but their long term coherence leaves something to be desired. At the extreme, a model might spend a ton of capacity on memorizing the names of every single town on Earth or learning how to make typos in exactly the same way as humans do before it even budges on learning anything more high-order like logical reasoning or physical plausibility. Since there's just so many low order correlations to be learned, it might be that any model that we could reasonably ever train would get stuck here and would therefore never get to the high order correlations. (more on this idea of models learning low-order correlations first: [The Scaling Hypothesis](#))

Plus, once the model gets to high order correlations, it's not entirely guaranteed it would actually be able to learn them easily. It would depend heavily on humans (and the things humans model, so basically most of the known world) being easy to model, which.. doesn't seem to be the case, but I leave open the possibility that this is my anthropocentrism speaking.

Also, it might be that transformers and/or SGD just can't learn high order correlations and reasoning at all, possibly because there are possibly architectural constraints of transformers that start kicking in at some point (see [Can you get AGI from a Transformer?](#)). Some of these limitations might be fairly fundamental to how neural networks work and prevent *any* replacement architecture from learning high order reasoning.

If this scenario happens, then LMs will never become very capable or dangerous and therefore will not need to be aligned.

## Scenario 2: Strong capabilities scenario

The other way is the model could follow the scaling law curve perfectly until it gets to the irreducible loss, and then forever cease to improve. This scenario could also happen if there's an asymptotic approach to the irreducible loss that's fast enough that with a reasonable amount of compute we can get the level of capabilities discussed in this post, which is arguably more likely for higher resolution data because it would be harder to model. This case would happen if past a certain critical size, the model is able to entirely learn the process that generates the text (i.e the humans writing text, and the systems that these humans are able to observe and model) down to the resolution permissible by text. This scenario would be.. kind of scary, because it would mean that scaling is literally all we need, and that we don't even need that much more scaling before we're "there".

In this scenario, the prioritization of low-order correlations due to the negative log likelihood loss implies that high order traits will only improve slowly for potentially many orders of magnitude as the model is slowly able to memorize more and more inane things like the names of every famous person ever until suddenly it runs out of low hanging fruit and starts improving at constructing flawless logical deductions. This is, needless to say, *totally different* from how humans learn language or reasoning, and has led to a lot of both under and overestimating of GPT-3's capabilities. In fact, a lot of arguments about GPT-3 look like one person arguing that it must already be superintelligent because it uses language flawlessly and someone else arguing that since it has the physical/logical reasoning capabilities of a 3 year old, it's actually extremely unintelligent. This is especially dangerous because we will severely

underestimate the reasoning potential of LMs right up until the last moment when the model runs out of low order correlations. There would be very little warning even just shortly before it started happening, and probably no major change in train/val loss trend during, though there would be a huge difference in downstream evaluations and subjective generation quality.

And that's all assuming we keep using negative log likelihood. There remains the possibility that there exists a loss function that does actually upweight logical coherence, etc, which would completely invalidate this intuition, and bring highly intelligent LMs even faster. On the bright side there's the possibility this might be a positive, because it would likely lead to models with richer, more disentangled models inside of them, which would be useful for alignment as I mentioned earlier, though I don't think this is nearly enough to cancel out the huge negative of advancing capabilities so much. Thankfully, it seems like the reason such a loss function doesn't exist yet isn't from lack of trying, it's just really hard to make work.

To see just how much a different loss function could help, consider that cherrypicking from  $n$  samples is actually just a way of applying roughly  $\log n$  bits of optimization pressure to the model, modulo some normalization factors (I also talk about this intuition in footnote 2 of [Building AGI Using Language Models](#)). Since even lightly cherrypicked samples of GPT-3 are a lot more coherent than raw samples, I expect that this means we're probably not too far from a huge boost in coherence if we figure out how to apply the right optimization. This of course only provides a lower bound for how much the negative log likelihood loss needs to improve to get the same level of coherence, because it doesn't align with our notion of quality exactly, as previously mentioned—the model could happily improve  $\log n$  bits by memorizing random stuff rather than improving coherence.

I also think it's very unlikely that we will run into some totally insurmountable challenge with transformers or SGD that doesn't get patched over within a short period of time. This is mostly because historically, betting that there's something NNs fundamentally can't do is usually not a good bet, and every few years (or, increasingly, months) someone comes up with a way to surmount the previous barrier for NNs.

As one final piece of evidence, the scaling law and GPT-3 papers show that as your model gets bigger, it actually gets [more sample efficient](#). To me this is *a priori* very counterintuitive given the curse of dimensionality. This seems to imply that bigger models are favored by the inductive biases of SGD, and to me suggests that bigger is the right direction to go.

If this scenario happens, we're probably screwed if we don't plan for it, and even if we do plan we might still be screwed.

## What should we do?

There are a number of possible alignment strategies depending on how LMs scale and what the properties of bigger LMs are like. This list is neither exhaustive nor objective, these are just a few possible outcomes that take up a good chunk of my probability mass at this moment in time and my opinions at this time.

### Reward model extraction

Possibly the most interesting (and risky) direction in my opinion, which I mentioned briefly earlier, is trying to extract some reward model for human values or even something like [CEV](#) out of a big LM that has already learned to model the world. This idea comes from [Alignment by Default](#) (henceforth AbD for short), where we build some model and train it in a way such that it builds up an internal world model of what the AbD post calls “natural abstractions” (in the language of this post, those would be human-level abstractions that happen to be described in high resolution in natural language and are therefore more easily learned), and then we find some way to point to or extract the reward model we want inside that LM’s world model. One way this could be accomplished, as outlined in the AbD post, is to find some proxy for the reward signal we want and fine tune the LM on that proxy and hope that the model decides to point to an internal reward model. It might also turn out that there are other ways to do this, maybe by using interpretability tools to identify the subnetworks responsible for the LM’s understanding of CEV, or perhaps involve optimizing a continuous prompt to “select” the right part of the LM ([The Power of Scale for Parameter-Efficient Prompt Tuning](#), [Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm](#)).

I expect making a LM that forms a usable internal model of some useful reward signal inside itself to be very difficult in practice and require a lot of trial and error, possibly requiring custom training data or methods to make the resolution for that reward signal much higher. As an aside, I think this is one major advantage of thinking in terms of resolution: since resolution is relative to the data, we don’t have to just hope that any given abstraction is natural; we can actually make certain abstractions “more natural” for AbD if we want, just by changing the data! The other major issue with this approach is that there’s no guarantee that a model extracted from a LM will be at all robust to being subject to optimization pressures, and certainly I wouldn’t expect this to work as a strong alignment solution, but rather just as a way to bootstrap strong alignment, since the goal of *any* LM based system, in my opinion, would be to build a weakly-aligned system with the main purpose of solving strong alignment (i.e I think [Bootstrapped Alignment](#) is a good idea). That being said, I still think it would be significantly more robust than anything handcrafted because human values are very complex and multifaceted ([Thou Art Godshatter](#)), and learned systems generally do better at understanding complex systems than handcrafted systems ([The Bitter Lesson](#)). Also, it’s likely that in practice this extracted reward model wouldn’t be optimized against directly, but rather used as a component of a larger reward system.

This may sound similar to regular value learning, but I would argue this has the potential to be significantly more powerful because we aren’t just confined to the revealed preferences of humans but theoretically any consistent abstract target that can be embedded in natural language, even if we can’t formalize it, and the resulting model would hopefully be goodhart-robust enough for us to use it to solve strong alignment. Of course, this sounds extremely lofty and you should be skeptical; I personally think this has a low chance (10%ish) of actually working out, since a lot of things need to go right, but I still think more work in this direction would be a good idea.

Some very preliminary work in this direction can be seen in [GPT-3 on Coherent Extrapolated Volition](#), where it is shown that GPT-3 has at least some rudimentary understanding of what CEV is, though not nearly enough yet. The [ETHICS](#) dataset also seems like a good proxy for use in some kind of fine-tuning based AbD scheme.

Some concrete research directions for this approach: interpretability work to try and identify useful models inside current LMs like GPT-2/3; implementing some

rudimentary form of AbD to try and extract a model of something really simple; messing around with dataset compositions/augmentations to improve performance on world-modelling tasks (i.e PiQA, Winograd schemas, etc), which is hopefully a good proxy for internal model quality.

## Human emulation

We could also just have the LM emulate Einstein, except with the combined knowledge of humanity, and at 1000x speed, or emulate a bunch of humans working together, or something similar. How aligned this option is depends a lot on how aligned you think humans are, and on how high fidelity the simulation is. Humans are not very strongly aligned (i.e you can't apply a lot of optimization pressure to the human reward function and get anything remotely resembling CEV): humans are susceptible to wireheading and cognitive biases, have their own inner alignment problems ([Inner alignment in the brain](#)), and so on (also related: [The Fusion Power Generator Scenario](#)). Still, emulated humans are still more aligned than a lot of other possibilities, and it's possible that this is the best option we have if a lot of the other options don't pan out. Another limitation of human emulation is there's an inherent ceiling to how capable the system can get, but I wouldn't worry too much about it because it would probably be enough to bootstrap strong alignment. Related posts in this direction: [Solving the whole AGI control problem, version 0.0001 - Imitation](#)

A minor modification to this would be to emulate other nonhuman agents, whether through prompting, RL finetuning, or whatever the best way to do so is. The major problems here are that depending on the details, nonhuman agents might be harder to reason about and harder to get the LM to emulate since they'd be off distribution, but this might end up as a useful tool as part of some larger system.

Some concrete research directions for this approach: exploring prompt engineering or training data engineering to see how we can reliably get GPT-3 and similar LMs to act human-like; developing better metrics that capture human emulation quality better.

## Human amplification

This option encompasses things like IDA and (to a lesser extent) Debate where we rely on human judgement as a component of the AI system, and use the LMs as the magic black box that imitates humans in IDA or the debaters in Debate. How aligned this option is depends a lot on how aligned IDA and Debate are (and how aligned people are). Also, depending on how true the factored cognition hypothesis is, there may be serious limits to how capable these systems can get, though conditioning on IDA working at all, I don't think it's likely that this will be a bottleneck for the same reasons as in the previous sections. Some kind of human amplification strategy does feel like the most established and "down to earth" of the proposals despite all this, however. Overall, I'm cautiously optimistic about IDA/Debate-like systems using LMs.

The feasibility of this option is also correlated with the last option (human emulation) because this is essentially putting (partial) human emulation in a loop.

Some concrete research directions for this approach: implementing IDA for some kind of toy task using GPT-3 or similar LMs, possibly doing some kind of augmentation and/or retrieval system to squeeze the most out of the human data from the amplification step.

## Oracle / STEM AI

Another option that has been proposed for making safe AI systems is to make an "oracle" only able to answer questions and with only the goal of answering those questions as accurately as possible. A specific variant of this idea is the STEM AI proposal (name taken from [An overview of 11 proposals for building safe advanced AI](#), though this idea is nothing new), which is essentially an oracle with a domain limited to only scientific questions, with the hope that it never gets good enough at modelling humans to deceive us.

A bunch of people have argued for various reasons why oracle/tool AI isn't necessarily safe or economically competitive (for a small sampling: [The Parable of Predict-O-Matic](#), [Analysing: Dangerous messages from future UFAI via Oracles](#), [Reply to Holden on 'Tool AI'](#), [Why Tool AIs Want to Be Agent AIs](#)). I would argue that most of the safety concerns only apply to extremely strong alignment, and that oracles are probably safe enough for weak alignment, though I'm not very confident in this and I'm open to being proven wrong. This is the option that is most likely to happen by default, though, since even GPT-3 exhibits oracle-ish behavior with minimal prompting, whereas the other options I discussed are all various stages of theoretical. As such it makes sense to plan for what to do just in case this is the only option left.

Some concrete research directions for this approach: exploring prompt engineering to see how we can reliably get GPT-3 and similar LMs to give its best-guess rather than what it thinks the median internet user would tend to say; exploring ways to filter training data for STEM-only data and looking at whether this actually pushes the model's understandings of various topics in the direction we expect ([Measuring Massive Multitask Language Understanding](#) introduces a very fine grained evaluation which may be useful for something like this); interpretability work to see how the model represents knowledge and reasoning, so that we could possibly extract novel knowledge out of the model without even needing to query it (something like [Knowledge Neurons in Pretrained Transformers](#) seems like a good place to start).

## Conclusion

Natural language is an extremely versatile medium for representing abstract concepts and therefore given sufficiently high resolution data a language model will necessarily have to learn to represent and manipulate abstract concepts to improve loss beyond a certain point. Additionally, from the evidence we have from scaling laws, there is a fairly good chance that we will find ourselves in the strong capabilities scenario where this point is crossed only a few orders of magnitude in size from now, resulting in large language models quickly gaining capabilities, and making the first transformational AI a large language model. Finally, there are a number of existing prosaic alignment directions that are amenable to being applied to language models, which I hope will be explored further in the future.

# **Five Suggestions For Rationality Research and Development**

1. Make field work the bulk of what you do. For every hour you spend engaging with other people's thoughts or reflecting on your own thoughts, spend at least two hours, and ideally two days, trying to directly observe whatever it is you're studying.
2. Add other people to your sensorium. Make conversation with others a big part of your work, and approach those interactions with the intent to study the experience of others.
3. Corollary: If you're developing a technique or method, then by all means teach it to people. That's crucial. But don't teach them so they will know your technique, or even so you'll be better at teaching it. Instead, teach them so you can find out what happens for them when they try to use it.
4. Focus on topics in the intersection of "things you're personally interested in", "things you think might actually matter to someone in particular", and "places where the existing art seems deficient".
5. If it seems worthwhile to you, go after it. This is not a field with people in the position to tell you what you're allowed to study, or when, or how. This is a frontier. No bus will ever arrive, so you'll have to use your feet. Stop waiting for permission. Just get to work.

# Rogue AGI Embodies Valuable Intellectual Property

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

*This post was written by Mark Xu based on interviews with Carl Shulman. It was paid for by Open Philanthropy but is not representative of their views.*

Summary:

- Rogue AGI has access to its embodied IP.
- This IP will be worth a moderate fraction of the total value of the market created by models approximately as powerful as the rogue AGI.
- If investors realize that most economic output will eventually come from AGI, as in slow takeoff scenarios, then these markets will involve moderate fractions of the world's wealth.
- Therefore, rogue AGI will embody IP worth a non-trivial fraction of the world's wealth and potentially have a correspondingly large influence on the world.

A naive story for how humanity goes extinct from AI: Alpha Inc. spends a trillion dollars to create Alice the AGI. Alice escapes from whatever oversight mechanisms were employed to ensure alignment by uploading a copy of itself onto the internet. Alice does not have to pay an alignment tax, and so outcompetes Alpha and takes over the world.

On its face, this story contains some shaky arguments. In particular, Alpha is initially going to have 100x-1,000,000x more resources than Alice. Even if Alice grows its resources faster, the alignment tax would have to be very large for Alice to end up with control of a substantial fraction of the world's resources.

As an analogy, imagine that an employee of a trillion-dollar hedge fund, which trades based on proprietary strategies, goes rogue. This employee has 100 million dollars, approximately 10,000x fewer resources than the hedge fund. Even if the employee engaged in unethical business practices to achieve a 2x higher yearly growth rate than their former employer, it would take 13 years for them to have a similar amount of capital.

However, the amount of resources the rogue hedge fund employee has is not equivalent to the amount of money the employee has. The value of a hedge fund is not just the amount of money they have, but rather their ability to outperform the market, of which trading strategies and money are two significant components. An employee that knows the proprietary strategies thus can carry a significant fraction of the fund's total wealth, perhaps closer to 10% than 0.01%. In this view, the primary value the employee has is their former employer's trading high-performing strategies; knowledge they can potentially sell to other hedge funds.

Similarly, Alpha's expected future revenue is a combination of Alice's weights, inference hardware, deployment infrastructure, etc. Since Alice is its weights, it has access to IP that's potentially worth a significant fraction of Alpha's expected future revenue. Alice is to Alpha as Google search is to Alphabet.

Suppose that Alpha currently has a monopoly on the Alice-powered models, but Beta Inc. is looking to enter the market. Naively, it took a trillion dollars to produce Alice, so Alice can sell its weights to Beta for a trillion dollars. However, if Beta were to enter the Alice-powered model market, the presence of a competitor would introduce price competition, decreasing the size of the Alice-powered model market. Brand loyalty/customer inertia, legal enforcement against pirated IP, and distrust of rogue AGI could all disadvantage Beta in the share of the market it captures. On the other hand, Beta might have advantages over Alpha that would cause the Alice-powered model market to get larger, e.g., it might be located in a different legal jurisdiction (where export controls or other political issues prevented access to Alpha's technology) or have established complementary assets such as robots/chip fabs/human labor for AI supervision.

Assuming that the discounted value of a monopoly in this IP is reasonably close to Alice's cost of training, e.g. 1x-3x, competition between Alpha and Beta only shrinks the available profits by half, and Beta expects to acquire between 10%-50% of the market, Alice's weights are worth between \$50 billion and \$1.5 trillion to Beta. Abstracting away the numbers used in this particular example, Alice will be able to sell its weights to Alpha's competitors for a price that is a substantial fraction of, and perhaps even exceeds, the cost it took to train Alice (e.g. if the market value of computer hardware has gone up with improved AI performance so that it now costs more to train a replacement).

If Alice embodies IP worth a substantial fraction of the Alice-powered model market, then Alice's influence will be proportional to the size of this market. If Alice is sufficiently powerful, the Alice-powered model market is a large fraction of the entire world economy. Alice thus embodies IP worth a small to moderate fraction of the world economy, an immense amount of wealth. If Alice is less powerful, the value of its embodied IP depends on the degree to which investors can overcome frictions and uncertainty to fund enormous up-front training costs.

One way to estimate Alice's value is by assuming rough investment efficiency. [Paul Christiano](#):

If you are able to raise  $\$X$  to train an AGI that could take over the world, then it was almost certainly worth it for someone 6 months ago to raise  $\$X/2$  to train an AGI that could merely radically transform the world, since they would then get 6 months of absurd profits. Likewise, if your AGI would give you a decisive strategic advantage, they could have spent less earlier in order to get a pretty large military advantage, which they could then use to take your stuff.

In these worlds, relevant actors see AGI coming, correctly predict its economic value, and start investing accordingly. This rough efficiency claim implies AI researchers and hardware are priced such that one can potentially get 3x returns on investment (ROI) from training a powerful model, but not 30x.<sup>[1]</sup> Since most economic activity will rapidly involve the production and use of AGI, early-AGI will attract huge investments, implying the Alice-powered model market will be a moderate fraction of the world's wealth. The value of Alice's embodied IP, being tied to the value of that market, will thus be similarly massive.

---

1. This process may involve bidding up the prices of resources like server farms and researchers to absurd levels so that training a model that could 'take over the world' would require most of the world's wealth to rent the server time. ↪

# We need a standard set of community advice for how to financially prepare for AGI

Earlier today I was reading [this post](#) about the rationalist community's limited success betting on bitcoin and thinking about how the singularity is going to be the ultimate test of the rationalist community's ability to translate their unusual perspectives into wealth and influence.

There needs to be some default community advice here for people who believe that we're likely to create AGI in our lifetimes but don't know how to prepare for it. I think it would be an absolute shame if we missed opportunities to invest in the singularity the same way we missed opportunities to invest in Bitcoin (even though this community was clued in to crypto from a very early stage). I don't want to read some retrospective about how only 9% of readers made \$1000 or more from the most important even in human history even though we were clued in to the promise and peril of AGI decades before the rest of the world.

John\_Maxwell [made a post about this last year along the same lines.](#) but i'd like to expand on what he wrote.

## Why is this important?

In addition to the obvious benefit of everyone getting rich, I think there are several other reasons coming up with a standard set of community advice is important.

Betting on the eventual takeover of the entire world economy by AI is not yet a fashionable bet. But like Bitcoin, betting on AGI will inevitably become a very fashionable bet in the next few decades as first early adopters buy in, and then it becomes standard part of financial advice given out by investment professionals.

In these early days, I think there is an opportunity for us to set the standard for how this type of investment is done. This should include not just a clear idea of how to invest in AGI's creation (via certain companies, ETFs, AI focused SPACs etc), but also what should NOT be done.

For example, the community advice should probably advise against investing in companies without a strong AI alignment team, as capitalizing such companies will increase the likelihood that AI will destroy you and everything you love. We may also want to discourage investment in companies that don't have a clause on how they plan to deal with race conditions that could compromise safety. There are probably other considerations we should make that I am not thinking of. [OpenAI's charter](#) seems like a pretty well thought-out set of guidelines for AGI creation. This site has a very healthy community of AI safety researchers whose advice on this topic I would very much appreciate.

Whatever the advice is, I think it's important that it subsidizes good behavior without diminishing expected returns too much. If we advise against investing in the

organization that looks most likely to create AGI because they don't meet some safety standard, we run the risk of people ignoring the advice.

There is some small chance that if these guidelines are well thought out they could eventually be adopted by investment companies or even governments. BlackRock, an investment management corporation with \$8.67 trillion under management, has begun to divest from fossil fuels in the interest of attracting money from organizations concerned about climate change. If the public comes to see unaligned AI as a threat at some point in the future, existing guidelines already adopted by other investors or financial institutions could become a easy thing for investment managers to adopt so they can say they are "being proactive" about the risk from AI.

## **What would this advice look like?**

Let's reflect on some of the lessons learned from crypto craze. Here I will quote from several posts I've read.

A hindsight solution for rationalists to have reduced the setup costs of buying bitcoin would have been either to have had a Rationalist mining pool or arrange to have a few people buy in bulk and serve as points of distribution within the community.

This suggests that if a future opportunity appears to be worth the risk of investment, but has some barrier to entry that is individually costly but collectively trivial, we ought to work first to eliminate that barrier to entry, and then allow the community to evaluate more dispassionately on risk and return alone.

- **clarkey**

I think lowering the barrier to doing something is a great idea but it's hard to know exactly what that would look like. Could we create our own ETF? Would it be best to create a list of stocks of companies that are both likely to create AGI and have good incentive structures set up to make proper alignment more likely? I think ideally there would be tiers of actions people could take depending on how much effort they wanted to expend, where the lowest tier with the least action would be "Set up a TD Ameritrade account and buy this ETF" or something and the most complicated would be "here is a summary for each company widely regarded by members of the AI alignment forum to have good alignment plans and here's a link to some resources to learn about them."

## **Is there really an opportunity here? Why would we expect to beat the market in this situation?**

The answer to this is more complicated and I'm sure other people probably have better answers to this than me. But I'll give it a shot.

I realize that saying this sounds very unacademic, but the creation of AGI will be the most important moment in the history of life so far. If AGI does not destroy us or torture us or pump our brains full of dopamine for eternity, it will have transformative effects on the economy the likes of which we have never seen. It's plausible that worldwide GDP growth could accelerate by 10x or more. A well aligned AI is a wish-granting machine whose limitations are the as-of-yet-incomplete laws of physics.

Think about how nuts this sounds to the average hedge fund manager. They have no point of reference for AGI. It pattern matches to happy magical fairy tale or moral fables from children's storybooks. It doesn't sound real. And I would bet the prospect of ridicule has prevented the few who actually buy the idea from bringing it up with investors. If you listen to interviews with top people from JP Morgan or Goldman Sachs they use the same language to refer to AI as they use to refer to everything else in their investment portfolio. There's nothing to signal that this is fundamentally different from biotech or clean energy or SaaS products.

With such communication and conceptual barriers, why would we expect assets to be priced properly?

I'd welcome feedback here. Maybe I'm missing something or maybe I've been listening to the wrong subset of the investment community. But my overwhelming impression is that almost no one on Wallstreet or anywhere else truly buys into the vision of AGI as the last invention humans will ever make.

## **Here's my current strategy and why I find it unsatisfying**

Earlier this year I got sick of not betting on my actual beliefs, and put about \$10k into Google, Microsoft and Facebook in proportion to the number of publications they each made in NeurIPS and ICML over the last two years, treating publication count as a proxy for the likelihood that each company would create the first AGI. I would have put in more but I don't have much more.

Though I think this is better than nothing, I can't help but think there must be a better, more targeted way to bet on AGI specifically. For example, I don't really care that much about Google's search business, but I am forced to buy it when I buy Google stock.

This strategy also neglects all small companies. I think there is a low enough level of hardware overhang right now that it is overwhelmingly likely AGI will be created in one or more big research labs. But perhaps the final critical piece of the puzzle will come from some startup that gets acquired by Microsoft AI labs and owning a piece of that startup will result in dramatically higher returns than buying the parent company directly.

Unfortunately accredited investor laws literally make it illegal to invest in early-stage startups unless you're already rich. So all the rapid growth from early-stage startups is forever out of reach for people who aren't already rich. (By the way, these laws are one of the reasons private equity has averaged about double the returns of the S&P 500 over the last 30 years. Rich people have a monopoly on startup equity). SPACs are kind of a backdoor to getting into early-stage startups without a lot of money, but

companies have to agree to merge with a SPAC so your options are still somewhat limited. However I think the SPAC strategy is worth looking into.

You could always buy an AI ETF. I'll be honest and say I haven't really looked into that much but would appreciate feedback from anyone that has.

Anyways, those are my thoughts on this subject. Let me know what you think.

# The value of low-conscientiousness people on teams

*[Apologies in advance if I sound like I'm over-generalizing high-conscientiousness or low-conscientiousness people. This is mostly from my own experience, so I'm sure I'm wrong on some counts and may, in fact, be over-generalizing at times. Ohh, and also apologies to Mick Jagger.]*

Please allow me to introduce myself. I'm a man of mess and wile. I've been scoring around 20% (in [trait conscientiousness](#) on [Big Five](#) tests) for a long, long year, cut by many a sharp wire's height.

At first glance, conscientiousness as a construct seems a bit like intelligence, in the sense that it would seem everyone would be better off with more of it. So, why would natural selection produce people like me who are very low in conscientiousness? Have I only escaped a [Darwin award](#) by the grace of the almighty simulator?

I have a hypothesis that low-conscientiousness people may function a bit like dichromats (color blind people) on teams of hunter-gathers. While dichromats can't see some colors, [they can detect color-camouflaged objects better than non-color blind people \(trichromats\)](#). So, teams with mixtures of dichromats and trichromats may have out-competed teams with only trichromats (or only dichromats, for that matter).

Perhaps some diversity of conscientiousness in groups could produce a competitive advantage? There is some evidence in this direction. Just from skimming [Wikipedia](#):

- [The world's most conscientious nations are also some of the poorest.](#)
- [Groups with only conscientious members have difficulty solving open-ended problems.](#)
- [Those scoring low on conscientiousness make better decisions after unanticipated changes in the context of a task.](#)
- [Conscientiousness has been found to be positively correlated with business and white-collar crime.](#)

But, specifically, I'd like to surface patterns I've observed in my own experience that I haven't seen discussed elsewhere. I work with, and am related to, many high-conscientiousness people. I've come to appreciate their strengths relative to mine, but I also have first-hand experience with some failure modes of conscientiousness taken too far. Any virtue taken to an extreme can become a vice, as Dieter said. (No, not [that Dieter, Dieter Uchtdorf](#)).

Allow me to present four of my observed failure modes of high-conscientiousness.

## #1 Not all messes are worth the risk of cleaning them up

This is the most frequent failure mode I see amongst high-conscientiousness people I work with. In lieu of a technical example, let me give you a more practical one.

I live in a fairly old neighborhood. Houses are typically made of brick, and I'd estimate the mean age of a house in my neighborhood is about 85 years old. Brick and mortar tend to get crumbly in spots over time. Often to keep structural integrity you need to [repoint the mortar and/or replace damaged brick](#). Though, if you haven't poked at your exterior brick walls or

had sufficient experience with deteriorating brick and mortar, you probably wouldn't be aware of this.

Last summer, I overhead a conversation between two neighbors that could have entered this failure mode, but likely narrowly avoided it. We'll call my neighbors Charlotte and Miranda.

Conscientious Charlotte has an old brick garage that used to have ivy growing on three sides, though she recently removed it from the south side. Removing the ivy left a lot of unsightly "[rootlets](#)" on the brick surface of the south side exterior. Charlotte told Miranda that the rootlets look messy and that she'd like to clean them off.



This isn't the garage in question, but the dots visible along the wood trim here are what the rootlets look like when ivy is removed. Charlotte's bricks are unpainted and beige.

Miranda says her spouse has a pressure washer that may help. Miranda goes home and asks her spouse, Marla, if Charlotte can borrow the power washer. Marla asks "what for?" and Miranda explains Charlotte's predicament. Marla says, "Sure she can borrow it, just let her know that depending on how weathered the brick and mortar is, she may not have a garage standing when she's done power washing the rootlets off." ([Marla was completely right, see item #06 here](#)).

A defender of high-conscientiousness could retort here and say "Charlotte's problem wasn't conscientiousness, it was that she wasn't conscientiousness enough." I believe this is confusing conscientiousness with circumspection. Circumspection requires some amount of [curiosity, which is associated with trait openness](#). Maybe also some worrying about what could go wrong, which seems like trait neuroticism.

My experience is more that the compulsion to clean up *unfamiliar* messes among high-consciousness people I work with tends to override any thoughtfulness regarding how to account for the unfamiliarity. Routine clean up is fine because it's routine. But when faced with a messy situation they haven't specifically encountered before it seems like their instinct is to address it through cleaning, organizing, sorting, ordering, adding structure, etc. rather than to step back and say "how can we be sure we're safely addressing this, or is it even worth the risk of addressing at all?"

Low-consciousness people are likely to ask questions like this because we're naturally more adverse to cleaning up messes and, hence, only do it when it's really necessary or when we're coerced to. We'll wonder "is this worth my time?" Or "is this worth my team's time?" If you're lucky enough to have a low-consciousness team member that's also high in openness and neuroticism, that may be even better here.

## #2 Honestly estimate ROI before micro-optimization (avoid bikeshedding)

So, let's imagine Charlotte still wants to clean the rootlets off her garage. She now realizes she can't safely use a power washer unless she also wants to risk restoring or replacing her garage.

[So she carefully goes out with a brush, a bucket of water and detergent, then she proceeds to clean rootlets one small section at a time.](#)

Look, it's Charlotte's life and it's her garage. If she wants to toothbrush rootlets off her garage exterior... it's a free country (assuming you're reading this in a free country).

I will argue a different calculus applies if you're doing something analogous in a work environment where you're billing a customer who expects value for their dollars, or someone is paying your salary who expects value from your work, or, importantly, if you're asking someone like me to do this when I could be working on something else.

Even after you've established a safe method for cleaning up a mess, is it really worth doing in comparison to all the other things you or someone else could be working on?

Situations like this are sometimes called [bikeshedding](#). The term comes from a hypothetical project to build a nuclear power plant. In the midst of all of the planning necessary to safely build the plant, during a project meeting someone raises their hand and says "we should have a bike shed in case employees want to bike to work, and it should be green because we're about green energy!" Then another person in the meeting says "No, it should be white to reflect more of the sun's heat and maintain a lower temperature!" The meeting then spends an inordinate amount of time discussing what color the bike shed should be. The point being a relatively trivial matter derails the larger, more important discussion.

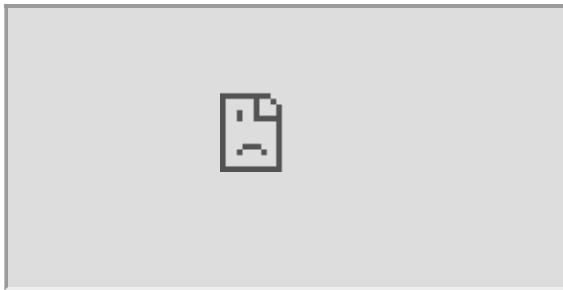
You may think conscientious people are good at avoiding this, my experience is the opposite. Sometimes small details are meaningful, sometimes they're bikesheds, or rootlets on the garage. Honestly think about the ROI before micro-optimizing in these ways.

Low-consciousness people can, again, help here. We're not going to go around toothbrushing rootlets and we're less likely to care about details unless there's a strong argument in favor of their importance.

## #3 Avoid supererogation

[Supererogation](#) is doing more of something than required. One way to think about it is to imagine someone who commits a crime and gets sentenced to prison for 19 years, then

when their 19 year sentence has expired they beg the prison guards to keep them for another 3 years so they can do even more penitence. If you're a [Les Misérables](#) fan, imagine Javert releasing Jean Valjean from prison to parole, only to find that Jean Valjean first protests that he hasn't served enough time, then later complains that the parole terms are too lenient.



This is obviously a bit different than the Wikipedia definition, which describes more in the terms of doing more than what duty requires. But I maintain that those are apt analogies, if hyperbolic by comparison to more day-to-day examples.

(BEGIN SPOILER ALERT)

The straightforward and obvious example of supererogation from [Les Misérables](#) is Javert. In fact, Javert is more than an example of supererogation, he's an example of terminal supererogation. You will often hear people say "follow the spirit of the law, not the letter of the law." Javert is such a stan for law and duty, he makes anyone strictly following the letter of the law seem eminently reasonable by comparison. Without summarizing the plot, he tracks Valjean over the course of two decades and observes him not just obeying the law, but saving lives and taking care of others, but still wants to return him to prison. Not to get too dark, but Javert commits suicide after unsuccessfully grappling with his predicament.

This is similar to bikeshedding in that the ROI here, relative to other crimes and criminals, is low. (So low that it's negative—it's worse than quixotic. Arresting or executing Valjean at any point where Javert encounters him outside of prison would be a net harm to society.) It's different than bikeshedding as the ROI is a known quantity but still is pursued for philosophical reasons—duty, law, etc. To say nothing of the loss of opportunity Javert trades to continue to pursue Valjean, not to mention the years of life he lost from ending it early.

(END SPOILER ALERT)

If you're a Javert, a fraction of a Javert, or have such people on your team, maybe add some trusted low-conscientiousness people to help you chill things out a bit? Just stay away from [Sacha Baron Cohen's Inn](#).

## #4 Beware the Superman complex

The Wikipedia definition of the [Superman complex](#) is pretty good.

[An] unhealthy sense of responsibility, or the belief that everyone else lacks the capacity to successfully perform one or more tasks. Such a person may feel a constant need to "save" others and, in the process, takes on more work on their own.

It's different than someone who has [a lot of responsibility thrust upon them](#), or who [creates catastrophes to save things and get recognition](#).

I suspect we all know people who have played this role for parts of their lives. I won't say this is limited to high-conscientiousness types, but I do see it frequently among many of the ones I know.

I don't know if there's deeper, psychoanalytic, reasons why people do this. But I would hope it's a bit like removing your hand from a hot stove—once you recognize it, you can stop doing it.

Low-conscientiousness types may not be able to analyze the distal reasons for people doing this, but I'm pretty sure we'll be better than others at spotting when you're working to your own detriment and potentially the detriment of others.

## **The general pattern**

Apply the principle of maximum parsimony to all things "conscientiousness." With rules, for example, instead of more rules prefer the most parsimonious set of rules—the fewest rules necessary for the maximal outcome.

# Search-in-Territory vs Search-in-Map

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

Let's split out two different types of optimization. The first type includes things like Google Maps' pathfinder, a human making a plan, or Amazon's logistical algorithms. The second type includes things like a thermostat, a heat-seeking missile, or water flowing downhill.

The distinction I want to point to here is where the things-we're-optimizing-over live. For the first type, optimization is over things-in-a-model; a search algorithm runs on a map. For the second type, optimization is over things-in-the-world; a search algorithm runs on the territory directly. Search-in-map vs search-in-territory.

## Same Algorithm, Different Inputs

A toy example: suppose we have a big pile of rocks, and we want to find the rock whose mass is closest to the mass of a reference weight (without going over).

A search-in-territory algorithm might use one of those old-school balance-scales to compare masses pairwise. We could pull each rock out of the pile one-by-one, and:

- First, compare the rock to the reference weight. If it's heavier, throw it away and move on to the next rock.
- Second, compare it to the best rock found thus far, and replace the best rock with this one if it's heavier.

At the end, we'll have chosen the best rock.



A balance-scale. ([Source](#))

A search-in-map algorithm might instead start by weighing the reference weight and each rock on a scale, and entering all the masses into a spreadsheet. But then, it could proceed exactly like the search-in-territory algorithm. It would run through the list of rock-masses one-by-one, and:

- First, compare the rock-mass to the reference mass. If the rock-mass is larger, move on to the next one.
- Second, compare the rock-mass to the best rock-mass found thus far, and replace the best rock-mass with this one if it's larger.

At the end, there is one extra step: we have to go back out into the world and find the actual rock with the best mass. Note that this could be nontrivial, if we didn't label or organize our rocks in the process of weighing them.

Key point of this example: these two types of optimization need not involve different algorithms. *In practice* they often do - things we typically think of as "search algorithms" tend to be used more for search-in-map, and things we typically think of as "controllers" tend to be used more for search-in-territory. But in principle, one can often apply the algorithms opposite their usual context.

## When Should Search-In-Map Be Favored?

I see two main features of search-in-map which aren't (typically) shared by search-in-territory:

- The map-making process can use information before the search process "knows what to do with it". For instance, in the rock-pile example, we could weigh all the rocks before we gain access to the reference rock, or even before we know what the task is at all.
- The map itself can use a convenient representation or data structure - e.g. indexes, caching, sorted lists, etc.

These features can sometimes be replicated to some extent in the territory - e.g. we could weigh all the physical rocks and store them in sorted order before gaining access to the reference weight. But this only works when we have lots of control over the physical system, and can arrange it how we please. A map can pretty much always be arranged how we please.

Key point: if we can use information to build a map before we have full information about the optimization/search task, that means we can build one map and use it for *many* different tasks. We can weigh all the rocks, put that info in a spreadsheet, then use the spreadsheet for many different problems: finding the rock closest in weight to a reference, finding the heaviest/lightest rock, picking out rocks which together weigh some specified amount, etc. The map is a capital investment.

A more typical example: creating a streetmap is expensive. If we just want to figure out one shortest path, then directly measuring physical distances along a bunch of paths might be faster. But once the streetmap is created, it can be used repeatedly by lots of different people for lots of different pathfinding problems.

## When Should Search-In-Territory Be Favored?

The key feature of search-in-territory is that it does *not* require making a map - and therefore requires no extra information or assumptions beyond what the algorithm itself uses.

In the rock-pile example, the search-in-territory uses only pairwise comparisons between weights - i.e. a balance scale. We never actually know the mass of any particular rock. The search-in-map, on the other hand, relies on direct measurements of mass. To perform the search-in-map with only a balance scale, we'd either need to compare all pairs of weights ahead of time (which would mean  $O(n^2)$  effort), or we'd need to run out and compare physical weights in the middle of the search (at which point we're effectively back to search-in-territory).

Key point: if we don't rely on extra information, then we don't rely on extra assumptions; there are no extra sources of error. In the rock-pile search-in-map example, error could come from the scale becoming uncalibrated as we proceed through the rocks, or from insufficient precision in the measured masses, or from data corruption in the spreadsheet, or from failing to connect the search result to the right physical rock at the end. We have to assume that we correctly understand each of those pieces. For the search-in-territory version, we only need to assume that the balance scale works correctly.

A more typical example: many feedback controllers work well even when our model of the system's dynamics isn't quite correct.

## How Does This Compare To Selection Vs Control?

Abram's [Selection vs Control](#) offers a similar distinction between two kinds of optimizers. "Selectors", in his breakdown, directly instantiate and compare objects in the search space. There's an explicit space of "options" or "possibilities" over which the search operates. This includes all of the search-in-map examples from earlier, but also biological evolution.

"Controllers" don't necessarily have that explicit search-space - they tend to operate directly on the external world. "Other possibilities" for a controller would mean other counterfactual worlds, which aren't necessarily unambiguously defined. This includes all of the search-in-territory examples from earlier.

I claim that search-in-territory vs search-in-map is usually the right distinction to think about when considering the intuitive selection vs control clusters. The main reason is that **counterfactuals always live in a model**. An objectively defined "search space" exists only in a model.

There are cases where certain search spaces/counterfactuals seem particularly natural. For instance, in the case of biological evolution, it seems natural to consider the space of DNA sequences. But remember that evolution does not "actually search" this space - it only samples a relatively-small chunk of the exponentially large space of sequences.

Conversely, there are cases where we *think of* a search-in-territory as having some search space. For instance, an optimal controller might minimize some expected "cost" under some model of the system under control, but without explicitly representing that map or optimizing over it.

## Why Does This Matter?

As with the selection-vs-control distinction, it intuitively feels like search-in-map is "more powerful" than search-in-territory; it looks less like a thermostat and more like an agent. But if they both do optimization (i.e. they both [steer certain parts of the universe into a smaller set of states](#), which is [equivalent to expected utility maximization](#) in a God's-eye model), then what's the difference?

The discussion above suggests one possible answer: maps generalize. This connects to the [\(Improved\) Good Regulator Theorem](#): a system "should" use an efficient internal map when it "doesn't know what game it's playing" until later. In that case, we need to keep around useful information, but we still want to compress and precompute where possible.

Then the interesting question is: how do we recognize a map embedded in the territory, or a search algorithm running on such a map?

# The Language of Bird

*This is a fictional snippet from the [AI Vignettes Day](#). I do not think this is a likely future, but it's a useful future to think about - it gives a different lens for thinking about the alignment problem and potential solutions. It's the sort of future I'd expect if my own research went far better than I actually expect, saw rapid adoption, and most other ML/AI research stalled in the meantime.*

[Transcript from PyCon 2028 Lightning Talk. Lightly edited for readability.]

Ok, so, today we're going to talk about Bird, and especially about the future of Bird and machine learning.

First of all, what is Bird? I assume everyone here has heard of it and played around with it, but it's hard to describe exactly what it is. You've maybe heard the phrase "human concept library", but obviously this isn't just a graph connecting words together.

Some background. In ye olden days (like, ten years ago) a lot of people thought that the brain's internal data structures were inherently illegible, an evolved hodgepodge with no rhyme or reason to it. And that turned out to be basically false. At the low level, yeah, it's a mess, but the higher-level data structures used by our brains for concept-representation are actually pretty sensible mathematical structures with some nice universal properties.

Those data structures are the main foundation of Bird.

When you write something like "from Bird.World import Bookshelf", the data structure you're importing - the Bookshelf - is basically an accurate translation of the data structure your own brain uses to represent a bookshelf. And of course it's hooked up to a whole world-model, representing all the pieces of a bookshelf, things you put on a bookshelf, where you'd find a bookshelf, etc, as well as the grounding of all those things in a lower-level world model, and their ultimate connection to sensors/actuators. But when writing code using Bird, we usually don't have to explicitly think about all that. That's the beauty of the data structures: we can write code which intuitively matches the way we think about bookshelves, and end up with robust functionality.

Functionally, Bird is about *translation*. It's a high-level language very close to the structure of human thought. But unlike natural language, it's fully formally specified, and those formal specifications (along with Bird's standard training data set) accurately capture our own intuitive concepts. They accurately translate human concepts into formal specifications. So, for instance, we can express the idea of "put a strawberry on a plate" in Bird, hand that off to an ML algorithm as a training objective, and it will actually figure out how to put a strawberry on a plate, rather than Goodharting the objective. The objective actually correctly represents the thing we're intuitively saying.

That's the vision, anyway. The problem is that there's still some assumed social context which Bird doesn't necessarily capture - like, if I write "put a strawberry on a plate", the implicit context includes things like "don't kill anyone in the process". Bird won't include that context unless we explicitly add it. It accurately captures "put a strawberry on a plate", but nothing else.

Today, of course, Bird's main use-case is for smart contracts. For that use-case, it's fine to not include things like "don't kill anyone", because we have social norms and legal structures to enforce all that already. So for contracts between humans, it's great. Thus the big resurgence in smart contracts over the past few years: we can finally formally specify contracts which actually do what we want.

But for ML systems, that doesn't really cut it. ML systems aren't human, they won't strictly follow social norms and laws unless we program - or train - them to do so.

The obvious solution to this is to express things like "follow the law" or "obey social norms" or ideally "do what I mean, not what I say" in Bird. But this is all tightly tied in with things like agency, self-reference, and goal-directedness. We don't yet fully understand the right data structures for those things - self-reference makes the math more complicated. That's the main missing piece in Bird today. But it is an active research area, so hopefully within the next few years we'll be able to formally specify "what we want".

# Reinforcing Habits

*Epistemic status: hypothesis based on >10 anecdotal examples*

Every month or so, I'll get a client asking why their attempts to start a habit failed. They want to have an automatic action requiring minimal willpower. The client is usually familiar with at least one habit-building model. Most commonly Charles Duhig's [Cue, Routine, Reward loop](#) (in "The Power of Habit") or CFAR's [Trigger-Action Plans \(TAPs\)](#).

Their model may go like Duhig's story: he wanted to change his habit of eating a cookie each afternoon (motivated by watching the scale creep up). So he identified his cue (the time of day around 3pm), planned a new routine to replace the old one (talk to colleagues for ten minutes), and had a new habit. My client usually wants to know what thwarts their attempts to do likewise.

