

Best of LessWrong: March 2021

1. [Rationalism before the Sequences](#)
2. [Politics is way too meta](#)
3. [Core Pathways of Aging](#)
4. [A whirlwind tour of Ethereum finance](#)
5. [Seven Years of Spaced Repetition Software in the Classroom](#)
6. [Dark Matters](#)
7. [Strong Evidence is Common](#)
8. [Fun with +12 OOMs of Compute](#)
9. [Jean Monnet: The Guerilla Bureaucrat](#)
10. [The case for aligning narrowly superhuman models](#)
11. [What's So Bad About Ad-Hoc Mathematical Definitions?](#)
12. [Another RadVac Testing Update](#)
13. [My research methodology](#)
14. [Trapped Priors As A Basic Problem Of Rationality](#)
15. [MIRI comments on Cotra's "Case for Aligning Narrowly Superhuman Models"](#)
16. [A Semitechnical Introductory Dialogue on Solomonoff Induction](#)
17. [Demand offsetting](#)
18. [How do we prepare for final crunch time?](#)
19. [Defending the non-central fallacy](#)
20. [Toward A Bayesian Theory Of Willpower](#)
21. [Thirty-three randomly selected bioethics papers](#)
22. [Open, Free, Safe: Choose Two](#)
23. [I'm still mystified by the Born rule](#)
24. [Book review: "A Thousand Brains" by Jeff Hawkins](#)
25. [Quotes from the WWMoR Podcast Episode with Eliezer](#)
26. [Covid 3/12: New CDC Guidelines Available](#)
27. [Coherence arguments imply a force for goal-directed behavior](#)
28. [Direct effects matter!](#)
29. [Covid 3/25: Own Goals](#)
30. [A Retrospective Look at the Guild of Servants: Alpha Phase](#)
31. [How You Can Gain Self Control Without "Self-Control"](#)
32. [The average North Korean mathematician](#)
33. [Chaos Induces Abstractions](#)
34. [MetaPrompt: a tool for telling yourself what to do.](#)
35. [How does bee learning compare with machine learning?](#)
36. [Logan Strohl on exercise norms](#)
37. [Kelly *is* \(just\) about logarithmic utility](#)
38. [Covid 3/4: Declare Victory and Leave Home](#)
39. [Multimodal Neurons in Artificial Neural Networks](#)
40. [Texas Freeze Retrospective: meetup notes](#)
41. [How To Think About Overparameterized Models](#)
42. [\[Lecture Club\] Awakening from the Meaning Crisis](#)
43. [Introducing Metaforecast: A Forecast Aggregator and Search Tool](#)
44. [The Point of Easy Progress](#)
45. [Epistemological Framing for AI Alignment Research](#)
46. [Generalizing POWER to multi-agent games](#)
47. [Clubhouse](#)
48. [The Flexibility of Abstract Concepts](#)
49. [Bureaucracy is a world of magic](#)
50. [Covid 3/18: An Expected Quantity of Blood Clots](#)

Best of LessWrong: March 2021

1. [Rationalism before the Sequences](#)
2. [Politics is way too meta](#)
3. [Core Pathways of Aging](#)
4. [A whirlwind tour of Ethereum finance](#)
5. [Seven Years of Spaced Repetition Software in the Classroom](#)
6. [Dark Matters](#)
7. [Strong Evidence is Common](#)
8. [Fun with +12 OOMs of Compute](#)
9. [Jean Monnet: The Guerilla Bureaucrat](#)
10. [The case for aligning narrowly superhuman models](#)
11. [What's So Bad About Ad-Hoc Mathematical Definitions?](#)
12. [Another RadVac Testing Update](#)
13. [My research methodology](#)
14. [Trapped Priors As A Basic Problem Of Rationality](#)
15. [MIRI comments on Cotra's "Case for Aligning Narrowly Superhuman Models"](#)
16. [A Semitechnical Introductory Dialogue on Solomonoff Induction](#)
17. [Demand offsetting](#)
18. [How do we prepare for final crunch time?](#)
19. [Defending the non-central fallacy](#)
20. [Toward A Bayesian Theory Of Willpower](#)
21. [Thirty-three randomly selected bioethics papers](#)
22. [Open, Free, Safe: Choose Two](#)
23. [I'm still mystified by the Born rule](#)
24. [Book review: "A Thousand Brains" by Jeff Hawkins](#)
25. [Quotes from the WWMoR Podcast Episode with Eliezer](#)
26. [Covid 3/12: New CDC Guidelines Available](#)
27. [Coherence arguments imply a force for goal-directed behavior](#)
28. [Direct effects matter!](#)
29. [Covid 3/25: Own Goals](#)
30. [A Retrospective Look at the Guild of Servants: Alpha Phase](#)
31. [How You Can Gain Self Control Without "Self-Control"](#)
32. [The average North Korean mathematician](#)
33. [Chaos Induces Abstractions](#)
34. [MetaPrompt: a tool for telling yourself what to do.](#)
35. [How does bee learning compare with machine learning?](#)
36. [Logan Strohl on exercise norms](#)
37. [Kelly *is* \(just\) about logarithmic utility](#)
38. [Covid 3/4: Declare Victory and Leave Home](#)
39. [Multimodal Neurons in Artificial Neural Networks](#)
40. [Texas Freeze Retrospective: meetup notes](#)
41. [How To Think About Overparameterized Models](#)
42. [\[Lecture Club\] Awakening from the Meaning Crisis](#)
43. [Introducing Metaforecast: A Forecast Aggregator and Search Tool](#)
44. [The Point of Easy Progress](#)
45. [Epistemological Framing for AI Alignment Research](#)
46. [Generalizing POWER to multi-agent games](#)
47. [Clubhouse](#)
48. [The Flexibility of Abstract Concepts](#)
49. [Bureaucracy is a world of magic](#)

50. [Covid 3/18: An Expected Quantity of Blood Clots](#)

Rationalism before the Sequences

I'm here to tell you a story about what it was like to be a rationalist decades before the Sequences and the formation of the modern rationalist community. It is not the only story that could be told, but it is one that runs parallel to and has important connections to Eliezer Yudkowsky's and how his ideas developed.

My goal in writing this essay is to give the LW community a sense of the prehistory of their movement. It is not intended to be "where Eliezer got his ideas"; that would be stupidly reductive. I aim more to exhibit where the drive and spirit of the Yudkowskian reform came from, and the interesting ways in which Eliezer's formative experiences were not unique.

My standing to write this essay begins with the fact that I am roughly 20 years older than Eliezer and read many of his sources before he was old enough to read. I was acquainted with him over an email list before he wrote the Sequences, though I somehow managed to forget those interactions afterwards and only rediscovered them while researching for this essay. In 2005 he had even sent me a book manuscript to review that covered some of the Sequences topics.

My reaction on reading "The Twelve Virtues of Rationality" a few years later was dual. It was a different kind of writing than the book manuscript - stronger, more individual, taking some serious risks. On the one hand, I was deeply impressed by its clarity and courage. On the other hand, much of it seemed very familiar, full of hints and callbacks and allusions to books I knew very well.

Today it is probably more difficult to back-read Eliezer's sources than it was in 2006, because the body of more recent work within his reformation of rationalism tends to get in the way. I'm going to attempt to draw aside that veil by talking about four specific topics: General Semantics, analytic philosophy, science fiction, and Zen Buddhism.

Before I get to those specifics, I want to try to convey that sense of *what it was like*. I was a bright geeky kid in the 1960s and 1970s, immersed in a lot of obscure topics often with an implicit common theme: intelligence can save us! Learning how to think more clearly can make us better! But at the beginning I was groping as if in a dense fog, unclear about how to turn that belief into actionable advice.

Sometimes I would get a flash of light through the fog, or at least a sense that there were other people on the same lonely quest. A bit of that sense sometimes drifted over USENET, an early precursor of today's Internet fora. More often than not, though, the clue would be fictional; somebody's imagination about what it would be like to increase intelligence, to burn away error and think more clearly.

When I found non-fiction sources on rationality and intelligence increase I devoured them. Alas, most were useless junk. But in a few places I found gold. Not by coincidence, the places I found real value were sources Eliezer would later draw on. I'm not guessing about this, I was able to confirm it first from Eliezer's explicit reports of what influenced him and then via an email conversation.

Eliezer and I were not unique. We know directly of a few others with experiences like ours. There were likely dozens of others we didn't know - possibly hundreds - on parallel paths, all hungrily seeking clarity of thought, all finding largely overlapping

subsets of clues and techniques because there simply wasn't that much out there to be mined.

One piece of evidence for this parallelism besides Eliezer's reports is that I bounced a draft of this essay off Nancy Lebovitz, a former LW moderator who I've known personally since the 1970s. Her instant reaction? "Full of stuff I knew already."

Around the time Nancy and I first met, some years before Eliezer Yudkowsky was born, my maternal grandfather gave me a book called "People In Quandaries". It was an introduction to General Semantics. I don't know, because I didn't know enough to motivate the question when he was alive, but I strongly suspect that granddad was a member of one of the early GS study groups, probably the same one that included Robert Heinlein (they were near neighbors in Southern California in the early 1940s).

General Semantics is going to be a big part of my story. Twelve Virtues speaks of "carrying your map through to reflecting the territory"; this is a clear, obviously intentional callback to a central GS maxim that runs "The map is not the territory; the word is not the thing defined."

I'm not going to give a primer on GS here. I am going to affirm that it rocked my world, and if the clue in Twelve Virtues weren't enough Eliezer has reported in no uncertain terms that it rocked his too. It was the first time I encountered really actionable advice on the practice of rationality.

Core GS formulations like cultivating consciousness of abstracting, remembering the map/territory distinction, avoiding the verb "to be" and the is-of-identity, that the geometry of the real world is non-Euclidean, that the logic of the real world is non-Aristotelian; these were *useful*. They *helped*. They reduced the inefficiency of my thinking.

For the pre-Sequences rationalist, those of us stumbling around in that fog, GS was typically the most powerful single non-fictional piece of the available toolkit. After the millennium I would find many reflections of it in the Sequences.