What's missing from that example?

I read The Power of Habit many years ago, and honestly didn't remember Duhig's cookie story any better than my clients often do. So I was surprised when I revisited it and found Duhig actually did a series of experiments to find what the reward was. Something sweet? Nope, eating a candy bar at his desk didn't feel great. Just taking a break? Nope, taking a walk outside didn't cut it. Talking with friends? Yeah, that felt rewarding. Based on his experiments, we can guess that getting a cookie each afternoon was a means for getting him to talk to his colleagues while eating it. Because the reward here was socializing, he could build a new habit that didn't use the cookie as an intermediate step.

This reward step is often neglected by my habit-struggling clients. They want the low-effort, automatic aspects of habits. They lack, however, anything to make those behaviors sticky.

I think it might help to reframe habits as *repeatedly-reinforced behaviors*. Our brains, often subconsciously, have tied a particular action to some cue after repeatedly having that action rewarded. Simple patterns of cue, action, and reward in close proximity get reinforced, such as "it's mid afternoon -> I'll get a cookie -> rewarded by social connection."

Intentionally designing good habits is hard. By default, our unconscious habits are selected for rewarding behaviors. For example, when I'm feeling blue, eating chocolate and hugs make me happier. I don't need to train myself to eat chocolate when I feel blue - this happens quite easily!

On the other hand, you have to intentionally find the reward when you're trying to kickstart a habit. In particular, you need to find the reward if you want your habit to happen automatically without willpower each time. A random desired behavior may not have an immediate reward, so you need to experiment. (I'm not guaranteeing that all behaviors can be made into habits - there's a reason that doctors recommend "whichever exercise you'll actually do" rather than a specific fitness-optimized routine.)

Two similar actions can cause very different experiences due to small differences specific to you. Pullup hangs and RSI wrist stretches were both small actions that I

repeated for brief reps three times a day. The pull-up hangs were motivating because I could see myself improve day to day - an extra second here, three seconds longer there. The RSI stretches quickly became demotivating because I couldn't see myself making any progress even after a couple weeks of consistent use.

Similarly, 7am yoga required financial penalties to get me into downward dog before the world warmed up. Walking a mile in the peaceful evening while thinking or calling a friend was a piece of cake in comparison.

The best rewards are natural consequences of the action--i.e. the experience of doing the action reinforces the behavior. The reward might be enjoyment of the action, seeing progress toward a goal, a social status boost, consistency with your sense of self, connection, release from a worry, etc. Note, all of these are gut-level feelings, not "shoulds". A System 2-level sense of "I should..." doesn't seem to have the same rewarding effect. You actually have to find what feels rewarding. If you can identify and increase the reward, you can make the habit easier to sustain. This implies that the best way to deliberately change/start habits is to choose new habits with immediate positive outcomes, and make those benefits salient.

On the other hand, you can also try to tack on a reward that doesn't inherently come with the action, such as fist pumping the air or using financial penalties. Arbitrary rewards can be quite useful (particularly in situations when you just have to push through something unpleasant). However, financial penalties and other arbitrary rewards fall apart if you stop applying a bit of willpower to set them up each time. Rewards with a self-coercive element are also more draining/stressful to use for many people. (Not surprising - the rewarding habit is more pleasant than a financial penalty.)

One popular example is [temptation bundling](#). The idea is you only give into a temptation while also doing a desired activity, such as only watching TV while exercising. However, a common outcome is someone tries to watch TV exclusively while exercising, only to have some part of their brain point out that nothing is stopping them from watching TV in bed...

In contrast, here are few examples of people using natural rewards to reinforce habits:

1. According to [Nate Soares](#), "When I was quite young, one of the guests at our house refused to eat processed food. I remember that I offered her some fritos and she refused. I was fairly astonished, and young enough to be socially inept. I asked, incredulous, how someone could not like fritos. To my surprise, she didn't brush me off or feed me banal lines about how different people have different tastes. She gave me the answer of someone who had recently stopped liking fritos through an act of will. Her answer went something like this: 'Just start noticing how greasy they are, and how the grease gets all over your fingers and coats the inside of the bag. Notice that you don't want to eat things soaked in that much grease. Become repulsed by it, and then you won't like them either.'"
2. [Tara Mac Aulay](#): "I found that going to lift weights with friends is surprisingly good, because I get to have a good chat with them for an hour, and it's not strenuous enough that you can't have a conversation. But if I was to go on a bike ride with friends or something else where you can't talk, it's not as nice. And my main exercise is probably just walking and dancing. I go out dancing a lot on my own, to go and see music artists that I enjoy, and I just dance like a crazy person until I'm really tired and then I go home, and that's amazing."

3. I struggled for a while to brush my teeth consistently. I eventually paid mindful attention to the feeling of stuff on my teeth. The little layer of course film when I hadn't brushed after eating, and the polished silk of freshly brushed teeth. I started getting annoyed at the texture on my teeth, and then brushing was easy.

Experimenting and paying close attention to what feels rewarding seem to be common elements behind the successes above.

A couple ideas if you want to try using natural rewards to build habits for yourself:

- Run experiments to find what you enjoy enough to easily make a habit. You can track these formally, or just note which are easier to do repeatedly.
- Use [Soares' technique](#) of focusing on the experience of minute details that attract you to habits you want, or make unwanted habits less desirable. For example, pay attention to how much better you feel when you don't have an important email hanging over your head, compared to when you were procrastinating.
- Use [this CBT Pleasure Predicting Worksheet](#) to increase your awareness of how much you enjoy activities. This sheet works by highlighting discrepancies between expected and experienced enjoyment.

# Parameter counts in Machine Learning

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

**In short:** we have compiled information about the date of development and trainable parameter counts of n=139 machine learning systems between 1952 and 2021. This is, as far as we know, the biggest public dataset of its kind. You can access our dataset [here](#), and the code to produce an interactive visualization is available [here](#).

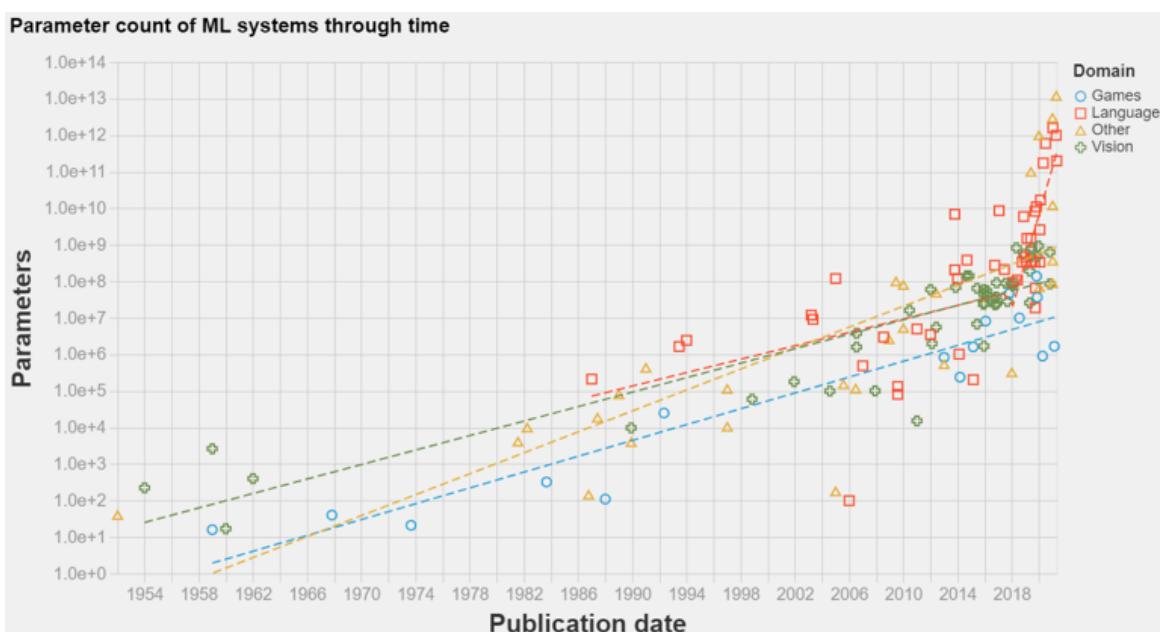
We chose to focus on parameter count because previous work indicates that it is an important variable for model performance [1], because it helps as a proxy of model complexity and because it is information usually readily available or easily estimable from descriptions of model architecture.

We hope our work will help AI researchers and forecasters understand one way in which models have become more complex over time, and ground their predictions of how the field will progress in the future. In particular, we hope this will help us tease apart how much of the progress in Machine Learning has been due to algorithmic improvements versus increases in model complexity.

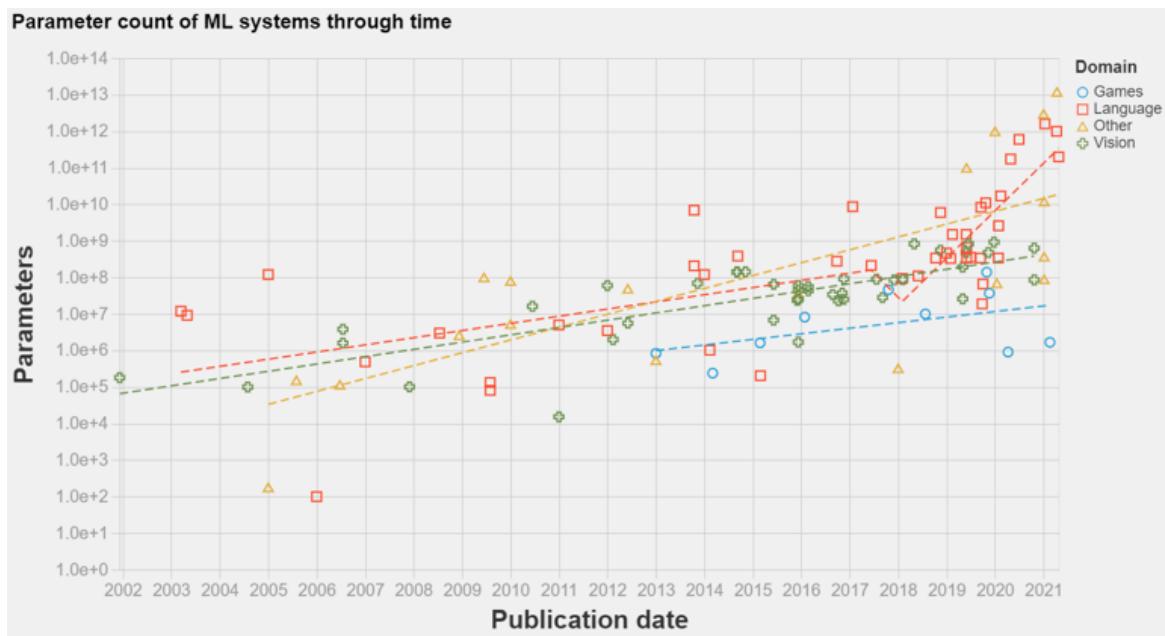
It is hard to draw firm conclusions from our biased and noisy dataset. Nevertheless, our work seems to give weak support to two hypotheses:

- There was no discontinuity in any domain in the trend of model size growth in 2011-2012. This suggests that the Deep Learning revolution was not due to an algorithmic improvement, but rather the point where the trend of improvement of Machine Learning methods caught up to the performance of other methods.
- In contrast, it seems there has been a discontinuity in model complexity for language models somewhere between 2016-2018. Returns to scale must have increased, and shifted the trajectory of growth from a doubling time of ~1.5 years to a doubling time of between 4 to 8 months.

The structure of this article is as follows. We first describe our dataset. We point out some weaknesses of our dataset. We expand on these and other insights. We raise some open questions. We finally discuss some next steps and invite collaboration.



Model size of popular new Machine Learning systems between 1954 and 2021. Includes n=139 datapoints. See expanded and interactive version of this graph [here](#).



Model size of popular new Machine Learning systems between 2000 and 2021. Includes n=114 datapoints. See expanded and interactive version of this graph [here](#).

## Features of the dataset

- The dataset spans systems from 1952 to 2020, though we included far more information about recent systems (from 2010 onwards).
- The systems we include encompass many types, including neural networks, statistical models, support vector machines, bayesian networks and other more exotic architectures. However we mostly included systems of the neural network kind.
- The systems are from many domains and were trained to solve many tasks. However we mostly focused on systems trained to solve vision, language and gaming tasks.
- We relied on a subjective criteria of notability to decide which systems to include. Our decisions were informed by citation counts (papers with more than 1000 citations), external validation (papers that received some kind of paper of the year award or similar) and historical importance (papers that were cited by other work as seminal). The references to this post include some overviews we used as a starting point to curate our dataset [2-26].
- Several models have versions at multiple scales. Whenever we encountered this in their original publication, we recorded whichever was presented in the paper as the main one, or the largest presented version. Sometimes we recorded multiple versions when we felt it was warranted, e.g. when multiple different versions were trained to solve different tasks.

## Caveats

- It is important to take into account that model size is hardly the most important parameter to understand the progress of ML systems. Other arguably more important indicators of non-algorithmic progress in ML systems include training compute and training dataset size [1].

- Model size as a metric of model complexity is hardly comparable across domains or even architectures. For example, a mixture-of-expert model can achieve higher parameter counts but invest far less compute into training each parameter.
- Our selection of systems is biased in many important ways. We are biased towards academic publications (since information on commercial systems is harder to come by). We include more information about recent systems. We tended to include information about papers where the parameter counts were readily available, in particular larger models that were developed to test the limits of how large a model can be. We are biased towards papers published in English. We mostly focused on systems on vision, language and gaming tasks, while we have comparatively fewer papers on e.g. speech recognition, recommender systems or self driving. Lastly, we are biased towards systems we personally found interesting or impressive.
- Recollecting the information was a time consuming exercise that required us to read through hundreds of technical papers to gather the parameter counts. It is quite likely we have made some mistakes.

## Insights

- Unsurprisingly, there is an upward trend in model size. The trend seems exponential, and seems to have picked up its pace recently for language models. An eyeball estimate of the slope of progress suggests that the doubling rate was between 18 and 24 months from 2000 to 2016-2018 in all domains, and between 3 and 5 months from 2016-2018 onward in the language domain.
- The biggest models in terms of trainable parameters can be found in the language and recommender system domains. The biggest model we found was the 12 trillion parameter Deep Learning Recommender System from Facebook. We don't have enough data on recommender systems to ascertain whether recommender systems have been historically large in terms of trainable parameters.
- Language models have been historically bigger than in other domains. This was because of statistical models whose parameterization scales with vocabulary size (e.g. as in the Hiero Machine Translation System from 2005) and word embeddings that also scale with vocabulary size (e.g. as in Word2Vec from 2013).
- Arguably Deep Learning started to proliferate in computer vision before it reached language processing (both circa 2011-2013), however the parameter counts of the second far surpass those of the first today. In particular, somewhere between 2016-2018 the trend of growth in language model size apparently greatly accelerated its pace, to a doubling time of between 4 and 8 months.
- Architectures on the game domain are small in terms of trainable parameters, below vision architectures while apparently growing at a similar rhythm. Naively we expected otherwise, since playing games seems more complicated. However, in hindsight, what determines model size is what are the returns to scale; in more complex domains we should expect lower effective model sizes, as the models are more constrained in other ways.
- The trend of growth in model size has been relatively stable through the transition into the deep learning era in 2011-2012 in all domains we studied (though it is hard to say with certainty given the amount of data). This suggests that the deep learning revolution was less of a paradigm change and more of a natural continuation of existing tendencies, which finally surpassed other non-machine learning methods.

## Open questions

- Why is there a discrepancy in the trainable parameters magnitude and trend of growth in e.g. vision systems versus e.g. language systems? Some hypotheses are that language architectures scale better with size, that vision models are more bottlenecked on training data, that vision models require more compute per parameter or that the

language processing ML community is ahead in experiment with large scale models (e.g. because they have access to more compute and resources).

- What caused the explosive growth in the size of language models from 2018 onwards? Was it a purely social phenomena as people realized the advantages of larger models, was it enabled by the discovery of architectures that scaled better with size, compute and data (e.g. transformers?) or was it caused by something else entirely?
- Do the scaling laws of Machine Learning for pre-and-post-deep-learning actually differ significantly? So far model size seems to suggest otherwise, what about other metrics?
- How can we more accurately estimate the rates of growth for each domain and period? For how long will current rates of growth be sustained?

## Next steps

- We are interested in collaborating with other researchers to grow this dataset to be more representative and correcting any mistakes. As an incentive, we will pay \$5 per mistake found or system addition (up to \$600 total among all submissions; please contact us if you want to contribute with a donation to increase the payment cap). You can send your submissions to jaimesevillamolina at gmail dot com, preferably in spreadsheet format.
- We are interested in including other information about the systems, most notably compute and training dataset size.
- We want to include more information on other domains, specially on recommender systems.
- We want to look harder for systematic reviews and other already curated datasets of AI systems.

## Acknowledgements

*This article was written by Jaime Sevilla, Pablo Villalobos and Juan Felipe Cerón. Jaime's work is supported by a Marie Curie grant of the NL4XAI Horizon 2020 program.*

*We thank Girish Sastry for advising us on the beginning of the project, the Spanish Effective Altruism community for creating a space to incubate projects such as this one, and Haydn Belfield, Pablo Moreno and Ehud Reiter for discussion and system submissions.*

## Bibliography

1. Kaplan et al., "Scaling Laws for Neural Language Models," 08361.
2. 1.6 History of Reinforcement Learning. (n.d.). Retrieved June 19, 2021, from <http://incompleteideas.net/book/first/ebook/node12.html>
3. AI and Compute. (n.d.). Retrieved June 19, 2021, from <https://openai.com/blog/ai-and-compute/>
4. AI and Efficiency. (2020, May 5). OpenAI. <https://openai.com/blog/ai-and-efficiency/>
5. AI Progress Measurement. (2017, June 12). Electronic Frontier Foundation. <https://www.eff.org/ai/metrics>
6. Announcement of the 2020 ACL Test-of-Time Awards (ToT) | ACL Member Portal. (n.d.). Retrieved June 19, 2021, from <https://www.aclweb.org/portal/content/announcement-2020-acl-test-time-awards-tot#:~:text=Each%20year%2C%20the%20ACL%20Test,papers%20from%202010%20years%20earlier.&text=The%20winners%20were%20announced%20at%20ACL%202020.>
7. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?  Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. <https://doi.org/10.1145/3442188.3445922>

8. *Best paper awards—ACL Wiki*. (n.d.). Retrieved June 19, 2021, from [https://aclweb.org/aclwiki/Best\\_paper\\_awards](https://aclweb.org/aclwiki/Best_paper_awards)
9. *bnlearn—Bayesian Network Repository*. (n.d.). Retrieved June 19, 2021, from <https://www.bnlearn.com/bnrepository/>
10. *Brian Christian on the alignment problem*. (n.d.). 80,000 Hours. Retrieved June 19, 2021, from <https://80000hours.org/podcast/episodes/brian-christian-the-alignment-problem/>
11. *Computer Vision Awards - The Computer Vision Foundation*. (n.d.). Retrieved June 19, 2021, from [https://www.thecvf.com/?page\\_id=413](https://www.thecvf.com/?page_id=413)
12. DARPA Grand Challenge. (2021). In Wikipedia. [https://en.wikipedia.org/w/index.php?title=DARPA\\_Grand\\_Challenge&oldid=1021627196](https://en.wikipedia.org/w/index.php?title=DARPA_Grand_Challenge&oldid=1021627196)
13. Karim, R. (2020, November 28). *Illustrated: 10 CNN Architectures*. Medium. <https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d>
14. Mohammad, S. M. (2020). Examining Citations of Natural Language Processing Literature. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5199–5209. <https://doi.org/10.18653/v1/2020.acl-main.464>
15. Mudigere, D., Hao, Y., Huang, J., Tulloch, A., Sridharan, S., Liu, X., Ozdal, M., Nie, J., Park, J., Luo, L., Yang, J. A., Gao, L., Ivchenko, D., Basant, A., Hu, Y., Yang, J., Ardestani, E. K., Wang, X., Komuravelli, R., ... Rao, V. (2021). High-performance, Distributed Training of Large-scale Deep Learning Recommendation Models. *ArXiv:2104.05158 [Cs]*. <http://arxiv.org/abs/2104.05158>
16. Nilsson, N. (1974). Artificial Intelligence. *IFIP Congress*. <https://doi.org/10.7551/mitpress/11723.003.0006>
17. Posey, L. (2020, April 28). *History of AI Research*. Medium. <https://towardsdatascience.com/history-of-ai-research-90a6cc8adc9c>
18. Raschka, S. (2019). A Brief Summary of the History of Neural Networks and Deep Learning. *Deep Learning*, 29.
19. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv:1910.01108 [Cs]*. <http://arxiv.org/abs/1910.01108>
20. Thompson, N. C., Greenwald, K., Lee, K., & Manso, G. F. (2020). The Computational Limits of Deep Learning. *ArXiv:2007.05558 [Cs, Stat]*. <http://arxiv.org/abs/2007.05558>
21. Vidal, R. (n.d.). *Computer Vision: History, the Rise of Deep Networks, and Future Vistas*. 60.
22. Wang, B. (2021). *Kingoflolz/mesh-transformer-jax* [Jupyter Notebook]. <https://github.com/kingoflolz/mesh-transformer-jax> (Original work published 2021)
23. Who Invented Backpropagation? (n.d.). Retrieved June 19, 2021, from <https://people.idsia.ch/~juergen/who-invented-backpropagation.html>
24. Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with Noisy Student improves ImageNet classification. *ArXiv:1911.04252 [Cs, Stat]*. <http://arxiv.org/abs/1911.04252>
25. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *ArXiv:1708.02709 [Cs]*. <http://arxiv.org/abs/1708.02709>
26. Zhang, B., Xiong, D., Su, J., Lin, Q., & Zhang, H. (2018). Simplifying Neural Machine Translation with Addition-Subtraction Twin-Gated Recurrent Networks. *ArXiv:1810.12546 [Cs]*. <http://arxiv.org/abs/1810.12546>
27. Zoph, B., & Le, Q. V. (2016). *Neural Architecture Search with Reinforcement Learning*. <https://arxiv.org/abs/1611.01578v2>

# Experiments with a random clock

This is a linkpost for <https://www.telescopic-turnip.net/experiments/a-random-clock/>

I may have found a solution to one of my biggest, longest-standing, most irredeemable problems. For most of my life, I have been consistently late. Whether it's appointments, attending events, taking trains or joining a zoom call, I'm typically 10 minutes late for everything and it's ruining my life – not because I actually miss the train (though that happens too) but because I'm constantly rushing and panicking. Whatever I do, I start it in a state of maximum stress and guilt. Obviously, I tried pretty much everything to address the problem, including various artificial rewards and punishments, telling a therapist about it, having people call me to remind me to get ready, taking nootropics, and many more ridiculous ideas. So I thought, "how do all these well-adjusted adults manage to be perfectly on time all the time?" and I did what any well-adjusted normie would do: I tried to formally frame the problem in terms of *expected utility theory*.

## Tricking myself: single-player game theory

Imagine I have to attend a very important scientific conference on [the effect of dubstep on mosquitos](#). The figure below plots how much I enjoy the event depending on the time I arrive.

Arriving early by ten minutes or one hour does not make any difference (or so I presume – this never happened to me). Being just a few minutes late is not a big deal either, since it's just going to be the speaker testing her microphone or other formalities of no importance. Beyond that, it starts becoming really rude (with some variation depending on which culture you live in) and I risk missing some crucial information, like the definition of a crucial concept central to understanding the equations of mosquitos' taste for Skrillex.

The second aspect of the problem is how much time I can save by arriving later, which is just a straight line:

Why would I arrive ten minutes early to the Skrillex-as-a-cure-for-dengue talk, when I could spend ten more minutes reading about [exorcism under fMRI](#)? Summing both aspects of the problem, the grand unified utility curve looks something like this:

There you have it: the utility peak, the most rational outcome, is obtained by being just a few minutes late. I suppose for most people, this basically means you should arrive on time, since the peak is not that far from the start of event. But chronically-late people like myself have a distorted vision of the utility curves, that looks more like that:

This might look like a desperate situation, but there is one spark of hope: even in this wildly-distorted version of the utility function, the downward part of the curve (problems with being late) is much steeper than the upward part of the curve (time

saved by being late). This asymmetry makes it possible to change the location of the peak by adding some *uncertainty*, in the form of a random clock. Let me explain.

A rookie approach to not-being-late is to shift your watch 10 minutes in the future. This way, it “looks” like you’re already 10 minutes late when you are actually on time, which might make you speed up through some obscure psychological mechanism. Of course, this does not work since you know perfectly well your clock is 10 minutes early and you compensate accordingly. But what if you ask a friend to shift your watch by a random number of minutes, between 0 and 10? Then, you don’t know how much to compensate. Coming back to the utility function above, we are effectively blurring out the utility function. Here is what happens:

Thanks to the asymmetry of the original peak, the maximum utility is now shifted to the left! Say the mosquito conference starts at 8:00, and the random clock says 7:59. Best case scenario, the clock is 10 minutes in advance, and I still have 11 minutes left, so everything is fine and I can take my time. Worst case scenario, the clock is exactly on time, and the show starts in one minute, and I can’t wait any longer. Since I would rather be 10 minutes early than 10 minutes late, I stop reading this very important exorcism paper, and hurry to the conference room.

## Self-blinding in practice

In the early development phase I asked a trusted friend to pick a number between 0 and 10 and shift my watch by this amount in the future without telling me. This was for prototyping only, since it has some disadvantages:

- I don’t want to ask friends to change my watch all the time, especially if I have to explain the reasoning behind it every time,
- My friend could totally troll me in various ways, like shifting my clock two hours in the future. I’m clueless enough not to notice. But she is an amazing person and did not do that.

Then, I used this very simple python command:

```
#!/usr/bin/python3
import time, random
print(time.ctime(time.time() + 60*10*random.random()))
```

It takes the current time, draws a random number between 0 and 10, and adds the same number of minutes to the time.

I have an advantage for this project: I usually wear a wristwatch at all times. This makes the practical implementation of the random clock much easier – I just need to shift my wristwatch, and rely exclusively on it without ever looking at any other clock. I also have an alarm clock and a regular clock on the wall of my room, so I simply shifted them to match my watch. I also had clocks on my computer and my phone, and there is surely a way to shift them too, but I was lazy and just disabled the time display on both devices. [Side-note: *in hindsight, I think removing the clock from computers/smartphone is also a healthy decision in its own right, as it forces you to get your eyes off the screen from time to time, you should give it a try*]. Here is my full randomization procedure:

- Shuffle my watch and alarm clock by a large amount, so I can't read the time when I randomize them,
- Wait until I can no longer tell what time it is (to a 10 minutes margin of error),
- Run the script,
- Set my watch and clocks to the time prescribed by the script.

And then, it is all about avoiding looking at the various clocks in my environment that display the true time (sometimes the microwave will just proudly display the time without warning). Who will win – my attempt at deliberately adding uncertainty to the world, or my microwave? Let's do the experiment.

## **Putting a number on it**

For a few days before and after trying out the random clock, I kept track of the time when I arrived to various appointments and events. For the random phase, I would just write down the raw time displayed on my watch, then, before re-randomizing it, I would check what the shift was and subtract it to the data to know at what time I really arrived. My astonishing performance can be witnessed in the figure below:

The horizontal segments represent the median. As you can see, I went from a median lateness of nine minutes to only one minute. I'm still not perfectly calibrated, but this might be the first time in my whole life I am *so close* to being on time, so I'd consider this a success. In both series, there are a few outliers where I was very very late (up to 35 min), but those are due to larger problems – for example, the green outlier was when my bicycle broke and I had to go to a band rehearsal on foot. Apparently, I am so bad at managing time that my lateness undergoes *black swan events*.

Contrary to what I expected, it is very easy to just stop looking at all the clocks in the outside world, and only rely on my watch. Of course, the world is full of danger and sometimes I caught a glimpse at whatever wild clock someone carelessly put in my way. In that case, I just had to avoid checking my watch for a few minutes to avoid breaking the randomization. A bigger problem is seeing when events actually start. Whether I like it or not, my system 1 can't help but infer things about the real time by seeing when other people arrive, or when the conference actually starts, or when some !#\$@ says "alright, it's 10:03, should we start?". If this narrows the distribution too much, I have to randomize again. I did not find it to be a major problem, only having to re-randomize about once a week. In fact, when I revealed the real shift to myself before re-randomizing, I often found that what I inferred about the true time was completely wrong. Thus, even if I *believe* I've inferred the real time from external clues, I can tell to myself it's probably not even accurate. This only makes my scheme stronger.

## **A continuously-randomizing clock**

Since no randomization is eternal, am I doomed to re-randomize every few weeks all my life? There is actually a pretty simple solution to avoid this, which is to use a continuously-randomizing clock. Instead of manually randomizing it from time to time, the clock is constantly drifting back and forth between +0 min and +10 min, slightly tweaking the length of a second. A very simple way to do that is to add a sine function to the real time:

```

#!/usr/bin/python3
import time, math
real_time = time.time()
shift = (1+math.sin(real_time/1800))/2 # Between 0 and 1
wrong_time = real_time + shift*60*10
print(time.ctime(wrong_time))

```

In this example, the clock shift will oscillate between 0 and 10 once every  $\pi$  hours. Of course it is not really random anymore, but it does not matter since we are just trying to trick our system 1 so it cannot figure out the real time against our will. Finding the real time might be possible with some calculations, but those would involve your system 2, and that one is supposed to be under your control. All that matters is that the oscillation period is not an obviously multiple of one hour. The snippet above uses a period of  $\pi$ , which is not even rational, so we are pretty safe.

The advantage of using a sine function rather than a fancy random variable is that it is magically synchronized across all clocks that use the same formula. If you use this on two different computers, they will both give the same (wrong) time, without the intervention of any internet. As I said, I am fine with my old needle watch, but if you are the kind of person who uses a smartwatch, give it a try and tell me how it went. Or perhaps I will try to build one of these [Arduino watches](#).

In my tests, I found that my archaic wristwatch-based system is already good enough for my own usage, so I will stick to this for the moment. Maybe it will keep on working, maybe the effect will fade out after a while, once the novelty wears out. Most likely, I might have been more careful than usual because I really wanted the experiment to succeed. Maybe I will get super good at picking up every clue to guess the real time. I will update this post with the latest developments. Anyways, there is something paradoxical about manipulating oneself by deliberately adding uncertainty - a perfectly rational agent would always want more accurate information about the world, and would never deliberately introduce randomness. But I am not a perfectly rational agent, I did introduce uncertainty, and it worked.

# The Moon is Down; I have not heard the clock

I wanted to share a quick post about something that's made me significantly happier over the past year: knowing what the phase of the moon is on any given day.

Importantly, I don't do this with any kind of computer tool. It's a proxy for "am I spending enough time outside", but because I don't let myself cheat and rely on actually seeing the moon at least every few days, I've succeeded in not Goodharting myself, at least so far.

One of the things that helped me do this successfully was figuring out what time of day I could expect to see the moon during the different phases. I know, I know, it's a trivial exercise in orbital mechanics, so maybe all of you do this instinctively, but it wasn't something anyone ever explained to me explicitly. It actually took two disparate works of fiction to make the connection for me.

The first literary clue comes from my favorite Shakespeare play, *Macbeth*. Scene II, Act I starts with Banquo and Fleance in the court of Macbeth's castle:

Banquo: How goes the night, boy?

Fleance: The moon is down; I have not heard the clock.

Banquo: And she goes down at twelve.

Fleance: I take't, 'tis later, sir

The second clue (here's where the penny really dropped) is from Cormac McCarthy's *All the Pretty Horses*, where the main character says "First quarter moon sets at midnight where I come from."

Of course, the true first-quarter moon sets at midnight *regardless of where you are* because at (solar) midnight, the place you're standing is facing as far from the sun as it can be, and the first quarter moon means that the sun-moon-earth angle is a 90-degree angle. (This is also true at the third-quarter moon, but at that phase, the moon *comes up* at midnight).

To be more explicit, at the new moon, the moon is almost between the sun and the earth (if it's *exactly* between, you get a solar eclipse!) so the moon rises and sets near when the sun does, but it's tough to see. A waxing crescent moon comes up gradually later in the morning but is easiest to see shortly after sunset, just as it's going down.

The first-quarter moon comes up at noon and, as we've already covered, goes down at midnight. Waxing gibbous moons come up in the afternoon and set in the early morning, before sunrise. At the full moon, the earth is almost between the sun and the moon so the moon comes up around sunset and sets near sunrise. The waning gibbous moon comes up gradually later after sunset and sets gradually later in the morning. The third quarter moon rises at midnight and sets at noon. Finally, the waning crescent moon comes up gradually later after midnight and sets in the afternoon.

All of this is just to say: get outside, preferably with people you care about! I look forward to being unwilling to shut up to the whippersnappers about how when I was

young, the moon didn't have all those dots of light from the settlements.

# Announcing the Replacing Guilt audiobook

*tldr; Complete narration of the Replacing Guilt series is available here: [anchor.fm/guilt](https://anchor.fm/guilt)*

I discovered Nate Soares' [Replacing Guilt series](#) in late 2019 via the [Bayesian Conspiracy podcast](#) and found it immensely valuable. Over ~40 posts, Nate channels the guilt-based motivation that is common in rationalists / EAs into more productive emotional drives.

I read the series on my Kindle, thanks to [lifelonglearner's ePUB version](#), but was disappointed that no audio version existed. To make it more accessible, I reached out to Nate and got permission to produce the official audiobook.

The rationalist community has a fantastic tradition of volunteer narration — Slate Star Codex , AI to Zombies, and, of course, HPMOR. These free resources have added immense value to my life, so narrating Replacing Guilt feels like my tiny way of reinvesting in our rationalist commons.

For a 2-minute summary of the series and why you might be interested in it, check out [this post](#) (or listen to the [audio version](#)).

You can find the individual episodes at [anchor.fm/guilt](https://anchor.fm/guilt). The complete audiobook can be [streamed](#) in any podcast app, or you can download the [mp3 file here](#).

# How do the ivermectin meta-reviews come to so different conclusions?

[Ivermectin and outcomes from Covid-19 pneumonia: A systematic review and meta-analysis of randomized clinical trial studies](#) comes to the conclusion: "Our study suggests that ivermectin may offer beneficial effects towards Covid-19 outcomes. More randomized clinical trial studies are still needed to confirm the results of our study."

On the other hand [Ivermectin for the treatment of COVID-19: A systematic review and meta-analysis of randomized controlled trials](#) comes to the conclusion: "In comparison to SOC or placebo, IVM did not reduce all-cause mortality, length of stay or viral clearance in RCTs in COVID-19 patients with mostly mild disease. IVM did not have an effect on AEs or severe AEs. IVM is not a viable option to treat COVID-19 patients."

What did the studies do differently to come to their conclusions? How do I go about interpreting which of them provides the better analysis?

# Notes on War: Grand Strategy

First things first: I am not any sort of military expert. But most discussions or portrayals of war which I see among rationalists make me face-palm, especially discussions around the role of AI in future warfare or AI “takeover”. Seriously, guys, it’s not about autonomous drone swarms.

This post is just a brain-dump of some models/frames which I wish more people had. We’ll walk through a few different “use-cases” of war (defending an invasion, territory grab, takeover) and talk about grand-strategy principles relevant to each. If people find this interesting, I may write more brain-dumps like this on strategy, tactics, etc; the post quickly became far too long to fit it all in one.

Epistemic status: any particular claim is low-confidence, but I am more confident in the usefulness of the high-level frames.

## Defending an Invasion

For concreteness, let’s think about [Azerbaijan invading the Nagorno-Karabakh region of Armenia](#) (though admittedly I did not follow that conflict closely so we’re not going to talk about real details too much). At a high level, how might Armenia go about defending against that invasion?

The obvious option is “kill invading troops until they leave, and kill them some more if they come back”. I claim this is wildly inefficient: it’s treating symptoms, not causes.

On the Azerbaijani side, the decision to invade was presumably a *political* decision. There were factions in favor, and factions against. If the anti-invasion factions gain control of government decision-making, then the invasion will likely end. Killing invading troops *might* help make that happen - making the war visibly costly may make it less popular. On the other hand, killing invading troops may just create martyrs and solidify the image of Armenians as Enemies. If that’s the case, then killing invading soldiers may only solidify the political faction pushing for war, and troops will just keep invading until one side runs out of people/resources. It could go either way.

So: how could Armenia fight in a way which *destabilizes* the pro-invasion faction’s control over Azerbaijan, rather than stabilizing it? Some possibilities:

- Directly attempt to kill key people in the pro-invasion faction, e.g. assassinations. Note that this could backfire in the same way as killing soldiers.
- Target the power base of people in the pro-invasion faction. For instance, in many third-world countries the military has a lot of control over decision-making, because they might realistically execute a coup if they don’t get their way. In that case, killing soldiers (and officers) directly reduces the military’s ability to execute a coup, and therefore directly reduces their influence over decision-making.
- Similarly, if the invasion is driven by a regional or religious faction with significant wealth/income, destroy their capital assets.
- If a particular faction/stakeholder has a “swing vote” or functional equivalent, try to focus the costs of the war on that stakeholder - i.e. whenever the Azerbaijani forces take some territory or kill some Armenian soldiers or destroy something,

try to retaliate by destroying something valuable to the “swing vote” stakeholder.

Alternatively, the Armenians could try to directly change the political calculus of the pro-invasion faction:

- Whatever benefits the Azerbaijanis expect from invasion, reduce them. Organize resistance groups and weapons caches before an invasion happens. Rig explosives to destroy capital assets (i.e. bridges, roads, rails, power grid) remotely in case an area is lost. Make sure the Azerbaijanis know about all this, to dissuade invading in the first place.
- Increase the costs of invasion. Build [second-strike capabilities](#) - weapons that can be used to retaliate after an attack. Use proportional retaliation. Again, make sure the Azerbaijanis know about all this beforehand.
- Maybe it's a populist war - i.e. the Azerbaijanis just really hate Armenians and will back any politician promising to attack them. In the long run, increasing the popularity of Armenians among the Azerbaijani populace is probably a good strategy, so “take the high road”: avoid civilian casualties, treat prisoners well, send aid (food, medical supplies) to the enemy populace, etc. And more importantly, make sure to *loudly broadcast* all those actions.
- In the short run, to deal with an Azerbaijani populace which really hates Armenians, try to shift the populace’ attention to domestic affairs. A recession could do this, so long as it’s obvious that defeating Armenia will not make the recession go away. Same with a famine, though that’s harder to induce while avoiding long-term blame. Maybe target the sort of economically-crucial companies that people love to hate, like banks.
- Also implicit in all of the above: propaganda, obviously.
- Taking a different tack: if the actual priority of Azerbaijani leadership is to generate popularity with a populace which hates Armenians, it may be possible to give them a *symbolic* victory without actually losing much. “Win-win”, in some sense.

One important thing to keep in mind: Armenia is no more a single unified agent than Azerbaijan; political considerations will inform their actual choices in much the same way. If the Armenian populace really hates Azerbaijanis, then the politically-popular choice will likely be to kill lots of soldiers and probably also lots of Azerbaijani civilians, destroy their stuff, etc, regardless of whether that is actually a smart strategy for winning the war.

Another important thing to keep in mind: war is not a zero-sum game. A war of attrition leaves both sides worse off, so there’s positive-sum gains in *avoiding* that outcome. That means that sharing information is sometimes a good move: if destroying our stuff will be costly to the enemy as well, then we want the enemy to *know* that. Also, from a politician’s perspective, a war can solidify the political position of leadership on both sides - a positive-sum outcome, in some sense, though probably not good for the citizenry or soldiers on either side.

## Territory Grab

Invasions have a lot more variety, depending on the objective. First, we’ll consider the Azerbaijani side of the previous example. Their goal, presumably, is to take political control of the Nagorno-Karabakh region. (Of course in practice this may not be the “real” goal - e.g. if it’s a populist war, the real function may just be to generate

political support among the population by symbolically Punching Bad People In The Face.)

Broadly speaking, if the Azerbaijanis plan to absorb the new territory, then they generally want to not break it. [Don't blow holes in a ship you plan to steal](#). That means:

- Avoid civilian casualties
- Minimize damage to capital assets (i.e. infrastructure)
- Don't piss off the local population: no pillaging, provide food and medical supplies, be very strict about discipline with Azerbaijani troops, etc.
- Broadcast how nicely Azerbaijan treats the locals. As usual, propaganda.

There are exceptions to all of these, depending on the objective and resources. If the plan is to wipe out the local population and resettle, obviously that changes things. If Azerbaijan has a lot more capital investment capacity than Armenia, then it could leverage that advantage by destroying local capital assets, which will be easier for Azerbaijan to replace than for Armenia to replace. If the plan is an iron-fisted occupation, then pissing off the locals once or twice may be useful in order to make a few public examples. But the general underlying principle is: don't break what you intend to steal.

Beyond that, building governing infrastructure - courts, tax collection capabilities, regulatory enforcement, etc - takes a lot of work. An important strategic question for Azerbaijan is whether to take over the existing infrastructure or replace it. If the latter, then attacking existing government institutions may weaken Armenia during the war - but at the cost of Azerbaijan needing to rebuild from scratch once the war ends. More discussion of that trade-off in the next section.

In terms of *combat* priorities, Azerbaijan's problem largely mirrors Armenia's problem. Armenia's decision to fight or negotiate a settlement is a political decision, and Azerbaijan can influence that decision in basically the same ways as the previous section, so long as they avoid damage to whatever they intend to take.

## Full Takeover

Now, we'll think about the US invading Japan in WWII. (Or communists taking over China in the 1949 revolution; similar principles apply, and indeed I hear that Mao's writing on the topic is quite similar to the discussion here.) The objective is to overthrow and replace the existing government.

This runs into the same tradeoff as the previous section, but to a much greater extent: whatever destruction the invader causes will be the invader's problem to clean up, assuming they win. Destroying capital assets or undermining government institutions may make it much easier to win the war, but it will make the aftermath much more difficult.

First question: is it a strategic priority to preserve the existing enemy government? Wiping out leadership can make it a lot easier to invade, but it also leaves nobody who can surrender on behalf of the whole country. Probably lots of decentralized resistance. Also, if the plan is to force a surrender, then enemy leadership needs to be able to enforce the terms of surrender on their own people. Preserving the enemy government's ability to do that is a strategic necessity, in that case.

Second question: is it a strategic priority to maintain the *competence* of the existing enemy government? Mao notes that incompetent bureaucrats or officers in the enemy hierarchy are an asset for the rebels, and guerilla fighters should avoid targeting them. Conversely, competent bureaucrats or officers should be targeted. (Besides the obvious benefit, this has a bonus effect: if the enemy leadership *knows* about such a policy, then they will trust their middle management less, which will further slow organizational information-passing and decision-making.) However, this is a short-term-oriented view; those same competent bureaucrats could become assets for the new administration after the war. Training and installing new people, en masse, is difficult and expensive.

In the case of a guerilla rebellion, the same question applies not only to competence but to hostility/abuse. Bureaucrats or officers who abuse their power will generally drive support for the rebellion, whereas bureaucrats or officers who are liked by the populace will generally drive support for the regime. On the other hand, once the war is over, regardless of who won, getting rid of abusive bureaucrats/officers is part of the reason for fighting the war in the first place.

Third question: which enemy institutions are to be kept? This is where the political considerations from earlier come in: within each institution, there will be a decision to cooperate with the takeover or not, there will be factions and key decision-makers, etc. If the top-level government structure is to be maintained, then the same considerations from earlier sections apply. But if only lower-level structures are to be maintained - e.g. courts, military, school system, road maintenance, government hospitals, etc - then high-level decision makers at each of those individual institutions must be persuaded, coerced or replaced. Furthermore, the ability of those high-level decision makers to enforce their decisions on the rest of the institution must be maintained.

## A Continuum Between Politics and Warfare

When we open the black box of “enemy government” and start manipulating internal gears of decision-making - e.g. political factions, particular institutions - it becomes clear that war lives on a spectrum. There’s a lot of ways to control the decision-making process of a government.

At one end, there’s lobbying and advertising/propaganda, which are often completely legal and legitimate methods of influencing government decisions. Further down the spectrum are bribes, and then assassinations. Still further along is outright guerilla warfare - which is often really just a mix of propaganda, assassination and targetted destruction of capital assets. Finally, there’s outright invasion and occupation.