This is not, however, meant to imply that GS is some kind of supernal lost wisdom that all rationalists should go back and study. Alfred Korzybski, the founder of General Semantics, was a man of his time, and some of the ideas he formulated in the 1930s have not aged well. Sadly, he was an absolutely terrible writer; reading "Science and Sanity", his magnum opus, is like an endless slog through mud with occasional flashes of world-upending brilliance.

If Eliezer had done nothing else but give GS concepts a better presentation, that would have been a great deal. Indeed, before I read the Sequences I thought giving GS a better finish for the modern reader was something I might have to do myself someday - but Eliezer did most of that, and a good deal more besides, folding in a lot of sound thinking that was unavailable in Korzybski's day.

When I said that Eliezer's sources are probably more difficult to back-read today than they were in 2006, I had GS specifically in mind. Yudkowskian-reform rationalism has since developed a very different language for the large areas where it overlaps GS's concerns. I sometimes find myself in the position of a native Greek speaker hunting for equivalents in that new-fangled Latin; usually present but it can take some effort to bridge the gap.

Next I'm going to talk about some more nonfiction that might have had that kind of importance if a larger subset of aspiring rationalists had known enough about it. And that is the analytic tradition in philosophy.

I asked Eliezer about this and learned that he himself never read any of what I would consider core texts: C.S. Peirce's epoch-making 1878 paper "How To Make Our Ideas Clear", for example, or W.V. Quine's "Two Dogmas of Empiricism". Eliezer got their ideas through secondary sources. How deeply pre-Sequences rationalists drew directly from this well seems to be much more variable than the more consistent theme of early General Semantics exposure.

However: even if filtered through secondary sources, tropes originating in analytic philosophy have ended up being central in every formulated version of rationalism since 1900, including General Semantics and Yudkowskian-reform rationalism. A notable one is the program of reducing philosophical questions to problems in language analysis, seeking some kind of flaw in the map rather than mysterianizing the territory. Another is the definition of "truth" as predictive power over some range of future observables.

But here I want to focus on a subtler point about origins rather than ends: these ideas were in the air around every aspiring rationalist of the last century, certainly including both myself and the younger Eliezer. Glimpses of light through the fog...

This is where I must insert a grumble, one that I hope is instructive about what it was like before the Sequences. I'm using the term "rationalist" retrospectively, but those among us who were seeking a way forward and literate in formal philosophy didn't tend to use that term of ourselves at the time. In fact, I specifically avoided it, and I don't believe I was alone in this.

Here's why. In the history of philosophy, a "rationalist" is one who asserts the superiority of a-priori deductive reasoning over grubby induction from mere material facts. The opposing term is "empiricist", and in fact Yudkowskian-reform "rationalists" are, in strictly correct terminology, skeptical empiricists.

Alas, that ship has long since sailed. We're stuck with "rationalist" as a social label now; the success of the Yudkowskian reform has nailed that down. But it's worth remembering that in this case not only is our map not the territory, it's not even immediately consistent with other equally valid maps.

Now we get to the fun part, where I talk about science fiction.

SF author Greg Bear probably closed the book on attempts to define science fiction as a genre in 1994 when he said "the branch of fantastic literature which affirms the rational knowability of the universe". It shouldn't be surprising, then, that ever since the Campbellian Revolution in 1939 invented modern science fiction there has been an important strain in it of fascination with rationalist self-improvement.

I'm not talking about transhumanism here. The idea that we might, say, upload to machines with vastly greater computational capacity is not one that fed pre-Yudkowskian rationalism, because it wasn't actionable. No; I'm pointing at more attainable fictions about *learning* to think better, or discovering a key that unlocks a higher level of intelligence and rationality in ourselves. "Ultrahumanist" would be a better term for this, and I'll use it in the rest of this essay.

I'm going to describe one such work in some detail, because (a) wearing my SF-historian hat I consider it a central exemplar of the ultrahumanist subgenre, and (b) I know it had a large personal impact on me.

"Gulf", by Robert A. Heinlein, published in the October–November 1949 *Astounding Science Fiction*. A spy on a mission to thwart an evil conspiracy stumbles over a benign one - people who call themselves "Homo Novis" and have cultivated techniques of rationality and intelligence increase, including an invented language that promotes speed and precision of thought. He is recruited by them, and a key part of his training involves learning the language.

At the end of the story he dies while saving the world, but the ostensible plot is not really the point. It's an excuse for Heinlein to play with some ideas, clearly derived in part from General Semantics, about what a "better" human being might look and act like - including, crucially, the moral and ethical dimension. One of the tests the protagonist doesn't know he's passing is when he successfully cooperates in gentling a horse.

The most important traits of the new humans are that (a) they prize rationality under all circumstances - to be accepted by them you have to retain clear thinking and problem-solving capability even when you're stressed, hungry, tired, cold, or in combat; and (b) they're not some kind of mutation or artificial superrace. They are human beings who have chosen to pool their efforts to make themselves more reliably intelligent.

There was a lot of this sort of GS-inspired ultrahumanism going around in Golden Age SF between 1940 and 1960. Other proto-rationalists may have been more energized by other stories in that current. Eliezer remembers and acknowledges "Gulf" as an influence but reports having been more excited by "The World of Null-A" (1946). Isaac Asimov's "Foundation" novels (1942-1953) were important to him as well even though there was not much actionable in them about rationality at the individual level.

As for me, "Gulf" changed the direction of my life when I read it sometime around 1971. Perhaps I would have found that direction anyway, but...teenage me wanted to be homo novis. More, I wanted to deserve to be homo novis. When my grandfather gave me that General Semantics book later in the same decade, I was ready.

That kind of imaginative fuel was tremendously important, because we didn't have a community. We didn't have a shared system. We didn't have hubs like Less Wrong and Slate Star Codex. Each of us had to bootstrap our own rationality technique out of pieces like General Semantics, philosophical pragmatism, the earliest most primitive research on cognitive biases, microeconomics, and the first stirrings of what became evolutionary psych.

Those things gave us the materials. Science fiction gave us the dream, the desire that it took to support the effort of putting it together and finding rational discipline in ourselves.

Last I'm going to touch on Zen Buddhism. Eliezer likes to play with the devices of Zen rhetoric; this has been a feature of his writing since Twelve Virtues. I understood why immediately, because that attraction was obviously driven by something I myself had discovered decades before in trying to construct my own rationalist technique.

Buddhism is a huge, complex cluster of religions. One of its core aims is the rejection of illusions about how the universe is. This has led to a rediscovery, at several points

in its development, of systematic theories aimed at stripping away attachments and illusions. And not just that; also meditative practices intended to shift the practitioner into a mental stance that supports less wrongness.

If you pursue this sort of thing for more than three thousand years, as Buddhists have been doing, you're likely to find some techniques that actually do help you pay better attention to reality - even if it is difficult to dig them out of the surrounding religious encrustations afterwards.

One of the most recent periods of such rediscovery followed the 18th-century revival of Japanese Buddhism by Hakuin Ekaku. There's a fascinating story to be told about how Euro-American culture imported Zen in the early 20th century and refined it even further in the direction Hakuin had taken it, a direction scholars of Buddhism call "ulmatism". I'm not going to reprise that story here, just indicate one important result of it that can inform a rationalist practice.

Here's the thing that Eliezer and I and other 20th-century rationalists noticed; Zen rhetoric and meditation program the brain for epistemic skepticism, for a rejection of language-driven attachments, for not just knowing that the map is not the territory but *feeling* that disjunction.

Somehow, Zen rhetoric's ability to program brains for epistemic skepticism survives not just disconnection from Japanese culture and Buddhist religious claims, but translation out of its original language into English. This is remarkable - and, if you're seeking tools to loosen the grip of preconceptions and biases on your thinking, very useful.

Alfred Korzybski himself noticed this almost as soon as good primary sources on Zen were available in the West, back in the 1930s; early General Semantics speaks of "silence on the objective level" in a very Zen-like way.

No, I'm not saying we all need to become students of Zen any more than I think we all need to go back and immerse ourselves in GS. But co-opting some of Zen's language and techniques is something that Eliezer definitely did. And I did, and other rationalists before the Yudkowskian reformation tended to find their way to.

If you think about all these things in combination - GS, analytic philosophy, Golden Age SF, Zen Buddhism - I think the roots of the Yudkowskian reformation become much easier to understand. Eliezer's quest and the materials he assembled were not unique. His special gift was the same ambition as Alfred Korzybski's; to form from what he had learned a teachable system for becoming less wrong. And, of course, the intellectual firepower to carry that through - if not perfectly, at least well enough to make a huge difference.

If nothing else, I hope this essay will leave you feeling grateful that you no longer have to do a decades-long bootstrapping process the way Eliezer and Nancy and I and others like us had to in the before times. I doubt any of us are sorry we put in the effort, but being able to shortcut a lot of it is a good thing.

Some of you, recognizing my name, will know that I ended up changing the world in my own way a few years before Eliezer began to write the Sequences. That this ensued after long struggle to develop a rationalist practice is not coincidence; if you improve your thinking hard enough over enough time I suspect it's difficult to avoid eventually getting out in front of people who aren't doing that.

That's what Eliezer did, too. In the long run, I rather hope that his reform movement will turn out to have been more important than mine.

Selected sources follow. The fiction list could have been a lot longer, but I filtered pretty strongly for works that somehow addressed useful models of individual rationality training. Marked with * are those Eliezer explicitly reports he has read.

Huikai, Wumen: "The Gateless Barrier" (1228)

Peirce, Charles Sanders: "How To Make Our Ideas Clear" (1878)

Korzybski, Alfred: "Science and Sanity" (1933)

Chase, Stuart: "The Tyranny of Words" (1938)

Hayakawa, S. I: "Language in Thought and Action" (1939) *

Russell, Bertrand: "A History of Western Philosophy" (1945)

Orwell, George: "Politics and the English Language" (1946) *

Johnson, Wendell: "People in Quandaries: The Semantics of Personal Adjustment" (1946)

Van Vogt, A. E: "The World of Null-A" (1946) *

Heinlein, Robert Anson: "Gulf" (1949) *

Quine, Willard Van Orman: "Two Dogmas of Empiricism" (1951)

Heinlein, Robert Anson: "The Moon Is A Harsh Mistress" (1966) *

Williams, George: "Adaptation and Natural Selection" (1966) *

Pirsig, Robert M.: "Zen and the Art of Motorcycle Maintenance" (1974) *

Benares, Camden: "Zen Without Zen Masters" (1977)

Smullyan, Raymond: "The Tao is Silent" (1977) *

Hill, Gregory & Thornley, Kerry W.: "Principia Discordia (5th ed.)" (1979) *

Hofstadter, Douglas: "Gödel, Escher, Bach: An Eternal Golden Braid" (1979) *

Feynman, Richard: "Surely You're Joking, Mr. Feynman!" (1985) *

Pearl, Judea: "Probabilistic Reasoning in Intelligent Systems" (1988) *

Stiegler, Marc: "David's Sling" (1988) *

Zindell, David: "Neverness" (1988) *

Williams, Walter John: "Aristoi" (1992) *

Tooby & Cosmides: "The Adapted Mind: Evolutionary Psychology and the Generation of Culture" (1992) *

Wright, Robert: "The Moral Animal" (1994) *

Jaynes, E.T.: "Probability Theory: The Logic of Science" (1995) *

The assistance of Nancy Lebovitz, Eliezer Yudowsky, Jason Azze, and Ben Pace is gratefully acknowledged. Any errors or inadvertent misrepresentations remain entirely the author's responsibility.

Politics is way too meta

... i.e., it doesn't spend enough time arguing about object-level things.

The way I'm using it in this post, "object-level" might include these kinds of things:

- While serving as Secretary of State, did Hillary Clinton send classified information to an insecure email server? How large was the risk that attackers might get the information, and how much harm might that cause?
- What are the costs and benefits of various communications protocols (e.g., the legal one during Clinton's tenure, the *de facto* one Clinton followed, or other possibilities), and how should we weight those costs and benefits?
- How can we best forecast people's reliability on security issues? Are once-off mistakes like this predictive of future sloppiness? Are there better ways of predicting this?

"Meta" might include things like:

- How much do voters care about Clinton's use of her email server?
- How much will reporters cover this story, and how much is their coverage likely to influence voters?
- What do various people (with no special information or expertise) *believe* about Clinton's email server, and how might these beliefs change their behavior?

I'll also consider discussions of abstract [authority](#), principle, or symbolism more "meta" than concrete policy proposals and questions of fact.

[This](#) is too meta:



Sam @sam_a_bell · Feb 20

crazy to me that the subhead on this is about Trump vs. about response to plague that has taken hundreds of thousands of American lives nytimes.com/2021/02/20/he...

...

Who Will Be the Next F.D.A. Chief?

Two leading contenders generate wider debate about the leadership needed to restore morale and scientific integrity to an agency battered by the politicized Trump administration.

1

5

20

↑



Matthew Yglesias

@mattyglesias

...

Replying to [@sam_a_bell](#)

I read this article and I have no idea what disagreements exist between the two contenders

4:56 PM · Feb 20, 2021 from Washington, DC · Twitter for iPhone

Meta stuff is *real*. Elections are real, and matter. Popularity, status, controversy, and Overton windows have real physical effects.

But it's possible to focus too much on one part of reality and neglect another. If you're driving a car while talking on the phone, the phone and your eyes are both perfectly good information channels; but if you allocate too little attention to the road, you still die.

When speaking the demon's name creates the demon

I claim:

There are many good ideas that start out discussed by blogs and journal articles for a long time, then get adopted by policymakers.

In many of these cases, you could delay adoption by many years by adding more sentences to the blog posts noting the political infeasibility or controversialness of the proposal. Or you could hasten adoption by making the posts just focus on analyzing the effects of the policy, without taking a moment to nervously look over their shoulder,

without ritually bowing to the Overton window as though it were an authority on immigration law.

I also claim that this obeys the same basic causal dynamics as:

My friend Azzie posts something that I find cringe. So I decide to loudly and publicly (!) warn Azzie "hey, the thing you're doing is cringe!". Because, y'know, I want to *help*.

Regardless of how "cringe" the average reader would have considered the post, saying it out loud can only help strengthen the perceived level of cringe.

Or, suppose Bel overhears me and it *doesn't* cause her to see the post as more cringe. Still, it might make Bel worry that *Cathy* and other third parties think that the post is cringe. Which is a sufficient worry on its own to greatly change how Bel interacts with the post. Wouldn't want the cringe monster to come after you next!

This can result in:

Non-well-founded gaffes: statements that are controversial/offensive/impolitic largely or solely because some people think they sound like the kind of thing that would offend, alienate, or be disputed by a hypothetical third party.

Or, worse:

Even-less-than-non-well-founded gaffes: statements that are controversial/offensive/impolitic largely or solely because some people are worried that a hypothetical third party might think that a hypothetical fourth party might be offended, alienated, or unconvinced by the statement.

See: [Common Knowledge and Miasma](#).

See: mimesis, herding, [bystander effect](#), [conformity instincts](#), [coalitional instincts](#), and [The World Forager Elite](#).

Regardless of how politically unfeasible the policy proposal *would have been*, saying that it's unfeasible will tend to make it more unfeasible. Others will pick up on the social cue and be more cautious about promoting the idea; which causes others to imitate *them* and be more cautious, and will cause the idea to spread less. Which makes people [nervously look around for proof](#) that this is a [mainstream-enough idea](#) when they first hear about it.

This strikes me as one of the main ways that groups can end up stupider than their members, and civilizations can end up [neglecting](#) low-hanging fruit.

Brainstorming and trying things

Objection: "Political feasibility matters. There may be traps and self-fulfilling prophecies here; but if we forbid ourselves from talking about it altogether, we'll be ignoring a real and important part of the world, which doesn't sound like the right way to optimize."

My reply: I agree with this.

But:

- I think mainstream political discourse is focusing on this way too much, as of 2021.
- I think people should be *more cautious* about when and how they bring in political feasibility, especially in the early stages of evaluating and discussing ideas.

- I think the blog posts discussing "would this be a good policy if implemented?" should mostly be *separate* from the ones discussing "how politically hard would it be to get this implemented?".

Objection: "But if I mention feasibility in my initial blog post on UBI, it will show that I'm a reasonable and practical sort of person, not a crackpot who's fallen in love with pie-in-the-sky ideas."

My reply: I can't deny that this is a strategy that can be helpful.

But there are *other* ways to demonstrate reasonableness that have smaller costs. Even just saying "I'm not going to talk about political feasibility in this post" can send an adequate signal: "Yes, I recognize this is a constraint at all. I'm treating this topic as out-of-scope, not as unimportant." Though even saying that much, I worry, can cause readers' thoughts to drift too much away from the road.

Policies' popularity (frequently) matters. But often the correct response to seeing a good idea that looks plausibly-unfeasible is to go "oh, this is a good idea", help bring a huge wave of energy and enthusiasm to bear on the idea, and try your best to get it discussed and implemented, *to find out* how feasible it is.

We're currently in a world where most good ideas fail. But... failing is mostly fine? See [On Doing the Improbable](#) and [Anxious Underconfidence](#).

Newer, weirder, and/or more ambitious ideas in particular can often seem too "out there" to ever get traction. Then a few years later, *everyone's* talking about UBI. That feeling of anxious uncertainty is not actually a crystal ball; better to try things and see what happens.

Ungrounded social realities are fragile

From [Inadequate Equilibria](#):

[...] The still greater force locking bad political systems into place is an equilibrium of silence about policies that aren't "serious."

A journalist thinks that a candidate who talks about ending the War on Drugs isn't a "serious candidate." And the newspaper won't cover that candidate because the newspaper itself wants to look serious... or they think voters won't be interested because everyone knows that candidate can't win, or something? Maybe in a US-style system, only contrarians and other people who lack the social skill of getting along with the System are voting for Carol, so Carol is uncool the same way Velcro is uncool and so are all her policies and ideas? I'm not sure exactly what the journalists are thinking subjectively, since I'm not a journalist. But if an existing politician talks about a policy outside of what journalists think is appealing to voters, the journalists think the politician has committed a gaffe, and they write about this sports blunder by the politician, and the actual voters take their cues from that. So no politician talks about things that a journalist believes it would be a blunder for a politician to talk about. The space of what it isn't a "blunder" for a politician to talk about is conventionally termed the "Overton window."

[...] To name a recent example from the United States, [this] explains how, one year, gay marriage is this taboo topic, and then all of a sudden there's a huge upswing in everyone being *allowed* to talk about it for the first time and shortly afterwards it's a done deal.

[...] We can say, "An increasing number of people over time thought that gay marriage was pretty much okay. But while that group didn't have a majority, journalists modeled a

gay marriage endorsement as a ‘gaffe’ or ‘unelectable’, something they’d write about in the sports-coverage overtone of a blunder by the other team—”

[...] The support level went over a threshold where somebody tested the waters and got away with it, and journalists began to suspect it wasn’t a political blunder to support gay marriage, which let more politicians speak and get away with it, and then the *change of belief about what was inside the Overton window* snowballed.

And:

What broke the silence about artificial general intelligence (AGI) in 2014 wasn’t Stephen Hawking writing a careful, well-considered [essay](#) about how this was a real issue. The silence only broke when Elon Musk [tweeted](#) about Nick Bostrom’s *Superintelligence*, and then made an off-the-cuff remark about how AGI was “[summoning the demon](#).”

Why did that heave a rock through the Overton window, when Stephen Hawking couldn’t? Because Stephen Hawking *sounded like* he was trying hard to appear sober and serious, which signals that this is a subject you have to be careful not to gaffe about. And then Elon Musk was like, “*Whoa, look at that apocalypse over there!!*” After which there was the equivalent of journalists trying to pile on, shouting, “A gaffe! A gaffe! A... gaffe?” and finding out that, in light of recent news stories about AI and in light of Elon Musk’s good reputation, people weren’t backing them up on that gaffe thing.

Similarly, to heave a rock through the Overton window on the War on Drugs, what you need is not state propositions (although those do help) or articles in *The Economist*. What you need is for some “serious” politician to say, “This is dumb,” and for the journalists to pile on shouting, “A gaffe! A gaffe... a gaffe?” But it’s a grave personal risk for a politician to test whether the public atmosphere has changed enough, and even if it worked, they’d capture very little of the human benefit for themselves.