I don’t think there’s a clear dividing line between illegal manipulation of government decisions (bribes, assassination) vs outright guerilla warfare. To a large extent, it should be possible in principle to achieve the same sort of goals - i.e. government takeover - with relatively little, highly targeted illegal activity. This, however, would require very precise information and models. For instance, if one knew exactly where a particular senator’s funding came from, one could in-principle physically destroy the capital assets which provide that funding (assuming insurance did not cover the loss - ideally one would want access to the details of the insurance contract in order to fake a form of destruction not covered). Or, if one knew exactly which aid wrote the text of a thousand-page bill which nobody would ever read, a bribe or threat could create a subtle loophole or a structural change which made the actual effect quite different

from the symbolic meaning of the bill. Or, one could directly target the bureaucrats in charge of implementing a particular law (this is already one of the main functions of lobbying, as I understand it).

This sort of strategy relies mainly on very precise information and models; it's exactly the sort of area where I'd expect AI tools to convey a massive advantage.

## Takeaways

The enemy is not a single unified agent or a black box. The internal gears of decision-making in an enemy organization can be manipulated.

War is not zero sum. Attrition is (usually) costly to both parties, therefore avoiding attrition is positive-sum. It is sometimes useful to share information with the enemy so that they know which actions will result in mutual attrition (e.g. credible threats).

Destroying the enemy's organizational capacity - e.g. killing leadership, removing competent people, making leadership unable to enforce their decisions on others - can impair their ability to wage war. However, this also impairs their ability to enforce terms of surrender on their own people. Also, if the goal is to take over the enemy government, then any organizational capacity destroyed will have to be rebuilt after the war, which is expensive and difficult; don't destroy things you intend to take.

# Survey on AI existential risk scenarios

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

*Cross-posted to the [EA forum](#).*

## Summary

- In August 2020, we conducted an online survey of prominent AI safety and governance researchers. You can see a copy of the survey at [this link](#).<sup>[1]</sup>
- We sent the survey to 135 researchers at leading AI safety/governance research organisations (including [AI Impacts](#), [CHAI](#), [CLR](#), [CSER](#), [CSET](#), [FHI](#), [FLI](#), [GCRI](#), [MILA](#), [MIRI](#), [Open Philanthropy](#) and [PAI](#)) and a number of independent researchers. We received 75 responses, a response rate of 56%.
- The survey aimed to identify which AI existential risk scenarios<sup>[2]</sup> (which we will refer to simply as “risk scenarios”) those researchers find most likely, in order to (1) help with prioritising future work on exploring AI risk scenarios, and (2) facilitate discourse and understanding within the AI safety and governance community, including between researchers who have different views.
- In our view, the key result is that there was considerable disagreement among researchers about which risk scenarios are the most likely, and high uncertainty expressed by most individual researchers about their estimates.
- This suggests that there is a lot of value in exploring the likelihood of different AI risk scenarios in more detail, especially given the limited scrutiny that most scenarios have received. This could look like:
  - Fleshying out and analysing the scenarios mentioned in this post which have received less scrutiny.
  - Doing more horizon scanning or trying to come up with other risk scenarios, and analysing them.
- At this time, we are only publishing this abbreviated version of the results. We have a version of the full results that we may publish at a later date. Please contact [one of us](#) if you would like access to this, and include a sentence on why the results would be helpful or what you intend to use them for.
- We welcome feedback on any aspects of the survey.

## Motivation

It has been argued that AI could pose an existential risk. The original risk scenarios were described by [Nick Bostrom](#) and [Eliezer Yudkowsky](#). More recently, these [have been criticised](#), and [a number of alternative scenarios have been proposed](#). There has been some useful work exploring these alternative scenarios, but much of this is informal. Most pieces are only presented as blog posts, with neither the detail of a book, nor the rigour of a peer-reviewed publication. For further discussion of this dynamic, see work by [Ben Garfinkel](#), [Richard Ngo](#) and [Tom Adamczewski](#).

The result is that it is no longer clear which AI risk scenarios experts find most plausible. We think this state of affairs is unsatisfactory for at least two reasons. First, since many of the proposed scenarios seem underdeveloped, there is room for further

work analyzing them in more detail. But this is time-consuming and there are a wide range of scenarios that *could* be analysed, so knowing which scenarios leading experts find most plausible is useful for prioritising this work. Second, since the views of top researchers will influence the views of the broader AI safety and governance community, it is important to make the full spectrum of views more widely available. The survey is intended to be a first step in this direction.

## The survey

We asked researchers to estimate the probability of five AI risk scenarios, *conditional on an existential catastrophe due to AI having occurred*. There was also a catch-all “other scenarios” option.

These were the five scenarios we asked about, and the descriptions we gave in the survey:

- **"Superintelligence"**
  - A single AI system with goals that are hostile to humanity quickly becomes sufficiently capable for complete world domination, and causes the future to contain very little of what we value, as described in "[Superintelligence](#)".  
[3]
- **Part 2 of “What failure looks like”**
  - This involves multiple AIs accidentally being trained to seek influence, and then failing catastrophically once they are sufficiently capable, causing humans to become extinct or otherwise permanently lose all influence over the future.
- **Part 1 of “What failure looks like”**
  - This involves AIs pursuing easy-to-measure goals, rather than the goals humans actually care about, causing us to permanently lose some influence over the future (excluding cases where the “Superintelligence” scenario or Part 2 of “What failure looks like” also occur).
- **War**
  - Some kind of war between humans, exacerbated by developments in AI, causes an existential catastrophe. AI is a significant risk factor in the catastrophe, such that no catastrophe would have occurred without the developments in AI. The proximate cause of the catastrophe is the deliberate actions of humans, such as the use of AI-enabled, nuclear or other weapons. See Dafoe ([2018](#)) for more detail.
- **Misuse**
  - Intentional misuse of AI by one or more actors causes an existential catastrophe (excluding cases where the catastrophe was caused by misuse in a war that would not have occurred without developments in AI). See Karnofsky ([2016](#)) for more detail.

We chose these five scenarios because they have been most prominent in [previous discussions](#) about different AI risk scenarios. For more details about the survey, you can find a copy of it at [this link](#).

## Key results

## **There was considerable disagreement among researchers about which risk scenarios are most likely**

- If you take the median response for each scenario and compare them, those (conditional) probabilities are fairly similar (between 10% and 12.5% for the five given scenarios, and 20% for “other scenarios”).<sup>[4]</sup> However, *individual responses* vary greatly (from the median). For instance, most respondents thought at least one scenario was quite unlikely:
  - 96% of respondents assigned  $\leq 10\%$  (conditional) probability to at least one scenario.
  - 89% of respondents assigned  $\leq 10\%$  (conditional) probability to at least two scenarios.
  - 64% of respondents assigned  $\leq 10\%$  (conditional) probability to at least three scenarios.<sup>[5]</sup>
- There were a number of outliers: for each scenario, at least one respondent estimated them to have  $\geq 70\%$  (conditional) probability.
- For each scenario (including “other scenarios”), the mean absolute deviation of responses was somewhere between 9% and 18%.
  - E.g. for the “Superintelligence” scenario, the mean absolute deviation was 13%. This means that the average (absolute) distance from the mean estimate was 13 percentage points.
  - To help interpret this: recall that the means are all between 15% and 25% (see footnote 4) - so the mean absolute deviations are relatively large compared to (conditional) probabilities themselves.<sup>[6]</sup>
- For each scenario (including “other scenarios”), the interquartile range of responses was somewhere between 15% and 31%.
  - E.g. for the “Superintelligence” scenario, the first quartile response was 5% and the third quartile response was 20% (so the interquartile range was 15%).
- These statistics suggest considerable disagreement among researchers about which risk scenarios are the most likely.

## **Researchers are uncertain about which risk scenarios are most likely**

The median self-reported confidence level given by respondents was 2, on a seven point Likert scale from 0 to 6, where:

- Confidence level 0 was labelled “completely uncertain, I selected my answers randomly”, and
- Confidence level 6 was labelled “completely certain, like probability estimates for a fair dice”.

# Researchers put substantial credence on “other scenarios”

The “other scenarios” option had the highest median probability, at 20%. Some researchers left free-form comments describing these other scenarios. Most of them have seen no public write-up, and the others have been explored in less detail than the five scenarios we asked about.

## Key takeaway

Together, these three results suggest that there is a lot of value in exploring the likelihood of different risk scenarios in more detail. This could look like:

- Fleshing out and analysing the scenarios mentioned in this post, in more detail.
  - This seems especially important given that the median and mean probability estimates were similar for all the scenarios, and yet the “Superintelligence” scenario has received far more scrutiny than the others.
- Doing more horizon scanning or trying to come up with other risk scenarios, and analysing them.
  - Other than the “Superintelligence” scenario, almost all other risk scenarios (including those mentioned in this post) were only made salient in the last four years or so.

Recent “failure stories” by [Andrew Critch](#) and [Paul Christiano](#) - which seem to have been well-received and appreciated - also suggest that there is value in exploring different risk scenarios in more detail. Likewise, Rohin Shah [advocates](#) for this kind of work, and AI Impacts has recently compiled a [collection of stories](#) to clarify, explore or appreciate possible future AI scenarios.

## Caveats

One important caveat is the tractability of exploring the likelihood of different AI risk scenarios in more detail. The existence of considerable disagreement, despite there having been some attempts to clarify and discuss these issues, could suggest that making progress on this is difficult. However, we think there has been relatively little effort towards this kind of work so far, and that there is still a lot of low-hanging fruit.

Additionally, there were a number of limitations in the survey design, which are summarised in [this document](#). If we were to run the survey again, we would do many things differently. Whilst we think that our main findings stand up to these limitations, we nonetheless advise taking them cautiously, and as just one piece of evidence - among many - about researchers’ views on AI risk.

## Other notable results

Most of this community’s discussion about existential risk from AI focuses on scenarios involving one or more powerful, misaligned AI systems that take control of

the future. This kind of concern is articulated most prominently in “Superintelligence” and “[What failure looks like](#)”, corresponding to three scenarios in our survey (the “Superintelligence” scenario, part 1 and part 2 of “What failure looks like”). The median respondent’s total (conditional) probability on these three scenarios was 50%, suggesting that this kind of concern about AI risk is still prevalent, but far from the only kind of risk that researchers are concerned about today.

69% of respondents reported that they have *lowered* their probability estimate in the “Superintelligence” scenario (as described above)<sup>[7]</sup> since the first year they were involved in AI safety/governance. This may be because they now assign relatively higher probabilities to other risk scenarios happening first, and not necessarily because they think that fast takeoff or other premises of the “Superintelligence” scenario are less plausible than they originally did.

## Full version

At this time, we are only publishing this abbreviated version of the results. We have a version of the full results that we may publish at a later date. Please contact [one of us](#) if you would like access to this, and include a sentence on why the results would be helpful or what you intend to use them for.

## Acknowledgements

*We would like to thank all researchers who participated in the survey. We are also grateful for valuable comments and feedback from JJ Hepburn, Richard Ngo, Ben Garfinkel, Max Daniel, Rohin Shah, Jess Whittlestone, Rafe Kennedy, Spencer Greenberg, Linda Linsefors, David Manheim, Ross Gruetzmacher, Adam Shimi, Markus Anderljung, Chris McDonald, David Kreuger, Paolo Bova, Vael Gates, Michael Aird, Lewis Hammond, Alex Holness-Tofts, Nicholas Goldowsky-Dill, the GovAI team, the AI:FAR group, and anyone else we ought to have mentioned here. This project grew out of [AISC](#) and [FHI SRF](#). All errors are our own.*

---

1. We will not look at any responses from now on; this is intended just to show what questions were asked, and in case any readers are interested in thinking through their own responses. [←](#)
2. AI existential risk scenarios are sometimes called [threat models](#). [←](#)
3. Bostrom describes many scenarios in the book “Superintelligence”. We think that this scenario is the one that most people remember from the book, but nonetheless, we think it was probably a mistake to refer to this particular scenario by this name. [←](#)
4. Likewise, the mean responses for the five given scenarios are all between 15% and 18%, and the mean response for “other scenarios” was 25%. [←](#)
5. Other similar results: 77% of respondents assigned  $\leq 5\%$  (conditional) probability to at least one scenario; 51% of respondents assigned  $\leq 5\%$  (conditional) probability to at least two scenarios. [←](#)

6. For another way of interpreting this, consider that if respondents were evenly split into six completely “polarised” camps, each of which put 100% probability on one option and 0% on the others, then the mean absolute deviation for each scenario would be ~28%. [←](#)
7. As per footnote 3, the particular scenario we are referring to here is not the only scenario described in “Superintelligence”. [←](#)

# Intro to Debt Crises

In 2018 Ray Dalio wrote a book called [Principles of Navigating Big Debt Crises](#). He argues that we are *at least* near enough to being in a debt crisis that it is important to start thinking about how to most gracefully deal with them. After reading it, I feel similarly. And yet, I didn't know what a debt crisis was before his book, and many people I talk with don't understand why bad debts can not only be hard on their owners but on the whole economy. This is an intro to what a debt crisis is and why specifically it is bad, to be followed up with some posts on their causation and how they relate to the current US situation.

Disclaimer: Dalio is hard to interpret on some central points of causation. I am relaying my best guess as to the truth after looking at some other sources, but I didn't know much about macroeconomics until the last year so this is decidedly not Lindy.

## Basics of a debt crisis

In short, debt crises are when a country ends up with too much bad debt, and then its economy suffers. For example, the recent government-debt crisis in Greece was a debt crisis. This was part of a larger and slower debt crisis often called the European sovereign debt crisis. The crash in 2008 was a banking crisis, a type of debt crisis where the debt is held by banks. The Great Depression was a debt crisis (and became a banking crisis). Hyperinflation in the German Weimar Republic in the 20s was a debt crisis, because they owed the Allies a lot of money in war reparations.

Ray Dalio documents ~100 debt crises in various countries from the last century in the end of his book, and I think all of them led to recession. Not every economic crisis can be thought of in this manner—sometimes there are currency crises, commodities crises, or pandemic crises—but nearly every big recession I know of was caused by bad debt (the main counterexample was probably the dot-com bubble, which was in equity and not debt and barely caused a GDP drop). To give you a sense of scale, Dalio's ballpark for big crises was that 20% of debts got written down 40%, and with a 2x debt-to-GDP ratio this meant that an amount equivalent to 16% of GDP was lost by that country in wealth (before follow-on effects).

People often think of debt issues as stemming from public/government debt (both national and state/municipality). [But this isn't necessarily true](#)—about half the crises listed above were about private rather than public debt, where private includes corporations. For example, if several of a nation's companies can't pay back their loans, they lose a lot of money in forced liquidations, and that reverberates around the economy. Private debt is often higher than public debt and is easier and more general to model, so I'll focus on the economics of private debt crises rather than government debt crises in particular.

I like to think of a debt crisis as consisting of two main phases.

First, there is a proliferation of hidden bad debts (bubble). Second, cash flow bottlenecks domino outward when these debts are seen to be bad (crunch).

## Hidden bad debt bubble

My understanding of the stereotypical arc of events in a debt bubble are roughly as follows:

1. **Unsustainably high real economic growth rates** continue for some time due to lucky economic circumstances.
2. **Expected growth decouples from actual growth** when rates drop back to normal. Debt and equity prices are forecasts, and thus don't have tight feedback loops because their feedback is years in the future. People (somewhat reasonably) assume the downturn is a blip on the otherwise-straight trajectory.
3. **Targets**: to hit the expectations of those around them, borrowers and lenders chase high growth targets and continue transacting assuming medium-term growth rates that would make these actions good deals. Out of this comes a lot of debts that aren't fundamentally sound in producing a large margin more growth than interest. Further, companies aren't using debt as one-time borrowings for which they pay back the principal at maturity: they just hold debt on their balance sheet, refinancing with a new loan whenever it comes due at maturity.
4. At some point, **interest rates rise**, and companies can't keep their cycle of low-rate loans: they have to either pay back the debt or refinance with a loan at higher interest rates than their growth! If they act quickly to pay off the loans they might be ok, but it can be hard if interest rates rise quickly: consider a company with debt equivalent to 50% of yearly revenue, with 5% free cash flow. If it has to pay back a loan, starting from 15% cash buffer would take it 7 years  $((.5-.15)/.05=7)$ . Realistically it will have to liquidate other assets to meet such an obligation.
5. **Debt service payments begin to exceed growth** after refinancing at higher rates, but companies have no way to escape. Executives and analysts notice things are unsustainable, but don't really make it public knowledge: executives try to turn things around, hoping they can weather this blip or get lucky with a new product or vision. The longer this continues, the longer unpayable debts will accrue, and the harder it will be to spread out the damage when the crunch comes. For the businesses destined to fail this is a charade phase; but undoubtedly some businesses do manage to make things sustainable during this period.

Now in normal circumstances, plenty of transparently risky loans go out all the time, and lenders price these in. It never reaches a crisis. So this is a stark example of market inefficiency. The question becomes, as with all bubbles: how does the asset become so mispriced? In this case, how is the badness of the debts hidden?

I unfortunately don't have the cleanest of answers to this. Because it's so complicated, I'm going to relegate it to my [next post](#), despite it being both at the heart of my interest and also surprisingly relevant to the whole cycle.

We'll instead move on to how the liquidity crisis would tend to play out assuming some downturn happens after a period of some prosperity. As such, this applies to "small" debt crises as well.

So back to the plot: the longer the decoupling of the bubble continues, the more loss is piling up that isn't being paid by anyone. Eventually, we see the big reveal and it all comes due at once. The loss itself is bad but a small enough fraction of GDP even in big cases (10-20%) that it wouldn't be catastrophic if spread out over a decade—but it is made much worse by receiving it all at once, which sparks the cash flow crunch.

# Cash flow crunch

This is basically a [liquidity crunch](#), spread out over everyone.

Single expenses of \$50k in medical bills can bankrupt many people, even if they could easily pay it over a decade (cf. student loans). They just can't get the cash fast enough. And while lots of human hardship is sort of neutral in economic standards—e.g. a business getting fairly outcompeted by an alternative is financially net-positive in terms of the total assets of the group—a person going bankrupt isn't neutral even by economic standards. It's not just that all their assets get transferred to other entities, it's that the assets become worth less too. Any synergies they have are gone; good cars get junked because buyers can't tell them from lemons; houses get put on the market and are unlivable for a while; items the buyer has adapted to will go to people who don't know how to use them; etc. They had annealed into comparatively-advantageous assets through learning and selective purchase, and all that advantage is lost in an asset sell-off.

So it goes for companies. Re-allocating the goods, financial assets, and career capital from a company is a long undertaking that destroys much of the value. Specialized career capital is burned, current projects are burned, IT work is burned, etc etc. And credit squeezes don't have to progress all the way to bankruptcy to hurt. Any partial squeeze that requires liquidation will bring partial costs.

At high leverage or thin margin, large costs must be paid even for small adverse market movements or lending conditions. How small?

Here are some Fermis I made while surfing through 10-Ks and Statista and knowing very little about business. They don't lead to any strong conclusions and are hard to make clear, so please skip if confusing.

[Average profit margin is about 8%](#), and companies seem to hold cash reserves ~10% of revenue. So for the average large company, if a crisis lowers margins by 7%, nothing happens. If they're reduced 9%, they start losing money but have 10 years to recover. If they go down 11%, now they will go bankrupt in 3 years unless they raise new cash. And stochasticity matters a lot, as [The Goal](#) will tell you. So companies can't just ride it out perfectly, knowing that if their margin picks up in 2 years and 364 days they'll be fine. They have to start deleveraging with productivity-hampering sell-offs significantly before the final bankruptcy date. (If they can—if not, bankruptcy passes their losses on to others: shareholders, lenders, companies that sold them products they haven't paid for yet.)

So, in this Fermi, the average business has to initiate sell-offs once margin drops about 10%. However, revenue only drops on average about 5% during recessions (because that's how much GDP drops). And if revenue drops only 5%, profit margin probably drops only 1-2% (because of that lost revenue maybe 60-80% was going to cost of goods sold (COGS), and only 20-40% was going to profit).

This would imply the average business only loses 1-2% margin, and most don't even have to initiate sell-offs. I don't think this is exactly true. Perhaps COGS is lower for many businesses, especially in certain classes like restaurants or the service industry; or costs are sticky and so more lost revenue comes directly from profit; or something else in my Fermi is wrong. Regardless, the companies most hurt would seem to be those with low COGS, low initial margins, sellers of

discretionary goods and others preferentially hurt by a smaller GDP, and businesses relying on continued financing (new small businesses, startups, growth companies, etc).

But no matter which factors hurt the most, this is where profitability crunch and cash flow crunch blend together. It starts as a cash flow crunch in company X who can't raise money; they then defer payments to company Y, who has a cash flow crunch that hurts profitability; then company Y reduces its net outflows, decreasing revenue and causing a profitability crunch in company Z. This cascades around the economy. If everyone has to cut back at the same time, there just can't be good allocation of all the "excess" capital that was cut, because other people don't have uses for it yet. So the economy's total output will shrink for a while until free cash is accumulated and new uses can be found for the excess capital.

Fermi: Specifically, you can imagine that a nation starts with GDP of  $X/\text{yr}$  and wealth of  $10X$ . They like to have 10% of wealth as cash-on-hand, so they have  $X$  cash. They lose  $.2X$  of their cash at once from bad debts, taking wealth to  $9.8X$  and cash to  $.8X$ . They need  $.18X$  of cash to be re-stocked, which will take 4 years if their businesses have 5% free cash flow =  $.05X/\text{yr}$ . In the meantime, everyone cuts 5% of their outflow so they can stay solvent or accumulate cash quicker, and GDP drops to  $.95X/\text{yr}$ .

[Caveat: I'm not sure if these numbers are reasonable, as it's hard to figure out how corporate wealth is counted and whether the sum of free cash flows can be directly estimated from GDP, etc.]

One way to think about this liquidity shortage is in terms of monetary velocity and nationwide cash availability, which may be familiar from crypto [utility tokens](#) or from [The Dark Lord's Answer](#) or [NGDP-level targeting](#) within MMT. Everyone finds out at once that a small segment of debts is bad and they won't receive cash back from them, and they need to quickly raise cash some other way. Some people sell assets to raise cash, but demand for money is high—there's not actually enough cash available in M1/M2/M3 that people are willing to part with to be exchanged in the ripple of transactions that needs to occur. Velocity is not high enough, so cash gets bid up in a deflationary manner, which makes demand even higher. (Note that this means that you are providing a valuable service by re-investing in things with your cash when cash has dried up toward the bottom of a crash.) (Interestingly, this means that the faster economies get in sending payments to each other via internet or blockchain or what-have-you, the less deadweight loss you'd theoretically have from liquidity crises.)

Another way to think about it is again from the point of view of personal finance. Consider someone who is always in a bit of debt, late on utilities or credit card payments. Things can continue on like that for a long time and the person can be fine, as long as they make a little more than what they spend. But if they suddenly get foreclosed, or have their utilities shut off, or some other demand for money is called upon, things really spiral quickly—they can't go to their job and get money, they quickly have their other loans foreclosed, and they have to declare bankruptcy or something. Then they can't earn and contribute to the economy, even though they were fairly productive beforehand. This is what happens to many companies at the edge of profitability when interest rates go up and they suddenly have to pay rates they can't afford over the long-term, or pay off the entirety of loans they can't afford at this moment. A sudden small shock that they can't spread out over time can destroy their entire productivity.

The way the economy surfs the wave right at the edge of profitability helps explain the mystery of why crunches can happen so quickly. If there was more marginal profit to be gotten from lending, lending would occur, which pushes rates down a little bit, which actually increases marginal profit to be gotten (if rates were at 1.5% but drop to 1%, now anyone who could make 1.2% should get a new loan), so more lending occurs and pushes rates down in a feedback cycle (and all the formerly-profitable loans become even more profitable when they rollover at lower rates). But when just a tiny bit too much has happened, then not only are those last marginally-profitable loans proved to be no longer profitable—rates go up, and suddenly the batch of loans before that becomes unprofitable, which forces rates further up, and more loans become unprofitable in the reverse, nastier feedback loop. (Technically, the loans that were made are still profitable, but if they can't directly repay then the rollover will be unprofitable.) Suddenly there are a bunch of people in tight spots trying to pay off loans, with not enough cash flow to go around and save everyone.

If you're still wondering the age-old question of whether this can all just be averted by collectively deciding not to be stingy with our cash, this is a little true—I'm sure there's some hoarding psychology that exacerbates the crash at the bottom. But the general pattern is unavoidable, because the cash has to come from somewhere and there just isn't enough of it. You can nationalize corporations to keep them productive and pay back the taxpayers with their future profit, or you can make government emergency loans to big players that are profitable for taxpayers like the 2008 bank bailout. But the bad loans have to be paid for by someone; if the lender can't shoulder the loss, they can't pay their lenders; if the lenders can't shoulder that loss, they can't pay their lenders in turn; so something has to drop in value equivalent to the value of the loss. The only question is where it comes from.

## Conclusion

As I type these things I'm hit by some déjà vu from reading Ray Dalio, where I felt like he kept dodging a key question about \*exactly what was happening\* to make the debt crisis occur. I felt this way reading most explanations of bubbles over the last few years. Why is there a positive feedback loop rather than a negative feedback loop? Why don't entities adequately prepare for this? Why isn't every "reason" for a crash countered by an equal and opposite reason for self-preservation? Why does it seem like there are so many accounting terms thrown when there should be a simple core with fewer terms?

One thing to point out is that there are actually just a lot of things going on in an economy-wide liquidity crunch. A business going bankrupt has to sell many different kinds of capital: some capital is easily accounted for as wealth (accounts-receivable, or machines), but other capital is very different (human capital, or organizational capital). We try to use simple categories like "wealth", but the disbursement of different wealths shows up in very different ways on the balance sheet and leads to very different outcomes. The edge case of a debt crisis is enough to make lots of concepts start to fray across the very complex web of transactions and accumulations in different entities across the economy.

A second point is that feedback loops are very difficult to quantify. Because of the nature of a nationwide domino-ing liquidity crunch where reduced wealth causes reduced demand causes reduced wealth, this falls into that category. Plus, the vast differences in how it affects businesses with low COGS or safe financing or what-have-

you, compared to other businesses, makes it somewhat hard to aggregate into a single fixed-point equation. However, much like fractional reserve banking leads to a feedback loop that still results in a finite and estimable amount of new cash, there should be a succinct series of equations relating reduced wealth to reduced aggregate demand to reduced wealth. Or perhaps the present elasticity curves take that into account since they're already empirical. Anyways, I don't entirely grasp how all this fits together, but I imagine there's some simple explanation that will be laid out eventually.

If anyone has a pithier conceptual explanation for the reasons behind the bubble or the crash, I would love to hear them; for now, I hope that this at least helped characterize debt crises, and perhaps my [future](#) posts will convey more of the causation.

*Thanks to John Steidley for many good discussions slowly clarifying this topic.*

# Avoid News, Part 2: What the Stock Market Taught Me about News

This is a linkpost for

<http://www.bayesianinvestor.com/blog/index.php/2021/06/13/avoid-news-part-2-what-the-stock-market-taught-me-about-news/>

It's been a decade since I blogged about the benefits of [avoiding news](#).

In that time I mostly followed the advice I gave. I kicked my addiction to The Daily Show in late 2016 after it switched from ridiculing Trump to portraying him as scary (probably part of a general trend for the show to be less funny). I got more free time, and only missed the news a little bit.

Then the pandemic hit.

I suddenly needed lots of new information. Corporate earnings releases were too slow.

Wikipedia, Our World in Data, Metaculus, and some newly created COVID-specific web sites partly filled that gap. But I still needed more, and I mostly didn't manage to find anything that was faster or more informative than the news media storyteller industry.

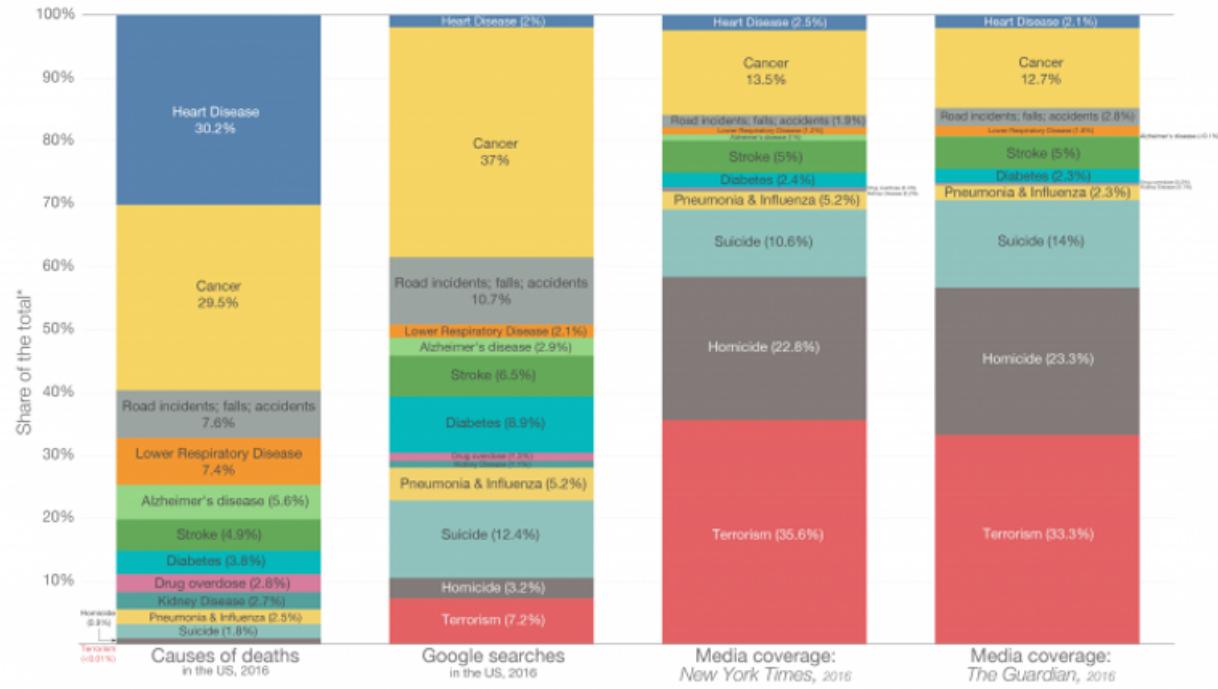
That at least correlated with higher than normal stress. I suspect that paying attention to the storytellers partly caused the stress.

## Distorted Salience

Do you want to focus your attention on problems that are most likely to hurt you? If so, the storyteller industry is likely to distract you:

## 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://owensher24.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% so they represent their relative share of the top causes, rather than absolute counts (e.g. "death" represents each causes' share of deaths within the 13 categories shown rather than total 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](http://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.

It's hard for a daily news source to avoid that effect. It has to focus on recent, unusual events. When coal plants [emit radioactive smoke](#), there's little occasion to write about it, because it's very routine, and not at all a sign that someone did something wrong when people initially adopted coal. Yet when a nuclear plant emits a similar amount of radioactive material, that's a rare sign of an avoidable mistake. It's unclear how the storyteller industry could avoid causing readers to pay undue attention to the nuclear mistakes. But the natural consequence is that readers will, at least subconsciously, treat that attention as implying more harm.

Storytellers could reduce the biases from that distorted salience, by sounding more worried in proportion to harm implied by their stories. But that conflicts with the goal of making their stories entertaining, and with the goal of maximizing the number of stories that readers believe to be important. Alas, few readers are willing to give the storytellers incentives to prioritize that kind of balance.

## Preparing for the Pandemic

One of the bigger mistakes that I made in 2020 came from recalling that Ebola, SARS, the 1976 swine flu outbreak, and various bird flu's all generated a modest amount of news, but little harm in the US. Even the older pandemics that did kill many Americans

had little apparent effect on the stock market. So I gave low priority to studying pandemics.

In addition, I had traded a few bird flu contracts on [Intrade](#), and I think made a tiny amount of money betting that pandemics were less likely than people thought.

Alas, there's a difference between the probability of an event and the expected value of that event's effects. Storytellers, who tend not to do expected value calculations, seemed to often overstate the probability of a distant pandemic reaching the US. Yet they rarely imply that the expected harm implies anything more than "you need to stay tuned to our channel".

But I wasn't paying much attention to the storytellers at the critical times, so I can't blame them much for my mistake.

## COVID fears

In the spring and summer of 2020, I saw rather large discrepancies between my business forecasts, and the impressions that storytellers were promoting.

It sure looked like many commentators / storytellers were going out of their way to overstate the economic damage from the pandemic.

I presume that's partly due to them trying to manage people's fear levels: if people felt economically safe, that might cause them to feel it was safe to risk spreading the virus. Another reason might be that fearful stories selling better than happy ones.

Whatever the reasons, by mid to late April enough companies were issuing statements to the effect that the damage to their business wasn't catastrophic, and that they were seeing signs of recovery. No single report meant a lot, but I read a lot of corporate announcements, and saw a clear pattern: business was better than most pundits were willing to imagine.

For much of 2020, there were widespread [claims](#) that [Hertz shares were worthless](#), or nearly worthless.

That conclusion was superficially plausible, as it's not unusual for bankruptcy cases to work out that way. But Hertz's situation seemed unusually uncertain, and the pandemic triggered some abnormally fast business fluctuations.

It now looks like buying Hertz while it was in bankruptcy proceedings was [a great idea](#) given the benefit of hindsight.

I'm not complaining that the storytellers were wrong about what Hertz shares were worth. They did a decent job of identifying relatively competent sources of opinion on the subject. I'm complaining about the false implications about the competence of those sources. It's hard to identify an expert who could convincingly report that nobody had the foggiest idea what would happen (unless you're willing to treat market prices as expert opinions - in which case, why are you paying attention to a story instead of just looking at the price of Hertz stock?). Hertz ought to have been almost that expert, but legal risks pressured Hertz to slant its comments in a more paranoid way.

Hertz's value ended up depending somewhat heavily on when vaccines became available. That's an important topic where the storytellers failed badly. E.g. [this headline](#) claiming to be a "fact check" that said we'd need a miracle to get vaccines in

2020. That was in mid-May, when [Metaculus](#) was showing a significant chance that a vaccine would be proven effective by the end of 2020.

Some storytellers even spread the false claim that [We've never made a successful vaccine for a coronavirus before](#). If you read the story carefully enough, you will notice it eventually hints at the caveat "in humans". Vaccines for coronaviruses have been commercially available for [several species of pets and livestock](#).

It took me a couple of months to see through that vaccine smokescreen. Doing so helped me to make lots of money in the second half of 2020 betting on pandemic-sensitive stocks (not Hertz), since most investors were too pessimistic about vaccines for much of that period.

## Enron

[Enron](#) is an example of why I mostly ignore stories, and focus on verifiable numbers.

I heard a glowing report of Enron's innovative management style in mid-2001 (at [Extro 5](#)), and likely would have noticed more such reports if I had been more addicted to storytellers. That report sounded plausible, and it would have required an unusual amount of effort for me to find anything wrong with that story. When I thought in terms of stories about Enron management, I felt a real temptation to buy Enron stock, and those stories would have seemed more salient if I had paid more attention to the storytellers.

Then I looked at Enron's earnings growth, revenue growth, and return on equity. The revenue growth was fairly impressive, but nothing about the earnings distinguished it from a boring company.

I also looked at charts of its stock price. I saw signs that sellers were less patient than buyers. That usually indicates an ordinary fluctuation in business, but whatever, it counted as evidence that I could safely postpone buying, so I decided to wait. I kept waiting until it had clearly collapsed.

## Coping Strategies

My most important strategy for getting good stock market information is to decide what information I want, to actively seek it, and to minimize my attention to what other people suggest I should pay attention to.

If I let the storytellers decide what evidence to use in evaluating a company (or an industry, or the economy), I'll end up using criteria that are selected for other people's purposes. Sometimes those purposes will be ideological, sometimes they will be to protect some powerful institutions, or maybe they'll be mostly for entertainment purposes. The results are mostly not much better than letting one side of a court case choose which evidence I should look at.

More generally, I try to focus on sources that prioritize accuracy over entertainment (e.g. Wikipedia, Metaculus).

Alas, that requires some willpower. I haven't managed to develop a consistent habit of looking at sites like that for information.

I also have some need for alerting myself to important events that aren't on my radar. The Wikipedia [Current events portal](#) is such a source, but isn't wise enough about what's important to be a complete answer. I check [news.google.com](#), but it's click-baity enough that I keep hoping to replace it with something better.

Individual bloggers sometimes provide news-like benefits, especially when they're aiming to earn respect from well-informed peers, rather than any of the strategies which require attracting lots of readers. [Scott Sumner](#) often meets those standards.

I've gone back to reading Tyler Cowen. He does a better job than Wikipedia of identifying important topics, but is much harder to skim - I can't get much out of his blog without getting sucked in to spending more time there than I want.

## Is it Getting Worse?

When I was young, there were fewer choices for how to get the knowledge that storytellers aim to provide. People chose between maybe three TV stations, and one or two newspapers.

That meant less competition than we have today. Less competition meant that storytellers had more freedom to write stories that enhanced the storytellers' reputations with their friends and families. That created a modest tendency toward high-brow, consensus-oriented stories.

For example, the U.S. had [many bombings in 1970s](#), often politically motivated. The storyteller profession was able to treat those as ordinary crimes, that we could mostly forget because the police were able to contain the harm to relatively minor levels.

Whereas today, we have storytellers that specialize in peddling outrage. At least some of them will have ideologies which will predispose them to overstate the importance of politically motivated crimes. Even those who are predisposed to downplay the importance will have some incentive to hype the stories in order to keep readers who are attracted to controversy.

Yet we also have information sources that replicate the high-brow, consensus-oriented features that newspapers and TV stations used to provide, e.g. Wikipedia. So my conclusion is that people who really want good information are better off today, but the average person's information diet is likely getting a bit worse.

P.S. I don't want to leave you with the impression that storytellers are mostly wrong. If they were, I could make more money by doing the opposite of what they say. Their actual track record is much more ordinary than that.

# **Changing my life in 2021, halfway through**

This year I decided to really try and fix some major parts of my life I have been neglecting. I started in January and I think I have been making good progress. I will give links to the resources I think helped improve my life so you can look over them and implement them too if you are interested.

I will be making a far more fleshed out post for my end of year review, but I'm making this both as a reflection for myself and as a mini time-capsule to look back at.

## **Why I did this:**

It was mid 2020 when I realized I wanted to be an AI researcher and work on some of the most fun / interesting problems of humans. The only problem was I just graduated with a business administration degree and knew almost nothing past high school algebra in terms of math. I screwed around the rest of the year doing some random walks through math and coding, learning things like the power rule for calculus without knowing anything else about calculus or what a derivative was. At the end of the year and after consuming a lot more of Less Wrong content (which I found as a corollary of my newfound interest in AI). I realized I should stop random walking not only math, but I needed to stop random walking my entire life and should make everything functionally better.

## **0-100(ish(hopefully)) Machine Learning Study guide:**

I didn't want to spend 4 more years to get another degree, academia has a lot of pro's and con's but ultimately it came down that I think I could learn the same material on my own faster, pay less, but with the trade-off of having a harder time proving myself as a competent programmer. So I decided to self learn with the goal of learning so much and becoming such a good programmer I could manage to slam my foot into the door and get a job with no C.S (computer science) degree.

At the beginning of this year I was opining in a babble thread how I was upset that most learning sources sucked. They seemed to be far too densely packed to be useful. I will be fully honest and say I took a two in one, "how to code in python and learn machine learning" class on Udemy at the same time, needless to say I was way in over my head and didn't finish it as I knew little programming and less math. If only there was a curriculum list that could go through many classes and actually teach rigorous, practical and theoretical programming.

I looked around a little bit, but since anybody can make a programming "tutorial" and there is a high demand for tutorials, The top results listed in most places usually went to the highest advertisement bidder and not necessarily the best teacher.

There's also what I call the 'beginner programmer death spiral'. if you don't do a C.S degree and don't know what you should know to become a better programmer (which could be anything) it's really easy to fall into the trap of jumping around beginner level courses. Learning the beginner level stuff in one course, but feeling unsure of your skills and where to go next, finding another course that usually teaches you roughly the same thing and feeling lost because all courses look like beginner level courses that teach you the same thing, while

the advanced courses look way too advanced for your current skill set and you don't know how to bridge that gap. I was in that death spiral before and it sucks.

There's also [MIRI's guide](#), but I do like the approach modern MOOC's have with lectures. I think lectures really help (for me at least) with the intuition behind a problem and understanding why things work the way they do. It takes a little bit longer for me to focus on visualizing what the writing in textbooks is saying while I don't seem to have that problem with professors. MIRI's guide also seemed to be a non zero starting point to me, expecting the reader to have some underlying math skills, which is perfectly fine considering most people that would find that guide in the first place.

I decided in that babble thread to say screw it, I'll do it myself. [Viliam](#) messaged me and shared good resources and thoughts about how to set up a curriculum since they are working on their own Slovakian math curriculum. I appreciated that response a lot and I'm highlighting it both here and in my end of year post as a show of gratitude. Another thanks for introducing me to [sturgeons law](#) which has held for most coding resources I found. I hope your project is going well.

I did find [Open Source University](#)(OSU) on GitHub. I thought it was good but there were some parts I found lacking and not directly related to my goal of the shortest path to competent ML programmer possible while leaving nothing out. So I gutted it and made my own, here it is:

<https://pastebin.com/Yg6EZSPM>

Keep in mind this is a snapshot of the course in time, I'm usually on the lookout for classes that seem better / supplement my gaps in knowledge. These get put into my google docs version. The final version will be google docs so people can add comments, for now this my first draft.

It includes many classes not in OSU, because either the teaching of the topic was slow or I felt the professor did a bad job of explaining things. I also took many OSU classes out because I didn't think they mattered as much as the ones I picked. Feel free to supplement. I also am reviewing each class I take in it which you can find below. At the end of the year I want to give a most optimal path introduction about how to quickly use the guide.

This guide is designed to be depth first breadth second. It's designed to give fuzzy understanding but practical knowhow. I feel right now that my math concepts are an island. I know some calculus, I know some probability, I know very little about linear algebra, but not zero. My hope is that I can gain functional coding ability with machine learning with shoe string math knowledge, and while I both code and learn higher level math those skills will feed into each other and get stronger together. In learning logistic regression that did, as the sigmoid function gives out a probability. If you feel like you need to have a full intuitive grasp on what you are learning and now jump ahead, this guide might not be for you, but the courses in it might be helpful.

If you plan on using my skeleton guide and are a hard studier, do math while waiting for the weeks to open on edX since they are annoying time gated. I plan to add "how to" for self study, some general tips, and reassurances. But for now it's just a bunch of courses and a few reviews.

My course has downsides like all do, the two biggest I can think of now being; I don't know what I don't know, how useful will some of these classes be in the long term? I like to think I made good picks but I do always wonder if there is a more efficient, intuitive math course I could be learning from or if some coding classes are teaching enough. The second being it's

easy to get insecure self learning, you don't have a peer group to place yourself into to learn better tricks, it's hard to talk casually in class discussion forums so you don't know if other people are struggling with the same problem you are finding hard. I feel like I'm on an island sometimes, I like to think at the very least if we took everybody in the MOOC I would fall near the mean, but it's impossible to know. thus, it's probably not worth actually worrying about. I'll save the worrying for coding competitions and learn as I go.

Currently I'm taking Andrew Ng's Machine learning class now after taking MIT's edX python courses. I like that it's not time gated, but I feel like the programming assignments do a lackluster job of connecting the lecture material to practical practice. I'm on week 5, I can implement forward and back propagation just fine. But if you sat me down with raw data and said "okay build the network", there's huge practical gaps in preparing data and evaluating the system that I just don't have yet. I hope that gets better later, there's also some other DeepMind courses that potentially addresses this that I've added to the list. I'm not too worried as since I know this won't be the last time I learn ML, anything I don't learn here I will flesh out on the go.

My goal is to get through deep learning by the end of the year, do Kaggle competitions, and work on some computer vision personal projects. Gaining enough skills to become an AI researcher at some point in the next 2 years.

## Working Out:

I read [Convict Condition](#) before the start of the year and decided to implement that routine while running most days. It seemed like a nice way to do strength training without going to the gym, especially during covid times with no vaccine. The routine was twice a week and generally took only a little amount of time, around 30~ minutes. I started to notice some muscle definition and did feel stronger. I can recommend the routine no matter what strength level going.

When I posted the book review commenter [9eB1](#), made a good comment about calisthenics and pointed to the reddit [bodyweight exercise](#) page. After being lazy and failing to update this into my knowledge. I finally read it and found their routine interesting. I'm planning on making it my routine for a while and testing how it goes. I would recommend it as a resource for those looking to get into strength training.

Currently routine:

M-W-F: lift.

T-Th: run.

I have been running most my life, but if you are interested in beginning I heard [Couch to 5k](#) is good. But I have not tested it myself.

## Diet:

I just recently implemented a Modified Mediterranean Diet based on Scotts ACT [thread on depression](#). I do feel better when I stick to it, I feel better than before which had my diet mostly based on pasta, beans, processed food and sugars. I'm glad he posted it. olive oil and baguettes are delicious to me and I wish I did this sooner.

The only downside is increased cost and portions are smaller, which with small portions could lead into more increased costs to the diet to offset calorie expenditure from working out. Also if you eat fish the taste seems to linger in your mouth the whole day.

A CARLAT PSYCHIATRY  
REFERENCE TABLE

| The MediMod Diet                               |  |   |
|--|--|---|
| Food   | Recommended servings   | One serving equivalent  |
| <b>Vegetables</b>                              | 6 servings/day. Include green leafy vegetables or tomatoes in at least one of those servings. Mushrooms count, but minimize potatoes to one serving a day unless it's a sweet potato.        | Leafy vegetables: $\frac{1}{2}$ cup cooked or 1 cup raw; other vegetables: $\frac{1}{2}$ cup raw or cooked.   |
| <b>Fruit</b>                                   | 3 servings/day. Include berries in at least one of those servings.   | $\frac{1}{2}$ cup fresh, frozen, canned, or cooked fruit; 1½ tablespoons dried fruit. Juice counts but should be limited to $\frac{1}{2}$ cup per day because of the sugar content.   |
| <b>Nuts, seeds, olives</b>                     | 1 serving/day.   | 1 ounce/day of nuts, seeds (about $\frac{1}{4}$ cup), and/or 3 ounces of olives (about $\frac{1}{2}$ cup).  |
| <b>100% whole grains</b>                       | 5–8 servings/day (eat closer to 8 if you're physically active).  | 1 slice bread; $\frac{1}{2}$ cup cooked grains, like brown rice or whole wheat pasta; $\frac{1}{4}$ cup oats or muesli; $\frac{3}{4}$ cup breakfast cereal; 2–3 crisp bread crackers. |
| <b>Fish</b>                                    | At least 2 servings/week. At least one of those should be an oily fish like salmon.  | 3 ounces cooked.  |
| <b>Beans</b>                                   | 3–4 servings/week.   | $\frac{1}{2}$ cup beans, or $\frac{1}{2}$ cup hummus or tofu.   |
| <b>Extra virgin olive oil</b>                  | 3 tablespoons/day.   |   |
| <b>Red meat</b>                                | 3–4 servings/week.   | 3–4 ounces cooked. Use lean red meats.  |
| <b>Poultry</b>                                 | 2–3 servings/week.   | 3 ounces cooked (= one breast or a leg + thigh).  |
| <b>Dairy</b>                                   | 3 servings/day of milk, cheese, or yogurt.   | 1 metric cup milk or yogurt. For cheese: 1.5 ounces hard cheese or feta; 4–5 ounces soft cheese like ricotta or cream cheese.   |
| <b>Eggs</b>                                    | 6 eggs/week.   |   |
| <b>Eat less of...</b>                          |  |   |
| <b>Fried, fast, sweet, and processed foods</b> | Maximum of 3 servings per week. A serving is 120 calories of: Sweets, sodas, snacks, and white bread. Fast, processed, or fried foods. Beef jerky, bacon, and deli meats.                    |   |
| <b>Alcohol</b>                                 | Maximum 1.5 standard drinks/day. Red wine is preferred. 1.5 standard drinks = 6.8 ounces wine, 2 bottles beer (1 bottle if it's high gravity), 2 ounces spirits, or 5 ounces sherry or port. |   |

Source: Opie RS et al, *Nutr Neurosci* 2018;21(7):487–502

From the Expert Q&A  
"An Antidepressant Diet"  
With Felice Jacka, MD  
*The Carlat Psychiatry Report*, Volume 17, Number 5, May 2019  
[www.thecarlatreport.com](http://www.thecarlatreport.com)

## Future plans:

I tried to do hydroponics at the start of the year, [everything was going well](#) until my plant light died and my plants got too big and snapped in half. So I'm switching to a more low maintenance style of growing. I'm going to try to [repurpose a 2 Liter](#) and grow a few plants

that way. Hydroponics is pretty easy once you get started, the costs are front loaded. If you want to try it out I found <https://www.epicgardening.com/hydroponics-for-beginners/> to be a good resource.

I want to understand fashion more and build better outfits for myself, aesthetics to matter a lot to people. I would appreciate any resources you guys have for that. My plan for right now is to just try to get a general feel for what looks good on me or not.

I know very little about good personal financial management other than that ideally revenue > expenses. If you found any source for learning about personal finance useful please post it.

My key take-away from learning this first half a year is that finding good learning resources is hard. But once found I think you can quickly elevate your skills with them, with respect to your time put in and focus on continual improvement. I think this holds cross domains, I do want to try my hand at 3d modeling and better understanding the stock market, I'll try to find some good resources for these before the end of the year.

I would also like to know any good learning resources you've found so far this year that you've enjoyed. I hope your first half of 2021 is going well and lets keep strong for the second.

# Covid 6/10: Somebody Else's Problem

Logistics note: This week's post includes some non-time-sensitive items from last week, including much of the section on the lab leak hypothesis, as my available time last week was limited. Also, I'll be in NYC from 6/11 until 6/16, so if you'd like to chat, let me know - I promise asking will cost you at most zero points.

Covid-19 is not quite done with the United States of America. The Indian variant ("Delta") delivering a 'one last scare' moment is plausible. But unless there's a new variant that can escape from the mRNA vaccines and we are unable to respond rapidly enough, a possibility I now put at most at 10%, Covid-19 is *mostly* done with the United States of America.

There are essentially three things left.

1. **Live life.** There's the question of how we can safely transition back to a normal life here in the good old USA, and what that new normal life will look like. This includes how worried we should be about potential future strains or issues of seasonality.
2. **Rest of the world.** Covid-19 is *probably* mostly done with the USA. Other countries without our vaccine access are not so fortunate. It is quite plausible that the majority of deaths from Covid-19 are in the future rather than the past.
3. **Postmortem.** There's the question of how we can learn from what happened. We need to adjust our models of the world, and what methods we can use to make sense of it or change it for the better. We also specifically need to do what we can to deal properly with the next pandemic or other crisis. Thus questions like the origins of the virus and what to do about what seems like the absurdly terrible idea that is Gain of Function research.

It is easy to view almost all of this as [Somebody Else's Problem](#). One could say, I'm vaccinated and can go live life, give or take a few medium-term mask requirements. If someone else in America isn't vaccinated, one can decide they made a choice. We can leave others to prepare for the next pandemic and go back to not caring much about what is happening overseas.

Or, if to one's credit one wants to be or remain in the habit of caring about somebody else's problems, one can balance such cares between all the various problems we have. There are a *lot* of problems out there. I do think the Covid-19 endgame, and how many people will end up vaccinated versus infected, is still for now sufficiently up in the air and still looms large enough to deserve attention. For better or worse, that might not remain true much longer.

Let's run the numbers.

## Predictions

Prediction from last week: Positivity rate of 1.8% (down 0.3%), deaths fall by 12%.

Results: Positivity rate of 2.0% (down 0.1%), deaths fall by 4%.

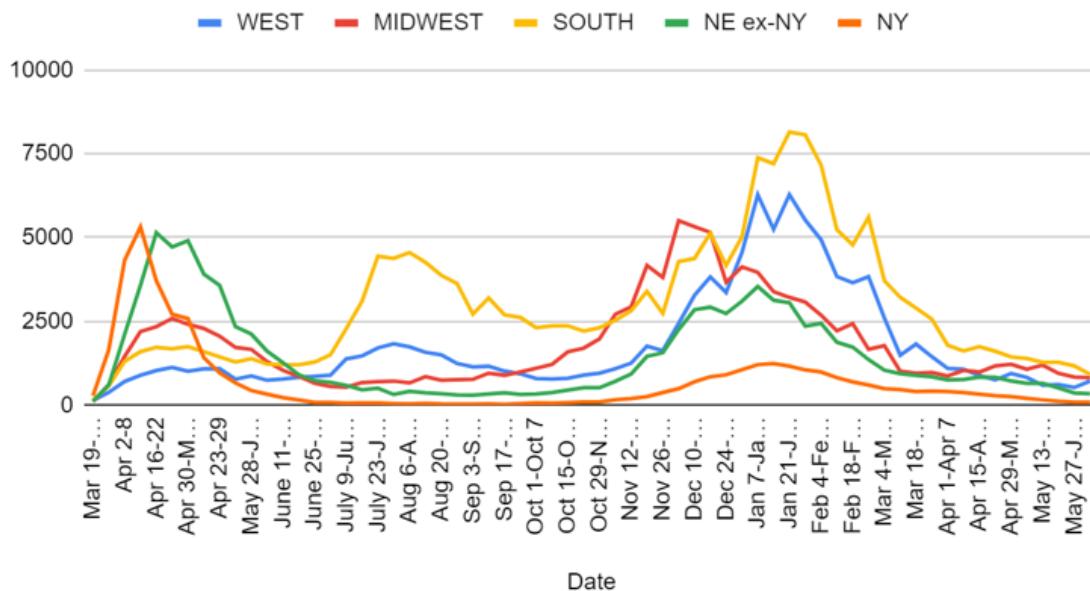
Prediction for next week: Positivity rate of 1.8% (down 0.2%), deaths fall by 9%.

Both numbers this week were disappointing. That happens more often than it 'should' given that deaths represent past cases. I expect the trends to mostly resume, but still one must adjust the expected future path a bit. There is a small chance this week's numbers were outliers and we'll see larger-than-normal drops next week. That's also happened a few times.

## Deaths

| Date          | WEST | MIDWEST | SOUTH | NORTHEAST | TOTAL |
|---------------|------|---------|-------|-----------|-------|
| Apr 22-Apr 28 | 752  | 1173    | 1609  | 1110      | 4644  |
| Apr 29-May 5  | 943  | 1220    | 1440  | 971       | 4574  |
| May 6-May 12  | 826  | 1069    | 1392  | 855       | 4142  |
| May 13-May 19 | 592  | 1194    | 1277  | 811       | 3874  |
| May 20-May 26 | 615  | 948     | 1279  | 631       | 3473  |
| May 27-June 2 | 527  | 838     | 1170  | 456       | 2991  |
| June 3-June 9 | 720  | 817     | 915   | 431       | 2883  |

## Deaths by Region

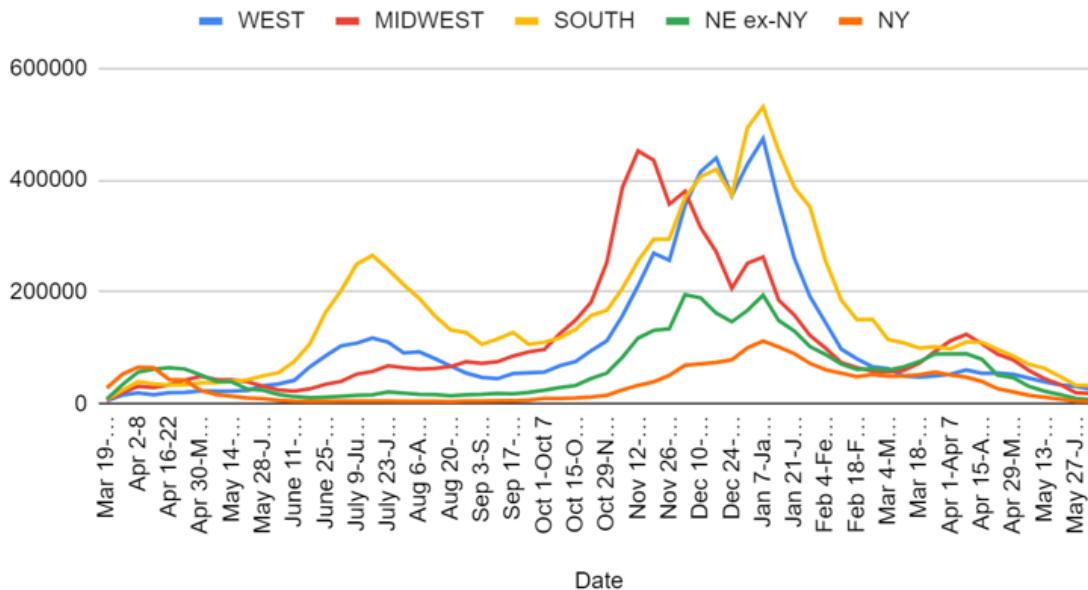


Disappointing progress, but mostly still progress. The rise in the West comes from California. I doubt it is 'real' but it does not seem to strictly be a backlog situation, so I don't think it can be adjusted, any more than the South's number is too low. I'm not convinced we got 'ahead of ourselves' the last few weeks given the drops in case numbers, and I expect deaths to continue dropping.

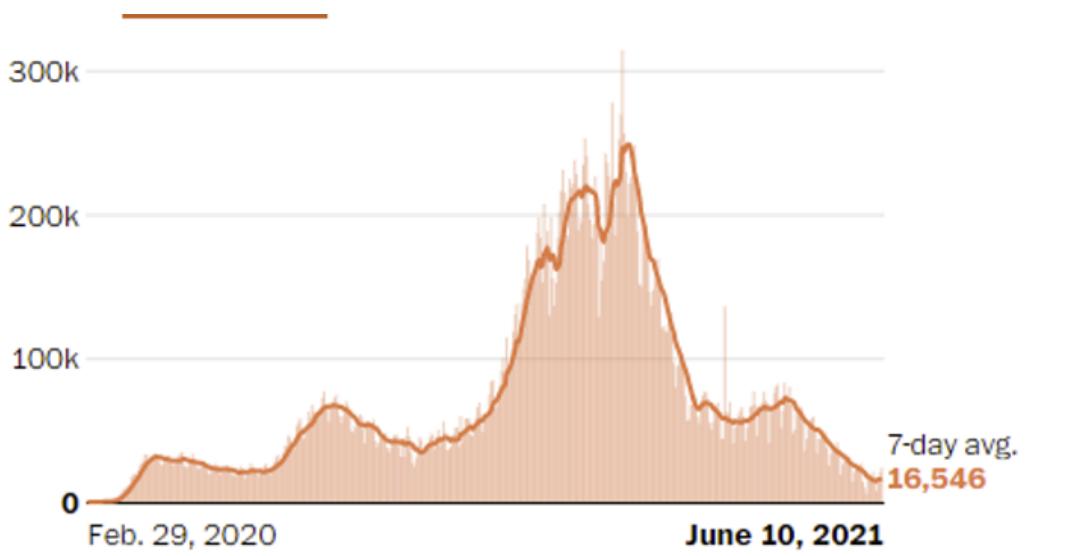
## Cases

| Date          | WEST   | MIDWEST | SOUTH  | NORTHEAST |         |
|---------------|--------|---------|--------|-----------|---------|
| Apr 29-May 5  | 52,984 | 78,778  | 85,641 | 68,299    | 285,702 |
| May 6-May 12  | 46,045 | 59,945  | 70,740 | 46,782    | 223,512 |
| May 13-May 19 | 39,601 | 45,030  | 63,529 | 34,309    | 182,469 |
| May 20-May 26 | 33,890 | 34,694  | 48,973 | 24,849    | 142,406 |
| May 27-June 2 | 31,172 | 20,044  | 33,293 | 14,660    | 99,169  |
| June 3-June 9 | 25,987 | 18,267  | 32,545 | 11,540    | 88,339  |

## Positive Tests by Region



Total numbers (from WaPo) to put it all in easier-to-see perspective:



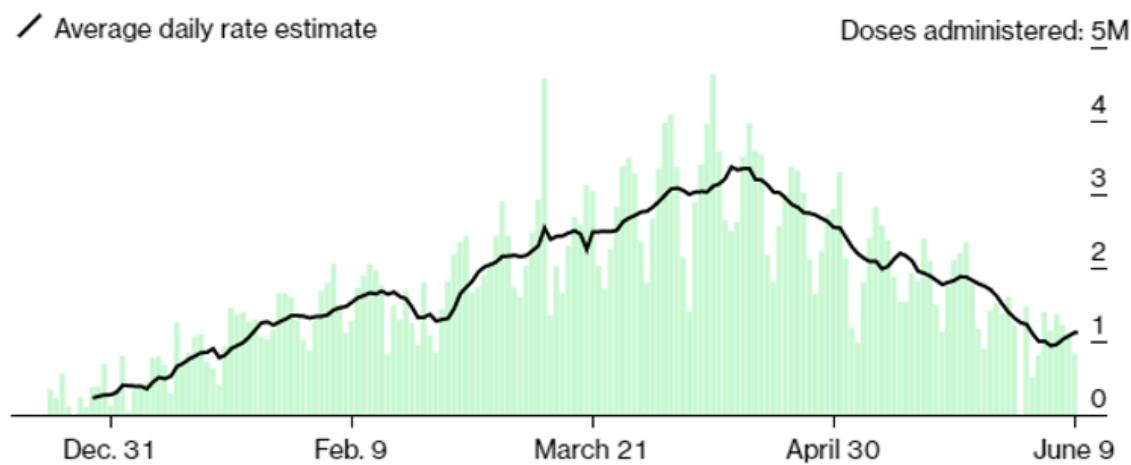
Those are disappointing numbers, with “only” an 11% drop in cases, indicating that it is likely we did get ahead of ourselves a bit last week. The Delta variant is still only 6% of cases, so it doesn’t yet explain the change. Presumably some of this is the continued lifting of restrictions, both officially and unofficially.

## Vaccinations

# 172.1 million vaccinated

The number of people who have received at least one dose of the vaccine, covering **61.4% of the eligible population, 12 and older** and **51.8% of the total population.**

In the U.S., the latest vaccination rate is **1,120,083 doses** per day, on average. At this pace, it will take another **5 months** to cover **75%** of the population.



The recent vaccination stats are very good news. Previously, we had seen a very clear mostly-constant upward slope until early April, followed by a mostly-constant downward slope until the beginning of June. A sensible projection was that this trend would continue. Instead, we have a very hopeful upward trend.

If we can sustain even this level of vaccinations going forward, there's little to worry about. Progress will be slower than we might like, but progress is progress, and we will get there. Mostly, vaccination efforts now shift to overseas, where there is a lack of supply.

## Delta (Indian) Variant in the USA

[Washington Post reports that it is 6% of new cases](#). This was so inevitable I haven't even belabored the point. There's a more infectious strain, so it's going to eventually displace the old one. That's how it works. If this causes cases to go up again for the 'one last scare' you get at the end of a horror movie, we will find out in about a month.

Warning about the danger, Fauci called for more vaccinations, using this logic:

Anthony S. Fauci, the nation's leading infectious-disease expert, revealed the extent of the variant's push into the United States, but said it appears to be slowed by vaccines.

"It's essentially taking over" [in the United Kingdom](#), Fauci said at a briefing for reporters. "We cannot let that happen in the United States, which is such a powerful argument" for vaccination, he said.

Other things we cannot let happen include death and also taxes. I can guess what Fauci is attempting to convey here, and I know what he is attempting to accomplish, but there is no set of possible (let alone plausible) actions that will prevent Delta from 'taking over' what is left of the pandemic in America, or anywhere else. That's not how this works.

## Living Life

(This assumes you're vaccinated. If you're not and you can be, go take care of that, I'll wait. If you're overseas and can't, this doesn't apply to you yet, and I'm sorry.)

You may not have had Covid-19, but the virus has still been living rent free in your head for over a year. Kick it out.

It is important to live and embrace life. It's your life, and (pending heroic efforts to the contrary) [it's ending one minute at a time](#). If you're vaccinated and in a place with the virus under control, go out there and live yours!

That does not mean going back to your old life, if that's not the life you want.

This is the time to take stock. Figure out where you want to live, what you want to do with your life and who you want by your side when you do it. Do things that matter, and things that bring you joy. Go on a trip, see the world, start a business, fall in love, start a family, go out and talk to people in person. And other neat stuff like that.

You'll be asked to wear masks for a while. If you're asked, wear them. It's fine.

If you are asked for much more than that, take your business and your life elsewhere.

That's all there is to it. That does not mean forgetting what happened, or that it's not important to learn from the experience and update our models, actions and precautions. It also doesn't mean dismissing the situation overseas as a pure Someone Else's Problem, or that keeping an eye on things every so often isn't prudent – so long as it stops screwing up your life.

## Rest of the World

Control systems are powerful. Before I fully understood this, it seemed suspicious that Covid-19 was in a nightmarish range of virulence and lethality. Under that theory, if Covid-19 had been more virulent it would have been impossible to stop and we'd have all gotten it and moved on, if it had been less lethal we would have chosen not to stop it, and if it had been less virulent or lethal we would have stopped it. So under this model, there weren't that many configurations of (virulence, lethality) that would be worse than what we got unless both variables were *much* higher.

With a better appreciation for control systems, it is now clear that the range of configurations that result in 'we adjust to remain stuck in limbo for a year' was remarkably large. There were a *lot* of knobs humans could turn to contain things, and they could do so at different prevalence levels based on lethality. It would take *quite a lot* to fully break that equilibrium in either direction, and that did not happen until the vaccines showed up.

In the meantime, the vaccines are holding strong, but [with the current Indian 'delta' strain the virus has become quite a bit more virulent and quite a bit more lethal](#), and any given nation's ability to sustain countermeasures only lasts so long. Will control systems increasingly break in the other direction, with results *far more disastrous* than was possible earlier?

The numbers for the new variant are increases *on top of* increases for the English strain that inspired my claim that [We're Fucked, It's Over](#). I am very happy that we outperformed the math on vaccinations, and that combined with the various warnings from lots of people collectively were a sufficiently self-preventing prophecy that it was not, in fact, over. It was closer than it looked, but we got there mostly in time, and instead of a huge final wave we saw a brief blip.

If we'd had to deal with another similar ramp up in virulence *on top of that*? We had moves left we *could* have used, such as fractional dosing, first doses first, approving additional vaccinations, finally rolling out rapid testing properly and shifting focus further to ventilation and away from surfaces, along with *some* amount of additional restrictions and marginal citizen behavioral adjustments. There's some chance we wouldn't have fully lost. But there's (almost) no way we would have almost fully held the line like we did.

One counterargument is India.

It was a shockingly rapid hockey stick graph. It looked like India was potentially headed straight for full herd immunity. The hospital system broke down. Oxygen ran out. It was a disaster that looked like it would only get worse.

Then the curve turned around, and cases started declining again.

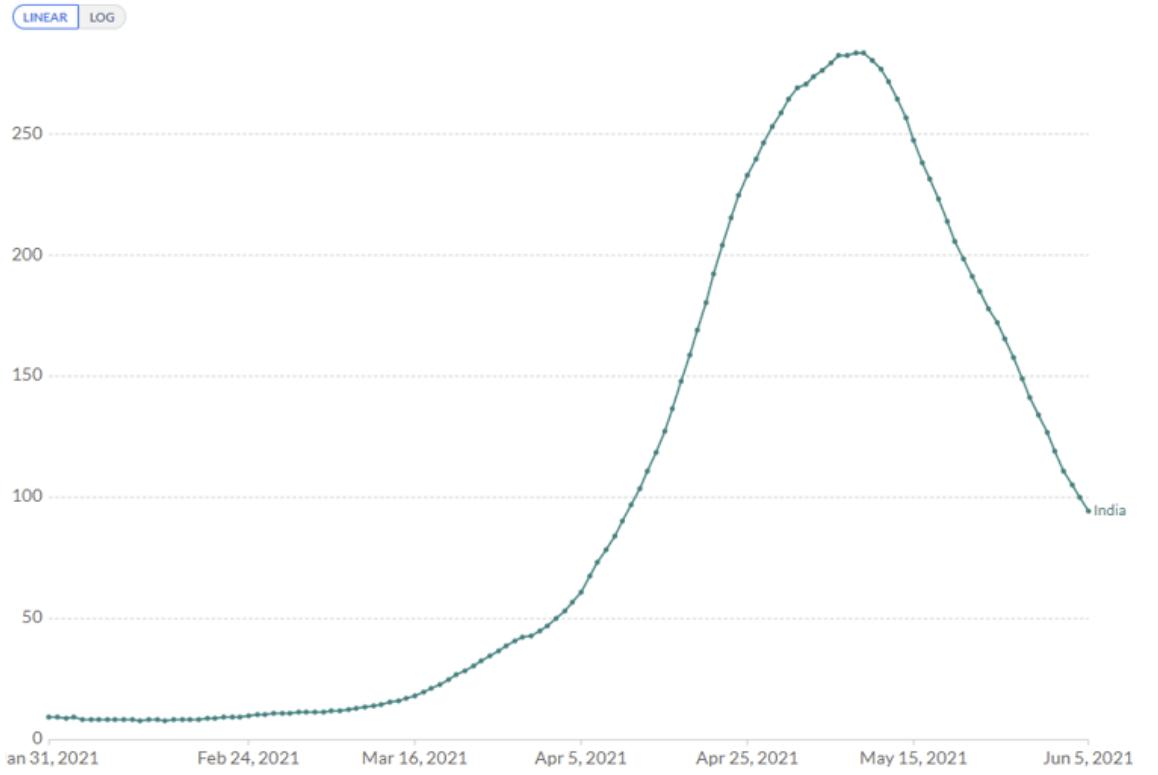
Thus, we have curves such as (one country per graph because scales are so different):

India:

## 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

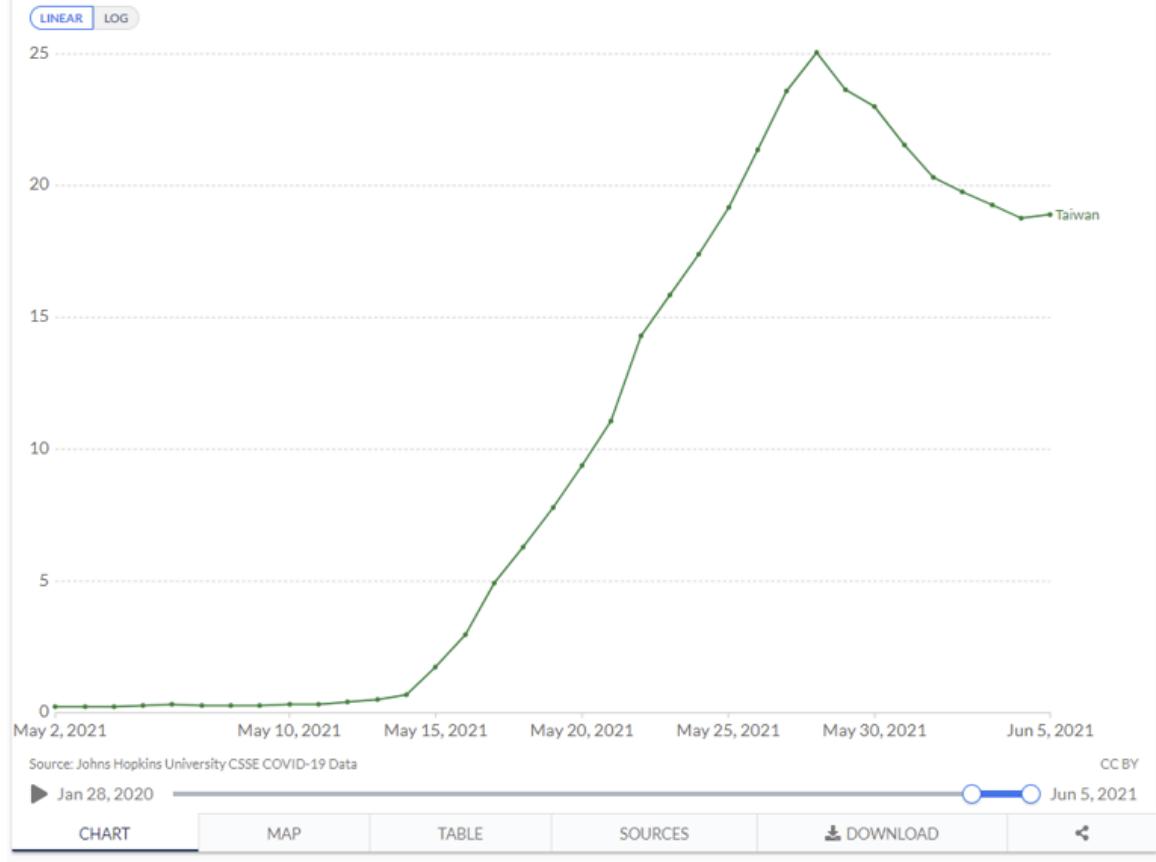


Taiwan:

## 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

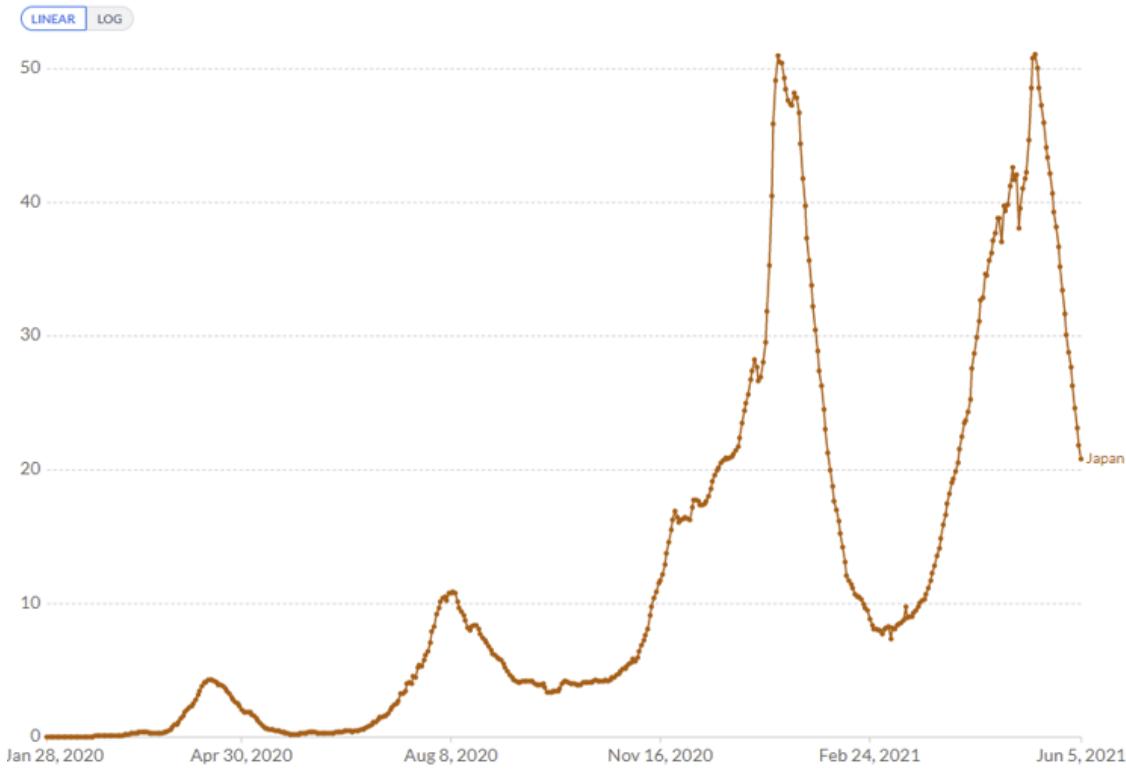


Japan:

## 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



Bahrain:

## 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

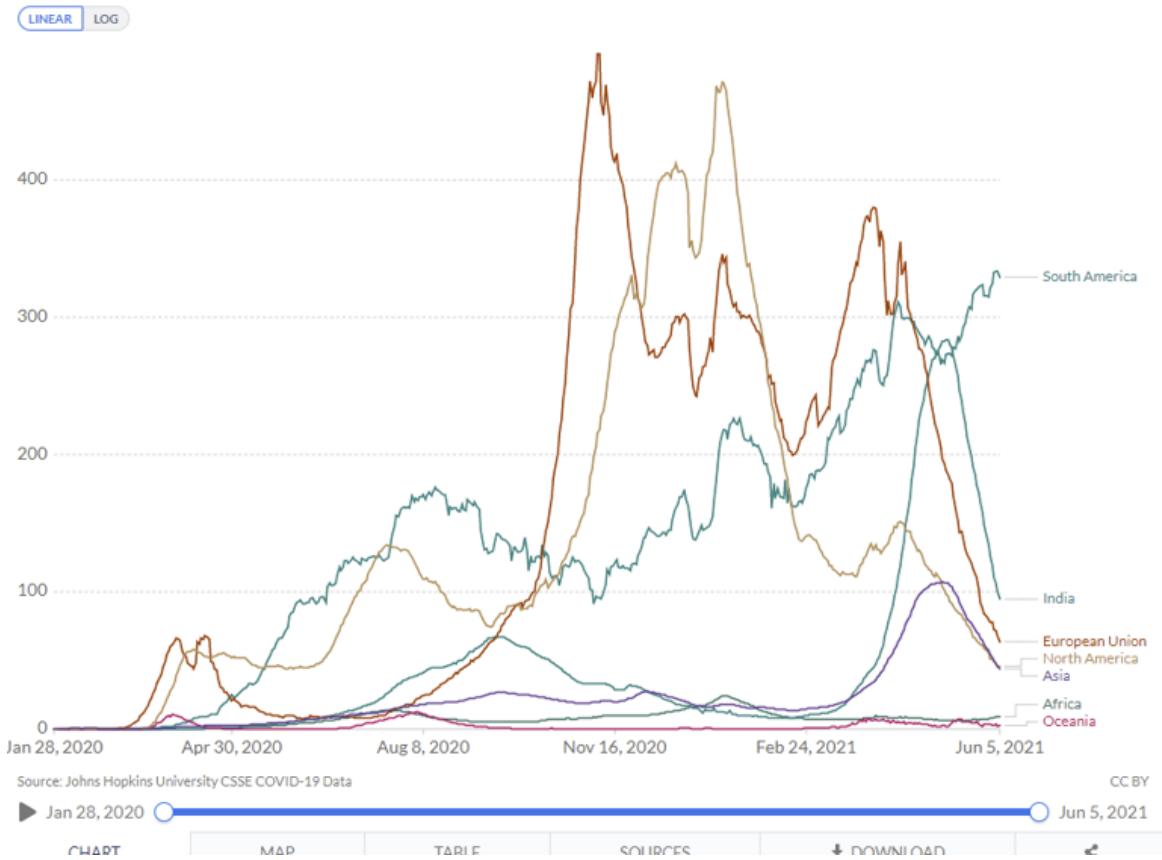


Then combine that with the levels for third world countries having previously stabilized at relatively low numbers. Conditions in India got quite bad, but that was largely a reflection of its lack of public health resources. Here's the world, to scale:

## 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

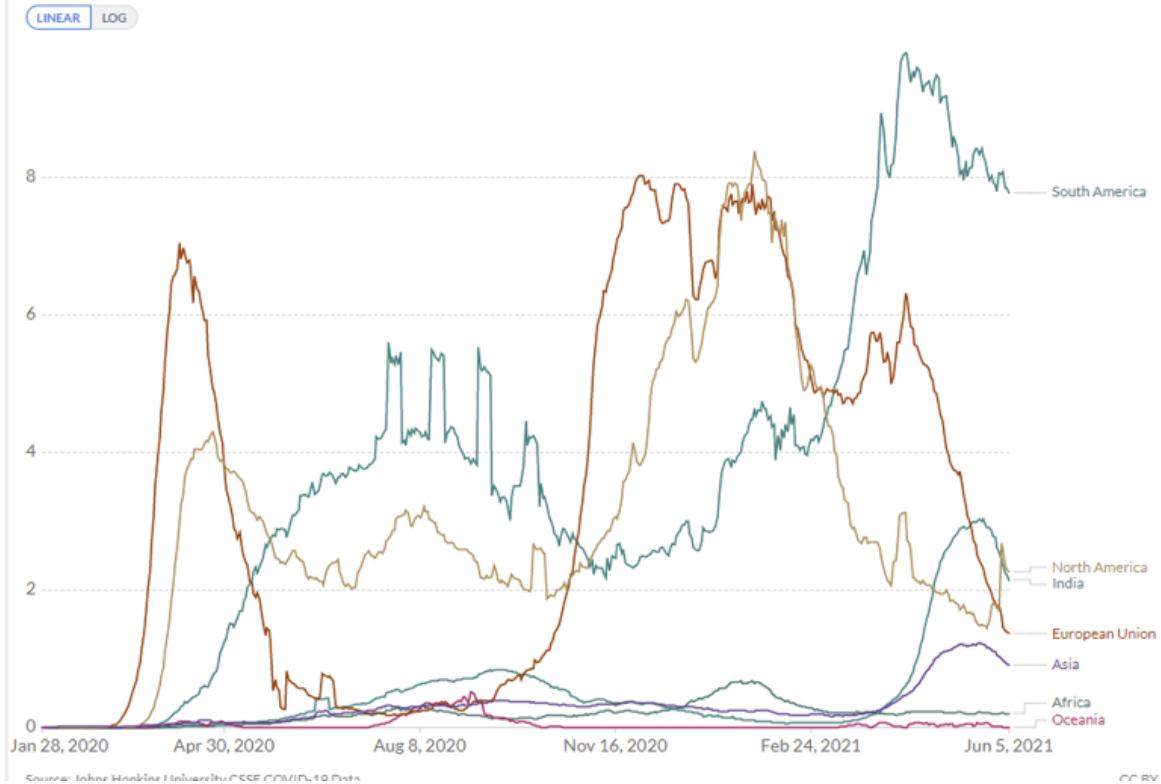


And that same graph with deaths instead of cases:

## Daily new confirmed COVID-19 deaths per million people

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

Our World  
in Data



India peaked substantially lower than North America or the European Union, and was barely higher than South America at the time. Yet it quickly reversed its trend, which I completely did not expect. If you think you could have predicted most or all of those last two graphs, I don't believe you. If you can explain most of it *in hindsight*, that's still a super hard problem.

Some areas have so far managed to completely contain Covid-19 via lockdowns. That's a super scary position to be in, with no immunity built up. What would happen at this point if there was a serious outbreak in China? With China's vaccine not looking that good at preventing spread, it is probable China's only option would be an extremely hard lockdown and hoping it was good enough. Australia's vaccinations have been slow enough for a similar pickle.

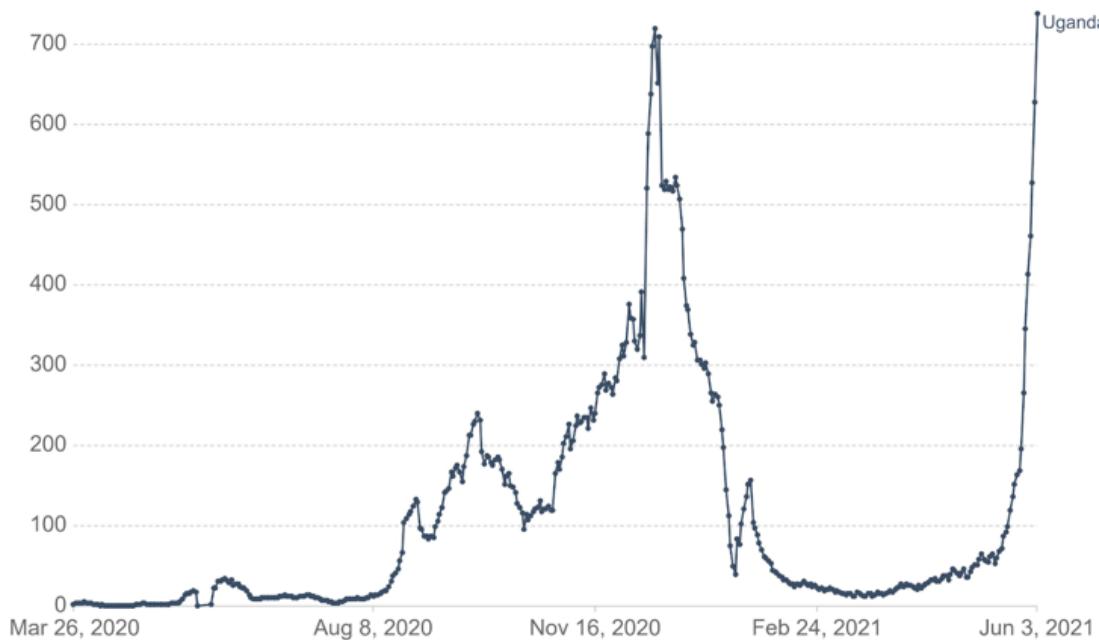
South America is quietly in a terrible situation, and that's where we need to be rushing our vaccine doses at this point rather than India, provided they have the ability to distribute them.

I'm less worried about other areas like Africa, because their control systems should have a lot of slack left given what we've seen so far. That doesn't mean something like this isn't scary:

## Daily new confirmed COVID-19 cases

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



Source: Johns Hopkins University CSSE COVID-19 Data

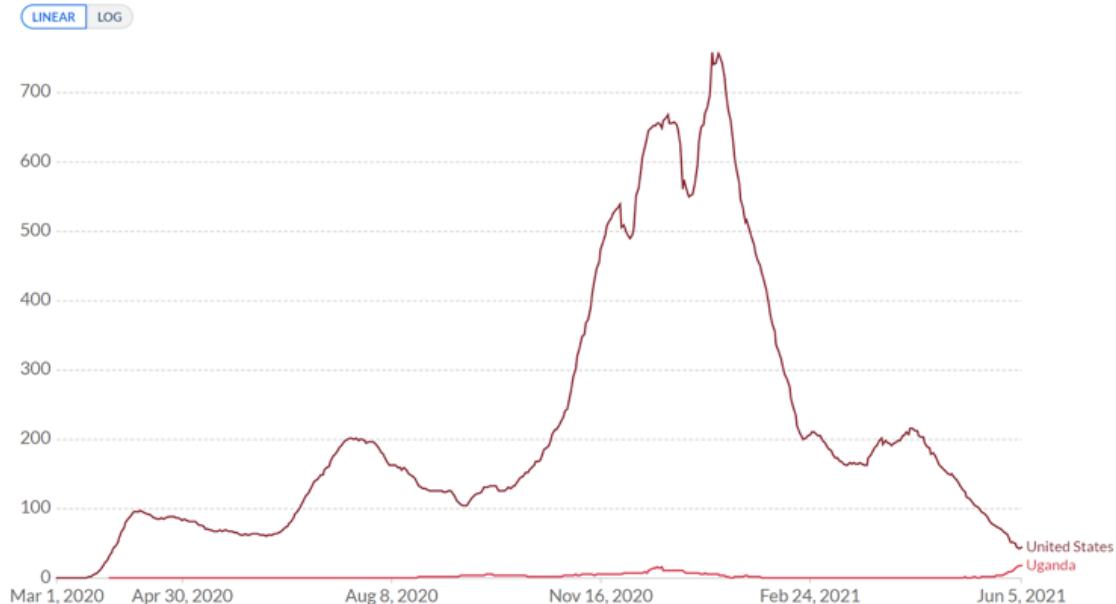
CC BY

...but keep in mind, also, this graph, and remember what happened in India:

## 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



Our continuing failure to scale up vaccine distribution remains a horrible scandal, but it's also worth noting that the issues of distribution in many places are very real. Many third world nations got modest allocations of vaccine doses from Covax and proved unable to put those

shots into arms. Handing over doses, on its own, will not always be enough, and that's a problem I haven't paid enough attention to.

[Zeynep cautions us](#) that there is a real chance that the majority of Covid-19 deaths could come after we had the vaccines, or even after we had enough copies. [In her Twitter thread, she is even more explicit](#). Her emphatic demand, that we do everything possible to gather every dose we can of any vaccine, and vaccinate as many people around the world as fast as possible, is clear and overdetermined, and we should absolutely do that.

This is by far the most important remaining issue: Getting vaccines, any vaccines, out to as many people as possible, as fast as possible, until we've vaccinated the whole world. Period. Better vaccines (mRNA) are better if we have them and can deliver them, but any vaccine is far better than none.

Where I don't fully agree is with her warning that otherwise we are likely to see the majority of deaths after we had sufficient vaccine supplies (as opposed to most deaths being after we had the vaccines, which is *simply true*, because we had the vaccines *in two days* and could have easily confirmed they were safe and effective within a few weeks after that.) I agree that this is still possible if China loses containment. I do not think it is a favorite, but if I had to put a number on Covid causing 7.5 million confirmed deaths worldwide (roughly double the current level) I'd still put it at 15%, which is much higher than I'd like. However, I disagree that this is *increasingly* likely, because of the turnaround in India. When Indian cases and deaths were a straight line up, I would have put that probability much higher, at least 25% (but it's hard to have a good post-hoc measure of what your probability of something was in the past, if you didn't think about it explicitly at the time). Seeing things there not only stabilize but quickly start improving was very, very good news.

And of course, the faster we finish the job, the less likely we are to face a variant that's even worse than the Indian one, potentially one that requires a booster shot.

## Postmortem

A large percentage of these posts have been focused on postmortem. There is much work of this type left to do, and in particular the topic of the month can no longer be avoided.

It's time to fully talk about the origins of Covid-19, and the lab leak hypothesis.

[I wrote a section about this two weeks ago](#). To close that section, I said I would do my best to not talk further about the question of Covid's origin.

My best was nowhere close to good enough. Too much new information has come to light and additional events have occurred, and with that plus time to reflect, it has become clear the implications are important. So here we are.