The public health response to COVID-19 has *with surprising consistency* been a loop of:

- Someone like Marc Lipsitch or Alex Tabarrok gives a good argument for X (e.g., “the expected value of having everyone wear masks is quite high”).
- Lots of people spring up to object, but the objections seem all-over-the-place and none of them seem to make sense.
- Time passes, and eventually a prestigious institution endorses X.
- Everyone who objected to X now starts confabulating arguments in *support* of X instead.

I think this is a pretty socially normal process. But usually it’s a *slow* process. Seeing the Overton window and “social reality” change with such blinding speed has helped me better grok this dynamic.

Additionally, something about this process seems to favor faux certainty and faux obliviousness over expected-value-style thinking.

[Zvi Mowshowitz](#) discusses a hypothetical where Biden does something politically risky, and suffers so much fallout that it cripples his ability to govern. Then Zvi adds:

Or at least, that’s The Fear.

The Fear does a lot of the work.

It’s not that such things would actually definitely happen. It’s that there’s some chance they might happen, and thus no one dares find out.

[...] W]hen you’re considering changing from policy X to policy Y, you’re being judged against the standard of policy X, so any change is blameworthy. How irresponsible to

propose First Doses First.

But...

A good test is to ask, when right things are done *on the margin*, what happens? When we move in the direction of good policies or correct statements, how does the media react? How does the public react?

This does eventually happen on almost every issue.

The answer is almost universally that the change is accepted. The few who track such things praise it, and everyone else decides to memory hole that we ever claimed or advocated anything different. We were always at war with Eastasia. Whatever the official policy is becomes the null action, so it becomes the Very Serious Person line, which also means that anyone challenging that is blameworthy for any losses and doesn't get credit for any improvements.

This is a pure '[they'll like us when we win](#).' Everyone's defending the current actions of the powerful in deference to power. Change what power is doing, and they'll change what they defend. We see it time and again. Social distancing. Xenophobia. Shutdowns. Masks. Better masks. Airborne transmission. Tests. Vaccines. Various prioritization schemes. First Doses First. Schools shutting down. Schools staying open. New strains. The list goes on.

There are two strategies. We can do what we're doing now, and change elite or popular opinion to change policy. Or we can change policy, and in so doing change opinion leaders and opinion.

This is not to say that attempts to shift (or route around) the Overton window will necessarily have a high success rate.

But it is to say:

- Some of the factors influencing this success rate have the character of a self-fulfilling prophecy. And such prophecies can be surprisingly fragile.
- The Overton window *distorts thinking* in ways that can *make it harder to estimate* success rates. The current political consensus just feels like 'what's normal', 'what's reasonable', 'what's acceptable', and the fact that this is in many ways a passing fad is hard to appreciate in one's viscera. (At least, it's hard for me; it might be easy for you.)

(I also happen to suspect that some very recent shifts in politics and culture have begun to make Overton windows easier to break and made "[PR culture](#)" less adaptive to the new climate. But this is a separate empirical claim that would probably need its own post-length treatment.)

Steering requires looking at the road

I claim that politics is too meta.

(I would also say that, e.g., effective altruist discourse spends too much time on meta. And a lot of other discourses besides.)

Going overboard on meta is bad because:

- It can create or perpetuate harmful social realities that aren't based in any external reality.

- It can create (or focus attention on) social realities that actively encourage false beliefs (including false beliefs that make it harder to break the spell).
- It's unnecessary. Often the idea that we need to spend a lot of time on meta is *itself* a self-perpetuating myth, and we can solve problems more effectively by just trying to solve the problem.
- It makes it harder to innovate.
- It makes it harder to experiment and make high-EV bets.

It's also bad because it's *distracting*.

Distraction is a lot more awful than it sounds. If you go from spending 0% of your time worrying about the hippopotamus in your bathtub to spending 80% of your time worrying about it, your other life-priorities will probably suffer quite a bit even if the hippo doesn't end up doing any *direct* damage.

And it would be quite bad if 15% of the *cognition in the world* were diverted to hippos.

'All the News That's Fit to Print'

The New York Times

Late Edition
Today, clouds and some sunshine, brief, isolated, high 63. Thursday, mostly cloudy, low 56. Thursday, cloudy, afternoon showers, high 66. Weather map appears on Page D6.

VOL. CLXVI... No. 57,400 + \$2.50

NEW YORK, SATURDAY, OCTOBER 29, 2016

Hillary Clinton with a top aide, Huma Abedin, the estranged wife of Anthony D. Weiner, aboard the campaign plane on Friday.

NEW EMAILS JOLT CLINTON CAMPAIGN IN RACE'S LAST DAYS

F.B.I. Looks at Messages Found During Inquiry Into Weiner's Texts

By ADAM GOLDMAN and ALAN RAPPAPORT

WASHINGTON — The presidential campaign was rocked on Friday after federal law enforcement officials said that emails pertained to the closed investigation into Huma Abedin, the estranged wife of small server were discovered on a computer belonging to Anthony D. Weiner, the estranged husband of a top Clinton aide.

In a letter to Congress, the F.B.I. director, James B. Comey, said the emails had surfaced in an unrelated case, which law enforcement officials said was an F.B.I. investigation into black text messages from Mr. Weiner to a 14-year-old girl in New Mexico. Mr. Weiner, a former Democratic congressman from New York, is married to Huma Abedin, the top aide.

Mr. Comey's letter said that the F.B.I. would review the emails to determine if they improperly contained classified information, which is tightly controlled by the government. Senior law enforcement officials said that it was unclear if any of the emails were from Mrs. Clinton's private server. And while Mr. Comey said in his letter that the emails "appear to have been collected before the election," he did not say whether they appeared to have been collected before the election.

Mr. Comey could immediately inform Congress about the emails, or he could wait until after the election, despite his pledge of "transparency" in the investigation.

Mr. Comey, a Republican ap-

lant, has been under fire for his handling of the investigation into Mr. Weiner's laptop, which the F.B.I. had obtained as part of its investigation into Mr. Weiner. About a month ago, a per-

With 11 Days to Go, Trump Says Revelation 'Changes Everything'

By AMY COGDECK and PATRICK HEALY

Everything was looking up for Hillary Clinton. She was riding high in the polls, even seeing an improvement on thewithstanding. She was sitting on \$15 million in cash. At 12:37 p.m. Friday, her aides announced that she planned to campaign in Arizona, a state that a Democratic presidential candidate has carried only once since 1948.

Twenty minutes later, October delivered its latest big surprise:

This article is by Eric Lichtenbaum, Michael S. Schmidt and Matt Apuzzo.

WASHINGTON — James B. Comey, the F.B.I. director, handed a dilemma on Thursday when department officials bristled over the discovery of a new trove of emails that might be connected to the donor controversy and Hillary Clinton's private email server.

Mr. Comey could immediately inform Congress about the emails,

would risk accusations that he was unfairly harming her presidential campaign less than two weeks before the election.

Or he could wait until after the election, despite his pledge of "transparency" in the investigation.

Mr. Comey, a Republican ap-

lant, has been under fire for his handling of the investigation into Mr. Weiner's laptop, which the F.B.I. had obtained as part of its investigation into Mr. Weiner. About a month ago, a per-

Decision Pulls F.B.I.'s Leader Back Down Into Political Fray

This article is by Eric Lichtenbaum, Michael S. Schmidt and Matt Apuzzo.

WASHINGTON — James B. Comey, the F.B.I. director, faced a dilemma on Thursday when department officials bristled over the discovery of a new trove of emails that might be connected to the donor controversy and Hillary Clinton's private email server.

Mr. Comey could immediately inform Congress about the emails,

would risk accusations that he was unfairly harming her presidential campaign less than two weeks before the election.

Or he could wait until after the election, despite his pledge of "transparency" in the investigation.

Mr. Comey, a Republican ap-

lant, has been under fire for his handling of the investigation into Mr. Weiner's laptop, which the F.B.I. had obtained as part of its investigation into Mr. Weiner. About a month ago, a per-

Editorial on A18

Donald J. Trump at a rally in Manchester, N.H., on Friday.

Note the headlines' framings. The *New York Times* is creating a social reality by picking which stories to put on its front page, but all the reporting is also *about* social realities: "How

will this information (that we are here reporting on) change what's controversial, what's popular, what's believed, etc., and thereby affect the election?"

Presidential elections are a time when it's especially *tempting* to go meta, since we're all so curious about *us*, about how we'll vote. But they're also a time when it's an especially *terrible idea* to go meta, because *our models and behavior are unusually important* at this time and *this is the exact time when we most need to be thinking about the object-level tradeoffs* in order to make a higher-quality decision.

Imagine trying to steer a ship by polling the crew members, but then covering over all the ship's windows and doors with real-time polling data.

I don't think the reporters and editors here *feel* like they're doing the "loudly and publicly warn someone that they're being cringe" thing, or the "black out the ship's windows with polling data" thing. But that's what's happening, somehow.

If the emails are important in their own right, then great! The headlines and articles can be all about their object-level importance. Help voters make a maximally informed decision about the details of this person's security practices and how these translate into likely future behavior.

Headlines can focus on object-level details and expert attempts to model individuals' rule-following behavior. You can even do head-to-head comparisons of *object-level differences* about the people running for president!

But the front-page articles really shouldn't be about the *controversy*, the *buzz*, the second-order perceptions and spin and perceptions-of-spin. The buzz is *built* out of newspaper articles, and you want the resultant building to stand on some foundation; you don't want it to be a free-floating thing built on itself.

The articles shouldn't drool over the tantalizingly influenceable/predictable beliefs and attitudes of *the people reading the articles*. Internal decisions about the topic's importance (and the angle and focus of the coverage) shouldn't rest on *hypothetical perceptions* of the news coverage. "It's important because it will affect voters' beliefs because we're reporting that it's important because it will affect voters' b..."

This is an extreme example, but I'm mostly worried about the mild examples, because small daily hippos are worse than large rare hippos.



i

NBCNEWS.COM

Congress passes \$1.9T Covid relief bill, with \$1,400 checks, in major win for Biden

"A major win for Biden" here isn't just trying to give credit where credit's due; it's drawing attention to one of the *big interesting things* about this \$1.9 trillion bill, which is its effect on Biden's future *popularity* (and thereby, maybe just maybe, the future balance of political power).

This is certainly *one of the interesting things about bills*, and I could imagine a college class that went through bill after bill and assessed it mainly through the frame of "who's winning, the Republicans or the Democrats?", which might teach some interesting things.

But "who's winning?" isn't the *only* topic you could teach in a college course by walking through a large number of historical bills.

So why has every major news outlet in the US settled on "who's winning" as the one true frame for public policy coverage? Is this what we'd pick if we were making a conscious effort to install good norms, or is it just a collective bad habit we've fallen into?

If you don't look at the road while you're driving, you get worse decision-making, and the polarization ratchet continues, and we all get dumber and [more sports-fan-ish](#).

My political views have changed in a big way a few times over the years. In each case, the main things that caused me to update weren't people yelling their high-level principles at me more insistently. I was persuaded by being hit over the head repeatedly with *specific object-level examples* showing I was wrong about which factual claims tend to be true and which tend to be false.

If you want to produce good outcomes from nations, try arguing more about *actual disagreements* people have. Policy isn't built out of vibes or branding any more than a car engine is.

Core Pathways of Aging

Most overviews of aging suffer from multiple problems:

- They dump a bunch of findings with no high-level picture.
- Many of the claims they make are outdated, purely theoretical, and sometimes even outright disproven by existing work.
- They are usually written by working academics, who are shy about telling us when their peers' work is completely wrong.
- They are shy about making strong claims, since this would also implicitly mean denying some claims of the authors' peers.

This post is a high-level brain-dump of my current best models of the core pathways of aging, as I currently understand them. I have no particular reason to avoid calling out claims I think are wrong/irrelevant, and I'm going to present high-level models without pages and pages of disclaimers and discussions about results which maybe disagree with them (but are probably just wrong/irrelevant).

Epistemic status: I would be surprised if none of it turned out to be wrong, but there are multiple lines of evidence supporting most claims. It is not highly polished, and references are included only when I have them readily on hand. My ideal version of this piece would have more detailed references, more double-checking behind the claims, and more direct presentation of the data which backs up each claim. Unfortunately, that would take enough time and effort that I'm unlikely to actually get to it soon. So... here's what I could produce in a reasonable amount of time. Hopefully it will be wrong/unhelpful in ways orthogonal to how most overviews are wrong/unhelpful.

Foundations

First, let's recap a couple foundational principles. I'll go through these pretty quickly; see the linked posts for more info.

[Homeostasis and "Root Causes" in Aging](#): the vast majority of proteins, cells, etc, in the human body turn over on a timescale from days to months. At any given time, their level (e.g. protein concentration, cell count, etc) is in equilibrium on the turnover timescale - i.e. the rate of creation approximately equals the rate of removal. For any X with turnover much faster than aging (i.e. decades), if we see the level of X increase/decrease on the timescale of a human lifetime, then that is *not* due to permanent "accumulation of X" or "depletion of X"; it is due to increase/decrease in the *rate* of creation/removal of X. For instance:

- DNA damage is typically repaired on a timescale of [hours or faster](#), depending on the type. If DNA damage levels increase with age, that is due to an increase in rate of damage or decrease in rate of repair, *not* permanent accumulation.
- Typical senescent cells turn over on a timescale of [days to weeks](#). If the number of senescent cells increases with age, that is due to an increase in rate of senescent cell production or decrease in rate of removal, *not* permanent accumulation.
- Elastin is believed to [not turn over at all](#) in humans. So if we see elastin deposits increasing with age (e.g. in [wrinkles](#)), then that *could* be permanent accumulation.

Furthermore: suppose we have a positive feedback cycle. Increasing A decreases the rate of production of B, so B decreases. But decreasing B decreases the rate of removal of A, so A increases. If both A and B individually turn over on a timescale of hours or faster then this feedback loop as a whole will also typically operate on a timescale of hours or faster - i.e. count/concentration of A will explode upward on roughly that timescale. More generally, a

feedback loop will usually operate on the timescale of its slowest component, exactly like the rate-limiting step of a chemical reaction.

Main upshot of all this: since aging involves changes on a timescale of decades, there must be some component which is out-of-equilibrium on a timescale of decades or longer (i.e. does not turn over significantly across a full human lifespan). These are the components which we'll call "root causes". Everything else which changes with age, changes only *in response to* the root causes. Reset the root causes to young-organism levels, and everything else will equilibrate to young-organism levels in response. Furthermore, a reset of the root causes only needs to happen once every few decades for humans - it fully resets the human to a youthful state, so ongoing treatment is not needed on short timescales.

[The Lens, Progerias and Polycausality](#): The lens of the human eye consists of fiber deposits which do not turn over significantly over the course of a human lifetime. New fiber layers are added over time, so the lens grows from around 3.5mm in infancy to 5.5mm in old age. The main clinical result is the near-universal need for reading glasses in old age.

This is a well-understood root cause of one symptom of old age. Furthermore, it is very likely independent of most other age related diseases - lens thickening is unlikely to cause cancer or heart disease, for instance. So, it's an existence proof: *there is more than one root cause of aging*.

That said, there's a fair bit of evidence that *most* symptoms of aging - including the major age-related diseases - share a common root cause, or at least a common core pathway. Some kinds of evidence of this:

- Most symptoms/diseases of aging are correlated - someone who has one early is likely to have others. Conceptually, if you do a factor analysis on aging symptoms, there's one big factor for a bunch of diseases, even after controlling for the number of years one has lived. ("Aging clock" is a relevant piece of jargon here.)
- At the cellular level, a lot of diseases of aging "look similar", and involve similar pieces. There's a decrease in cell count, increase in damaged proteins/DNA/fats, and inflammation. We see roughly this pattern in Alzheimers, atherosclerosis, muscle loss, and many others.
- Certain simple interventions reliably produce many diseases of aging - for instance, progerias are single mutations which produce a whole "early aging" phenotype
- Conversely, certain simple interventions reliably delay many diseases of aging - e.g. calorie restricted diets.

... so these all point to shared underlying causes. This post will be about the "core pathways" most likely involved.

Major Diseases

These subsections will talk about various specific age-related diseases, mainly highlighting how they connect to the cellular processes we'll talk about later. Two main themes to watch: reactive oxygen species and senescent cells.

Reactive oxygen species (ROS) aka free radicals are produced in greater numbers in old age. These are short-lived, highly reactive molecules. They react with all sorts of things, oxidatively "damaging" whatever they hit, including proteins, fats, and DNA.

Senescent cells are cells which have partially shut down in a programmatic way, triggered by some sort of "stress" on the cell (e.g. DNA damage, exposure to harsh chemicals or radiation, etc). They pump out inflammatory signals (called the senescence-associated secretory phenotype, or SASP). Eventually, they're removed by the immune system.

In later sections, we'll see that these two are tightly coupled. For now, we'll talk about how they seem to underlie a variety of age-related diseases.

Atherosclerosis

If you dissect young and old mammals, one of the most obvious internal differences is in the blood vessels:

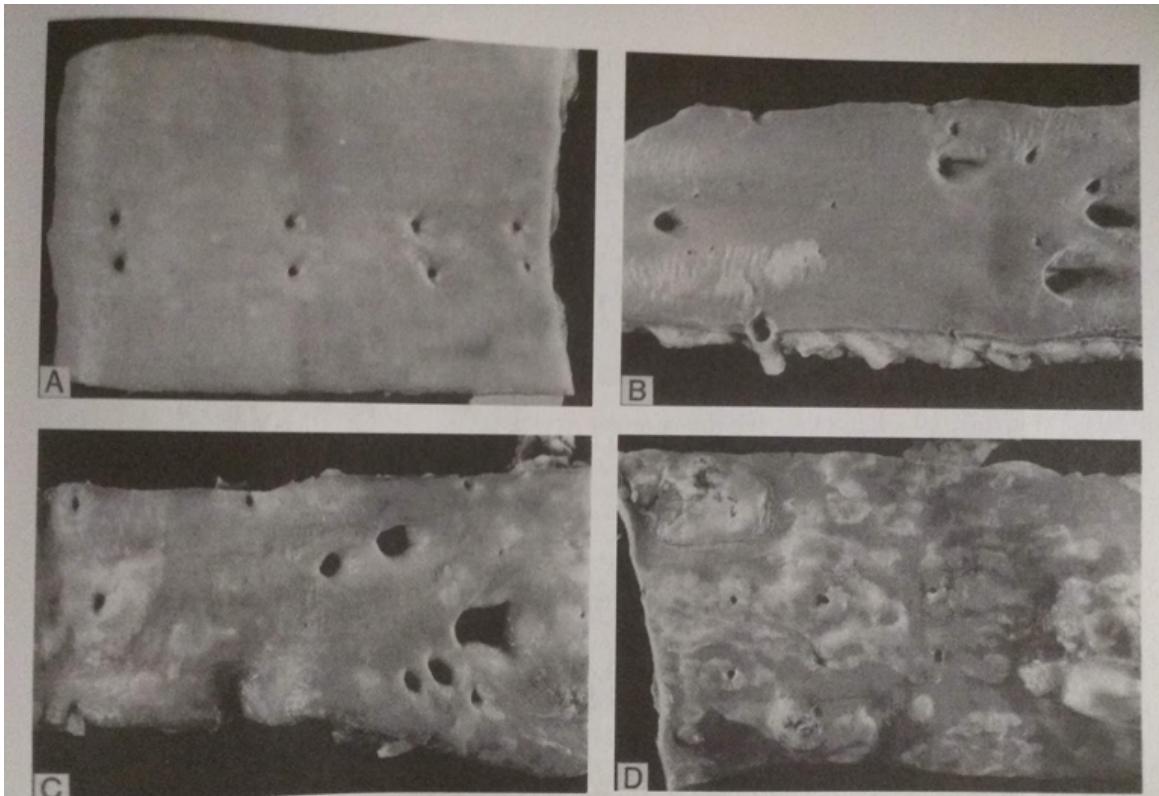


FIGURE 16.4 Progression of morphologic changes in human aorta from early to advanced atherosclerosis. Aortas were split open and the intima exposed for photography. (a) Aorta (thoracic) of a 32-year-old male showing early fatty plaques (represented by lighter Aortas from a 32 year old and a 24 year old on top, a 55 year old and a 65 year old on bottom. ([source](#))

Mammals of any age have “fatty streaks” along the walls of the vasculature, which are exactly what they sound like. (These are the slightly lighter patches in the pictures above - more obvious in the older aortas, but faintly visible in the young pair as well.) In older mammals, the fatty streaks tend to be larger, until in old age they necrotize (aka die) in the middle and turn into thick “atherosclerotic plaques” (the dark patches in the lower right picture above). These can block blood circulation, and sometimes a chunk of the plaque can break off and block circulation in smaller vessels; either of these can cause e.g. heart attack or stroke.