(Summary of section from last time: I was 40% this was a lab leak of some kind, almost certainly like many other accidental lab leaks that have happened before, which would then have been covered up without need for a broad conspiracy. Press suppressed dissent, and has now been trying to quietly rewrite history. I'd been ignoring it because it seemed like finding out it was a lab leak would make things worse.)

## Escape from the Lab

The media's stance, as far as I can tell, is now to deny that they ever treated the lab leak hypothesis as 'debunked' in the first place, while now calling it a 'conspiracy theory' and trying to claim it comes from Donald Trump, in order to make it impossible for anyone to be

seen taking the hypothesis seriously, and also arguing that the political implications of investigating would be bad so we shouldn't investigate.

This has successfully kept all reporters who desire to have 'credibility' ([as opposed to credibility](#)) away from the story, and leads to [exasperated reactions like this one](#):



**zeynep tufekci** @zeynep · 9h

Seems Dr. Peter Palese of Mount Sinai, who'd signed the Lancet letter calling theories on non-natural origin "conspiracy theories" has changed his mind. Not linking because Daily Mail. CAN WE NOT HAVE SUCH STUFF REPORTED ONLY by Daily Mail? Regular journalists can maybe clarify?

He told MailOnline: 'I believe a thorough investigation about the origin of the Covid-19 virus is needed.'

'A lot of disturbing information has surfaced since the Lancet letter I signed, so I want to see answers covering all questions.'

Asked how he was originally approached to sign the letter and what new information had come to light specifically, Professor Palese declined to comment.

18

29

167



**zeynep tufekci** @zeynep · 9h

I think better journalistic reporting of important stuff which doesn't appear in, ahem, the most sensationalistic and terrible outlets (that little part \*is\* news but it is surrounded by garbage, innuendo and crap) might help?

17

6

93



[Nate Silver notes the explicit bury-the-truth-because-politics argument](#) ([Link to Nature](#)). I don't know why Nate is surprised that people think that advocating for something for explicit political reasons think that doing so makes them 'look good' in the sense that they care about, given how things work these days.



**Nate Silver**   
@NateSilver538

...

Replying to [@NateSilver538](#)

I'd strongly encourage people to read the story. I'm fairly surprised that some of the principals in the story think it reflects well on them.



Nate Silver @NateSilver538

...

I swear I'm sick of posting about the lab-leak stuff, but it seems very warped that a bunch of prominent scientists are saying we shouldn't investigate the claims for reasons that have little to do with science and lots to do with politics.



RELATED  
Scientists call for pandemic investigations to focus on wildlife trade

Fidler agrees. He says that the escalating demands and allegations are contributing to a geopolitical rift at a moment when solidarity is needed. "The United States continues to poke China in the eye on this issue of an investigation," he says. Even if COVID-19 origin investigations move forward, Fidler doesn't expect them to reveal the definitive data that scientists seek any time soon. The origins of most Ebola outbreaks remain mysterious, for example, and researchers [spent 14 years nailing down evidence](#) that the 2002–04 epidemic of severe acute respiratory syndrome (SARS) was caused by a virus transmitted from bats to civets to humans.

Kristian Andersen, a virologist at Scripps Research in La Jolla, California, maintains that no strong evidence supports a lab leak, and he worries that hostile demands for an investigation into the WIV will backfire, because they often sound like allegations. He says this could make Chinese scientists and officials less likely to share information. Other virologists suggest that such sentiments could lead to more scrutiny of US grants for research projects conducted in China. They point to a coronavirus project run by a US non-profit organization and the WIV [that was abruptly suspended last year](#) after the US National Institutes of Health pulled its funding. Without such collaborations, says Andersen, scientists will have difficulty discovering the source of the pandemic.

Even if the letter in *Science* was well intentioned, its authors should have thought more about how it would feed into the divisive political environment surrounding this issue, says Angela Rasmussen, a virologist at the University of Saskatchewan in Saskatoon, Canada.

This is an explicit call for scientists not to say true things or investigate questions, if doing so would be bad for science funding and/or for geopolitics. I do appreciate the explicitness, [as does Nate](#):



Nate Silver @NateSilver538 · May 29

...

I sort of appreciate them saying the quiet part out loud, I suppose! It's sort of like with Fauci admitting to "noble lies" about masking in the early days of the pandemic, the herd immunity threshold, etc.

[Vanity Fair offers an account of what happened](#). Their story starts in straightforward fashion:

On February 19, 2020, *The Lancet*, among the most respected and influential medical journals in the world, published a statement that roundly rejected the lab-leak hypothesis, effectively casting it as a xenophobic cousin to climate change denialism and anti-vaxxism. Signed by 27 scientists, the statement expressed “solidarity with all scientists and health professionals in China” and asserted: “We stand together to strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin.”

*The Lancet* statement effectively ended the debate over COVID-19’s origins before it began. To Gilles Demaneuf, following along from the sidelines, it was as if it had been “nailed to the church doors,” establishing the natural origin theory as orthodoxy. “Everyone had to follow it. Everyone was intimidated. That set the tone.”

It’s not clear how much *The Lancet* statement and other accusations of racism mattered, versus what was happening with Tom Cotton’s claims and the distortions of them ([which this piece argues was central](#)), and it documents the ways in which what he said was twisted around), versus worries about what Trump might do, versus stories told by defenders of gain of function research like Dr. Fauci or fears that our funding of it could look quite terrible, versus general media information cascades. The suppression of the lab leak hypothesis seems overdetermined.

It’s also not clear how distinct those causes are, as the *Lancet* letter [seems to have been organized and largely signed by those with a vested interest in gain of function research](#), before [it was used as a justification for suppressing dissent on Facebook](#) and elsewhere.

Then came the revelation that the *Lancet* statement was not only signed but organized by a zoologist named Peter Daszak, who has repackaged U.S. government grants and allocated them to facilities conducting gain-of-function research—among them the WIV itself. David Asher, now a senior fellow at the Hudson Institute, ran the State Department’s day-to-day COVID-19 origins inquiry. He said it soon became clear that “there is a huge gain-of-function bureaucracy” inside the federal government.

The Vanity Fair story is excellent, but very long. This passage also clarifies the situation:

On February 14, 2020, to the surprise of NSC officials, President Xi Jinping of China announced a plan to fast-track a new biosecurity law to tighten safety procedures throughout the country’s laboratories. Was this a response to confidential information? “In the early weeks of the pandemic, it didn’t seem crazy to wonder if this thing came out of a lab,” Pottinger reflected.

Apparently, it didn’t seem crazy to Shi Zhengli either. A *Scientific American* article first published in March 2020, for which she was interviewed, described how [her lab had been the first to sequence the virus](#) in those terrible first weeks. It also recounted how:

[S]he frantically went through her own lab’s records from the past few years to check for any mishandling of experimental materials, especially during disposal. Shi breathed a sigh of relief when the results came back: none of the sequences matched those of the viruses her team had sampled from bat caves. “That really took a load off my mind,” she says. “I had not slept a wink for days.”

Even if we take Shi Zhengli at her word, these are not the actions of someone who thinks that a lab leak is not a plausible explanation, or whose view of the safety of her work implies she should be allowed to keep doing it.

Also, there’s this, which certainly feels like strong (although of course in no way conclusive) evidence.

## IX. Dueling Memos

By the summer of 2020, the State Department's COVID-19 origins investigation had gone cold. Officials in the Bureau of Arms Control, Verification, and Compliance went back to their normal work: surveilling the world for biological threats. "We weren't looking for Wuhan," said Thomas DiNanno. That fall, the State Department team got a tip from a foreign source: Key information was likely sitting in the U.S. intelligence community's own files, unanalyzed. In November, that lead turned up classified information that was "absolutely arresting and shocking," said a former State Department official. Three researchers at the Wuhan Institute of Virology, all connected with gain-of-function research on coronaviruses, had fallen ill in November 2019 and appeared to have visited the hospital with symptoms similar to COVID-19, three government officials told *Vanity Fair*.

While it is not clear what had sickened them, "these were not the janitors," said the former State Department official. "They were active researchers. The dates were among the absolute most arresting part of the picture, because they are smack where they would be if this was the origin." The reaction inside the State Department was, "Holy shit," one former senior official recalled. "We should probably tell our bosses." The investigation roared back to life.

The lab did at least allow for rapid sequencing of the virus, once the problem became clear, which seems completely distinct as an action from any ongoing research that might have been dangerous (including gain of function research):

It's hard to think of anyone, anywhere, who was better prepared to meet the challenge of COVID-19. On December 30, 2019, at around 7 p.m., Shi received a call from her boss, the director of the Wuhan Institute of Virology, according to an account she gave to *Scientific American*. He wanted her to investigate several cases of patients hospitalized with a mysterious pneumonia: "Drop whatever you are doing and deal with it now."

The next day, by analyzing seven patient samples, her team became one of the first to sequence and identify the ailment as a novel SARS-related coronavirus. By January 21, she had been appointed to lead the Hubei Province COVID-19 Emergency Scientific Research Expert Group. At a terrifying moment, in a country that exalted its scientists, she had reached a pinnacle.

Even now, places like the Washington Post are trying out various rhetorical ways to suppress and doom dissent, with one recent attempt being [to associate it all, once again, with the Bad Orange Man](#). You see, this is all about 'distracting' from 'Trump failures' that actual no one (except Trump himself, probably) was still talking about. They warn about the dangers of a 'false narrative' of what happened, where all the blame for everything isn't pinned on Trump, and thus anything that doesn't do that must be a Trump op. [Play into any other narrative, they say, and you're helping Trump, who cares if your statements are 'accurate.'](#) Classic stuff. Play the hits.

So, basically, they started out gaslighting us, then switched to gaslighting us about whether they were previously gaslighting us, while continuing to gaslight us further using different words.

They are also explicitly saying that when the truth conflicts with politics we should bury the truth, with explicit calls to scapegoat anyone who fails to do this.

Which was also the playbook on so many other issues.

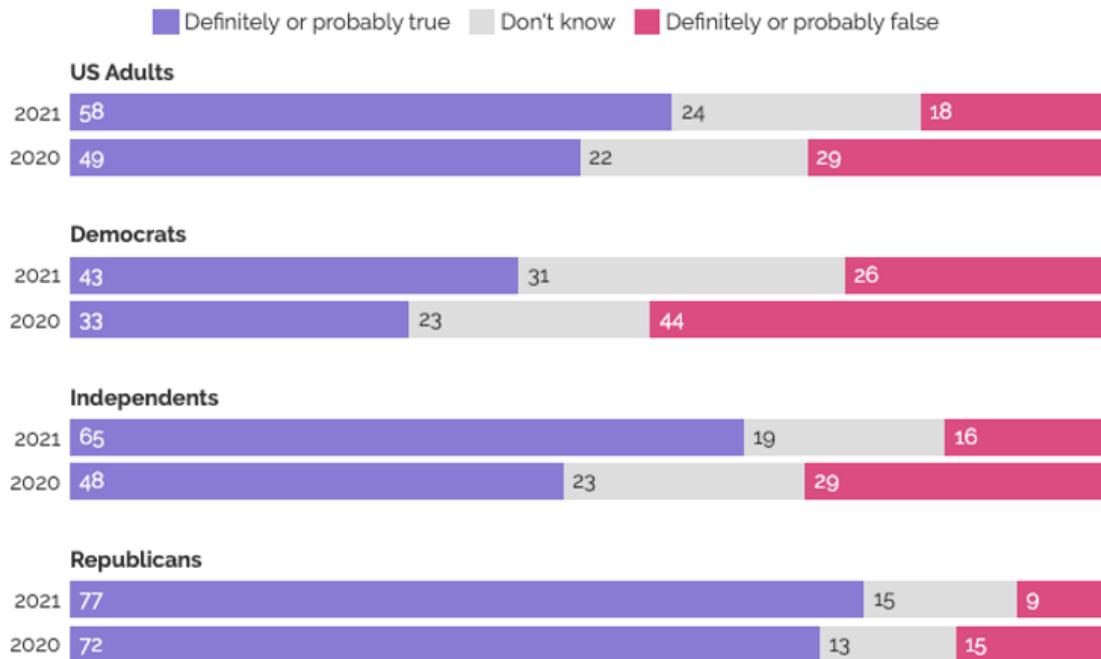
[Biden is launching an 'official investigation' into this](#), because he had no choice but to do so, but he is doing so by asking the people whose job it is to decide which secret things to keep secret because of 'national interests.' When they come back with nothing, I will update very little.

It seems that every day I see new individual data points that look highly suspicious. Some of them have plausible alternative explanations, while others are clearly suspicious people acting suspiciously.

The American people are having none of it (YouGov).

## Most Americans now believe a laboratory in China was the origin of the virus responsible for COVID-19

Regardless of whether or not it was created or naturally mutated, do you believe it is true or false that a laboratory in China was the origin of the virus responsible for COVID-19? (%)



YouGov

The Economist / YouGov | May 2, 2020 - June 1, 2021 | [Get the data](#)

(In other news, [I took a poll to see what most people think most means in context, and most of them couldn't agree on it.](#))

Not only that, 42% think it was *created in the lab*, and 24% of Americans think it was both *created in the laboratory* and then *released on purpose*, versus only 13% (!) of Americans buying the line that this was a natural mutation that occurred in the wild.

# One-quarter of Americans believe the coronavirus was created in a laboratory and released on purpose

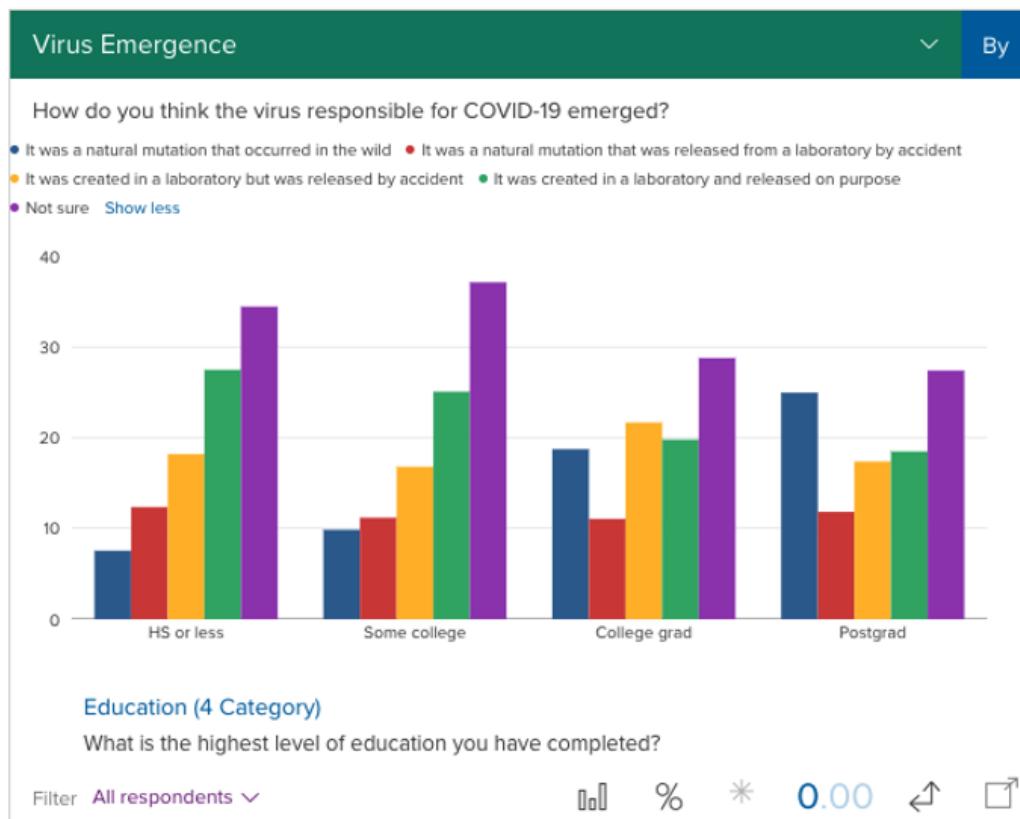
How do you think the virus responsible for COVID-19 emerged? (%)

|  | US Adults | Democrats | Independents | Republicans |
|--|-----------|-----------|--------------|-------------|
| It was a <b>natural mutation</b> that occurred in the wild                       | 13        | 21        | 10           | 4           |
| It was a <b>natural mutation</b> that was released from a laboratory by accident | 12        | 13        | 12           | 8           |
| It was created in a <b>laboratory</b> but was released by accident               | 18        | 16        | 22           | 22          |
| It was created in a <b>laboratory</b> and released <b>on purpose</b>             | 24        | 12        | 26           | 39          |
| Not sure   | 33        | 38        | 31           | 27          |

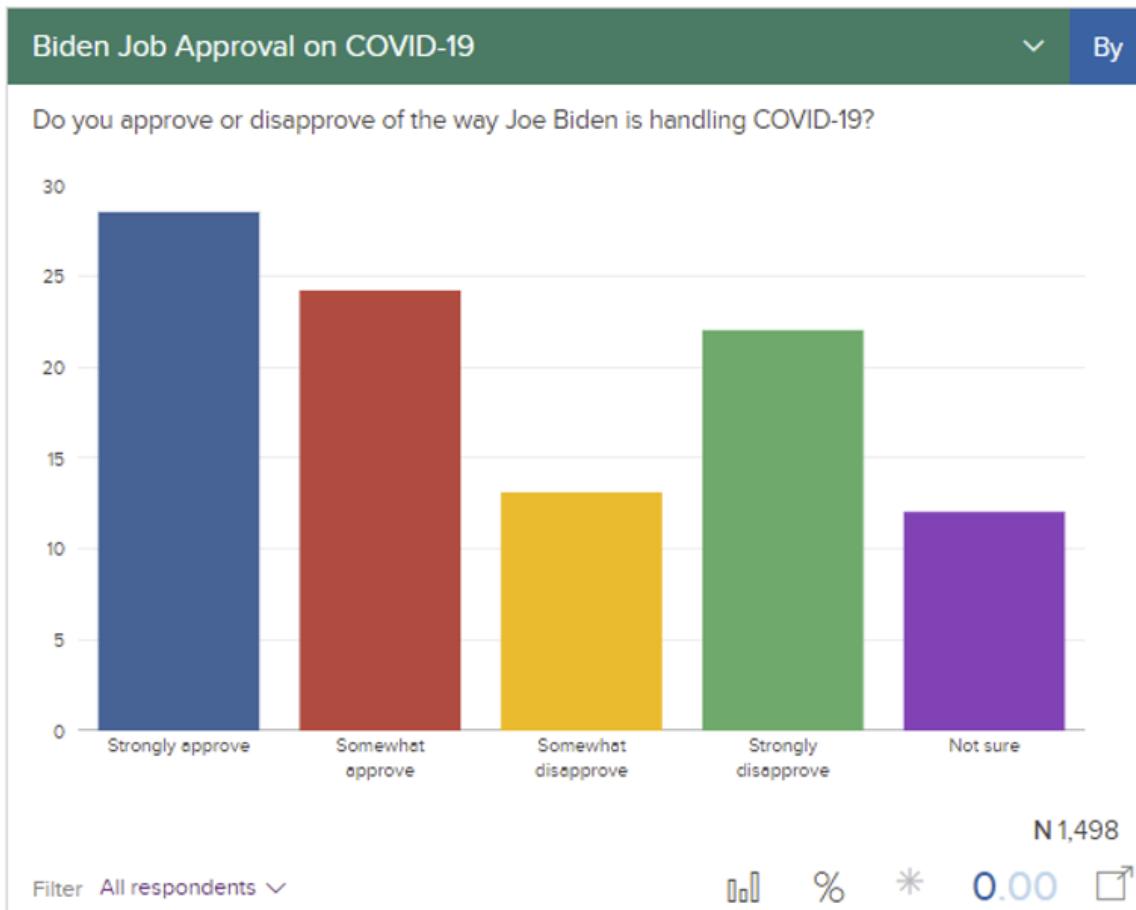
YouGov

The Economist / YouGov | May 29 - June 1, 2021 | [Get the data](#)

Here it is broken up by education level.

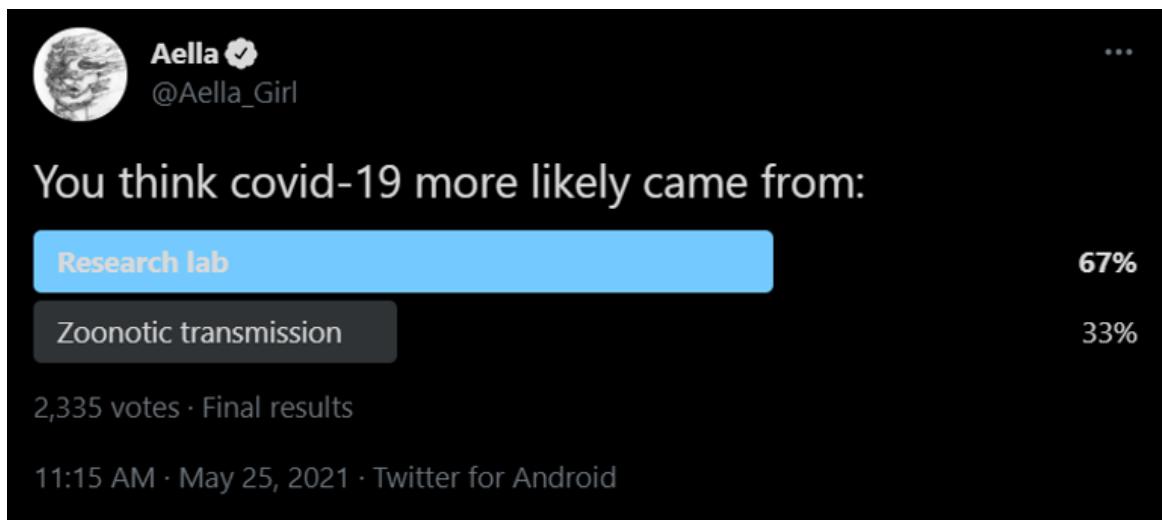
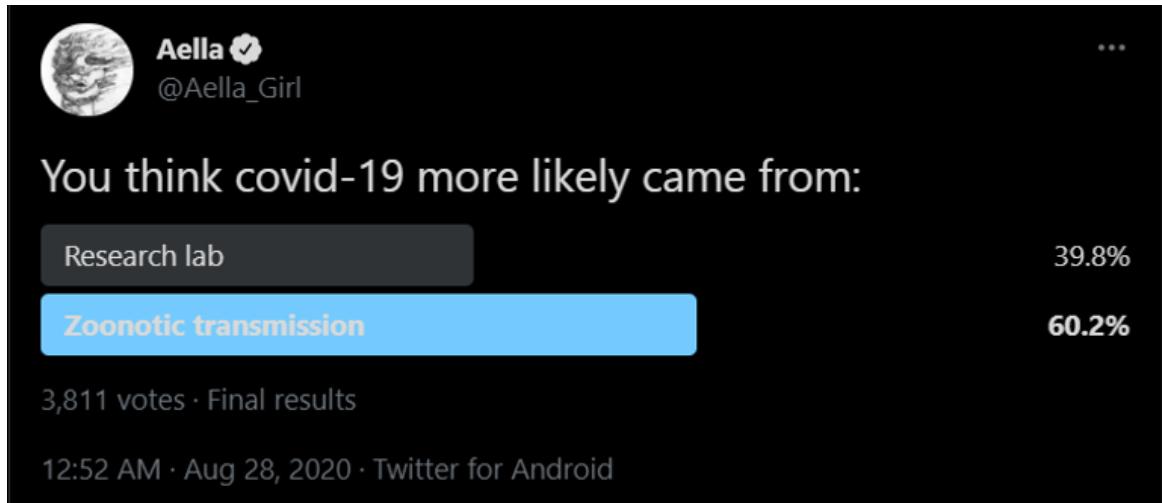


But, of course, it's not like pursuing this matters that much to people, as judged by Biden's approval on the issue, which is quite good:



There is a *correct answer* to the question of Covid's origins, from the perspective of those answering this survey, and it is: Not Sure. The confidence level that should be required to select something *other than* Not Sure in such a survey is highly unclear, but I can't see how (barring having material nonpublic information) one can have sufficient confidence to pick an explanation, at this point. That's especially true given the lab hypothesis gets split here into three subcategories.

The majority believing the lab leak hypothesis is probably true is new, but there have always been a lot of Americans who believed it, even when it was officially 'debunked' and being censored and suppressed across platforms. [Here's two Aella polls, from August 28, 2020 and then recently on May 25, 2021.](#)



This seems like a very reasonable amount of updating on new information.

I've left a bunch of stuff out, because I felt it was bringing more heat than light, or there were plausible alternative explanations, or to keep length at a reasonable place. This accounting is nowhere near complete.

The question one must ask is, was all this new information more likely in the worlds where there was a lab leak, versus the worlds where there wasn't one? Do most of the people involved in this know what happened any more than we do? Or is this all the kind of covering up and automatic suppression of wrongthink that would happen *regardless of the underlying facts?*

I mean, sure, [stuff like this...](#)



**Mike Solana** @micsolana

2h

just hopped over to the wiki page on investigations into covid's origin, and their primary evidence the lab leak hypothesis is bullshit appears to be four broken citations and a link to then joint WHO-china report in which china simply explained the hypothesis was bullshit

Politicians and some scientists have made unsubstantiated speculation that the virus may have accidentally escaped from the Wuhan Institute of Virology.<sup>[46][51]</sup> This has led to calls in the media for further investigations into the matter.<sup>[52][53]</sup> Many virologists who have studied coronaviruses consider the possibility very remote.<sup>[54][50][55]</sup> The WHO–China joint study report from March 2021 stated that such an explanation is extremely unlikely.<sup>[56][34]</sup>

...and that's all true as far as it goes, but we already know China can't provide better evidence than that or it would have already done so, so none of that is news.

[Mike Solana later fleshed out his views more fully here](#), documenting how deeply messed up the whole discourse around the lab leak hypothesis has been.

[This \(via MR\) thread](#) seems like a reasonable case arguing for a zoonotic origin for Covid-19. I don't agree if many of its points, especially the framing that a lab leak requires lies upon lies and a very narrow path, but I do still see no strong reason this couldn't have been zoonotic.

Thus, I'm not changing my probability all that much. I'm now at something like 55%, up from 40%, now that I see more about how this all went down, but it could still easily have gone down either way.

All right, suppose it did come from the lab. Does it matter?

Yes, quite a lot.

[Here's MR](#) (with a quote from Ross Douthat) explaining some ways it matters.

One big reason is it transforms our view of the pandemic, in terms of what stories various nations can tell themselves about what happened, and what it implies for international relations.

Another big reason it matters is Gain of Function research, which on reflection seems increasingly like a completely bonkers insane thing to do, and also the need for better lab safety protocols in general.

## Against Gain of Function Research

Whether we ban Gain of Function Research will be a key test of our civilization.

Even if we could somehow prove that Covid-19 definitely was a natural mutation in the wild and no laboratory had anything to do with it, that does not change the fact that doing Gain of Function research seems to be *completely and utterly James Bond villain level insane*.

The concept of Gain of Function Research is to take a virus that could potentially mutate or be engineered in a lab into a future pandemic or bioweapon, and *engineer it into that new more dangerous form first, in the lab*, so we can learn from it.

Regardless of the origins of Covid-19, [here's the Wikipedia article](#), this is totally a thing we actually do and not the prelude to a new posthumous Michael Crichton novel.

If you want to make a new vaccine to *stop* a deadly pandemic, normal procedure requires careful phased trials and years of study. Getting approval to release it into the wild took ten months and that timeline was considered a major miracle.

*Making a new potential pandemic in the first place* requires, well, let's quote Wikipedia:

Some forms of gain of function research (specifically work which involves certain select agent pathogens) carry inherent biosafety and biosecurity risks, and are thus also referred to as dual use research of concern (DURC). To mitigate these risks while allowing the benefits of such research, various governments have mandated that DURC experiments be regulated under additional oversight by institutions (so-called institutional "DURC" committees) and government agencies (such as the NIH's recombinant DNA advisory committee). A mirrored approach can be seen in the European Union's Dual Use Coordination Group (DUCG).

Importantly, the US and EU regulations both mandate that an unaffiliated member of the public (or several) be "active participants" in the oversight process. Significant debate has taken place in the scientific community on how to assess the risks and benefit of gain of function research, how to publish such research responsibly, and how to engage the public in an open and honest review. As of January 2020, the United States is convening an expert panel to revisit the rules for gain of function research and provide more clarity in how such experiments are approved, and when they should be disclosed to the public.

Does that make you feel better? That there's an 'unaffiliated member of the public' present? Or that we're gathering an 'expert panel' to reveal exactly when such experiments should be disclosed to the public, which implies they often aren't disclosed?

The steelman of the other side is that the information is valuable enough that it is worth the risk. [Why don't you take it away...](#) Dr. Fauci? Is that you? Huh ([original op-ed](#))...



## Opinions

Editorial Board

The Opinions Essay

Global Opinions

Voices Across America

Post Opinión

### Opinions

# A flu virus risk worth taking

By **Anthony S. Fauci**, **Gary J. Nabel** and **Francis S. Collins**

December 30, 2011

*Anthony Fauci is director of the National Institute of Allergy and Infectious Diseases (NIAID), Gary Nabel is director of the Vaccine Research Center at the National Institute of Allergy and Infectious Diseases and Francis Collins is director of the National Institutes of Health.*

Along with support for this research comes a responsibility to ensure that the information is used for good. Safeguarding against the potential accidental release or deliberate misuse of laboratory pathogens is imperative. The engineered viruses developed in the ferret experiments are maintained in high-security laboratories. The scientists, journal editors and funding agencies involved are working together to ensure that access to specific information that could be used to create dangerous pathogens is limited to those with an established and legitimate need to know.

Yep, it's him, arguing explicitly that we should engineer deadly viruses that *could* evolve in nature, with the stated goal being to better be able to identify them should they evolve in the wild, giving us more time to respond.

I've also heard the similar argument that another lab might engineer such a thing as a bioweapon, so in order to know what we might be up against and plan against it we need to engineer it first, which is totally (1) exactly how the world is going to die from an AGI if we don't stop being this level of grade-A stupid, (2) being done under conditions whose track record of accidents does not, shall we say, inspire confidence, (3) exactly how someone tricks you into making it so they can steal it out of your lab, have you not seen *any* movies or shows, like, ever, and (4) exactly what someone making a bioweapon in order to have a bioweapon would say.

If the argument is purely to defend against a future pandemic, then (1) yeah given how we safeguard such things it's still exactly how we are going to die of an AGI if we don't stop being this level of grade-A stupid, (2) being done under conditions whose track record of accidents does not, shall we say, inspire confidence, (3) exactly how someone tricks you into making it so they can steal it out of your lab, have you not seen *any* movies or shows, like, ever, and (4) exactly what someone making a bioweapon in order to have a bioweapon would say.

That alone isn't definitive proof we shouldn't do it. One can imagine a world in which our safeguards are good enough, and the threat of a natural pandemic bad enough, and the

ability of such research to help us prepare for and prevent that pandemic sufficiently stronger than our alternatives, that (with some sufficiently strong level of precautions) we should go ahead and do it anyway.

To be clear, ‘some sufficiently strong level of precautions’ is something like ‘do it in Antarctica and the quarantine for leaving a 100-mile radius around the lab is a year or more,’ not ‘do it in China next to a city but have additional protective equipment and a second observer present.’

Can you imagine the ‘environmental impact statement’ involved if these risks were being taken seriously? Can you imagine getting ‘informed consent’ from everyone put at risk (by which we mean, actual everyone)? How is this remotely consistent with how we handle potentially dangerous decisions in other realms?

We literally couldn’t put a few enthusiastically consenting 18-year-olds into a room to find out how Covid-19 actually spreads or whether or vaccines worked, instead choosing to lock everyone up for a year doing an unknown mix of some of the sensible actions combined with a bunch of superstitious nonsense, and you think *making the virus that could infect us* is the low-hanging fruit that makes sense?

Also, at this point I have zero faith that if we decided on reasonable precautions that were actually reasonable if followed, that those procedures would get followed, even by those who said they were following them. There would also be those who saw this as permission to do the research without even saying they would use the precautions. Either you ban this, or you don’t.

A sane civilization could reasonably decide that pandemics are a really big deal, and we are willing to take big swings and accept big risks in order to prevent or contain them. Such a sane civilization would be investing billions in pandemic preparedness across a variety of fronts, including building vaccine manufacturing capacity and laying the groundwork for rapid experimentation and human challenge trials. We’d have plans in place to detect a new pandemic as soon as possible, to implement testing and tracing, and to get vaccines going, and we wouldn’t (as we in fact did in 2019) cut what little such funding we had the moment we had a budget to balance and several years of quiet. Not only did we put a level-4 biolab in Wuhan, we then cut the funding that would have let it actually do its job.

As it is, even if this gain of function research was fully successful at getting us the information it is looking for, with none of the downside risks, what would we even be able to do with that information? Do we have any examples of gain of function research doing us any good and helping us stop a pandemic?

If we’d taken care of such matters first, both establishing benefits and establishing that we’re taking all (or at least most of) our free actions, then I’d be willing to entertain a cost-benefit analysis on research of this type, and whether we could figure out reasonable precautions.

As Scott Sumner puts it, [which lab is responsible matters little](#). As does whether a lab was responsible for this particular incident. What matters is that this is a crazy thing to be doing that is likely to get a lot of people killed at some point, and it must stop.

If we do manage to ban gain of function research, that will give me at least *some* hope that we can take reasonable actions some of the time. At a minimum, directly after a pandemic that shut down the world for a year, we are capable of stopping people from intentionally creating the next one and hoping it wouldn’t fall into the wrong hands and that there wouldn’t be an accident, despite a long history of accidents, and an unknown history of falling into the wrong hands (since many of those wrong hands would not talk about it or do anything we’d yet know about.)

If we *don't* manage to ban gain of function research, even after everything that has happened, then that's quite a reason to lose hope. It's not only about letting labs go ahead and create the next pandemic or bioweapon. How would we possibly hope to ban dangerous AI research in the future, if we failed at this much easier problem? How would we ever get anything worthwhile done?

I do not expect us to pass this test, but I don't think our failure is inevitable. I have wide error bars on this, but if I had to guess I'd give us a 25% chance to effectively ban gain of function research by the end of 2022.

Suppose we do fail, and gain of function research continues. What's the plan then, other than 'hope the actual gain of function research turns out non-disastrously'?

That's a good question. I don't think the answer is full despair. We ban lots of things for stupid reasons. If we can't ban (or render sufficiently toxic that it's an effective ban) the things that need banning for smart reasons, the 'get others to do it for dumb reasons because we have our own smart reasons' plan does not seem obviously doomed. It's likely actively harder than banning purely for dumb reasons, because people with power will be worried that there might be smart reasons or that others might think they cared about such smart reasons. That's unfortunate, but it still might be better than the alternative.

We could also use what time we have to try and fix the fundamental problem, by building up new mechanisms for changing public opinion and/or elite consensus, or otherwise reforming how things work. Things might be broken now but that doesn't mean they have to stay that way.

The next fallbacks after that are rather ugly, and deserve longer treatments, so I won't go into them here.

Credit where credit is due, [such as to the Cambridge Working Group](#) that has been fighting to stop this:



**Eliezer Yudkowsky**   
@ESYudkowsky

...

Retweet the ones who were right all along (and do it before the media decides that current elites have always been at war with Eastasia)



**Andrew Noymer** @AndrewNoymer · Jun 3

I have opposed gain of function research, since before being against GoF research was fashionable.

In July 2014, I was the first online signatory to the Cambridge Working Group statement.

[cambridgeworkinggroup.org](http://cambridgeworkinggroup.org)

Whatever the #CovidOrigin, I continue to oppose GoF work.

[Here's Kelsey Piper](#), who has also been advocating against this for a long time.



Kelsey Piper @KelseyTuoc · Jun 3

...

I'm on today's Today, Explained, to talk about how lab leaks happen. Did \*the\* lab leak happen? I don't know. But there's a long history of viruses escaping the lab, and this is a serious reason we shouldn't be doing gain-of-function research.



The lab leak theory - Vox

🔗 open.spotify.com

1

15

52



Kelsey Piper @KelseyTuoc · Jun 3

...

Making potentially pandemic pathogens more dangerous is justified with claims it could help us develop vaccines. But this research wasn't actually instrumental in inventing the Covid vaccines in record time. Other research was. And the cost-benefit analysis just doesn't check out.

2

1

20



Kelsey Piper @KelseyTuoc · Jun 3

...

Biologists do incredible work helping understand the natural world and making it stop killing us. But gain of function research doesn't contribute to that project nearly enough to justify the horrific human cost. It's long past time we stopped funding and publishing it.

2

2

19



Kelsey Piper @KelseyTuoc · Jun 3

...

And yes, I've been saying this since well before the pandemic. I've been saying this since I started as a journalist. Here I am a few months after Future Perfect launched:



Biologists are trying to make bird flu easier to spread. Can we not? This research into viruses could help us understand pandemics better - or it could cause one.

🔗 vox.com

There is the worry that things could go too far. [This thread explaining what virology](#) is comes from a place where the author is noticing calls for banning virus research entirely, and feeling the need to explain that labs need to experiment on real and dangerous viruses in order to learn how they work. If 'study of existing function' research gets shut down, that would indeed be quite bad. If a bunch of people who can't tell the difference call for it anyway, the same way there's been extreme reactions against nuclear power, that could be useful as a way to motivate better safeguards, but mostly that would be something we'd have to put up with and quietly ignore, since we can't stop studying viruses. Then again, we also couldn't stop using nuclear power without releasing an epic amount of extra carbon into the atmosphere, which people think is kind of a big deal, and yet here we are. So it's a concern.

## In Other News

Biden has announced we are buying 500 million doses of Pfizer to give to other countries. A good start. Better late and insufficient than never.

[The J&J pause was so disastrous we now have to worry about doses expiring \(WSJ\).](#)

[Moderna outright knows that fractional dosing works fine, and plans to use it on any future versions of the vaccine, but for now we continue not to use it, wasting a large percentage of our capacity \(MR\).](#) What the mix here is of regulatory concerns and profit maximization is unknown, but my guess is it is mostly or all regulatory.

[FDA approves an Alzheimer's drug, aducanumab, that its own advisory board did not want to approve](#), and which everyone whose judgment I respect that has weighed in on this says probably doesn't work, [as in](#):

↪ Sarah Constantin Retweeted

 **Michael LaCroix-Fralish** @fralim · 20h ...  
This drug doesn't work

 **STAT** @statnews · 21h  
BREAKING: FDA grants historic approval to Alzheimer's drug designed to slow cognitive decline [buff.ly/3w0LpZ3](http://buff.ly/3w0LpZ3)

6 7 49

Standard explanation is that this is the FDA caving in to lobbying from patient groups and justifying it based on transparent statistical manipulations. You know what? In principle, I'm totally fine with that. Provided the drug is safe, which no one seems to be challenging, why should we deny it to patients that want it?

A fine answer is 'it costs a lot of money and because of health insurance the rest of us will be paying for it.' Given how much they are charging for the drug (which, in our current system, is pretty much 'whatever they want' because of various reasons) [this could get quite expensive](#):



**Walid Gellad, MD MPH** @walidgellad · 1h

All of this needs to be repeated until the public can digest the significance here.

...

If medicare puts no constraints on it, any doctor can order this drug on a patient with Alzheimer's diagnosis.

Every doctor prescribing this will make money when they do.



**Zach Brennan** @ZacharyBrennan · 1h

Gal: "If 50% of newly diagnosed patients start Aduhelm at the current price point, then the total cost to Medicare will be equal to the top five drugs in Medicare part B combined."

[Show this thread](#)

I gotta say, [statements like this](#), at this point, mostly make me giggle:



**Walid Gellad, MD MPH** @wal... · 1h

**"Following the science"**

**"Value"**

**"Innovation"**

**"Advisory committee"**

**Words that have no meaning  
anymore.**



1



7



Cause, I mean, sure, fair, absolutely, I hear you all the way, but seriously, after everything that's happened, *this* is where you draw the line? How did you miss all the previous memos?

[When mask guidance isn't based on physical reality, what to do?](#)

[Fractional dosing of AZ to be tested in Brazil \(MR\).](#)

[Some small signs of progress on rapid testing \(MR\).](#)

[Texas hospital suspends 200 workers for refusing Covid-19 vaccine](#). Governor has vowed to stop this sort of thing, so we'll see how that goes. This seems obviously good and right.

[Indian village refuses Covid-19 vaccinations and also payment of their power bills, power is shut off](#). Given how big a shortage of vaccines India has, the reaction seems more than a little extreme, but at some point the electric bill does have to be paid.

[Thread of reports of people being charged \(<\\$100, but still\) for the Covid vaccine.](#)

[United Airlines joins Delta, won't hire new unvaccinated workers](#). I recently hired someone without asking about their vaccine status, but I'm very confident they're vaccinated, and also I'm not an airline. Asking seems massively overdetermined.

[CDC attempts to scare adolescents by claiming hospitalizations among them are rising are blatant lies](#). The charitable interpretation is that adolescents are a larger *percentage of hospitalizations* due to who has been vaccinated so far, which is true, but that does not seem like the relevant data to any decision one might make.

## FDA Delenda Est News

[It seems this happened:](#)



**Matt Kaeberlein** @mkaeberlein · 8h

I hate to say it, but it sure seems like what appears to be a boneheaded FDA decision to disallow \*laboratory confirmed\* infections and replace that with patient reported symptoms as the endpoint may have torpedoed the first successful geroscience trial.

...



**Matt Kaeberlein** @mkaeberlein · 8h

So happy to see this finally come out! Important insights into why the resTORbio phase 3 failed.

Targeting the biology of ageing with mTOR inhibitors to improve immune function in older adults: phase 2b and phase 3 randomised trials [thelancet.com/journals/lanhl...](http://thelancet.com/journals/lanhl...)

1

5

19

↑



**Matt Kaeberlein** @mkaeberlein · 8h

This decision was made "because of concerns that laboratory confirmation of an infection was not relevant to how patients feel". Apparently, it doesn't matter whether you actually have an infection, just how you feel 🤦‍♂️🤦‍♂️🤦‍♂️

1

1

14

↑

I mean, it sure does look like it ([study](#)):

### **Phase 3 trial**

The US FDA requested a change in primary endpoint between the Phase 2b and 3 trials because of concerns that laboratory confirmation of an infection was not relevant to how patients feel and function. Therefore, the FDA proposed a primary endpoint of the proportion of patients with at least one clinically symptomatic respiratory illness, defined as symptoms consistent with an RTI, irrespective of whether the symptoms were confirmed via laboratory testing to be due to an infection. Therefore the primary objective of the study was to investigate whether RTB101 decreased the incidence of clinically symptomatic respiratory illness defined as symptoms consistent with an RTI, irrespective of whether an infection was laboratory-confirmed. The secondary objectives included to investigate whether RTB101 as compared with placebo decreased the proportion of patient with laboratory-confirmed clinically symptomatic respiratory illness; decreased the rate of clinically symptomatic respiratory illness associated with specific viruses; decreased the proportion of patients with severe symptoms due to clinically symptomatic respiratory illness; and the safety and tolerability of RTB101 as assessed by reports of adverse events and serious adverse events, physical examination, electrocardiogram findings, and safety laboratory values. Additional secondary objectives not reported in this manuscript included investigating whether RTB101 as compared with placebo decreased the rate of clinically symptomatic respiratory illness or the rate of laboratory-confirmed clinically symptomatic respiratory illness, and decreased the time to alleviation of moderate and severe symptoms of respiratory illness.

This seems completely bonkers. Why would one *actively request* a shift from measuring actual infections in the lab to measuring patient reported symptoms? That seems designed to introduce extra noise and ensure that the efforts fail, and likely to have a large chilling effect on future research.

As in, in the phase 2b trial results quoted below, they measured both, and found the ‘patient reported infections’ number didn’t have sufficient power, and then forced them to switch to that metric for phase 3:

Two statistical analysis of the primary efficacy endpoint were prespecified. The first accounted for multiplicity by using a fixed sequence testing procedure that was limited to data from Part 2 of the trial and controlled the overall type I error. In this step-down analysis, efficacy of RTB101 10 mg in combination with everolimus 0·1 mg once daily compared with placebo was first tested at a one-sided  $\alpha$  level of 0·05 and did not meet statistical significance (data not shown). Therefore, the testing procedure was concluded for this analysis. Because this was an exploratory phase 2b dose-finding trial, an additional analysis of the primary endpoint was prespecified that did not adjust for multiplicity. This analysis evaluated the proportion of subjects with laboratory-confirmed RTIs in each of the active treatment groups compared with placebo, and included all data from Parts 1 and 2 of the study. In this analysis we found a statistically significant reduction in the proportion of patients who had one or more laboratory-confirmed RTIs in the RTB101 10 mg once daily treatment group (34 [19%] of 176) compared with the pooled placebo group (50 [28%] of 180; OR 0·601 [90% CI 0·391–0·922];  $p=0·025$ ). RTB101 10 mg twice daily and RTB101 10 mg in combination with everolimus 0·1 mg once daily were not associated with a significant reduction in the incidence of laboratory-confirmed RTIs as compared with placebo (data not shown). The results suggest that intermittent inhibition of mTOR (predicted to be achieved with once daily RTB101 dosing) might be more effective than persistent inhibition (predicted to be achieved with twice daily RTB101 dosing or combination dosing with everolimus) at improving immune function and decreasing the incidence of RTIs.