At any age, a lower-fat diet is associated with smaller fatty streaks and lower chance of atherosclerosis, though the streaks universally grow with age holding diet constant.

The big breakthrough in atherosclerosis came in the 80's/90's. It was known that a certain type of cell (called a macrophage) would hoover up fat from the bloodstream, then adhere to the cell wall when full, forming the fatty streaks. The missing puzzle piece was that the macrophages don't just hoover up any random fats (there's rather a lot of fat in the bloodstream and not *that* many macrophages; it would be like sweeping sand off the beach).

They specifically hoover up partially oxidized fats. Steinberg has a [great review](#) of various experiments feeding into this discovery.

The upshot: streaks and plaques grow in old age primarily because the concentration of (partially) oxidized fats in the bloodstream increases in old age. These probably don't have a very slow turnover time, so either the rate of (partial) oxidation of fats increases in old age, or the rate of removal decreases. Which is it, and what causes it?

The main candidate here is ROS: increasing ROS levels in old age increase the rate of oxidative damage to fats. (One interesting question: how does the relative increase in oxidative damage to fats compare to DNA/proteins? Do the numbers actually line up for these to share a source? I haven't seen a solid Fermi estimate on this, I'd be interested to see one.)

Vascular Stiffening

Even aside from atherosclerosis, the walls of blood vessels change in old age. In particular they become stiffer.

In normal operation, the heart pumps out blood in discrete chunks - each heartbeat pumps out some amount of blood into the arteries. The arteries expand a bit to accumulate this blood, much like a balloon. And, like an inflated balloon, the arteries are at slightly higher pressure from this extra blood, until it's pushed out through the capillaries and the cycle starts again with the next heartbeat. It's sort of like a capacitor: the heart sends out blood at high pressure in sudden waves, and the arteries expand and contract to store the extra blood and smooth out the pressure variance.

If the walls become stiffer, then the arteries are less able to smooth out this pressure variance. From the heart's perspective, it's trying to pump blood into a hard container, which works about as poorly as you'd expect; this is one of the major causes of heart failure (alongside atherosclerosis). The vessels also become more likely to burst (i.e. aneurysm).

What causes the loss of elasticity?

The main answer seems to be oxidative damage to proteins in the wall of the blood vessels. The key experimental evidence backing this up is the effect of aminoguanidine (AG), a powerful scavenger of certain ROS: AG administration [reverses](#) age-related stiffening of the arteries. This is only temporary - i.e. the arteries go back to their stiff state shortly after AG administration stops - so it isn't a root cause, but it is a link in the causal chain. Once again, oxidative damage likely caused by increased ROS seems to be the key causal intermediate.

(Please don't go running out to take aminoguanidine. The side effects are serious, and the benefits are short-term - you'll go right back to where you would have been as soon as you stop taking it.)

An interesting side note: many sources claim that a shift in collagen:elastin ratio is involved in stiffening of the vasculature. This seems to be based on a purely theoretical paper from decades ago, and the actual data [doesn't back it up](#), yet review articles and textbooks continue to mention it.

Alzheimer's

The first and most important thing to know about Alzheimer's (aka dementia, aka old folks losing their memory) is that it is *not* caused by accumulation of amyloid beta.

Decades ago, people noticed that if you look at the brains of old people with dementia, they usually have lots of plaques, and these plaques are made of a particular protein fragment called amyloid beta. Therefore *clearly* amyloid beta causes dementia. Pretty soon people

were using amyloid beta plaques to *diagnose* dementia, which made it really easy to show that the plaques cause dementia: when the plaques are how we diagnose “dementia”, then by golly removing the plaques makes the “dementia” (as diagnosed by plaques) go away.

As far as I can tell, there has never at any point in time been compelling evidence that amyloid beta plaques *cause* age-related memory problems. Conversely, I have seen at least a few studies suggesting the plaques are not causal.

Meanwhile, [according to wikipedia](#), 244 Alzheimer’s drugs were tested in clinical trials from 2002-2012, mostly targeting the amyloid plaques. Of those, only 1 drug made it through. Bottom line: they don’t work.

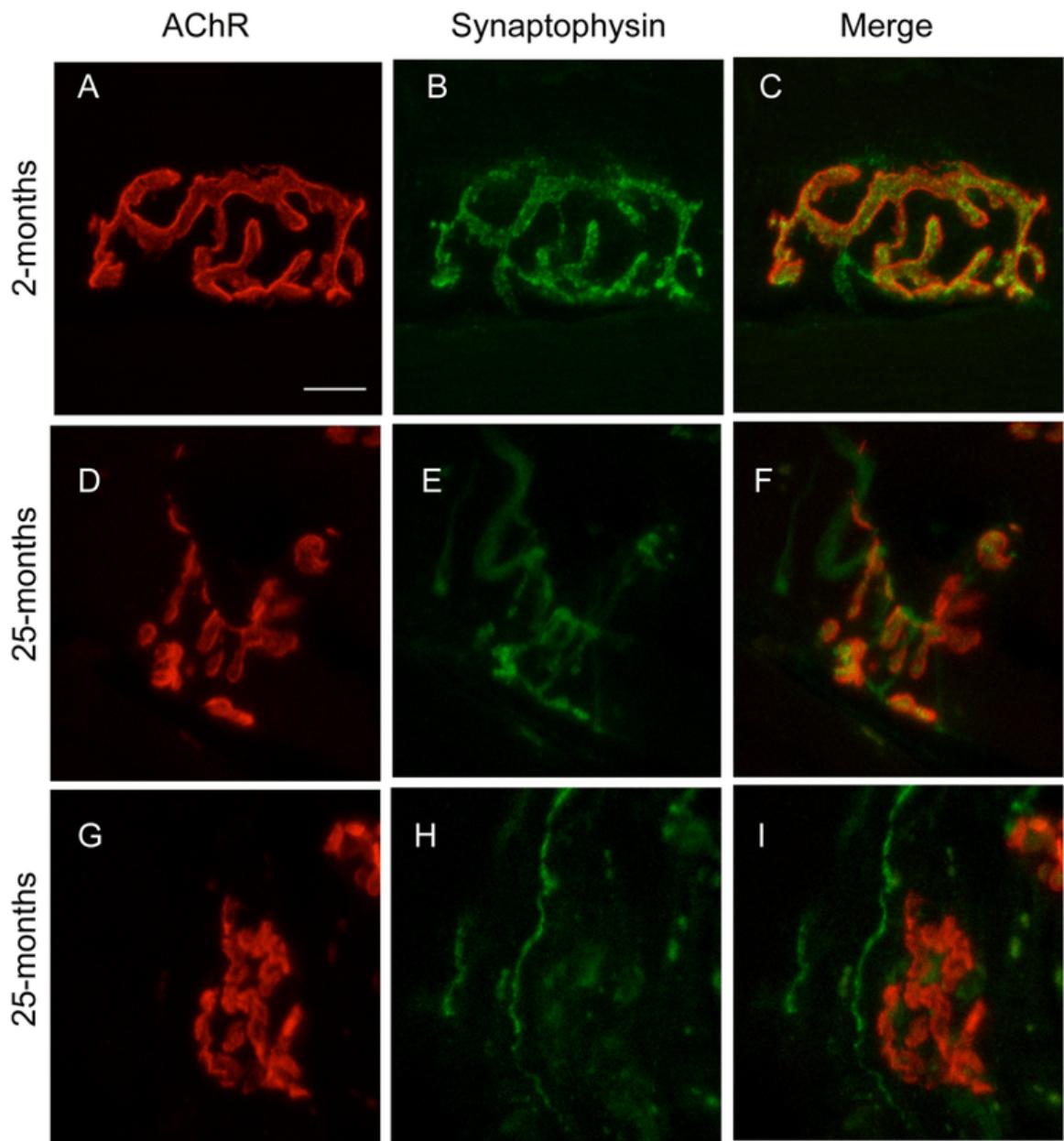
So what does cause Alzheimer’s? I don’t know; it’s not a disease I’ve studied in depth. I know plenty of studies find the usual culprits involved - inflammation, damaged proteins, etc.

I do know of [one particularly interesting cluster of studies](#) which found that the brain “opens up” during sleep, increasing flow of cerebral fluid to clear out whatever junk accumulated during the day - including amyloid beta. The flow [takes place in](#) “paravascular” spaces, i.e. spaces around the blood vessels, which widen during sleep. Given the age-related changes in blood vessels (e.g. thickening & stiffening of walls), it would make sense for this paravascular space to not open up as much in older people, and indeed [that seems to be the case](#). Whether this has anything to do with dementia, I don’t know, but it is my current best guess for the main cause of the amyloid plaques. If true, this would mean that the plaques share their main root causes with changes in the vasculature.

Sarcopenia

Sarcopenia is age-related loss of muscle mass - i.e. old people becoming physically weak. Overviews often say it can be (at least partially) “reversed” via exercise, which seems to be fairly obvious bullshit aimed at making people exercise more. Bottom line: for any fixed amount of exercise, muscular strength will fall with age. (Obviously more exercise can still increase muscle strength at any age.)

Many sources will claim that lots of research has shown sarcopenia to be caused by loss of muscle innervation - i.e. problems with the neuromuscular junctions (NMJs). As with amyloids and Alzheimers, as far as I can tell there was never at any point in time any compelling evidence that NMJ changes are actually causal for muscle loss, and my current best guess is that they are not causal. However, research on age-related changes in NMJ structure did produce very shiny images, which I think is probably the main reason they received so much attention in the first place.



Neuromuscular junctions from young (2-month) and old (25-month) mice. ([source](#))

So what does cause sarcopenia? I don't yet have a full picture here, but I can sketch out my current best guess.

Muscle cells are not normal cells; they're mega-cells, a hundred times longer than a normal cell, with hundreds of nuclei all within the same cell. They need so many nuclei because molecules which would diffuse quickly from one end to the other of a normal cell would take far too long to diffuse across a muscle cell (diffusion time increases faster-than-linearly with distance); so, things need to be produced locally.

This setup also makes it difficult for an entire muscle cell to turn over. Instead, the nuclei turn over (along with all the usual protein/membrane/etc turnover mechanisms); they're regularly replaced from "satellite cells", a type of stem cell nestled in next to the muscle whose job is basically to crank out new nuclei from time to time.

In sarcopenia, one cross-section of the long muscle cell will fail first - a “ragged red” section - and then failure gradually spreads along the length. The failing section involves the usual culprits: ROS, mitochondrial deficiency (related to ROS; more on that later), inflammation signals, etc.

My best guess is that this is basically just how cellular senescence manifests in a muscle cell. One particular satellite cell is “close to senescence”, and produces nuclei which rapidly “senesce”. But since muscle cells are really a bunch of relatively-isolated components along their length (due to long diffusion time), this only results in one section of the cell failing. Eventually, the “senescence” spreads to adjacent sections via the usual mechanisms of senescence-induced-senescence (more on that later). Key characteristics of “ragged red” sections of the muscle cell - like high ROS, mitochondrial deficiency, or inflammation - are the usual characteristics of senescent cells, and sarcopenia probably shares its root cause(s) with cellular senescence more generally.

Other Diseases

This section briefly mentions a few other age-diseases which I’ve read a little bit about, but haven’t studied in as much depth. I’ll just give a few comments on how they tie in to the cellular processes discussed later.

Arthritis: arthritis is basically inflammation in the joints. I’ve heard plenty of claims that it’s caused by increasing numbers of senescent cells. This would make sense; senescent cell counts are firmly established to increase with age, and senescent cells secrete inflammatory factors, so put 2 and 2 together. On the other hand, I also heard a high-profile clinical trial based on this hypothesis recently failed; I don’t consider that especially strong evidence, since that could easily be due to something specific to the trial. I don’t know enough to have a strong belief on whether senescent cells are the main factor here, but it’s my most likely current model.

Osteoporosis: calcium regulation goes completely bonkers with age. I’ve looked into this only briefly, and my main conclusion is that it’s a confusing mess and I have no idea what’s going on. The most promising direction I’ve stumbled on was in a physiology class, unrelated to aging, when the professor mentioned that osteoblasts (bone-making cells and a major calcium regulator) are derived from immune cells and respond to inflammatory signals. Given that aging in general tends to involve low-grade chronic inflammation (most likely due to increasing numbers of senescent cells), “calcium regulation goes completely bonkers” seems like the sort of thing which would result.

Cancer: the key requirement for cancer is cells with oncogenic mutations. As usual, increasing cancer rates could be caused in two main ways: increasing production rate of mutations, or decreasing removal rate of mutant cells. There are plausible age-related mechanisms for both of these. On the production side, we’ll later discuss how genomic instability relates to ROS and senescence, and in particular the role of transposons. On the removal side, the immune system weakens with age, likely for the same reasons as everything else weakens with age. Also, in this case it’s plausible that an increase in the production rate of precancerous cells and/or senescent cells would slow the removal rate as well, simply because the immune system has limited capacity.

Cataracts: as with hardening of the vasculature, cataracts can be reversed by aminoguanidine, so the same considerations apply.

Core Intermediates

Now we get to the meaty part. At the microscopic level, there’s a handful of pieces which pop up again and again in age-related diseases:

- Oxidative damage to DNA, proteins, fats, etc.
- ROS
- Senescent cells
- Inflammation
- Mitochondrial dysfunction

Some of these have obvious connections. For instance, more ROS presumably lead to more oxidative damage to DNA/proteins/fats/etc. Some key questions here:

- Where are the ROS produced? Mitochondria are the top candidate - there's a known mechanism for ROS production by mitochondria, as well as experimental evidence that mitochondrion-targeted antioxidants specifically reduce ROS-induced damage.
- How do the ROS and/or damaged molecules move between compartments, e.g. nucleus/cytoplasm/extracellular? I have seen very little on this, and consider it a major blindspot. I'm not sure if it's a blindspot for the field or if I just haven't found the right cluster of papers.
- Are the quantitative changes in DNA/protein/fat damage compatible with a single underlying cause? Do they match plausible estimates of ROS from dysfunctional mitochondria? Again, I haven't seen Fermi estimates here, but I'd like to.

We do have evidence (from aminoguanidine and similar drugs; good overview [here](#) for protein damage) that ROS are causal for various types of damage.

Another connection: we've already mentioned that senescent cells release inflammatory factors, the so-called "senescence-associated secretory phenotype" (SASP). The one question here for which I haven't seen a clear answer is whether increasing numbers of senescent cells quantitatively match age-related increases in inflammation. Drugs to remove senescent cells are a hot area right now, and should provide more evidence on whether senescent cells are causal for age-related inflammation.

So we have two clusters of probably-causally-connected processes. One of these involves dysfunctional mitochondria producing excess ROS which damages DNA/proteins/fats, and the other involves senescent cells inducing inflammation.

The really big discovery of the past twenty years was the connection between cellular senescence and ROS/mitochondrial dysfunction. Turns out, ROS, DNA damage, and mitochondrial dysfunction are all tied together in one bistable feedback loop, and cellular senescence is basically a state-change in that feedback loop.

The DNA Damage <-> Mitochondrial ROS Feedback Loop Underlying Senescence

The two key papers here are "[Feedback between p21 and reactive oxygen production is necessary for cell senescence](#)" and "[Mitochondrial Dysfunction Accounts for the Stochastic Heterogeneity in Telomere-Dependent Senescence](#)". [Another group](#) found basically the same phenomenon, but they found some weird differences compared to normal senescence which I think were probably artifacts of how they blocked mitochondrial function, so they're less directly relevant to cellular senescence in the wild.

Here's how the feedback loop works:

- A cell's DNA is damaged, inducing a damage response. (p21 is the main signalling molecule which activates the damage-response genes.)
- As part of this damage response, mitochondria are shifted into a lower-efficiency state, producing less energy and more ROS.
- The ROS then further damage DNA.

At low levels of activation, the ROS do not produce enough DNA damage to make this loop self-sustaining. The cell stays in a low-ROS, low-damage state - the “normal” state. But once it passes some threshold, the positive feedback takes off, and the cell transitions into a high-ROS, high-damage state - the “senescent” state.

Notably, after a few days, the cell changes in some other (not yet fully understood) manner, locking in the senescent state. After this point, even if ROS are suppressed, the cell will remain in senescence rather than switching back to the normal-cell state. Transposons are one plausible candidate for this lock-in; more on that shortly.

The experimental evidence for this process is beautiful (I definitely recommend those papers, especially the second). And it neatly unifies all of the pieces we listed above: cellular senescence is a positive feedback loop between ROS, damage and mitochondrial dysfunction, and the SASP connects it all to inflammation.

There's still one big question: why is this positive feedback loop active more often, for more cells, in old age? The whole feedback loop is fast, senescent cells are removed on a timescale of days to weeks, so there must be some upstream change either increasing the rate of triggering senescence (presumably by somehow damaging DNA, possibly via ROS, or inducing mitochondrial dysfunction) or decreasing the rate of removal of senescent cells. In fact, [both of these do occur](#), although personally I think the increase in rate of triggering senescence is much more likely to be causal.

Further Up The Causal Chain

There are lots of stories about how various plausible root causes could, perhaps, trigger the senescence feedback loop. I think transposons are the most likely candidate, though mitochondrial mutations are a plausible mechanism as well. We'll talk about both of those in the next two subsections. Most other proposed root causes can, I think, be ruled out at this point - we'll talk about some of those in the “Other Root Causes” subsection below.

Transposons

A transposon is a gene whose main function is to copy itself. The LINE-1 family of transposons (most common active transposons in humans) consist of a protein which snips the DNA, and another protein which reverse-transcribes the transposon's mRNA into DNA attached to the snipped end. (I'm glossing over some details here.)

These things are extremely common in the genome - [more than half of human DNA consists of dead transposons](#). (“Dead” here means that the transposon mutated at some point in evolutionary history, so it's no longer functional.) Fortunately, the number of non-dead transposons in the human genome is much smaller - even the highest estimates I've seen put the number typically below 100.

We do have mechanisms to repress transposon activity, most notably epigenetic mechanisms. Most DNA is usually tightly coiled up around little cylindrical proteins (called histones), where it can't be easily transcribed. “Epigenetics” typically refers to modifications of the DNA and/or histones which make the coils tighter or looser, making the DNA difficult or easy to access. Most transposons are epigenetically tagged so that they're kept tightly coiled most of the time. Indeed, an argument could be made that this is the *primary* role of epigenetics - it's certainly what most epigenetic modifications are doing most of the time, since transposons fill so much of the genome.

There's an obvious story by which transposons could be a root cause of aging. Most of the time they're repressed, but every once in a while, one of them manages to copy itself. Once it's copied, there's no undoing it - the transposon count will only go up over time. Eventually, there's enough active transposons that the repression mechanisms aren't so reliable. At that

point, the transposon protein which snips the DNA will be expressed quite a bit, resulting in lots of DNA damage - those snips are exactly the sort of damage which can trigger senescence. (Note: it's a lot easier to snip the DNA than to reverse-transcribe, so I generally expect there to be a lot more snips than successful transposon-copy events.)

Setting aside this root-cause story for a moment, there's also evidence that [cellular senescence causes derepression of transposons](#). We'll talk more about the details of that later, but the key idea here is that there's a trade-off at the cellular level between repairing damage and repressing the transposons. When DNA damage is high, the cell temporarily shifts resources to repair that damage, deregulating the transposons. In senescence, the level of DNA damage is constantly high, so transposons go wild.

Of course, transposons themselves cause DNA damage (i.e. the snips), so eventually this can lock in senescence. That's my current best understanding of why senescence gets "locked in" after a few days: with transposons driving the DNA damage, the cell will remain senescent even if ROS are suppressed. It's a second senescence feedback loop, slower to start up, but permanent - once the transposons are active enough, the cell cannot leave senescence.