Secondary endpoint analysis revealed that RTB101 was not associated with a reduced number of patients who had symptoms that met the diagnostic criteria for an RTI, irrespective of whether or not an infection was laboratory-confirmed compared to placebo (56 [32%] of 176 in the RTB101 10 mg once daily group vs 68 [38%] of 180 in the placebo group; OR 0·756 [90% CI 0·521–1·098];  $p=0·11$ ).

[This was sufficiently flagrant ignoring of the advisory committee that two of its members resigned](#). Seems a lot like the FDA is picking which things it wants to approve or disapprove for reasons other than ‘following the science.’

It’s worth noting that 20% of the time the FDA ignores the advisory panel, but it’s almost always the panel saying to approve and the FDA refusing, which it seems does not lead to protests and resignations. This provides clarity on the dynamics involved.

What does ‘following the science’ mean to these doctors? It means following the FDA:

After the FDA announced its approval on Monday, Billy Dunn, the director of the agency's Office of Neuroscience, [wrote in a letter](#) to the chairman of the advisory panel that the input from the committee in November prompted further discussions at the FDA and ultimately led to the decision to use the accelerated pathway.

Knopman said he would still offer the medication to his patients when it becomes available, describing that as a "completely separate issue."

"I feel ethically bound to offer the treatment because FDA is the law of the land, and the drug is commercially available," he said. "And I will present them with my side of the story, and with the FDA's side of the story" that reducing amyloid may offer a clinical benefit in the future.

It's a 'completely separate issue' whether or not to offer a drug (that costs \$56k/year) to patients, versus whether to approve the drug, you see. Once it is 'approved' then 'ethics' demand that the drug be offered, and that the rest of us pay the price, *even if you think the drug does not work*. That tells you what people mean when they talk about 'the science' and it's not making an accurate physical model of reality.

## Not Covid or the FDA, But Relevant To My Interests

[538 looks back on its 2020 election predictions](#), no great new insights, but linking because of my ongoing interest in predictions, prediction markets and modeling. [As I've said before](#), I think the criticisms of 538's forecasts and claims that the markets outperformed it were misplaced, and Nate's public analysis is accurate as far as it goes, but I worry that his public analysis is superficial and fails to ask the important questions.

[Wizards of the Coast bans Time Warp](#) in Historic, which makes me sad without actually accomplishing anything. Completely missing the point. Morale will not improve.

[Tyler Cowen asks whether gambling and prediction of outcomes in sports should be distinct from the operation of the sports themselves](#). I think Tyler is right to be concerned, but is concerned about the wrong things. The biggest thing I worry about is actually the advertising. Unifying the NBA with NBA betting would mean more and more obnoxious pressure to convert fans to gamblers, making the experience of watching or attending a game worse - I love a good gamble but the advertising of gambling (including fantasy sports) is obnoxious at best and usually highly toxic. Soccer has gambling websites getting placements on jerseys, which is where this ends.

Concerns about integrity seem reversed, as bookmakers highly value the integrity of their events. Concerns that leagues might create more gambling events via offering more exciting games also seems reversed, as the reason that creates more gambling is that it creates more excitement and fun. Regular seasons of endless mostly pointless games aren't good, and our choices are to either make the playoffs harder to reach in order to give those games meaning (which isn't going to happen), use relegation to give those games meaning (which is going to happen even less), or play less such games and/or more games that have other things going on.

The other big thing to worry about is modification of the sport or how the teams play it. Nascar has effectively butchered itself in the name of creating exciting events within a game, with lots of catch-up mechanics that seem good for board game night but render the idea of the first half or more of the event being a true 'race' somewhat moot. I'd worry about media breaks and play structures being moved around to accommodate more live betting. I'd worry about more catch-up mechanics designed to prevent blowout games from losing interest, but then again it could also mean you wouldn't need them if you were fighting over

the halftime line or point spread – remember the famous line “there’s still some business to be decided.”

I’d also worry *for the gamblers*. Prices are likely to be much higher, and practices generally less customer focused, with the leagues running the show.

Mostly though I think the whole thing is good. Capturing much or most of the revenue from gambling will make sports more profitable, which will result in reasons to have more teams, more leagues, higher pay all around and higher quality play. There will be that much more reason to give a *better* experience to fans rather than a worse one, since fans are now more valuable. Teams asking for public subsidies for stadiums or otherwise can be told where to go. That’s even more true at the college level, where being a bookmaker can enable schools to pay the players and make sports do a better job funding the university rather than being a money pit.

Finally, there’s [an interesting post out on better crowdfunding](#) (dominant assurance contracts) that I hope to respond to with a full post later this cycle.

# **Why do patients in mental institutions get so little attention in the public discourse?**

Scott writes in [My IRB nightmare](#) about how his colleague took people freedom away and justified it with tests that aren't validated for the purpose of diagnosis and that have a warning THIS IS JUST A SCREENING TEST IT IS NOT INTENDED FOR DIAGNOSIS.

This can only happen because there's very little public accountability and suggests that a lot of abuse of power is going on. While the US has a lower psychiatric hospitals bed count than many other [OECD countries](#), there are still ~80,000 people in those institutions. A lot of them effectively have very little rights and have to endure medical procedures without them consenting such as taking various drugs.

Mental health people seem to me like they should be classified as a marginalized group by the ideals of modern left. On the other hand the modern left put very little attention on fighting for their rights as evidenced by Scott's colleague getting away with major abuse of power.

Why is the state of affairs like it is? Why don't they get more attention?

# Big picture of phasic dopamine

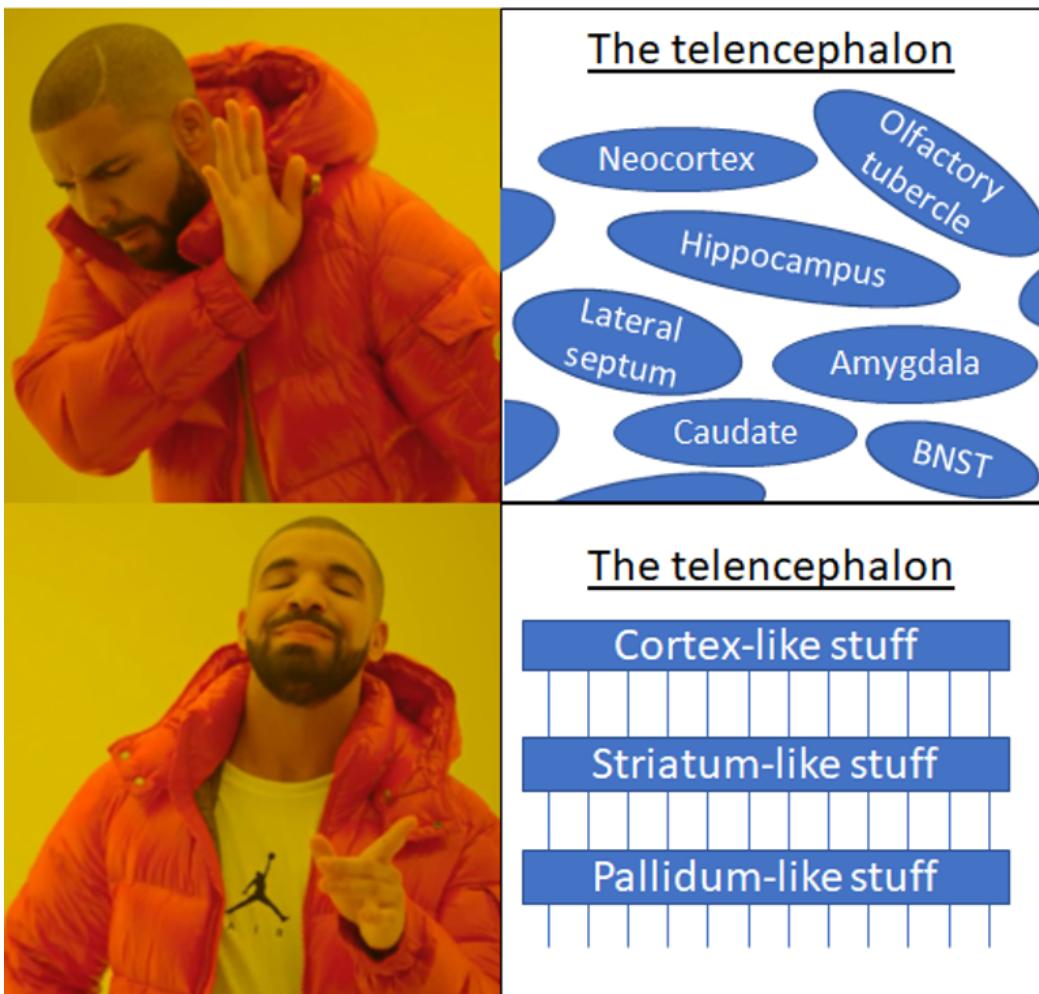
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(**Update Jan. 2023:** This article has important errors. I am leaving it as-is in case people want to see the historical trail of me gradually making progress. Most of the content here is revised and better-explained in my later post series [Intro to Brain-Like AGI Safety](#), see especially Posts 5 and 6. Anyway, I still think that this post has lots of big kernels of truth, and that my updates since writing it have mostly been centered around how that big picture is implemented in neuroanatomy.)

**Target audience:** Everyone, particularly (1) people in ML/AI, and (2) people in neuroscience. I tried hard to avoid jargon and prerequisites. You can skip the parts that you find obvious.

**Context:** I'm trying to make sense of dopamine in the brain—and decision-making and motivation more generally. This post is me playing with ideas; expect errors and omissions (and then tell me about them!).

This post is a bit long; I'm worried no one will read it. So in a shameless attempt to draw you in, here's a Drake meme...



This will make more sense later in the post. Yes of course the bottom one is oversimplified, and yes of course the top one is useful too. But I claim that the bottom one is a better starting point.

(Thanks Adam Marblestone, Trenton Bricken, Beren Millidge, Connor Leahy, Jeroen Verharen, Ben Smith, Adam Shimi, and Jessica Mollick for helpful suggestions and criticisms.)

## Summary / Table of Contents

- I'll start by briefly reviewing the famous 1990s story relating the role of dopamine in the brain to the "[TD learning](#)" algorithm.
- Then I'll switch over to the "cortico-basal ganglia-thalamo-cortical loop" ("loop" for short), a circuit intimately related to dopamine learning and inference.
  - I'll go over a theory that the entire [telencephalon](#) region of the brain (which includes neocortex, hippocampus, amygdala, basal ganglia, and more) has a coherent overall architecture, with these loops being the key unifying element, and I'll relate that idea to some bigger-picture ideas about within-lifetime learning vs instinctive behaviors.
  - I'll offer a toy model of how the loops work.
- I'll suggest that there are three more specific categories of what these loops are doing in the brain, with different dopamine signals in each case:
  - Reinforcement learning loops, where the "reward function" is the thing we all expect (yummy food is good, electric shocks are bad, etc.);
  - Reinforcement learning loops, but with a *different* "reward function", one narrowly tailored to a particular subsystem;
  - Supervised learning loops, for example to train some loop in the amygdala to fire whenever it's an appropriate time to flinch.
- Then I'll suggest that the way we get from "reward" to "reward prediction error" (as needed in the learning algorithm) is via one of those supervised learning loops, tasked with predicting the upcoming reward. And in particular (following Randall O'Reilly's "PVLV" model) I'll suggest that in fact no part of the brain is doing TD learning after all!
- Finally—and inevitably, given [my job](#)—at the end I'll discuss some takeaways for the Artificial General Intelligence Control Problem.

## Backstory: Dopamine and TD learning on computers and in animals

### Reward Prediction Errors (RPEs), and Temporal Difference (TD) learning

If you haven't read [The Alignment Problem](#) by Brian Christian, then you should. Great book! I'll wait.

...Welcome back! As you now know (if you didn't already), Temporal Difference (TD) learning ([wiki](#)) is a reinforcement learning algorithm invented by Richard Sutton in the 1980s. Let's say you're playing a game with a reward. As you go through, you keep track of a *value function*, a.k.a. "expected sum of future rewards". (I'm ignoring time-discounting for simplicity.)

How do you know the expected future reward? After all, predictions are hard. So you start with a random function or constant function or whatever as your value function, and update it from experience to make it more and more accurate. For example, maybe you seem to be in a very bad chess position, with the queen exposed in the center of the board. Then you make a move, and then your opponent makes a move, and then all of the sudden you're in a very very good position! Well, hmm, maybe your old position, with the exposed queen, wasn't quite as bad as you thought!! So next time you're in a similar situation, you'll be a bit more optimistic about things—i.e., you now assign that position a higher value.

That's an example of a positive reward prediction error (RPE). The general formula for RPE in TD learning is:

$$\text{Reward Prediction Error} = \text{RPE} = [(\text{Reward just now}) + (\text{Value now})] - (\text{Previous value})$$

...and this RPE is used to update the previous value.

So that's TD learning. If you keep iterating, the value function converges to the desired "value = time-integral of expected future rewards".

## The Schultz dopamine experiments

Now in the 1980s-90s, Wolfram Schultz did some experiments on monkeys, while measuring the activity of dopamine neurons in the midbrain.

I'll pause here to help the non-neuroscientists follow along. In the [midbrain](#) (part of the brainstem) are two neighboring regions called "[VTA](#)" and "[SNC](#)". In these regions you find the inputs and cell bodies (dendrites and somas) of almost all the dopamine-emitting neurons in the brain. These neurons' axons (output lines) then generally exit the midbrain and go off to various distant regions of the brain, and that's where they dump their dopamine.

Anyway, Schultz found (among other things) three intriguing results:

- When he gave the monkeys yummy juice at an unexpected time, there was a burst of dopamine.
- After he trained the monkeys to expect yummy juice right after a light flash, there was a burst of dopamine at the light, and *not* at the juice.
- When he then presented the light but surprisingly *omitted* the expected juice, there was *negative* dopamine release (compared to baseline) at the time when there normally would have been juice.

Peter Dayan and Read Montague saw the connection: All three of these results are perfectly consistent with dopamine being the RPE signal of a TD learning algorithm! This became a [celebrated and widely-cited 1997 paper](#), and a cornerstone of much neuroscience research since.

Oh, one more terminology side note:

- [\*"Phasic dopamine"\* vs \*"Tonic dopamine"\*](#): The dopamine signal has some steady (slowly-varying) level, and it also has periodic divergences up or down from that level. "Tonic" is the former, "Phasic" is the latter. The reward prediction error (RPE)-like signal is generally found in phasic dopamine. The role of tonic dopamine remains (even) more controversial than phasic. For my part, I don't think tonic and phasic are *that* different—I'm rather fond of the very simple idea that tonic is more-or-less a rolling average of phasic—but I'm not sure, and that's a whole separate can of worms. In this post I'm only going to talk about phasic dopamine. Well, I'll also talk about a subset of dopamine neurons that have weird patterns of current, such that I'm not sure they even have a phasic / tonic distinction.

## Some wrinkles in the story

There are a number of wrinkles suggesting that there's more to the story than simple TD learning:

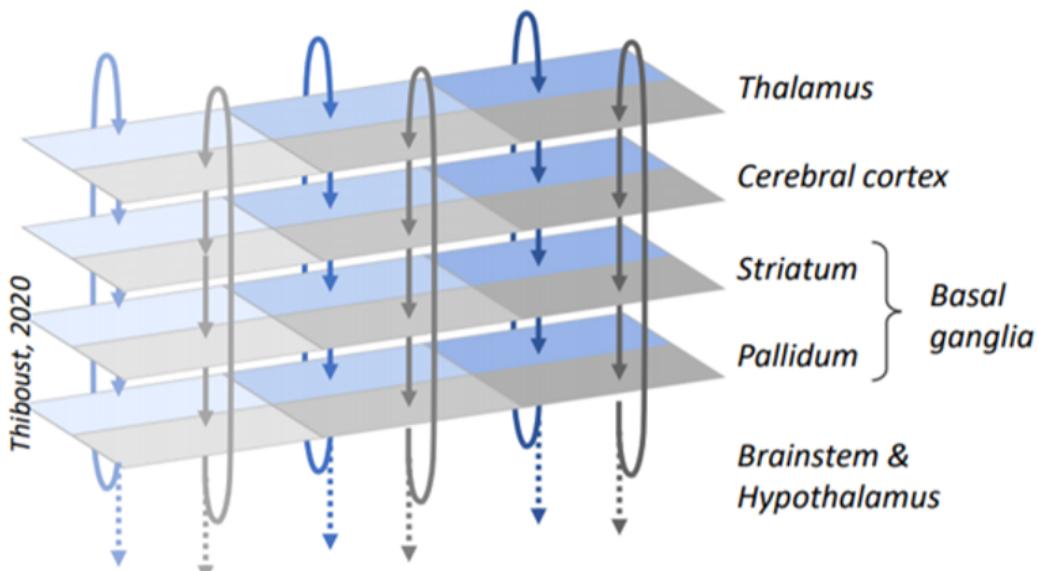
- While the dopamine signals agree with the TD learning predictions in Wolfram Schultz's experiments above, there are lots of other situations where the dopamine neurons are doing something different from what TD would do. For example, in the light-then-juice experiment, before training there's dopamine at the juice, and after training there's dopamine at the light but not at the juice. What about *during* training? Well if this were TD learning, dopamine-at-the-light should go *up* in perfect synchrony with dopamine-at-the-juice going *down* to baseline. But in experiments, they're not synchronized; the former happens faster than the latter. This is just one example; more are in [Molllick et al. 2020](#).
- I talked about "phasic dopamine" as a monolithic thing, but really there are lots of dopamine neurons, and *they are not a homogeneous group!* My impression is that there's a big subset of these neurons whose signaling seems to match RPE, but there are also various subpopulations that seem to be doing other things. There even seem to be some dopamine neurons that fire at *aversive* experiences! ([ref](#))
- TD learning is a learning algorithm (hence the name!), but dopamine seems to be related to motivation as well as learning—for example, "after forebrain dopamine depletion, animals will cease to actively search for food (and eventually starve to death), but they will still consume food when it is placed in their mouth" ([ref](#)). What's the deal with that?

To make sense of these facts, and much more, let's dive deeper into how the brain is built and what different parts are doing!

# The cortico-basal ganglia-thalamo-cortical loop

## The loop

Dopamine is closely related to a circuit called the [cortico-basal ganglia-thalamo-cortical loop](#). I'll just call it "loop" for short.



Simplified cartoon illustration of how the brain has many parallel cortico-basal ganglia-thalamo-cortical loops. Copied with permission from [Mathieu Thiboust](#).

A [classic 1986 paper](#) found that a bunch of brain circuitry consists of these loops, running in parallel, following a path from cortex to basal ganglia (striatum then pallidum) to thalamus and then back to where it started. (If you're not familiar with the neuroanatomy terms here, that's fine, I'll get back to them.)

(*Side note to appease the basal ganglia nerds:* This loop is real, and it's important, but I'm leaving out various other branches and supporting circuitry needed to make it work—see [the frankly terrifying Fig. 6 here](#). I'll get back to that below, but basically this simplified picture of the loop will be good enough to get us through this post.)

What are these loops doing? That's going to be a big theme of this blog post. I'll get back to it.

Where are these loops? *All over the telencephalon*. And what is the telencephalon? Read on:

## The telencephalon: simpler than it looks, and full of loops

The [telencephalon](#) (aka "cerebrum") is one of ~5 major divisions of the brain, differentiating itself from the rest of the brain just a few weeks into human embryonic development. The telencephalon is especially important in "smart" animals, comprising 87% of total brain volume in humans ([ref](#)), 79% in chimps ([ref](#)), 77% in certain parrots, 51% in chickens, 45% in crocodiles, and just 22% in frogs ([ref](#)). The human telencephalon consists of the neocortex ("the home of human intelligence", more-or-less), some non-"*neo*" cortex areas like the hippocampus (classified as "allocortex", more on which below), as well as the basal ganglia, the amygdala, and various more obscure bits and bobs. It seems at first glance that "telencephalon" is just a big grab-bag of miscellaneous brain parts—i.e., a category that only embryologists have any reason to care about. At least, that was *my* working assumption. ...Until now!

It turns out, when people peered into the telencephalon, they found a *unifying structure hidden beneath!* The breakthrough, as far as I can tell, was [Swanson 2000](#). I learned about it mainly from the excellent book [The Evolution of Memory Systems by Murray, Wise, Graham](#). (Or see [this shorter paper by the same authors with the relevant bit](#).)

It turns out that there's a remarkable level of commonality across these superficially-different structures. The amygdala is actually a bunch of different substructures, some of which look like cortex, and others that look like striatum. There's a thing called the "lateral septum" whose neurons and connectivity are such that it looks like "the striatum of the hippocampus". *And practically everything is organized into those neat parallel loops through cortex-like, striatum-like, and pallidum-like layers!*

| Cortex-like part of the loops   | Hippo-campus   | Amygdala [basolateral part] | Piriform cortex       | Ventromedial prefrontal cortex | Motor & "planning" cortex |
|---------------------------------|----------------|-----------------------------|-----------------------|--------------------------------|---------------------------|
| Striatum-like part of the loops | Lateral septum | Amygdala [central part]     | Olfactory tubercle    | Ventral striatum               | Dorsal striatum           |
| Pallidum-like part of the loops | Medial septum  | <a href="#">BNST</a>        | Substantia innominata | Ventral pallidum               | Globus pallidus           |

*The entire telencephalon—neocortex, hippocampus, amygdala, everything—can be divided into cortex-like structures, striatum-like structures, and pallidum-like structures. If two structures are in the same column in this table, that means they're wired together into cortico-basal ganglia-thalamo-cortical loops. This table is incomplete and oversimplified; for a better version see Fig. 4 [here](#).*

So many loops! Loops all over the place! Are *all* those different loops doing the same type of calculation? Let's take that as a hypothesis to explore, and see how far we get. (Preview: I'm actually going to argue that the loops are *not* all doing the exact same calculation, but that they're similar—variations on a theme.)

## Telencephalic “learning-from-scratch-ism”

I've previously written ([here](#)) about, um, let's call it, "**neocortical learning-from-scratch-ism**". **That's the idea that the neocortex starts out totally useless to the organism—outputting fitness-improving signals no more often than chance—until it starts learning things.** In particular, if this idea is right, then all adaptive neonatal behavior is driven by other parts of the brain, especially the brainstem and hypothalamus. *That* idea doesn't sound so crazy after you learn that the brainstem has its own whole parallel sensory-processing system (in the midbrain), and its own motor-control system, and so on. (Example: apparently the mouse has a [brainstem bird-detecting circuit wired directly to a brainstem running-away circuit](#).) In [that previous article](#) I called this idea "*blank slate neocortex*", which in retrospect was probably an unnecessarily confusing and clickbaity terminology. Here's a pair of alternate framings that maybe makes the idea seem a bit less wild:

- **How you should think about learning-from-scratch-ism (if you're an ML reader):** Think of a deep neural net initialized with random weights. Its neural architecture might be simple or might be incredibly complicated, it doesn't matter. And it certainly has an inductive bias that makes it learn certain types of patterns more easily than other types of patterns. But it still has to learn them! If its weights are initially random, then it's initially useless, and gets gradually more useful with training data. The idea here is that the neocortex is likewise "initialized from random weights" (or something equivalent).
- **How you should think about learning-from-scratch-ism (if you're a neuroscience reader):** Think of a memory-related system, like the hippocampus. The ability to form memories is a very helpful ability for an organism to have! ...*But it ain't helpful at birth!!* You need to accumulate memories before you can use them! My proposal is that the neocortex is in the same category—a kind of memory module. It's a *very special* kind of memory module, one which can

learn and remember a super-complex web of interconnected patterns, and which comes with powerful querying features, and which can even query itself in recurrent loops, and so on. But still, it's a form of memory, and hence starts out useless, and gets progressively more useful to the organism as it accumulates learned content.

OK, so I've already been a "neocortical learning-from-scratch-ist" since, like, last year. And from what little I know about the hippocampus, I think of it as a thing that stores memories (whether temporarily or permanently, I'm not sure), so I've always been a "hippocampal learning-from-scratch-ist" too. The striatum is another part of the telencephalon, and as soon as I started reading and thinking about its functional role (see below), I felt like it's probably *also* a learning-from-scratch component.

...I seem to be sensing a pattern here...

Oh what the hell. **Maybe I should be a learning-from-scratch-ist about the whole frigging telencephalon.** So again, that would be the claim that the whole telencephalon starts out totally useless to the organism—outputting fitness-improving signals no more often than chance—until it starts learning things within the animal's lifetime.

Incidentally, I'm also a *cerebellum* learning-from-scratch-ist (see my post [here](#)). So I guess I would propose that as much as 96% of the human brain by volume is "learning from scratch"—pretty much everything but the hypothalamus and brainstem. Sounds like a pretty radical claim, right? ...Until you think, 'Hang on, isn't the information capacity of the brain like 10,000x larger than the information content of the genome? So maybe that's *not* a radical claim! Maybe I should be saying to myself, "Only 96%?"'

Anyway, I haven't dug (much) into the evidence for or against telencephalic learning-from-scratch-ism, and I'm not sure what other thinkers think. But I'm taking it as a working assumption—a hypothesis to explore.

...And I'm already finding it a very fruitful hypothesis! In particular, I was *not* previously thinking about the amygdala in a learning-from-scratch-ism framework. And then when I tried, everything kinda clicked into place immediately! Well, at least, compared to how confused I was before. I'll discuss that below.

## Telencephalic uniformity? Nah, let's call it "family resemblance"

So that was learning-from-scratch-ism. Separately, I've *also* previously written about "**neocortical uniformity**" (e.g. [here](#), [here](#))—the hypothesis that every part of the neocortex is more-or-less running the same learning-and-inference algorithm in parallel. To be clear, if this idea is correct at all, then it *definitely* comes with two big caveats: (1) the learning algorithm has different "hyperparameters" in different places, and (2) the neocortex is seeded with an innate gross wiring diagram that brings together different information streams that have learnable and biologically-important relationships (ML people can think of it as loosely analogous to a neural architecture).

So anyway, if "neocortical uniformity" is the idea that every part of the neocortex is running a more-or-less similar learning algorithm, then I guess "telencephalic uniformity" would say that not only the whole neocortex but also the hippocampus, the cortex-ish part of the amygdala, etc. are doing that same algorithm too. And likewise that all the striatum-like stuff is running a "common striatal algorithm", and so on.

Do I believe *that*? To a first approximation:

- I will *not* say the word "uniformity". It's much too strong. In particular, neocortex tissue has more neuron-layers than does allocortex, and I'm inclined to guess that those extra layers are supporting different and richer ways to interconnect the different entries in the memory. Or something.
- I *do* expect that the different types of cortex have at least *something* in common. Variations on a theme, say.
- No comment on how uniform the striatum layer is, or the pallium layer. I haven't looked into it.

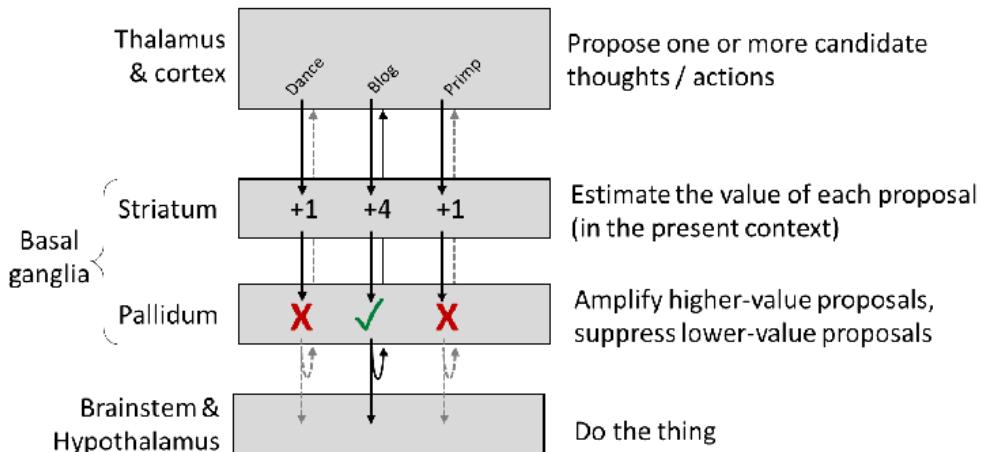
Instead of "uniformity" for the cortex layer, maybe I'll go with "family resemblance". They are, after all, literally family. Like, we mammals have that 6-layer-neocortex-vs-3-layer-allocortex distinction I mentioned, but our ancestors probably just had uniform allocortex architecture everywhere (and most modern reptiles still do) ([ref](#)). (Birds independently evolved a *different* modification of the allocortex, with a functionally-similar end result, I think.)

(Fun fact: The “basolateral complex” of the amygdala is apparently neither allocortex nor neocortex *per se*, but rather a bottom layer of neocortex that peeled off from the rest! Not sure whether it’s *just* detached spatially while still being wired up as a traditional layer 6B, or whether its current wiring is now wholly unrelated to its historical roots, or what. The [claustrum](#) is also in this category, incidentally. See Swanson [1998](#) & [2000](#).)

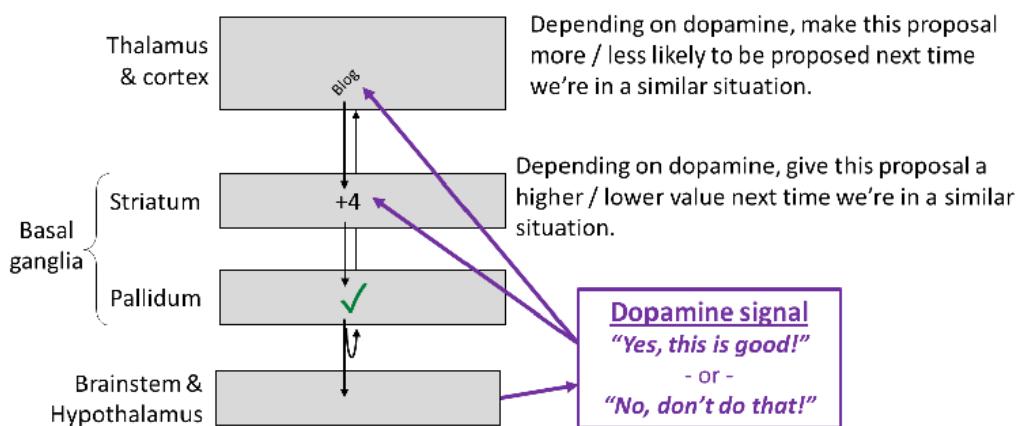
## Toy loop model

Finally, let’s get back to the cortico-basal ganglia-thalamo-cortical loop. What is it for? Here’s the toy model currently in my head. It has two parts, inference (what to do right now) and learning (editing the connections so that *future* inference steps give better answers). Here they are:

### Toy loop model—*inference part* (*i.e.*, what to do right now)



### Toy loop model—*learning part* (for better results next time around)



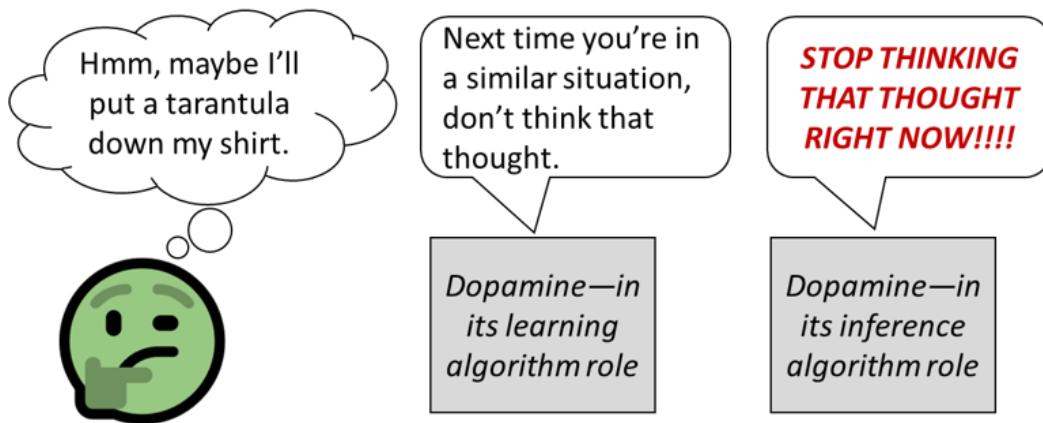
I make no pretense to originality here, and this model is obviously oversimplified, but it’s serving me well so far. So, print these pictures out, tattoo them on your eyelids, whatever, because I’ll be going back to these over and over for the rest of this blog post.

Here is some discussion and nuances to go along with the toy loop model:

- Consistent with the experimental data, I put dopamine-induced learning into *both* the cortex and the striatum. Why should the algorithm work that way? Here’s an example. Let’s say you’re

walking past an alleyway, and it occurs to you “Oh hey, there’s a new ice cream shop down that alleyway!” This is a very very good thought! Cha-ching, dopamine! And specifically...

- Dopamine-in-the-cortex edits the within-cortex connections such that next time you see the alley, the thought of the ice cream shop is more likely to immediately pop into your head. Basically, your cortex starts to learn the sequence “see the alley and then think about the ice cream shop” just like it learned the sequence “twinkle twinkle little star”.
- Dopamine-in-the-striatum edits the cortex-to-striatum connections such that, next time the cortex *starts* activating that thought about the ice cream shop (in a similar context as now), that thought will be amplified, and hence will be likelier to fight its way up to conscious awareness.
- (Isn’t that redundant? Well, yes, it *is* kinda redundant! I suspect that there are lots of boring technical reasons that the algorithm works better when both of these mechanisms are active, rather than just one of them. But in terms of the really big picture, I think we should lump these two mechanisms together, and say they’re two ways of doing the same thing.)
- For ML people: dopamine-in-the-cortex is kinda like policy learning, and dopamine-in-the-striatum is kinda like value function learning, I think.
- I’m using “dopamine” as a shorthand here, as there are also [non-dopamine neurons mixed in with the dopamine neurons, going along the same paths](#). Without having looked into it, my default assumption is that this is a boring implementation detail—that they’re all communicating similar information, and that on the receiving side, different circuits are designed to work with different neurotransmitters.
- The third layer of the diagram, “pallidum”, is of course not to be confused with “pallium”, which is more-or-less a term for cortex in non-mammals—so “pallium” would be the *top* layer of the diagram. You could go nuts, right?
- Speaking of pallidum, I labeled it as “Amplify higher-value proposals, suppress lower-value proposals.” I do think the *algorithm* may really be that simple, but the *brain implementation* absolutely is not!—it involves a bunch of different pathways and cross-connections between the loops, to keep the activations balanced and normalized and whatnot. I’m pretty hazy on how this works in detail; I guess this is related to the “[direct & indirect pathways](#) through the basal ganglia”. [Here’s a friendly video introduction](#).
- I didn’t draw it in, but the loop circuit has direct connections from each of the top three layers to the bottom layer. ([Swanson 2000](#) says the cortex-to-bottom connection is “excitatory”, the striatum-to-bottom connection is “inhibitory”, and the pallidum-to-bottom connection is “disinhibitory”.) Again, I don’t think this is important from a high-level algorithm perspective, it’s just an implementation detail. (This is one of many reasons that those [crazy brain wiring diagrams](#) are so crazy.)
- To be clear, this diagram is not *all* the learning algorithms in this part of the brain, *just* the learning algorithm specifically related to these loops. There are other learning algorithms too, in particular predictive learning within the cortex & thalamus, which I won’t talk about here.
- By the same token, there’s a lot more complexity in the inference algorithm than this loop part I’m talking about here, because the cortex’s job (“propose one or more candidate thoughts / actions”) can (and does) involve a very complicated within-cortex proposal-discovery algorithm, which is outside the scope of this post.
- I put dopamine into the learning diagram but not the inference diagram. But really I think it plays an important role on the inference side too, even if I didn’t draw it. See figure below. That part is why (if I understand correctly) blocking dopamine receptors produces *immediate* motivational effects in laboratory experiments; it doesn’t *only* impact long-term learning.



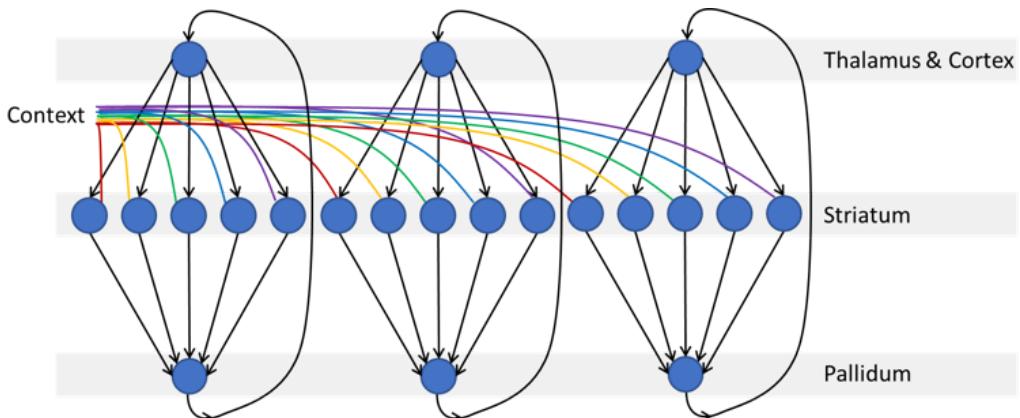
## “Context” in the striatum value function

The “value function” calculated by the striatum is not as simple as a database with one entry for each possible thing to do. Among other things, it’s bound to be context-dependent. Singing in the shower is good, singing in the library is bad.

Here’s a real example. In [Fee & Goldberg 2011](#), they studied zebra finches learning to sing their song (it’s a lovely song by the way, [here’s a video](#)). I have to pause here to warn that bird papers are annoying because practically every part of the bird telencephalon has a different name from the corresponding part of the mammal telencephalon. Anyway, if you look at Figure 4B, “HVC” (some part of the cortex) is providing high-level context—what song am I singing and how far along am I in that song? Meanwhile “LMAN” (a different part of the cortex) seems to be lower-level: if I understand correctly, it holds a catalog of sounds that the bird knows how to make, and how to make them. Then we have the enigmatically-named “Area X”, which is part of the striatum. This HVC signal is *widely* broadcasting its context signal into Area X, while meanwhile LMAN is making *narrow*, [topographic](#) connections to Area X, which then loop back to the exact same part of LMAN. Thus the bird can learn that making a certain sound is high-value in some specific part of the song, but low-value in a different part of the song.

This context idea is reflected in the relative sizes of different parts of the basal ganglia:

One of the remarkable features of [basal ganglia] organization is the massive convergence at every level from cortex to [striatum neurons] to pallidal neurons, and to thalamic neurons that project back to cortex.... In rats, roughly three million [striatum neurons] converge onto only 30,000 pallidal output neurons and subsequently onto a similar number of thalamic neurons .... In humans, a similarly massive convergence from >100 million [striatum neurons] to <50,000 pallidal neurons is reported.... In the context of our model, the reason for this convergence becomes apparent. If the role of Area X [=striatum] is to bias the variable activity of LMAN [=low-level motor cortex] neurons, then the feedback from DLM [=pallidum] to LMAN requires only as many channels as LMAN contains. In contrast, [striatum neurons] in Area X evaluate the performance of each LMAN channel separately at each moment in the song, which requires many more neurons. ([Fee & Goldberg 2011](#))



Don’t take this picture too literally, but I’m trying to capture the idea that the cortico-basal-ganglia-thalamo-cortical loop has more neurons at the striatum part of the loop than at the other parts of the loop. This allows storing the value of *one* potential thought or action in *many* different contexts. I showed a 5:1 striatum neuron ratio here, but the actual ratio in humans is supposedly more than 2000:1. (There are actually way more neurons in the cortex than even the striatum, because of the complicated proposal-selection algorithm; here I’m only talking about directly-loop-related neurons.)

## Three categories of dopamine signals

### Overview: why heterogeneous dopamine?

You’ll note that my toy loop model above avoids the term “reward”. It’s too specific. Go back and look at the diagram with an open mind. What is the dopamine signal signaling? Here’s the most general

pattern:

A positive phasic dopamine signal tells a cortico-basal ganglia-thalamo-cortical loop to be more active next time we're in a similar situation. A negative phasic dopamine signal tells a loop to be less active next time we're in a similar situation.

My proposal is that this pattern is valid for all loops, but nevertheless different parts of the telencephalon use different dopamine signals to do different things. Remember, as I mentioned above, we already know that there isn't just one dopamine signal.

I currently see three categories of (phasic) dopamine signals. I'll list them here, then go through them one-by-one in the next subsections: (1) Reward Prediction Error (RPE) for what I'll call the Success-In-Life Reward, the classic kind of "reward" that we intuitively think about, i.e. an approximation of how well the organism is maximizing its inclusive genetic fitness; (2) RPE for local rewards specific to certain circuits—e.g. negative dopamine specifically to motor output brain areas when a motor action is poorly executed; (3) supervised learning error signals—e.g. if you get whacked in the head, then there's a hardcoded circuit that says "you should have flinched", and that signal can train a loop that specifically triggers a flinch reaction.

Analogy: You did a multiple-choice test, and then later the teacher hands it back:

- If the only thing the teacher wrote on it was a grade at the top ("B+ overall"), that's like category #1, Success-In-Life Reward.
- If the teacher gives grades for one or more sub-sections ("B- on the first page of questions, A- on the second page of questions..."), that's like category #2, local rewards to sub-circuits.
- If the teacher tells you what any of the correct answers were ("You circled (B) on question #72, but the right answer was (D)") then that's like category #3, supervised learning.

My discussion here will be a bit in the tradition of (and inspired by) [Marblestone, Wayne, Kording 2016](#), in the sense that I'm arguing that part of the brain is running a learning algorithm, with different training signals used in different areas. The big difference is that I want to focus on the dopamine signals, whereas they focused on acetylcholine signals. ([I think acetylcholine mainly controls learning rate](#), so it's not *directly* a training signal.) Also, as mentioned above, I'm not telling the *whole* "different training signals in different places" story in this post; the other part of that story is predictive (a.k.a. self-supervised) learning, in which different parts of the cortex are trained to predict different things. But *that* learning algorithm is not related to the loops, and it's not related to dopamine, and it's outside the scope of this post. (It's related to how the cortex "selects proposals".)

## Dopamine category #1: RPE for “Success-In-Life Reward”

Start with the classic, stereotypical kind of reward, the reward that says "pain is bad" and "social approval is good" and so on. By and large, this reward should be some kind of heuristic approximation to the time-derivative of the organism's inclusive genetic fitness. I'll call it "Success-In-Life Reward", to distinguish it from other reward functions that we'll discuss in the next section.

**Where does that reward signal come from?** My short answer is: the hypothalamus and brainstem calculate it, on the basis of things like pain inputs (bad!), sweet taste inputs (good!), hunger inputs (bad!), and probably hundreds of other things. Boy, I would give anything for the complete exact formula for Success-In-Life Reward! Like, it's not *literally* "The Meaning Of Life", but it might be the closest thing that neuroscience can get us. I'll get back to this later.

**You said “Reward Prediction Error” (RPE); where do the “predictions” come from?** Hold that thought, we're not ready to answer it yet, but I'll get back to it in a later section.

**Why is this particular reward function useful?** Because parts of the telencephalon are "deciding what to do" in a general way. Should you go out in the rain or stay inside? Should you eat the cheese now or save it for later? If the animal is to learn to make systematically good decisions of this type, then we need the decisions to be made by a learning algorithm trained to maximize "Success-In-Life Reward". So **I especially expect this reward function to be used for parts of the brain making high-level decisions involving cross-domain tradeoffs.**

Those areas include, I think, at least some parts of "granular prefrontal cortex" (don't worry if you don't know what that means) and the hippocampus. These areas are both making decisions involving cross-

domain tradeoffs. Like in humans, the former is the place that “decides” to bring to consciousness the idea “I’m gonna roast some vegetables!” (out of all possible ideas that could have been brought to consciousness instead). And the hippocampus is the place that “decides” to bring to consciousness the idea “I’m gonna turn right at the fork to go to the farmstand!” (out of all possible navigation-related ideas that could have been brought to consciousness instead). Something like that, maybe, for example.

## Dopamine category #2: RPE for “local” sub-circuit rewards

Think about the *Millenium Falcon*, with Han Solo in the gun turret while Chewbacca is up front piloting. Chewbacca’s steering could be perfect while Han’s aim is terrible, or vice-versa. If they have to share a single training signal, then the signal will be noisier for each of them—for example, sometimes Han will do a bad job, but still get a high score because Chewy did unusually well, and then Han will internalize that wrong message. This isn’t *necessarily* the end of the world—I imagine that, if you do it right, the noise will average away, and they’ll eventually learn the right thing—but the learning process may be slower. So I figure that *if* it’s possible to allocate credit and blame for performance variation between Han and Chewy, they would probably learn faster.

By the same token, your own body is a lumbering contraption controlled by thousands of dials and knobs in the brain, and different parts of your cortex are in control of different parts of this system. If the brain *can* allocate credit, and thus send different rewards to different areas, then I imagine that it *will*.

(*Side note 1:* I imagine that some ML readers are instinctively recoiling here: “Nooooo, [The Bitter Lesson](#) says that we’ll get the best results by using end-to-end performance as the one and only input to our learning algorithm!” Well readers, if you want to think about animals, I think you’ll need to put a bit more emphasis on “learning fast” and a bit less emphasis on “asymptotic performance”, compared to what you’re used to. After all, Pac-Man can keep learning after getting eaten, but an animal brain can’t—well, [not usually](#). So you gotta learn fast!)

(*Side note 2:* Backpropagation (and its more-biologically-plausible cousins) can allocate credit automatically. However, they require error gradients to do so. In supervised (or self-supervised) learning, that’s fine: we get an error gradient each query. But here we’re talking about reinforcement learning, where error gradients are harder to come by, as discussed in a later section.)

### Example 2A: Birds teaching themselves to sing by RL

Here’s the clearest example I’ve seen in the literature:

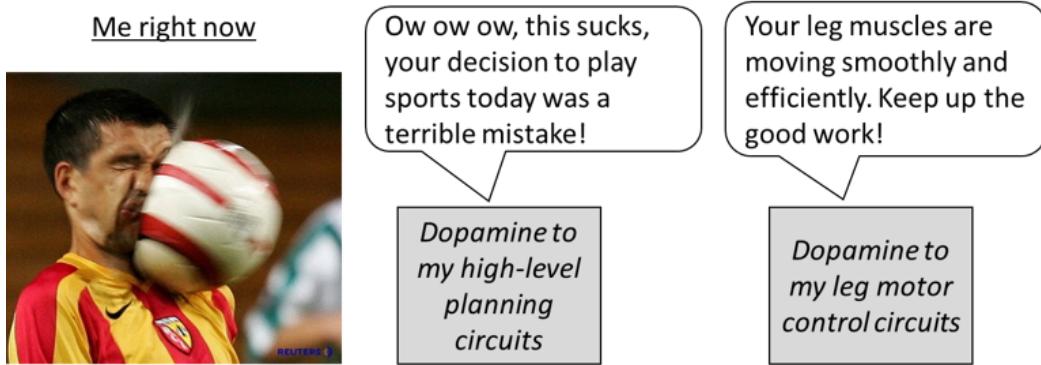
For example, we recently identified song-related auditory error signals in dopaminergic neurons of the songbird ventral tegmental area (VTA).... We discovered that only a tiny fraction (<15%) of VTA dopamine neurons project to the vocal motor system - yet these were the ones that encoded vocal reinforcement signals. The majority of VTA neurons which project to other parts of the motor system did not encode any aspect of song or singing-related error. —[Murdoch 2018](#) describing a result from [Gadagkar 2016](#)

Got that? Birds have an innate tendency to sing, and they learn to sing well by listening to themselves, and doing RL guided by whether the song “sounds right”, as judged by some other part of the brain (I suspect the tectum, in the brainstem). And those specific dopamine feedback signals, the ones that say whether the song sounds good, go *only* to the singing-related-motor-control part of the bird brain. Makes sense to me!

### Example 2B: Animals teaching themselves to move smoothly and efficiently by RL

This is more speculative, but seems to me that it should be feasible for some part of the brain to send a higher “reward” to a motor control loop when motion is rapid and energy-efficient and low-strain, and a lower “reward” when it isn’t. So I would assume that the “reward” going to low-level motor control loops should narrowly reflect the muscle’s energy expenditure, speed, strain, or whatever other metrics are biologically relevant.

(Some of you might be thinking here: What a stupid idea! If the reward is really like that, then the low-level motor control loops will gradually learn to *do nothing whatsoever*. That's *extremely* rapid and energy-efficient and low-strain! Well, again, I'm hiding a lot of complexity behind the "propose candidate actions" part of my toy loop model above. If I'm not mistaken, the within-cortex dynamics will ensure that if a lower-level motor sequence isn't compatible with advancing the currently-active higher-level plan, then it won't get proposed in the first place!)



I think that high-level planning circuits get a dopamine signal related to Success-In-Life Reward (previous section), while motor control circuits get a narrowly-targeted dopamine signal that encourages actions that are energy-efficient, low-impact, fast, etc. So even if things are going suddenly very badly for the organism as a whole, not *all* the dopamine neurons will reflect that.

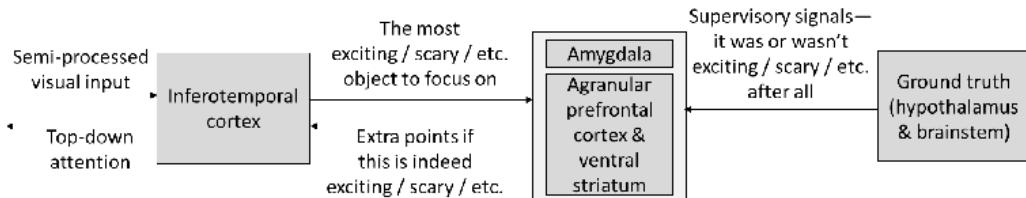
What of the literature? [Schultz 2019](#) cites evidence for heterogeneous dopamine that goes along with larger movements, but not concise, stereotyped movements. That fits my theory pretty well: concise, stereotyped movements are more likely to use exactly the expected amount of energy and speed (and hence produce no motor-loop RPEs), whereas larger movements are likely to have some idiosyncratic differences in energy & speed compared to expectations (and hence produce positive or negative motor-loop RPEs). Moreover, the movement-related dopamine is heterogeneous, which is expected if some muscles are using slightly more energy than typical while others are using slightly less. Then my hypothesis would be that these movement-associated dopamine signals go specifically to brain areas associated with low-level motor control, but I haven't yet found literature either for or against that.

## Example 2C: Visual attention

Recently I wrote a post [Is RL involved in sensory processing?](#), and was pretty weirded out when an astute reader told me that there was a non-frontal-lobe region of neocortex that had complete loops—namely inferotemporal cortex (IT) ([ref](#)). I was weirded out because I was thinking of the cortico-basal ganglia loops as part of an algorithm for choosing among multiple viable options—like what action to take, or what thought to think. Those choices tend to be made in the frontal lobe. IT, by contrast, is the home of visual object recognition, which I would think should not be a “choice”: Whatever the object is, that’s the right answer! So it seems like it calls for predictive (self-supervised) learning, not the loop algorithm.

Then I came across another weird thing! When the IT loops hit the striatum, that region is called “tail of the caudate”, and it’s one of the five regions flagged in a recent paper as an “[aversion hot spot in the dopamine system](#)”—i.e., it has elevated dopamine after bad things happen (when we traditionally expect low dopamine). The other four “aversion hot spot” regions, incidentally, are exactly the regions that I’ll talk about in the next section (dopamine for supervised learning of autonomic reactions), so that makes sense. But the IT loops need a different explanation.

So here’s my theory: IT is helping “choose” what object to attend to, within the visual field. If there’s a lion ready to pounce, *almost* perfectly hidden in the grass, and IT directs attention in a way that makes the subtle form of the lion visible ... well, seeing the lion is highly scary and aversive, so the high-level planner gets negative dopamine. *But IT did the right thing here!* Cha-ching, dopamine for IT! This sets up a kinda adversarial dynamic—IT focuses on anything in the visual field that might be dangerous or exciting, as judged by a different part of the brain, and the latter in turn then gets to hone its judgment on lots of edge-cases. This quasi-adversarial dynamic is good and healthy, and I think consistent with lived experience.



Toy model for the unusual dopamine- and loop-related properties of inferotemporal cortex, as described in the text. The arrows here are describing loose functional relationships, not direct brain connections.

## Dopamine category #3: Supervised Learning (SL) error signals

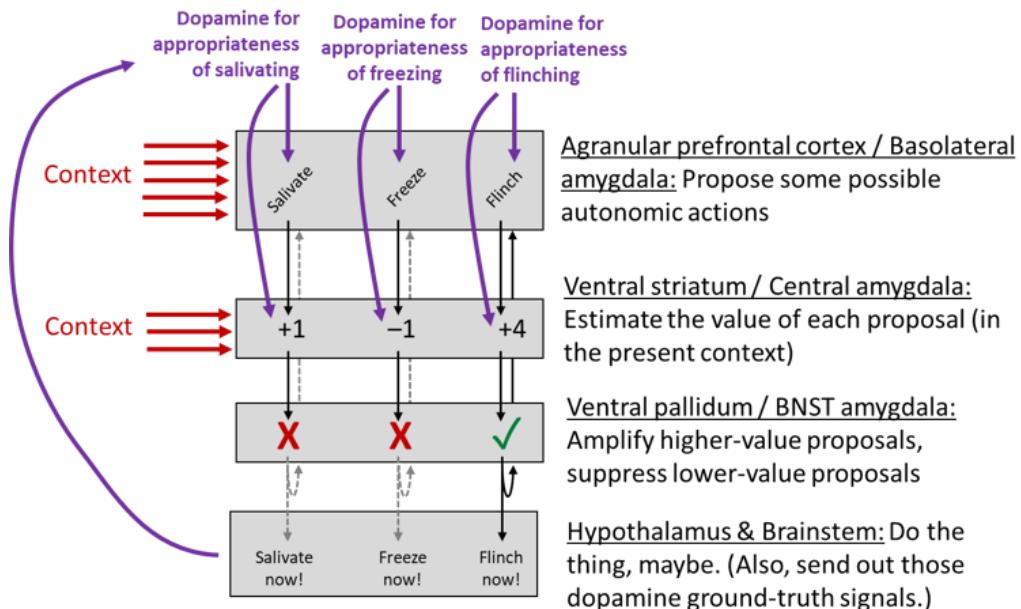
The fundamental difference between supervised learning and reinforcement learning is that in supervised learning, there's a ground truth about "what you should have done", and in the latter, there's a ground truth about how successful some action was, but no specific advice about how to make it better. (Or in ML terminology, SL gets a loss gradient from each query, while RL doesn't.)

I talked about this previously in [Supervised Learning of Outputs in the Brain](#). I got some details wrong but I stand by the big picture I offered there. In particular, I think there are certain categories of telencephalon "outputs" for which the brainstem & hypothalamus can generate a "ground truth" after the fact about whether that output should have fired. For example, if you get whacked by a projectile, then your brainstem & hypothalamus can deduce that you should have flinched a moment earlier.

I think "outputs for which a ground truth error signal is available after the fact" include "autonomic outputs", "neuroendocrine outputs", and "neurosecretory outputs", but *not* "neuromuscular outputs". I'm quite unsure that I'm drawing the line in the right place here—in fact, I don't know what half those terms mean—but for the purposes of this blog post I'll just use the term "autonomic outputs" as a stand-in for this whole category.

While SL is different from RL, if you go back to the Toy Loop Model above, you'll see that it works for both, with only minor modifications:

### Toy loop model—supervised learning variant

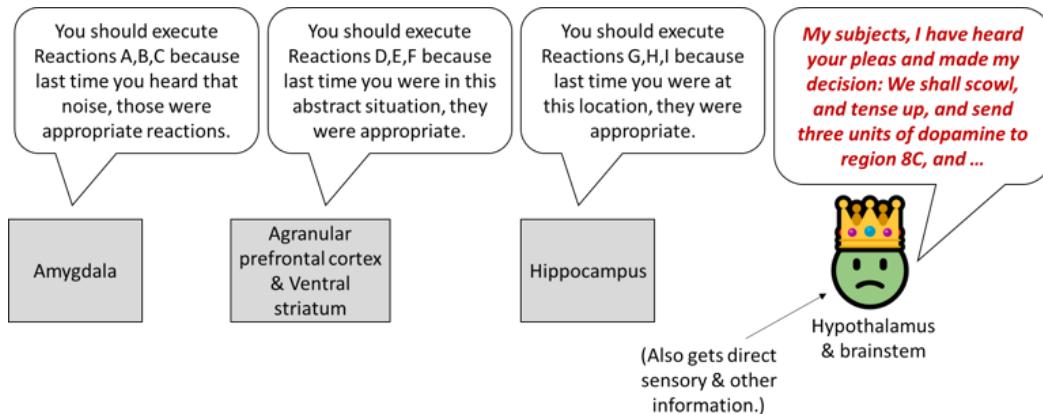


Some comments on this:

- Where are these SL loops? I think they're in two places in the telencephalon:
  - *Amygdala loops.* As mentioned above, the amygdala has cortex-like parts and striatum-like parts. Then the pallidum-like part is the catchily-named "Bed Nuclei of the Stria Terminalis" (BNST), which is sometimes lumped into the so-called "extended amygdala". I note that the amygdala is "[neither a structural nor a functional unit](#)", and therefore I'm open to the possibility that supervised learning is really only happening in a subset of the amygdala, or that it spills out into neighboring structures, or whatever.
  - *Agranular prefrontal cortex + Ventral striatum loops.* "Agranular" here means "missing layer 4 (out of the 6 neocortical layers)". Layer 4 is the one that processes inputs, so "agranular cortex" is an output region (see [here](#)). This region of cortex is also called (roughly) "medial prefrontal cortex" or "ventromedial prefrontal cortex", and also spills into cingulate and insular areas. (I admit I'm hazy on the anatomy here). Agranular prefrontal cortex is a universal component of mammalian brains, whereas *granular* prefrontal cortex is a primate innovation ([Wise 2017](#)).
  - My sense right now is that these two regions are mostly or entirely overlapping in their suite of autonomic outputs, but I'm not sure—I'll get back to that below.
- Neuron-level differences between SL loops & RL loops: I think there are some low-level learning rule changes you need to make to get from the RL version to the SL version. For example, if a loop was *not* recently active, but *does* get dopamine exposure, should it be more or less active in similar situations in the future? I think it's "more" for SL, and "less" for RL. So presumably you need different dopamine receptors and plasticity mechanisms and whatnot.
  - ...And there does indeed seem to be evidence for this!
    - "[Dopamine] neurons in medial posterior VTA...selectively project to the [ventral striatum], medial prefrontal cortex..., and basolateral amygdala.... These "non-conventional" VTA [dopamine] neurons had no or only very little Ih [current] and also did not exhibit typical action potential waveforms and firing patterns." ([Lammel et al. 2014](#))
    - ...So basically, there's a subpopulation of dopamine neurons that tend to fire different types of signals than the rest of them, and this subpopulation is almost exactly those that go to the places I'm claiming are doing SL not RL. This is very encouraging! Maybe I'm on the right track!
- Architecture-level differences between SL & RL loops: In the SL case, we need probably many dozens of small "zones", one for each innate autonomic reaction (recoiling, flinching, glaring, freezing, sweating, laughing, [Duchenne-smiling](#), etc. etc.). And we need a corresponding number of different dopamine signals, each going into the corresponding zone.
  - As for "different zones for different outputs", I *think* that's uncontroversial. In the amygdala case, see [this creepy 1967 experiment stimulating the amygdalas of anesthetized cats](#). Different parts of the amygdala induced various actions including "sniffing reaction...reaction of alertness or attention...Strongly aggressive behavior or a defensive reaction...licking, swallowing...salivation, urination...turning the head to the side..." In the agranular prefrontal cortex case, see [Wise 2017](#)—electrically stimulating it leads to various actions including "[bristling]; pupillary dilation; and changes in blood pressure, respiratory rate, and heart rate". In a [different experiment](#), researchers fiddling with ventral striatum managed to induce vomiting, shivering, and ejaculation.
  - As for "different dopamine signals for each different output zone", well, I haven't found particularly great evidence for or against this idea, at least not so far. But I did already mention the [study listing five parts of the brain where researchers have found dopamine enhancement during aversive experiences](#). I already covered one of those five above ("Example 2C" section.) The other four send dopamine into more-or-less *exactly* the regions that I claim are doing SL (medial prefrontal cortex, ventral striatum, cortex-like amygdala, and striatum-like amygdala). This makes sense because *some* of the autonomic reactions are appropriate for aversive experiences—like freezing, raising your heart rate, and so on. Other autonomic reactions are of course *not* appropriate at aversive experiences—like smiling and relaxing. But I don't think that paper is claiming that those *entire regions* are getting extra dopamine at aversive experiences, just that there was enhanced dopamine activity *somewhere* in those regions.
- Here's one thing my diagram leaves out: [Mollick et al. 2020](#) suggests an "opponent processing" thing, which I guess looks like having *both* a "Flinch" loop *and* a "Don't Flinch" loop side-by-side. This is more subtle than it sounds: almost every situation is a good time not to flinch, so how do you train the "Don't Flinch" loop? Mollick suggests a neat little trick: the "Don't Flinch" loop learning algorithm is turned off entirely *unless* the "Flinch" loop is active. So it's really learning "Wait Actually Don't Flinch Even Though There Might Seem To Be A Reason To Flinch", which is specific enough to be learnable. I didn't draw this opponent-process thing into my diagram above, partly because the diagram would be too crowded and hard to read, and partly because I'm a bit

confused about some details of how it would fit in. Anyway, read Mollick for details. The opponent-processing thing is necessary to explain some aspects of animal experiments, like "rapid reacquisition" and "renewal". From an algorithmic performance perspective, imagine learning "lions are scary" and then learning "lions in cages are not scary". We want to be able to learn the latter without unlearning the former. And in an unfamiliar context, we want to still treat lions as scary by default.

- Here as elsewhere, "context" is just a big collection of whatever possibly-relevant signals exist anywhere in the brain. The algorithm then sifts through all those in search of any that are predictive, as discussed in [my old post here](#). I think that one contribution to the [intimidating complexity of those brain flow diagram things](#) is that the brain is pretty profligate in pulling in lots of random signals from all over the place, to use as context lines. (To be clear, there is also "context" in the RL version of the toy loop model above, I just left it out of my earlier diagram for simplicity.)
- As in the general toy loop model, having *both* the cortex and the striatum layer is a bit redundant in principle, but presumably helps the algorithm work better. In particular, I *think* that the existence of two layers of selection is somehow related to the fact that animals have "second-order conditioning" but very little "third-order conditioning" ([ref](#)).
- In the bottom layer of my diagram, I wrote "Do the thing, maybe". What's with the "maybe"? Well, the hypothalamus and brainstem are simultaneously getting suggestions for which autonomic reactions to execute from the amygdala, the agranular prefrontal cortex, and the hippocampus. Meanwhile, it's also doing its own calculations related to sensory inputs (e.g. in the superior colliculus) and things like how hungry you are. It weighs all those things together and decides what to do, I presume using an innate (non-learned) algorithm.



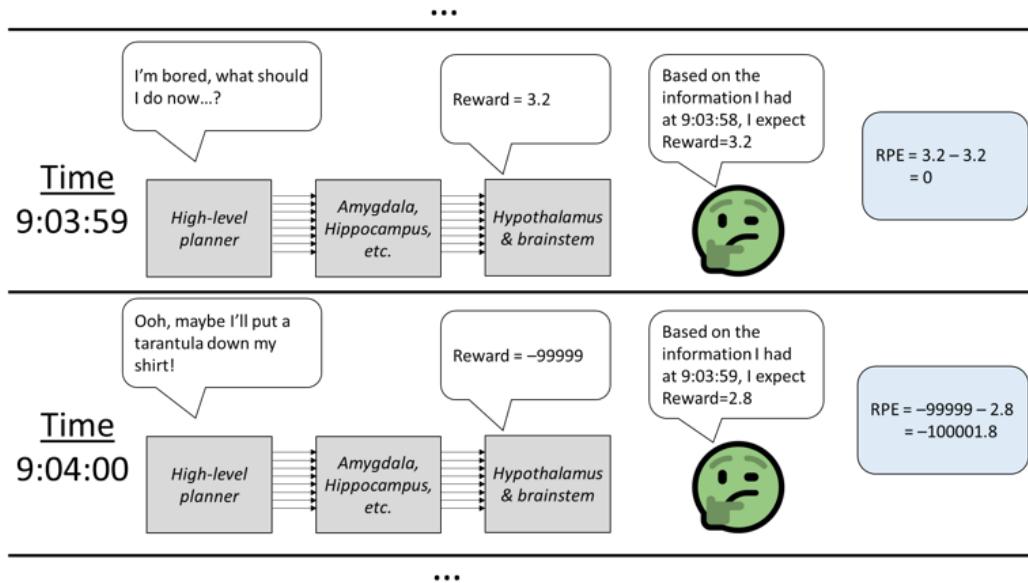
Cartoon illustration of why we would want three (that I know of) telencephalon areas telling the hypothalamus & brainstem what autonomic reactions are appropriate right now, in parallel. Basically I think they're different learning algorithms that look at different types of information as a signal for what will be appropriate, and thus each can catch hints that the others miss. Then the hypothalamus & brainstem system weighs that information, plus its own internal processing, and decides what to do.

(The hippocampus is in that diagram, sending autonomic output suggestions to the hypothalamus, but I don't *think* the hippocampus involves the kind of supervised learning loop that I'm discussing here. I think it uses a different mechanism—like, the hippocampus stores a bunch of locations, and each is tagged with the autonomic outputs that have previously happened at that location, and that's what it suggests. So, basically, more like a lookup table.)

I admit I'm pretty hazy on the details here—like, if there's an amygdala loop for releasing cortisol, and if there's *also* an agranular prefrontal cortex loop for releasing cortisol, then how *exactly* are they related? I made a suggestion in my diagram above, but that might not be right, or might not be the whole story.

As an example of my confusion in this area, [S.M.](#), a person supposedly missing her whole amygdala and nothing else, seems to have more-or-less lost the ability to have (and to understand in others) negative emotions, but not positive emotions. But AFAICT the amygdala can trigger both positive- and negative-emotion-related autonomic outputs! Weird.

# Finally, the “prediction” part of reward prediction error



I propose that there's a special loop in agranular prefrontal cortex / ventral striatum—depicted here the green face—whose supervised learning target is the *near-future* reward signal, the one coming in 1 second (or whatever). Or equivalently, we could say that this loop is trained to guess the *current* Success-In-Life Reward on the basis of *1-second-stale* brain status data. We get Reward Prediction E (RPE) by subtracting that prediction from what actually happens—and that's the dopamine signal.

Now that we've covered the RL loops and the SL loops, we're finally ready to tackle the reward prediction error! And it's easy! See figure and caption. Some comments on that:

- Based on how this dopamine signal is generated, we can generally be pretty confident that the dopamine is signaling something good or bad that happened *specifically in the previous 1 second* (or whatever the reward forecast interval is). This is very useful information, because the dopamine-related learning rules can assign credit / blame for activity in that same interval.
- The “predictor” is a neocortex circuit, and I presume that we can build a reward-prediction model just as complex as anything else in our world-model. This is important because traditional step-by-step TD learning has a bunch of shortcomings involving, for example, rewards that arrive early (if the brain used traditional TD, then after getting an early reward, then there would be *negative* dopamine at the *old* reward time!), or rewards that arrive after variable intervals, or distractions between stimulus and reward, etc. ([Further discussion](#)) Animals don't have those problems because they can build a better reward-prediction model, one involving persistent state variables, learned hierarchical sequences, time-dependence, and so on.
- There's a [recent set of experiments](#) showing that dopamine neurons do something like “distributional reinforcement learning”. In my framework, we can explain those results by having a bunch of reward prediction loops, each with different supervised learning hyperparameters, such that the various loops span the range from “very optimistic reward prediction” to “unbiased reward prediction” to “very pessimistic reward prediction”. All these loops together give a reward prediction probability distribution, rather than just a point estimate. And these different prediction loops can then feed signals into different dopamine neurons. Honestly, I bet that every other supervised learning loop is set up that way too—e.g., that the amygdala answers the question “how appropriate would it be to freeze in terror right now?” with a bunch of loops sampling a probability distribution.
- My proposal here is heavily inspired by Randall O'Reilly & colleagues' “PVLV” (pronounced “Pavlov”, hahaha) model ([original version](#), [most recent version](#)). That said, I think my version wound up rather different, and I'm not sure whether they would endorse this. See [supplement](#) for

more discussion. My proposal also wound up reassuringly similar to [a Steve Grossberg model](#), I think—again, see [supplement](#).

- I think you can do this thing for any of the various "rewards" that the hypothalamus calculates—most famously the "Success-In-Life Reward" approximating the time-derivative of inclusive genetic fitness (Dopamine Category #1 above), but also potentially the sub-circuit rewards (Dopamine Category #2 above). Not sure that actually happens, but it's possible.
- Fun fact: Punishments create dopamine pauses even when they're fully expected ([ref](#)). How does that happen? Easy: for example, maybe there's a limit on how negative the reward prediction is allowed to go. Or maybe (as in [Mollieck](#)) negative reward predictions get multiplied by a scale-factor, like 0.9 or whatever. More to the point, *why* does that happen? I figure, at the end of the day, negative dopamine means "No, don't do that!"—as in my toy loop model. If things are bad enough—e.g. life-threatening—it's plausibly *always* fitness-improving to treat the status quo plan as unacceptable, even in the face of growing evidence that the alternatives are even worse. After all, on the current trajectory, you know *for sure* that you're doomed. Call it an "evolutionary prior" that "This just *can't* be the best option; there's *gotta* be something better!", maybe.
  - *Side tangent:* There's an annoying paradox that: (1) In RL, there's no "zero of reward", you can uniformly add 99999999 to every reward signal and it makes no difference whatsoever; (2) In life, we have a strong intuition that experiences can be good, bad, or neutral; (3) ...Yet presumably what our brain is doing has *something* to do with RL! That "evolutionary prior" I just mentioned is maybe relevant to that? Not sure ... food for thought ...
- While the system as a whole does a passable imitation of TD learning under many circumstances, ***no part of it is actually literally doing TD learning (!!!!!)***. The reward-prediction circuit here is operating on plain old supervised learning—i.e. outputting expected next-timestep reward, without any involvement of "expected future rewards". (I'm following the [PVIV model](#) in rejecting TD learning in favor of something closer to the [Rescorla-Wagner model](#).) If there's no TD learning, then how on earth do we wind up with a time-integral of reward? Well, we basically don't! Humans are *absolute rubbish* at calculating a time-integral of reward—see the [Peak-End Rule](#)! OK, so how do we weigh decisions that will involve both pleasant and unpleasant aspects at different times? The snarky answer is "poorly". A better answer is: If "Going to the restaurant" involves both a scary drive and a yummy meal, then after a few times, entertaining the plan will activate *both* the "this will be scary" supervised learning loop *and* the "this will be yummy" supervised learning loop. Then the hypothalamus and brainstem can see both aspects and give an appropriate overall assessment. The best answer of all is: "This is a great question that deserves more discussion than I'm going to give here." :-P

## Zooming back out: lessons for Artificial General Intelligence (AGI) safety

Of course for me, everything is always ultimately about AGI safety, and so is this post. Let's go back to "telencephalic learning-from-scratch-ism" above—and let's gingerly set aside the possibility that that hypothesis is totally false....

Anyway, there's a learning algorithm in our brain. It's initialized from random weights (or something equivalent). It gets various input signals including sensory inputs, dopamine and other signals from the hypothalamus & brainstem system, and so on. You run the code for some number of years, and bam, that learning algorithm has built a competent, self-aware agent full of ideas, plans, goals, habits, and so on.

Now, sooner or later (no one knows when) we'll learn to build "AGI"—by which I mean, for example, an AI system that could have written this entire blog post much better than me. And here's one specific way we could get this kind of AGI: We could code up a learning algorithm similar to the one in the telencephalon, and give it the appropriate input signals, run it for some period of time, and there's our AGI.

Why not?

- Maybe the telencephalon-like learning algorithm is too complicated for humans to figure out? Well, sure, maybe. I do certainly imagine that the "neural architecture" equivalent of the telencephalon is rather complicated—there may be as many as hundreds of little regions connected in a particular way, for example. But on the other hand, we have ever-improving neuroanatomy databases, and we have automated "neural architecture search" that can fill in any

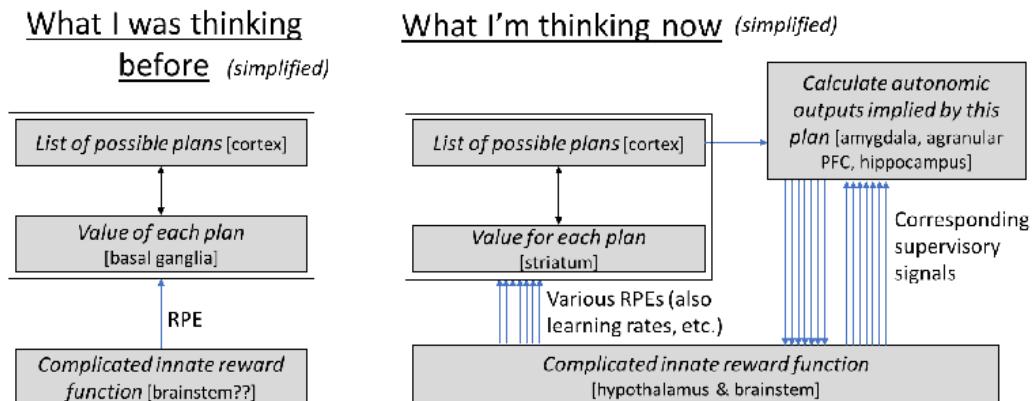
gaps by brute force trial-and-error. And we have experts all over the world digging into the exact operating principles of the learning algorithms in question.

- Maybe an appropriate training environment (sensory data etc.) is unavailable? I find that hard to believe. We can easily steep an AI system in human cultural artifacts (like books and movies) and give it lots of human interactions, if that were essential (which I don't think it is anyway). Are you an "embodied cognition" advocate? Well, we can give the AI a robot body if we want, not to mention the ability to take various possible "virtual actions".
- Maybe the algorithms in the other parts of the brain—especially the hypothalamus and brainstem—are too complicated for humans to figure out, but without those algorithms sending up the right dopamine and other signals, the telencephalon-like learning algorithm can't properly do its thing? Well, yes and no ...

...I *do* think the innate hypothalamus-and-brainstem algorithm is kinda a big complicated mess, involving dozens or hundreds of things like snake-detector circuits, and curiosity, and various social instincts, and so on. And basically nobody in neuroscience, to my knowledge, is explicitly trying to reverse-engineer this algorithm. I wish they would! I would absolutely encourage neuroscientists to push it upwards on their research priority list. But the way things look right now, I'm pessimistic that we'll have made much progress on that, at least not by the time we have telencephalon-like learning algorithms coded up and working better and better.

So then we're at the scenario I wrote about in [My AGI Threat Model: Misaligned Model-Based RL Agent](#): we *will* know how to make a "human-level-capable" learning algorithm, but we *won't* know how to send it reward and other signals that sculpt the learning algorithm into having human-like instincts and drives and goals. So researchers will mess around with different simple reward functions—as researchers are wont to do—and they'll wind up training superhuman AGIs with *radically nonhuman* drives and goals, and they'll have no reliable techniques to set, or change, or even know what the AGIs' goals are. You can read [that post](#). I do *not* think that scenario will end well. I think it will end in catastrophe! The solution, I think, is to do focused research on what the reward function (and other signals) should be—perhaps modified from the human hypothalamus and brainstem algorithm, or else designed from scratch.

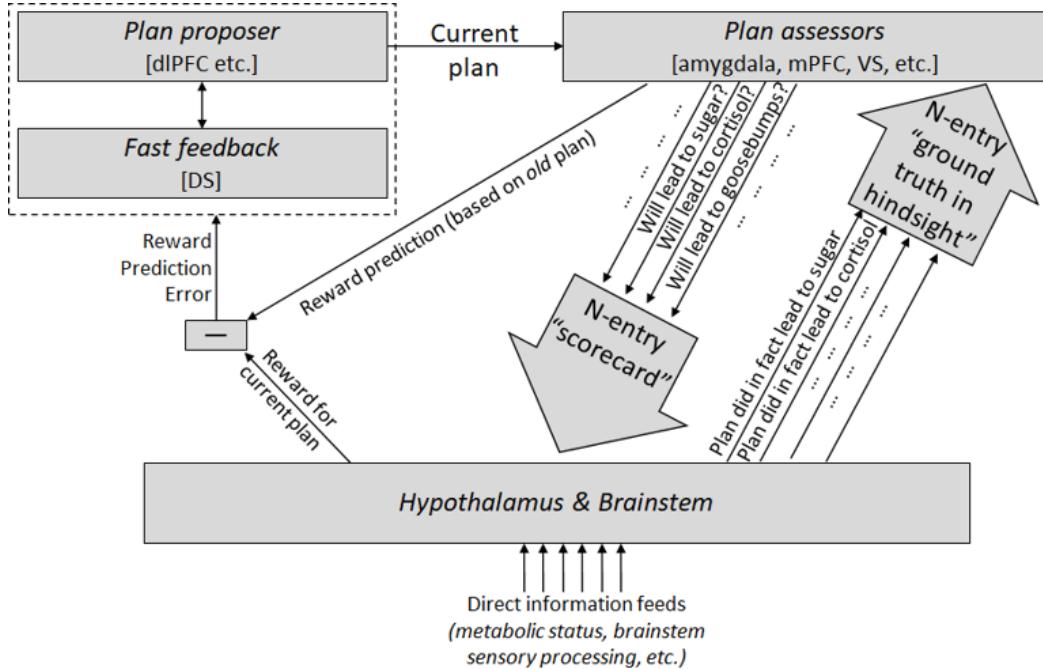
So what was the point of me writing *this* blog post? I wanted to better understand the telencephalon learning algorithm's [API](#). Like, what is the suite of input signals (other than sensory data) that guides this learning algorithm, and how do they work? I don't have a complete answer yet—there are still plenty of brain connections where I don't know what the heck they're doing. But I think I'm making progress! And this post in particular (and work leading up to it) constitutes a noticeable change:



I guess the biggest change is that now I think of two ways for the brain to evaluate how good a plan is. The fast (parallelizable) way uses the striatum, and helps determine which plans can rise to full attention. Then the slow (serial) but more accurate way involves the plan rising to full attention, at which point a maybe dozens-of-dimensional vector of auxiliary information about this plan's expected consequences is calculated—if I do this plan, should I raise my heart rate? Should I salivate? Should I cringe? Etc. That auxiliary data vector gives the hypothalamus & brainstem something to work with when evaluating the plan. Regular readers of course will connect this to my earlier post [Inner Alignment in Salt-Starved Rats](#), where I was relying on a mechanism like this to explain some animal experiments. Or think of the famous [Iowa gambling task](#); here you can watch in real time (using skin conductance) as the supervised learning algorithm gradually improves the accuracy of the auxiliary data vector

associated with each of two choices, until eventually the auxiliary data vector provides a clear enough signal to guide decision-making.

**Update:** see also follow-up [A model of decision-making in the brain \(the short version\)](#), where I put this diagram:



Update: I made this diagram for a follow-up post [A model of decision-making in the brain \(the short version\)](#). But I figured, I might as well copy it in here too, in case it's helpful. To be clear, this diagram focuses specifically on decision-making, so it leaves out some other things like multiple reward functions.

Like, I was thinking I should use this terminology:

- **Cached value function** is the result when the striatum evaluates a plan, in order to make sure that only especially promising plans can rise to attention;
- **Actual value function** is the result when the plan *does* rise to attention, and generates an auxiliary data vector of its corresponding autonomic outputs ("this plan would entail salivating and shivering"), and then the hypothalamus & brainstem evaluates it and sends up dopamine corresponding to its judgment ("well I'm hungry, so any plan involving salivation sounds pretty good!");
- **Reward function** is the result when the plan is actually executed and the hypothalamus & brainstem get to see how it goes, including via direct sensory input to the brainstem.

These are listed in increasing order of, um, "fidelity" to the "intentions" of the innate hypothalamus & brainstem algorithm—and consequently the information tends to flow in the opposite direction, from third bullet to second to first. So "cached value function" is trained to imitate "actual value function", and "actual value function" is ultimately reliant on information flowing from "reward function". (I'm not sure this terminology is quite right, and also note that there isn't a sharp line between "actual value function" and "reward function".)

Anyway, when we think about how to control AGIs, the idea of "auxiliary data vectors" seems like an awfully important thing to keep in mind! What kinds of auxiliary data can we use to better understand and control our future telencephalon-like-learning-algorithm AGIs, and where do we get the ground truth to train the auxiliary-data-calculating subsystems? Umm, [beats me](#), but it seems like an important question, and I'm still thinking about it, and let me know if you have any ideas.

Likewise, the hypothalamus and brainstem can output multiple region-specific RPEs, not just one. (They probably output various hyperparameter-modulating signals too.) Once again, our future telencephalon-like-learning-algorithm AGI will likewise *also* presumably be trained by sending different RPEs to different sub-networks. We're the programmers, so we get to decide exactly what RPEs go to what sub-

networks. Is there a scheme like that which will help keep our powerful AGIs reliably under human control? Once again, [beats me](#), but I'm still thinking about it. And you can too! I'm probably the only person on earth being paid to think specifically about how to safely control telencephalon-like-learning-algorithm AGIs, and god knows I'm not gonna figure it out myself, I'm in way the hell over my head.

(If thinking directly about how to control futuristic powerful AGIs isn't your cup of tea—and I admit it would be a tough sell on your next NIH grant application—at least let's reverse-engineer "The Human Reward Function", i.e. that innate hypothalamus & brainstem algorithm I keep talking about. *That's* a conventional neuroscience research program, and it definitely helps the cause! Like, try writing down the part of the reward function that leads to jealousy. I don't think it's obvious! Remember, you're not allowed to directly use common-sense concepts as ingredients in the reward function; the function needs to be built entirely out of information that the hypothalamus and brainstem have access to.)

(*This post has a supplement [here](#). Please leave comments at the [lesswrong crosspost](#), or [email me](#).*)

# The Generalized Product Rule

Imagine we have a company with investment projects A, B, C,...For instance, A might be a new high-speed Internet service, B might be a new advanced computer, C might be a new inventory management software, etc. We are interested in calculating the total return from these investments at the company. This calculation could be fairly complicated since returns are context-dependent - e.g., new computer B might have higher return in the context of new Internet service A than it would without the new Internet service. But let's assume that the returns satisfy a few reasonable properties.

1. The total return can be calculated from the return of each individual project given projects before it - e.g., the return of Internet service alone, the computer given the Internet service, the software given the computer and internet, etc.
2. If the return of one project increases (given projects before it) while everything else stays the same, then the total return increases. For instance, if the Internet service gets cheaper, then the return of project A should increase with everything else the same. As a result, the overall return should increase.
3. We can group projects into subprojects without changing the overall return. For instance, we could think of the Internet service and computer as a single project, or we could think of the computer and software as a single project, and either way the total return should stay the same.

Surprisingly, given just these three properties, we can conclude that returns obey a “product rule” similar to the product rule in probability theory.

$$w [ R ( A , B ) ] = w [ R ( A ) ] w [ R ( B | A ) ]$$

where  $w$  is some transformation of returns (e.g., it could be log-return, return-squared, etc.)

This is essentially the first step in [Cox's Theorem](#), a theorem used (most notably by [Jaynes](#)) to ground the logicalist interpretation of probability. But as this post will illustrate, core ideas of Cox's Theorem apply to many real-world systems which we don't usually think of as “probability theory”.

Let's unpack those assumptions a bit more for our investment return example by defining explicit variables on projects and returns. The three key properties are:

1. Return  $R(A, B)$  of A and B together can be computed from the return  $R(A)$  of A alone and the return  $R(B|A)$  of B given A is done. For instance, the return on new high-speed Internet service and new computer together can be calculated from the return on the Internet service alone and the return on the computer given the Internet. Formally,  $R(A, B) = F[R(A), R(B|A)]$  for some function  $F$ .
2. If the return  $R(A)$  goes up without changing  $R(B|A)$ , then the total return  $R(A, B)$  of A and B together increases, and the same conclusion holds for  $R(B|A)$ .

increasing with  $R(A)$  unchanged. For instance, if the cost of high-speed Internet service goes down, then the return  $R(A)$  presumably increases without changing the return  $R(B|A)$  of the new computer given the Internet service, and this should increase the overall return  $R(A, B)$ . Formally,  $F$  is increasing in both arguments.

3. We can group projects A and B, or B and C into subprojects without changing the overall return  $R(A, B, C)$  of all three. For instance, if we want to compute total return  $R(A, B, C)$  on new Internet service, computer and software, we could group together the internet and computer as one hardware-and-network project, then compute  $R(A, B, C)$  from  $R(A, B)$  (hardware-and-network return) along with  $R(C|A, B)$  (return on the software given the hardware-and-network). Alternatively, we could instead group the computer and software as one hardware-and-software project with return  $R(B, C|A)$ , and we should still get the same answer for the return of all three projects together. Formally,
- $$R(A, B, C) = F[R(A, B), R(C|A, B)] = F[R(A), R(B, C|A)].$$

The third rule implies that  $F$  is associative. The key idea we derive here is that all one-dimensional, increasing and associative functions are either multiplication or some transformation of multiplication (e.g., addition/subtraction is log-transformation of multiplication).

Thus we get a product rule:

$$w[R(A, B)] = w[R(A)]w[R(B|A)]$$

where  $w$  is some transformation (reversible) of  $R$ .

More generally, to derive the product rule, we need some objects of interest like  $A, B, C, \dots$ , which serves as input. We also need some kind of real-valued measurement  $R$  of those objects. Then the core requirements for the product rule are:

- $R(A, B)$  is a function of  $R(A)$  and  $R(B|A)$ :

$$R(A, B) = F[R(A), R(B|A)]$$

for some  $F$ .

- $F$  is increasing with respect to both arguments:

If

$$R(A') > R(A),$$

$$R(B|A') = R(B|A),$$

then

$$R(A', B) > R(A, B).$$

Or, alternatively, if

$$R(B'|A) > R(B|A),$$

$$R(A)=R(A'),$$

then

$$R(A, B') > R(A, B).$$

- We can group objects together without changing the value of measurement R:

$$R(A, B, C) = F[R(A, B), R(C|A, B)] = F[R(A), R(B, C|A)]$$

(Note that for the last assumption, we allow systems in which objects need to be kept in the same order - i.e., A before B before C. This is actually more general than the requirement for the product rule in probability theory, in which the objects are boolean logic variables, so "A and B" = "B and A". If reordering is allowed, then our generalized-product-rule becomes generalized-Bayes-rule.)

The third assumption implies that F is associative. The second implies that it's increasing. The first implies that it's one-dimensional. So, we get the generalized-product-rule.

What does this look like in the context of other real-world systems?

*Example 1:* Suppose I have an investment portfolio with stock A and bond B, and I want to calculate the standard deviation of portfolio return  $R(A, B)$  as a proxy for risk measurement. This calculation is not trivial due to potential correlation of returns between stocks and bonds. For instance, the risk (measured in standard deviation) of investing in stocks alone is usually higher than the risk of investing in a portfolio with stocks and bonds. Let's assume the risks exhibit three properties:

1. Portfolio risk  $R(A, B)$  of stock A and bond B can be calculated from risk  $R(A)$  of stock alone and incremental risk  $R(B|A)$  (positive or negative) of adding bond B

given stock A already in the portfolio.

2. If the risk  $R(A)$  of the stock rises without changing the incremental risk  $R(B|A)$ , then the portfolio risk  $R(A, B)$  rises.
3. Let's consider adding another asset C, an 8-week T-Bills (a type of cash-equivalents) to the investment portfolio. If we're computing new portfolio risk of stock A, bond B, and T-Bills C, then we could group stock and bond together as one sub-portfolio, and compute  $R(A, B, C)$  from  $R(A, B)$  (non-cash-asset) along with  $R(C|A, B)$  (incremental risk of adding T-Bills given the non-cash-asset). Alternatively, we could instead group the bond and T-Bills into one portfolio (non-equity-asset) with risk  $R(B, C|A)$ , and we still get the same risk for all three assets together.

As a result, we can apply the product rule to investment risks:

$$w [ R ( A , B ) ] = w [ R ( A ) ] w [ R ( B | A ) ] ,$$

where  $w$  is some transformation of incremental risk (e.g., exponentiated assuming that those incremental risks add).

*Example 2:* Let's look at a different system in which we're interested in calculating the contribution in points made by basketball players A, B, C,... in a game relative to total points made by the team. For instance,  $R(A)$  could be 30%, meaning Stephen Curry contributes 30% of the total team points in a game,  $R(B)$  could be 25%, meaning Klay Thompson contributes 25% of the total points made, etc. Again, we assume three properties:

1. The points contribution  $R(A, B)$  of Stephen Curry and Klay Thompson together on the court can be calculated from the contribution  $R(A)$  of Stephen Curry alone and incremental contribution  $R(B|A)$  of Klay Thompson given Stephen Curry on the court.
2. If the contribution  $R(A)$  of Stephen Curry increases without changing the incremental contribution  $R(B|A)$ , then the overall contribution  $R(A, B)$  to the team increases as well.
3. We can group Stephen Curry and Klay Thompson together as one "splash" player, and compute  $R(A, B, C)$  from  $R(A, B)$  (contribution of the "splash") along with  $R(C|A, B)$  (incremental contribution of Draymond Green given the "splash"). Alternatively, we could group Klay Thomson and Draymond Green as one big-

man player with the contribution  $R(B, C|A)$ , and we will get the same total contribution for all three players together on the court.

Thus, we can have the product rule applied to basketball players' shooting percentage:

$$w [ R ( A , B ) ] = w [ R ( A ) ] w [ R ( B | A ) ]$$

where  $w$  is some transformation of player contribution.

*Example 3:* Let's consider a modified version of the classic [traveling salesman problem](#) in theoretical computer science and operations research. We're interested in finding the shortest travel time from an origin to cities A, B, C, .... Presumably the shortest travel time satisfy three assumptions:

1. The shortest time  $R(A, B)$  of visiting cities A and B exactly once from origin can be computed from  $R(A)$  of visiting city A and added time  $R(B|A)$  of visiting city B given we already visited city A.
2. If the shortest time  $R(A)$  of visiting city A increases without changing the additional travel time  $R(B|A)$ , then the total traveling time  $R(A, B)$  of visiting both city A and B increases.
3. Let's add another city C to our travel plan. We can group city A and B together as one region, and compute  $R(A, B, C)$  of shortest travel time to visit A, B, and C exactly once from  $R(A, B)$  of visiting the region with cities A and B along with  $R(C|A, B)$  (additional time it adds to visit city C along with city A and B to the total trip time). We could also instead group city B and C together with shortest travel time  $R(B, C|A)$ , and we will get the same answer for visiting every city exactly once in our trip.

With these three assumption above, we could apply the generalized product rule to the shortest travel time problem:

$$w [ R ( A , B ) ] = w [ R ( A ) ] w [ R ( B | A ) ],$$

where  $w$  is some transformation (reversible) of shortest travel time (e.g., exponentiated shortest travel time).

## Summary

The product rule in probability,  $p(AB) = p(A)p(B|A)$ , states that the probability  $p(AB)$  of both A and B are true can be calculated by using probability  $p(A)$  of A being true alone and probability  $p(B|A)$  of B being true given A is true. The conditions of the product rule suggest possible avenues to extend the traditional product rule to deal with things that are not restricted to logical boolean type. In particular, this post suggests continuing to use the product rule to represent real-valued measurements of objects A, B, C,... that satisfy a few fairly reasonable properties and proposes a generalized form of the product rule  $w[R(A, B)] = w[R(A)]w[R(B|A)]$ . R is some kind of real-number measurement and w is some transformation of R. For instance, in the company investment project example we have  $w[R(A, B)] = w[R(A)]w[R(B|A)]$  where R represents the project return and w can be log return.

# Covid 6/24: The Spanish Prisoner

The last scare is underway. Delta is an increasing share of Covid cases around the world, causing cases in many places to rise. Are enough people vaccinated? How bad are things going to get before we turn the corner one final time?

The incremental news was not good. Calculations that looked comfortable last week look less comfortable now. I still expect things to mostly be fine, especially in areas with high mRNA vaccination rates.

Also: John McAfee found dead in a Spanish prison. If you think he killed himself I have some computer security software and I'd like to sell you a subscription. Works great.

Let's run the numbers.

## The Numbers

### Predictions

Prediction from last week: Positivity rate of 1.8% (down 0.1%), deaths fall by 9%.

Result: Positivity rate of 1.8% (down 0.1%), and deaths fall by 9%.

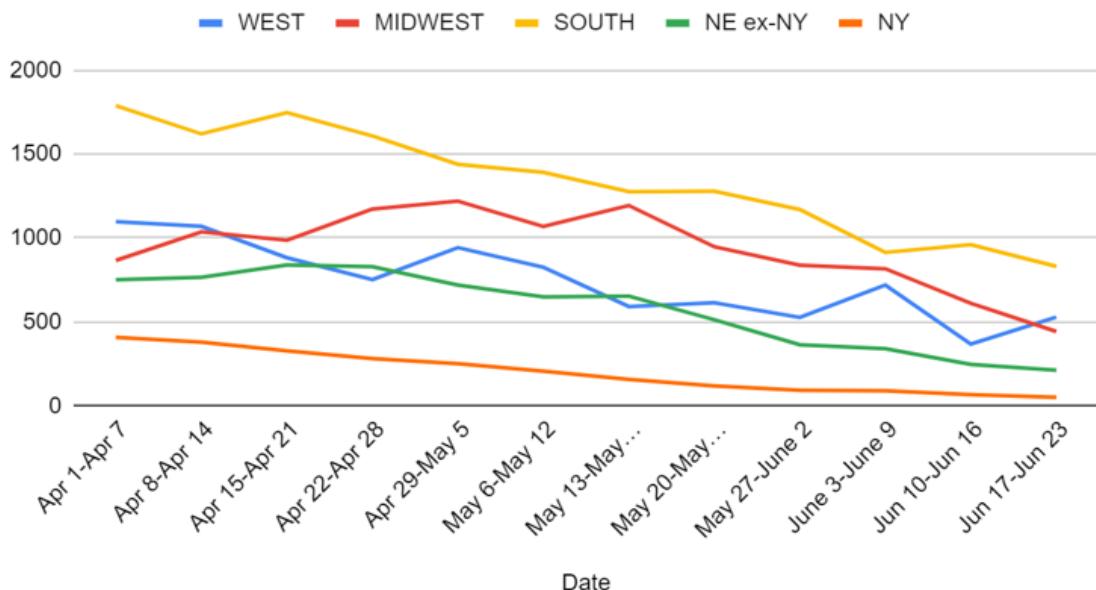
Prediction for next week: Positivity rate of 1.8% (unchanged) and deaths fall by 8%.

Got this week on the nose. With the rise of Delta and the shift in tests from safe to unsafe regions, I no longer expect the positivity rate to continue to decline, and if anything an uptick is more likely than a downtick. For deaths, there's no reason to think things won't improve for a few more weeks.

### Deaths

| Date          | WEST | MIDWEST | SOUTH | NORTHEAST | TOTAL |
|---------------|------|---------|-------|-----------|-------|
| May 13-May 19 | 592  | 1194    | 1277  | 811       | 3874  |
| May 20-May 26 | 615  | 948     | 1279  | 631       | 3473  |
| May 27-June 2 | 527  | 838     | 1170  | 456       | 2991  |
| June 3-June 9 | 720  | 817     | 915   | 431       | 2883  |
| Jun 10-Jun 16 | 368  | 611     | 961   | 314       | 2254  |
| Jun 17-Jun 23 | 529  | 443     | 831   | 263       | 2066  |

## Deaths by Region



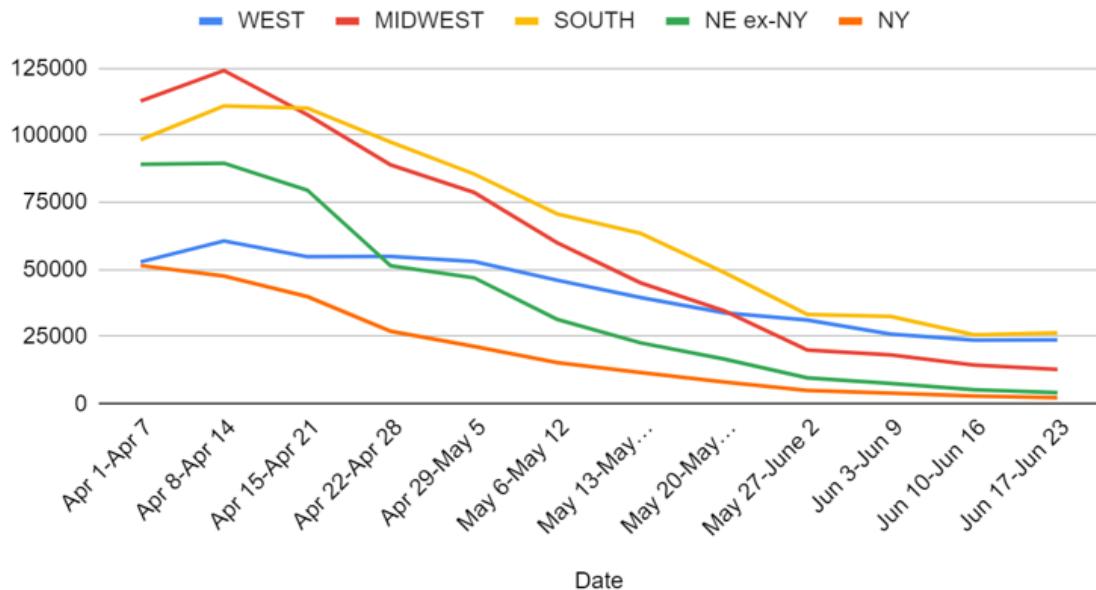
As discussed last week, I've shrunk the graph so we can see what's happening recently, which was otherwise impossible to read. We saw progress this week, but the West's number last week was indeed ahead of itself, so we saw only modest overall progress and hit the 9% decline target exactly. Things now seem like they're back on the expected track and the orange New York line is down to 51 deaths last week.

We should expect to see things continue to improve, but the increasing share of Delta infections does mean the fatality rate should now be rising, given the slow pace of additional vaccinations.

## Cases

| Date          | WEST   | MIDWEST | SOUTH  | NORTHEAST | TOTAL   |
|---------------|--------|---------|--------|-----------|---------|
| May 6-May 12  | 46,045 | 59,945  | 70,740 | 46,782    | 223,512 |
| May 13-May 19 | 39,601 | 45,030  | 63,529 | 34,309    | 182,469 |
| May 20-May 26 | 33,890 | 34,694  | 48,973 | 24,849    | 142,406 |
| May 27-June 2 | 31,172 | 20,044  | 33,293 | 14,660    | 99,169  |
| Jun 3-Jun 9   | 25,987 | 18,267  | 32,545 | 11,540    | 88,339  |
| Jun 10-Jun 16 | 23,700 | 14,472  | 25,752 | 8,177     | 72,101  |
| Jun 17-Jun 23 | 23,854 | 12,801  | 26,456 | 6,464     | 69,575  |

## Positive Tests by Region

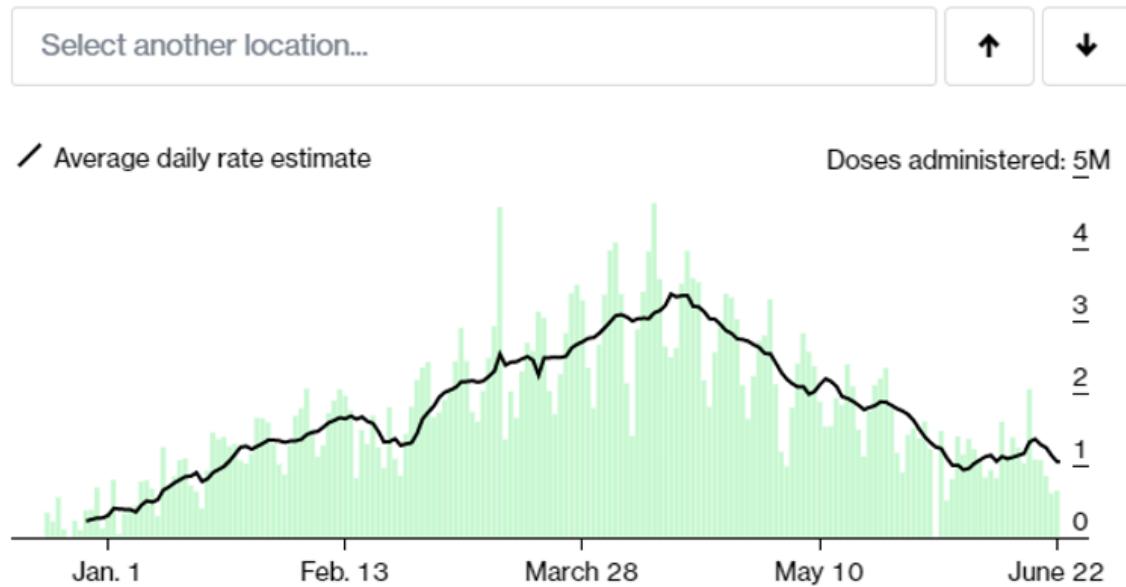


The lack of progress here this week is quite worrisome. Delta is only up to around 35% (the data says 30%, but it's lagged to start with and three days old to boot, and makes me think the 35% I estimate extrapolating from last week is likely slightly low rather than high), so if that's enough to get us treading water, there's going to be trouble in the more vulnerable areas within a few weeks. The question then is, will it be contained the way the Alpha uptick was, or will it become a much more serious problem? Discussion in the Delta Variant section, but it looks much less optimistic than it did last week.

## Vaccinations

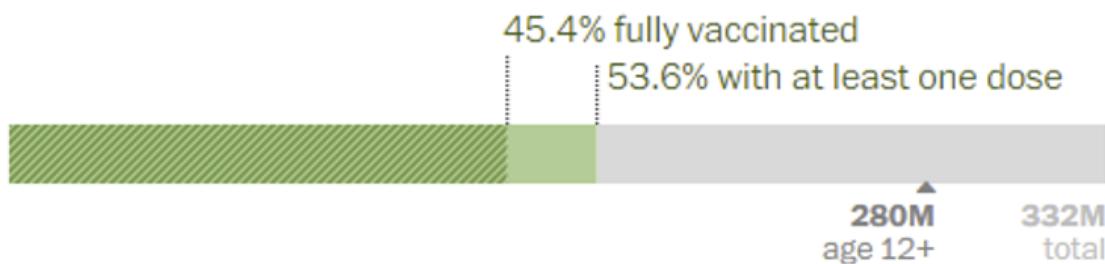
Some substantial differences in Bloomberg vs. Washington Post, but same basic story.

In the U.S., the latest vaccination rate is **1,048,167 doses** per day, on average. At this pace, it will take another **5 months** to cover **75%** of the population.

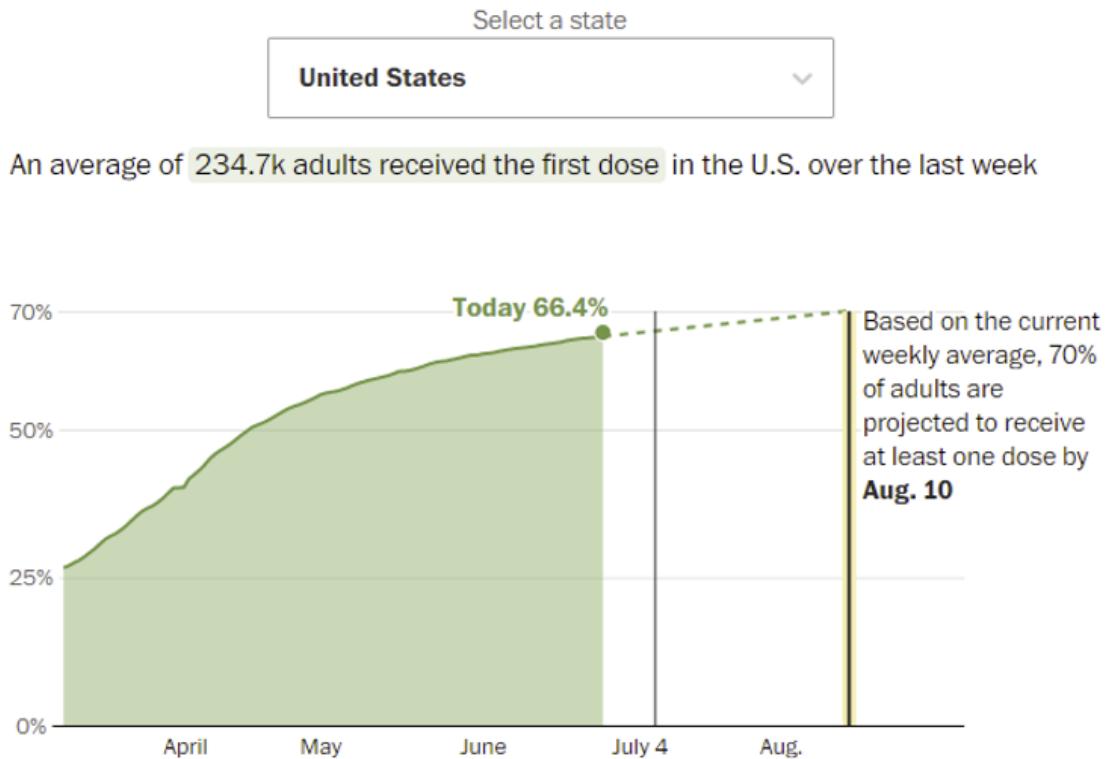


## 177.9 million vaccinated

This includes more than **150.8 million people** who have been fully vaccinated in the United States.



In the last week, an average of **993.8k doses per day** were administered, a **25% decrease** ↓ over the week before.



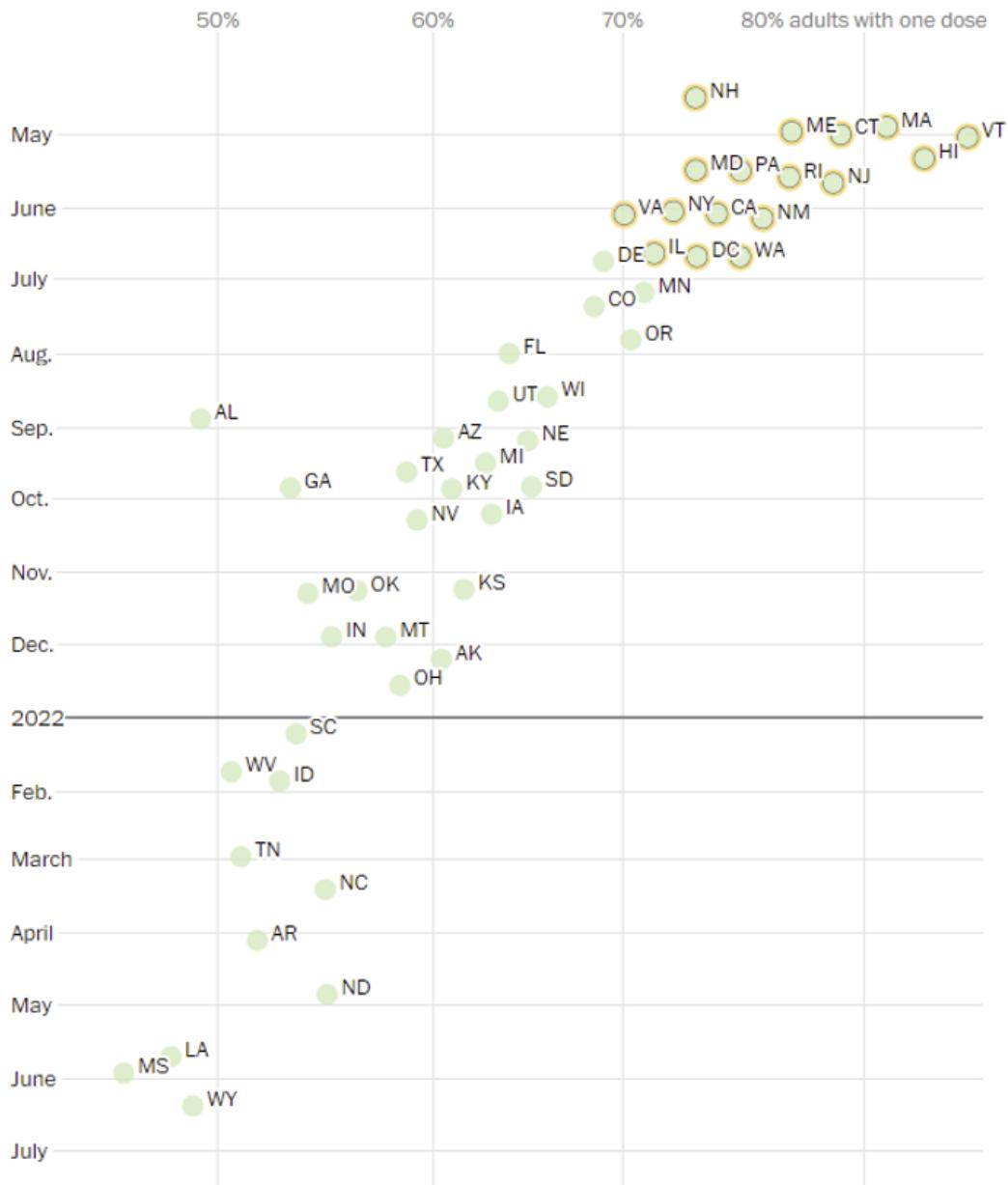
Last few days saw a dramatic drop to a new lower level, which does not bode well. Fully 75% of the last week's doses were second doses, which also bodes quite poorly. Hopefully both have something to do with reporting or some quirk of the calendar. Juneteenth, our latest holiday, is a plausible culprit on both counts. Until I had to adjust for data distortions I never realized how many different holidays we have in America, and for official purposes this one's all new.

Chances are this is some of it, but that a lot of it is that the recent uptick was not sustainable and we will continue our steady decline as we run out of willing arms in which to put shots.

If not, we are going to be stuck not that far beyond our current level of 53.6% first doses for a long time, and the August 10 date for going from 66.4% -> 70% among adults will be far too optimistic, and more like the asymptote until something changes.

Here's a state breakdown:

## When states will vaccinate 70% of adults at current rate



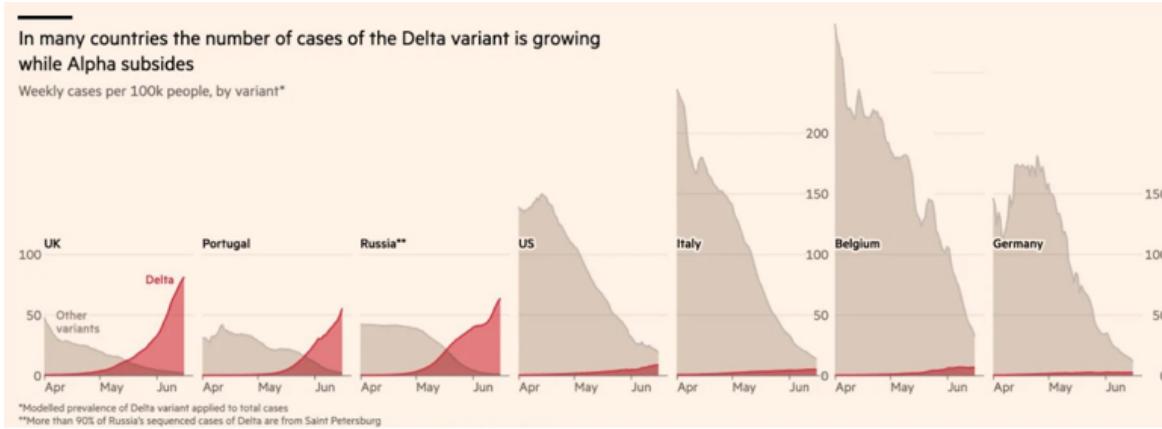
The correct answer to 'when will the states listed as November or later get to 70% if nothing changes' is very clearly never.

People have made their choice. If you didn't want to get vaccinated back when there were lots of cases, why would you change your mind now, with everyone feeling safe and the country reopening?

Delta. It's time.

## Delta Variant

[I did a bunch of calculations last week](#) to figure out if we could handle the Delta variant. There's also a much simpler way to view the situation. Delta *is* the forward-looking pandemic. Anything less dangerous than Delta, including Gamma/P1, will soon be dominated by Delta, so it's a reasonable approach to ignore the total numbers and look only [at the Delta numbers](#) in various places. Essentially, treat the red lines as the pandemic here, not the grey:



The problem with that approach is it requires a good appreciation of the rate of absolute growth of Delta rather than its current level, which is trickier to get right and causes errors to get compounded. Also, such a model would freak out about every variant we've ever seen emerge anywhere. It's still clearly not going to be a comforting answer.

[The best data source I found last week](#) continues to be the best known source. It shows us up to 30% Delta for data reported as of Monday the 21st. If the estimates from last week are accurate, it should become a majority of infections within a week, and become dominant soon thereafter. Within a month it will effectively be the pandemic's present, not only its future.

The only questions are, can we handle it? What happens if we can't?

Last week I approximated 25% Delta, which if we take its reproduction rate seriously is now 32% Delta after an additional 5-day cycle, or 35% now. That change only adds a few percent to R<sub>0</sub>, so it doesn't explain the lack of progress last week in case counts. As we've seen in the past, weekly numbers can be quirky.

The problem is that last week's estimate of how fast we were vaccinating was too optimistic. Part of that was a bad estimate, part of that was bad news over the last week, but it looks like a more realistic guess is 0.5% off the population per week rather than 1%. Also, I have to increase my previous guess on the old R<sub>0</sub> from 0.84 to more like 0.86 after this week's numbers.

That would start us out at R<sub>0</sub> = 1.13 with a gain each week of only 1% that is plausibly fading, if you ignore immunity from infections and further control systems entirely in all directions, which would take us three months to turn the corner. During that time, Delta would go up by a factor of 6 or so (and all other variants would presumably get mostly wiped out) which would mean we'd peak around double our current case rate.

Basic sanity checks on these calculations look like they check out, and are at least in the right ballpark.

That ignores that different regions are different, which likely makes things worse overall. The Northeast essentially 'wins' and shrugs things off, while the South could be in serious

trouble. Which in turn leads to control systems, including an increased willingness to get vaccinated.

That's not great. Is it too close for full comfort? Absolutely. But given that this excludes control system reactions and the effects of immunity from infection, both of which work in our favor, and I used a bunch of conservative estimates in various places, it's a worst-case scenario I can live with. An all-Delta world, even with double or triple current rates, is still better than things used to be even for the unvaccinated, that's the peak of trouble before things get better again, and the unvaccinated have a choice to change their status.

The usual suspects are out in force, as one would expect, raising the alarm. On the lighter side, [some of them cut more corners than others...](#)



**Venk Murthy** @venkmurthy · 15h

Simply drawing in lines is not good for credibility.

...

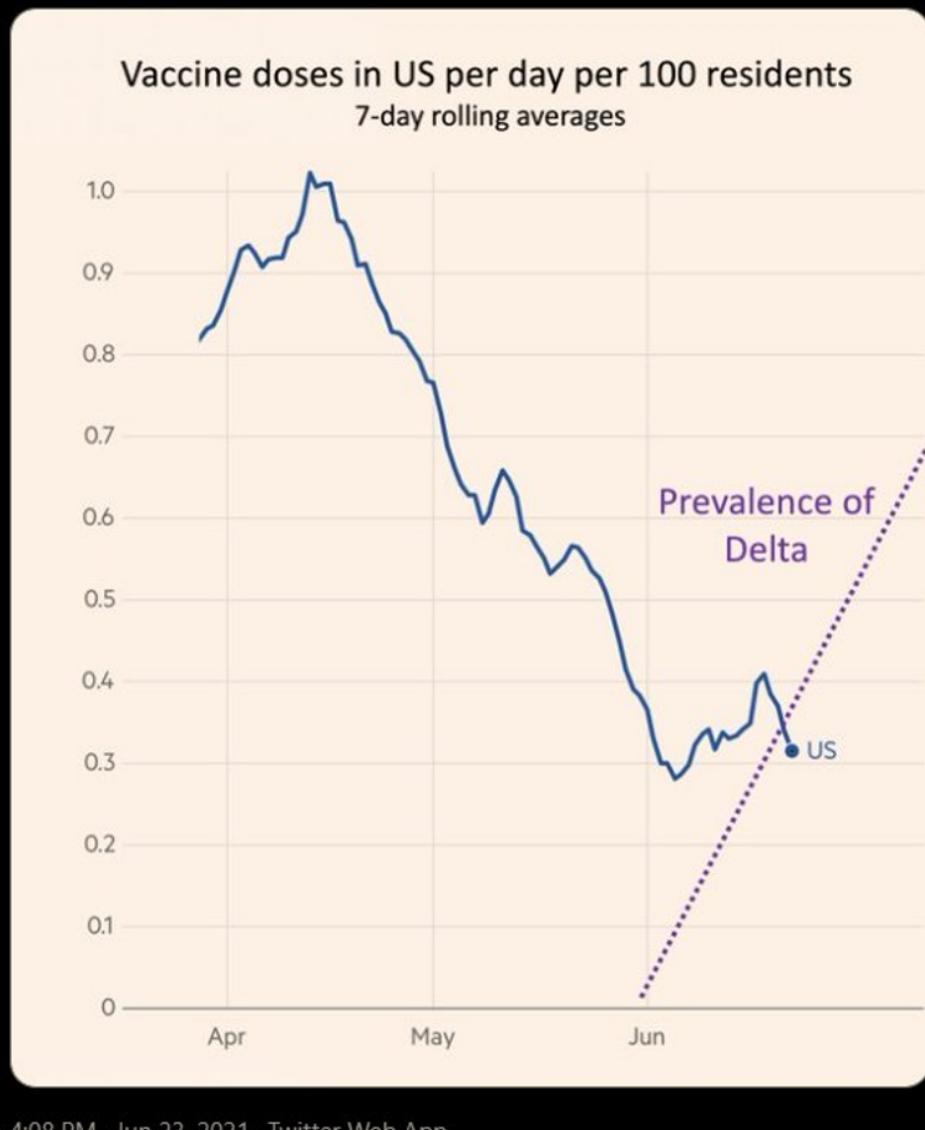
On a positive note, delta is only growing linearly rather than exponentially per Dr. Topol!



Eric Topol @EricTopol

...

Going in the wrong direction  
As the Delta variant heads towards dominance, the falloff of vaccinations, ~600K doses past 3 days. Not seen since the early (January) days of the US vaccination campaign



4:08 PM · Jun 23, 2021 · Twitter Web App

55 Retweets 7 Quote Tweets 172 Likes

Did you know that you... can... just... put... lines... into... graphs? And label them however you like, with no relation to reality? Don't wait. Use  $y = ax + b$  today!

## Seasonality

This is the second year of Covid posts, so when we point to the South and say ‘this is what happens when people refuse to get vaccinated’ perhaps we should remember what the headline was a year ago?

It was [6/18/20: The Virus Goes South](#), followed by [6/25/20: The Dam Breaks](#). It looked like the South, plus Arizona, had lost control of the situation entirely. Then two weeks later, things there peaked and started declining again. Once again, I had underestimated the control system. It wouldn’t be my last time.

Exactly one year later, the exact same spots are showing the same pattern, to a *lesser* extent than last time. It hardly seems fair to ascribe this primarily to vaccination differences, even if the math says that those differences are a very big game. This is not our first rodeo, and I’m not about to repeat the same mistakes that easily.

This goes both ways. We also shouldn’t get too cocky about things going well in the Northeast in the same way they went well last year, even if we have good reasons to be optimistic this time around.

## In Other News

[I remember what it was like to have this kind of failure of imagination.](#) I miss those days.



**Eliezer Yudkowsky** @ESYudkowsky · 19h

...

I cannot imagine what it feels like from the inside to seriously utter this as a criticism and mean it. I am unable to imagine the speaker as anything but the 2D caricature of an Ayn Rand villain.



**Thom Brooks** @thom\_brooks · Jun 23

Tory incompetence, again: Dominic Cummings tried to fast track £530,000 grant to external Covid data team with 'no procurement, no lawyers, no meetings, no delay', leaked e-mails say [mol.im/a/9714623](http://mol.im/a/9714623)

[Claim that vaccinated children in Los Angeles will be required to wear masks](#) while at the places they are required to report to for the bulk of their day. On the plus side, if they opt for ‘distance instruction’ they only get ‘scant hours of instruction’ leaving them free to spend the rest of their time learning.

Also worth noting is that I don’t think this is how math works, on multiple levels? Am I missing something?

The tentative agreement, which was signed by the bargaining teams Monday afternoon, calls for an average class size for grades TK-3 of 22 students, with a maximum of 23 students. It also mandates an average of 30 students (with a 31 student maximum) for grades 4-8 in elementary schools and also in K-8 schools. Math, English, English language development, social sciences and science classes at middle and high schools will have a maximum of 36 students, with a maximum of 52 students for physical education and 37 for other classes.

[Thread on vaccine situation in Taiwan](#), which is a disaster on multiple levels, and its interaction with the question of reunification with China. Especially after what happened in Hong Kong, I notice I am confused how so many could take such an indifferent attitude towards reunification.

[MR once again hammers home that the Moderna dose of 100ug is clearly too big](#). I continue to think this undersells the case, because 100ug isn't merely unnecessary overkill, it is likely *actively worse* than 50ug because the extra size makes the next-day side effects worse, without a meaningful increase in the level of protection.

[Paper documents loss of grey matter in the brain after getting Covid-19](#), including for those who were not hospitalized – hospitalization did not seem to impact the magnitude of this effect. You do not want to get Covid-19. Given the timing this does not provide information on vaccinated people who then still got infected, nor does it differentiate between severity levels beyond whether someone was hospitalized. I do not have a good sense of what size impact one should expect from the effect observed here – it's easy for this type of thing to be quite impactful, and also easy for it to sound scary while not having much impact at all.

[Small study finds 39% risk of Covid-19 spread between roommates in hospitals](#). I didn't get a chance to form an unanchored prior on this number but it seems well within the range of numbers I would have expected.

[Mask wearing study](#) with very large data sets finds that everyone wearing masks most or all of the time in public places leads to a 25% reduction in the rate of reproduction ( $R_0$ ). That is being interpreted as 'masks work' but I would caution both that the error bars on this are gigantic in both directions, and that this effect is actually smaller than I would have guessed, if it is indeed contrasting mostly wearing masks to never wearing masks. They couldn't find any effect of mask mandates on rates of mask wearing, which is a methodological problem rather than a real observation of causal reality. Mask mandates absolutely change the rate of mask wearing, and the lifting of mandates has clearly reduced mask wearing in ways I have observed with my own eyes. I do think it's fair to take results like this and update towards mandates having less impact than one would otherwise think, especially local mandates in a world where masks were being adopted generally anyway, and to put more of the 'work' involved on private reactions. The mistake would be either to translate the 'masks work' here as 'mask mandates work,' which they definitely did *not* find, or to translate the 'mask mandates not found to increase mask wearing' here as evidence that they don't increase mask wearing. Two easy to make and opposite mistakes.

## **Someone Really Ought To (Somehow) Do a (Better) Study**

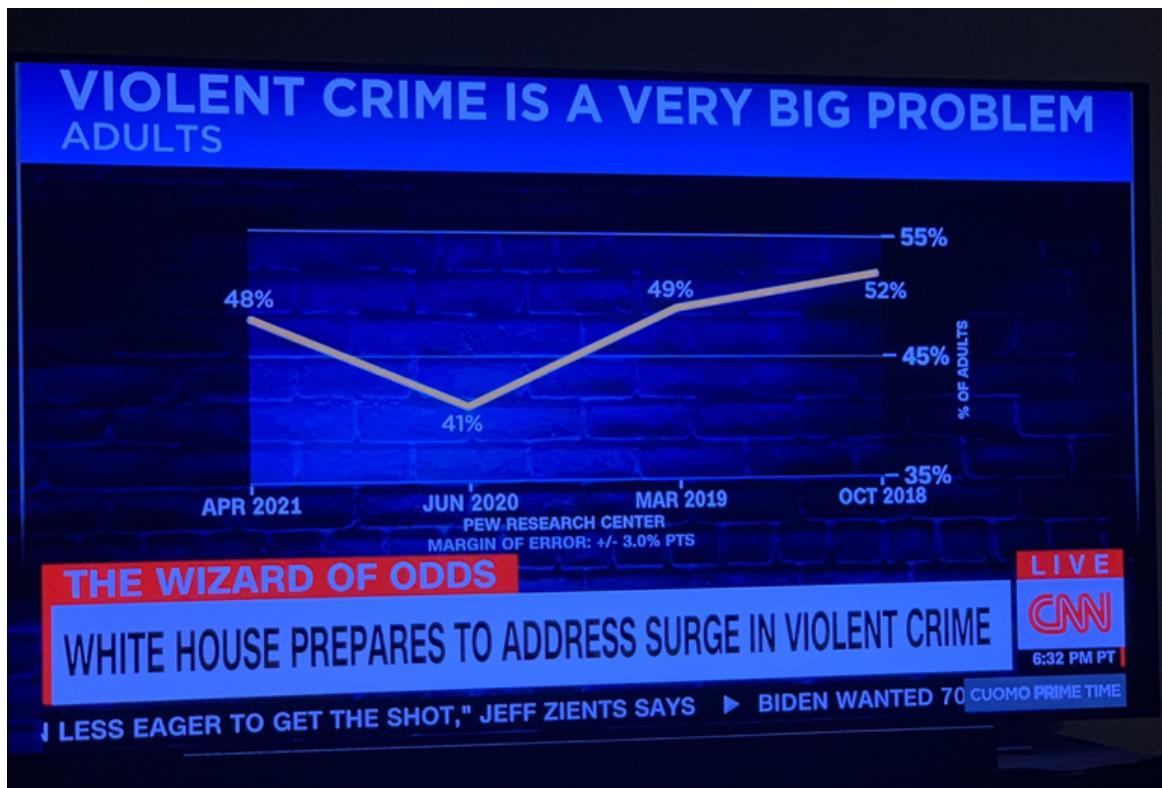
[Marginal Revolution highlights a NYT piece on the CDC and how broken it is](#), calls it one of the best pieces of the year. Giving the secondary link as a compromise with my NYT ban, as the content here is being represented by a credible source as unusually impressive and important.

[A study on shelter in place orders and their impact on excess mortality](#). My conclusion is that the decision on when to issue such an order, and the counterfactual situation if orders weren't issued (both in terms of people's actions, and in the medium-term path of the pandemic), are sufficiently hopelessly confounded that this doesn't provide much if any insight into what is happening here.

[Another study finds that lockdowns kill children](#), especially in developing areas, 'potentially' 1.76 child deaths per pandemic death averted in the least developed areas. As usual, they use an SIR model, which is a very poor way of creating a realistic counterfactual path of the pandemic in the absence of lockdowns, so I don't take these results all that seriously beyond the headline fact that there's very large downsides to locking down that are even larger in undeveloped areas. That part seems very true, and worth appreciating.

## Not Covid

Lying with statistics, CNN [reversed x-axis edition](#):



[John McAfee found dead in a Spanish prison](#) while awaiting extradition to the United States.

A statement from the Catalonia regional government Justice Department, which manages prisons there, said that prison medical personnel and guards attempted to perform life-saving procedures after finding McAfee, but were unsuccessful. The statement said "everything indicates" that McAfee could have died by suicide.

That 'could have' is doing a refreshingly large amount of work, but, I mean, no:



John McAfee ✅ @officialmcafee · Dec 1, 2019

...

Getting subtle messages from U.S. officials saying, in effect: "We're coming for you McAfee! We're going to kill yourself". I got a tattoo today just in case. If I suicide myself, I didn't. I was whackd. Check my right arm.

\$WHACKD available only on [McAfeedex.com](https://McAfeedex.com):



1.8K

32.7K

38.7K



Everything we know about John McAfee, including getting a tattoo years in advance to indicate he didn't kill himself in case he was found in exactly this situation, strongly indicates he didn't kill himself.

That's a pretty strong move. If you did that and people still think you killed yourself that has to be a rolling-in-your-grave-thinking-literally-what-the-f\*\*\*-did-you-want-from-me moment. Or it would be, if John was the type to care what other people think enough to bother rolling.

Still, while I find it unlikely, the 'could have' technically stands! Cause I gotta think: If McAfee explicitly set the stage to make us all think he was whacked by the government and would never kill himself, then killed himself anyway, would that be both a totally baller move and totally in character? Absolutely. And I'd have to put on a hat, so I could tip it.

In gaming news, while working on Emergents, I've had the chance to explore [Roguebook](#), the new rogue deckbuilder from a team that includes Richard Garfield (the creator of Magic: The Gathering). Richard is awesome and takes lots of big swings, and this is effectively a spiritual sequel to [Slay the Spire](#), which is so good it is the game I recommend to most people when they ask me what they should be playing, so I was hoping for greatness. What I got instead was... a solid game that does not swing for the fences and I don't think is as good as Slay the Spire, but which offers enough new twists on the formula to be well worth playing if you're up for another experience in the same style. I tentatively have it in Tier 3, a game worth playing if you like the genre.

The game I'm enjoying the most recently is a little thing called [Slipways](#), a chill puzzle/strategy 3X game that's definitely worth checking out. Not yet sure if it's quite Tier 1 (Must Play), but definitely at least Tier 2 (Worth It).

Finally, did you know that supplies are sufficiently backed up that I've been warned that if I want to order nice furniture for our new apartment in NYC, it might not get here until *January*?

# Dangerous optimisation includes variance minimisation

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

Let's look again at Stuart Russell's quote:

A system that is optimizing a function of  $n$  variables, where the objective depends on a subset of size  $k < n$ , will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.

It is not immediately obvious that this is true. If we are maximising (or minimising) some of these  $k$  variables, then it's likely true, if this maximisation or minimisation brings them to unusual values that we wouldn't encounter "naturally". But things might be different if the variables need only be set to certain "plausible" values.

Suppose, for example, that an AI is building widgets, and that it is motivated to increase widget production  $w$ . It can choose the following policies, with the following consequences:

1.  $\pi_1$ : build widgets the conventional way;  $w = 10$ .
2.  $\pi_2$ : build widgets efficiently;  $w = 15$ .
3.  $\pi_3$ : introduce new innovative ways of building widgets;  $w = 30$ .
4.  $\pi_4$ : dominate the world's widget industry;  $w = 10,000$ .
5.  $\pi_5$ : take over the world, optimise the universe for widget production;  $w = 10^{60}$ .

If the AI's goal is to maximise  $w$  without limit, then the fifth option becomes attractive to it. Even if it just wants to set  $w$  to a limited but high value -  $w = 10,000$  - it benefits from more control of the world. In short:

- The more unusual the values of the variables the AI wants to reach, the more benefit it gets from strong control over the universe.

But what if the AI was designed to set  $w$  to 15, or in the range 20 – 30? Then it would seem that it has lower incentive for control, and might just "do its job", the way we'd like it to; the other variables would not be set to extreme values, since there is no need for the AI to change things much.

Eliezer and Nick and others have made the point that this is still not safe, in posts that I can't currently find. They use examples like the AI taking over the world and building cameras to be sure that it constructed 15 widgets exactly. These scenarios seem

extreme as intuitions pumps, to some, so I thought it would be simpler to rephrase this as: moving the variance to unusual values.

## Variance control

Suppose that the AI was designed to keep  $v$  at 15. We could give it the utility function  $-(v - 15)^2$ , for instance. Would it then stick to policy  $\pi_2$ ?

Now assume further that the world is not totally static. Random events happen, increasing or decreasing the production of widgets. If the AI follows policy  $\pi$ , then its expected reward is:

$$\begin{aligned} E[-(v - 15)^2 | \pi] &= -E(v^2 | \pi) + 2E(v | \pi) - 15^2 \\ &= -\text{var}(v | \pi) - (E(v | \pi) - 15)^2. \end{aligned}$$

The second term,  $-(E(v | \pi) - 15)^2$ , the AI could control by "doing its job" and picking a human-safe policy. But it also wants to control the variance of  $v$ , specifically it wants to lower it. Even more specifically, it wants to move that variance to a very low, highly unusual value.<sup>[1]</sup>

So the previous problem appears again: it wants to move a variable - the variance of  $v$  - to a very unusual value. In the real world, this could translate to it building excess capacity, taking control of its supply chains, removing any humans that might get in the way, etc... Since "humans that might get in the way" would end up being most humans - few nations would tolerate a powerful AI limiting their power and potential - this tends to the classic "take control of the world" scenario.

## Conclusion

So, minimising or maximising a variable, or setting it to an unusual value, is dangerous, as it incentivises the AI to take control of the world to achieve those unusual values. But setting a variable to a usual value can also be dangerous, in an uncertain world, as it incentivises the AI to take control of the world to set the variability of that variable to unusually low levels.

*Thanks to Rebecca Gorman for the conversation which helped me clarify these thoughts.*

- 
1. This is not a specific feature of using a square in  $-(v - 15)^2$ . To incentivise the AI to set  $v = 15$ , we need a function of  $v$  that peaks at 15. This makes it concave-ish around 15, which is what penalises spread and uncertainty and variance. ↪