Unfortunately, this makes it difficult to distinguish the chicken from the egg. We know that senescence can cause transposon activity, but we also suspect that transposon activity comes first and causes senescence. We can't test that hypothesis just by looking at transposon activity in senescent cells. In principle, we could test it by looking for an age-related increase in transposon count in *non-senescent* cells, but that turns out to be actually pretty-difficult in practice. (Modern DNA sequencing involves breaking the DNA into little pieces, sequencing those, then computationally reconstructing which pieces overlap with each other. That's a lot more difficult when the pieces you're interested in have millions of near-copies filling most of the genome. Also, the copy-events we're interested in will vary from cell to cell.)

One more thing to note about this model: suppose that a stem cell has a high transposon count, but not high enough to undergo senescence itself. That stem cell will pump out new cells, as stem cells typically do, and those cells will themselves be close to senescence. We should therefore see little clusters of senescent cells, each derived from one stem cell. This is where I expect most senescent cells come from in aged tissue. The "ragged red" sections of aging muscle cells are a good example - one satellite cell is near senescence, so it pumps out nuclei which rapidly senesce, and the section of the muscle cell near that satellite ends up senescing.

This picture also offers an obvious story for cancer - transposons are a major driver of mutations.

Mitochondrial Mutations

Mitochondrial mutations are the center of the "mitochondrial free radical theory of aging" (MIFRA); [Aubrey de Gray's MIFRA book](#) provides an excellent (though out-of-date) summary of the model. Some parts of the model have been pretty well nailed down by evidence at this point - e.g. the idea that most core diseases of aging are mainly driven by ROS, which in turn are produced by mitochondria. The positive feedback loop underlying senescence even further connects symptoms of aging to mitochondrial ROS.

The main piece of the MIFRA model which still seems up-in-the-air is the idea that the root cause is mitochondrial mutations.

Background: mitochondria have their own separate DNA. There's only a handful of genes on it; most necessary mRNA/protein sequence is supplied by the nuclear DNA. But the mitochondria's little DNA is particularly prone to mutation - it doesn't have the nucleus keeping it safely separated from the highly-energetic processes of the rest of the cell. The

flip side is that mitochondria turn over frequently, separate from turnover of the cell itself, and they have quality-control mechanisms - e.g. if a mitochondrion isn't producing energy (as indicated by low transmembrane potential) then it's broken down.

A key idea hypothesis for the mitochondrial mutation model is that mutant mitochondria which are only *partially* defective aren't broken down. In fact, under this hypothesis, such mitochondria have a replicative advantage, and can expand to take over the whole cell. This cell then becomes a "hotspot", pumping out lots of ROS.

Today, we would expect that such hotspots are senescent cells - this degree of mitochondrial dysfunction should certainly be enough to induce senescence.

There is some evidence for this - for instance, mitochondrial mutants tend to take over whole cells, rather than being spread evenly across cells. However, there just aren't very many of them - senescent cells outnumber mutant-mitochondria-dominant cells by a wide margin in old age (order of magnitude: think 10% vs 1%). Also, as we mentioned earlier, senescent cells don't reproduce and do turn over, so even if mutant mitochondria took over one cell, they'd need to somehow expand to others.

This still doesn't rule out the model entirely. Some possibilities:

- Maybe a small fraction of senescent cells are long-lived, and these produce enough ROS to induce senescence in a much larger population of cells.
- Maybe mutant mitochondria spread before the cell is cleared (there is evidence for exchange of mitochondria between neighboring cells).
- Maybe stem cells with some mutant mitochondria produce new cells which go on to rapidly senesce, much like we hypothesized for transposons.

On the other hand, it's also plausible (I think more plausible) that defective mitochondria are negatively selected in healthy cells, and they only expand to take over in already-senescent cells, where all the mitochondria are already pumping out less energy and more ROS anyway.

Other Root Causes?

Telomeres

DNA has repetitive regions called telomeres at the ends . Each time a cell divides, the copying starts a little ways in from the end, so the telomeres get a bit shorter. Eventually, the telomere runs out, and the bare DNA end is interpreted as damage, inducing senescence. This has long been known to cause cellular senescence *in vitro*, and was hypothesized as a root cause of aging.

Indeed, telomeres are known to shorten with age. On the other hand, upregulating telomerase seems to do approximately nothing to prevent cellular senescence or aging more generally. What's going on here?

Stem cells produce a protein (telomerase) which extends their telomeres. Since most cells are replaced regularly by stem cells, that should be enough to prevent telomere-induced senescence. But then why do we observe shorter telomeres in old age? The key result here is that DNA damage, and ROS in particular, shorten telomeres. *In vivo*, telomere length is mainly a measure of ROS damage, not a measure of age directly. So, while telomere loss can cause senescence, the main thing which causes telomere loss is not gradual shortening as cells divide, but rather ROS-induced damage.

So, telomere loss is likely involved as an intermediate cause (as a type of DNA damage induced by ROS), but not a root cause.

AGEs

Advanced glycation end-products (AGEs) are proteins which have somewhat-randomly reacted with a sugar in a [Maillard reaction](#) - the same type of reaction which browns foods and gives them flavor when cooking. These products are hypothesized to accumulate long-term in the body, since they can't be broken down.

I don't know whether there's any evidence that these molecules *actually* accumulate long-term. (Just because they're not broken down doesn't mean they're not simply excreted.) I haven't seen direct evidence, but I haven't searched very carefully either, and I haven't seen direct evidence against.

The main argument I've seen in favor of AGEs as a root cause of (some) diseases of aging was from aminoguanidine. It seems like people thought it prevented AGE formation before they realized that it interrupted oxidative damage more broadly, and thus interpreted results from aminoguanidine as indicative of a causal role for AGEs. In hindsight, even if this had shown a causal role for AGEs, it would have ruled them out as a root cause: the effects of AG rapidly wear off once administration of the drug ceases, so it's definitely not blocking any root cause.

Senescent Cells as Root Cause

For a while, people hypothesized that senescent cells accumulate with age without turning over, acting as a root cause. As mentioned earlier, the actual evidence suggests that senescent cells turn over on a timescale of days to weeks, which would mean this theory is wrong - senescent cell accumulation is not a root cause.

However, there is a saving throw: maybe a small subset of senescent cells are longer-lived, and the experiments measuring senescent cell turnover time just weren't capturing the long-lived subset in particular. Results from senolytics (drugs which kill senescent cells) suggest this is also wrong: the effects of senolytics rapidly wear off once the drug stops being administered, whereas reversing a root cause should set an organism back to a youthful state longer-term.

Protein Damage, DNA Damage, Etc as Root Cause

Sometimes people suggest protein damage, DNA damage, etc, as root causes. These generally turn over on fast timescales, so... no. I expect that most of these people do not have any thought-out notion of what "root" means in "root cause", and are just using it as a synonym for "really important".

Other Pieces of The Pathway

This last section will briefly dive into a couple other aspects of the core pathways of aging which I expect people might be interested in, and talk about how they fit into the picture.

Sirtuins and NAD

David Sinclair published [a popular book](#) on aging a couple years ago, mainly talking about his own research areas. The sirtuins are one of the main key pieces of that research.

We mentioned earlier that, when damage is detected, the cell redirects resources from repressing transposons to repairing damage. Sirtuins are one such resource. They directly trade off genomic stability (including transposon repression) for repair capacity.

Notably, sirtuins consume the energy carrier NAD as part of their repair role. Lots of things use NAD as an energy carrier, switching it between a high-energy and low-energy state, but sirtuins actually *consume* it - the whole molecule is incorporated into a new structure. Usually the cell has rather a lot of NAD, but if the damage load is high and the sirtuins are

doing lots of repairs, then it can be depleted. That leaves less of it for various cellular processes which use NAD as an energy carrier - including mitochondrial energy production.

This all fits neatly into our main model: high ROS-induced damage draws sirtuins away from transposon repression, so they become active. Meanwhile, the sirtuins' consumption of NAD can interfere with mitochondrial function, resulting in more ROS production.

This also adds in one more piece: at younger ages, certain kinds of cellular stress (like radiation exposure or chemical damage from an inflammatory response to an infection) can also damage DNA, temporarily reducing repression of transposons. This probably won't result in immediate senescence in most cells, but a few transposons may copy before everything goes back to normal. The aging clock ticks forward a little faster than usual, due to these events.

Damaged Proteins Connect To Everything

We've mentioned oxidatively damaged proteins many times now. What we didn't mention was [the numbers](#): in young organisms, perhaps 10% of protein is damaged. In old organisms, it's more like 20-30%.

That is a percentage of *all proteins*. And almost everything our cells do is done by proteins.

Natural conclusion: efficiency of a very wide variety of processes will be thrown off in aging.

In most cases, this shouldn't be too noticeable - we're only talking about a 10-20% change, and random noise does that for most protein species anyway. And the body has [lots of feedback loops in place to handle exactly this sort of thing](#). By and large, biological networks evolve to be robust to 10-20% changes in protein concentrations. But, it does make things difficult for science: if there's a 10-20% change in everything, then there are always going to be statistically significant age-related changes in everything, even though most of them aren't actually all that relevant.

So, be warned: there's lots of 10-20% age-related changes all over the place which mostly aren't that relevant.

Recap

Here's the core positive feedback loop again:

- A cell's DNA is damaged, inducing a damage response.
- As part of this damage response, mitochondria are shifted into a lower-efficiency state, producing less energy and more ROS.
- The ROS then further damage DNA.

Since this is a positive feedback loop, it has two stable states: one state with low damage and ROS (the "normal" cell state), and one with high damage and ROS (the "senescent" cell state).

A few days after senescence is triggered, the cell's transposons become active, copying themselves and further damaging the DNA. At this point, the transposon activity alone is enough to maintain the senescent state, and senescence is locked in. The cell will remain senescent until it's cleared out by the immune system, a few days to weeks later.

What causes more senescent cells as we age? The transposon model says that transposon count increases with age - they are the root cause which permanently accumulates over time. Once the transposon count in a cell is high enough, it produces enough damage to trigger the senescence feedback loop. More specifically, we end up with stem cells with

enough transposons to be just below the senescence trigger, and these stem cells produce new cells which rapidly senesce.

Senescent cells release inflammatory factors (the SASP) as well as ROS. These cause the bulk of age-related diseases. ROS damage to fats causes buildup of fatty streaks and eventually plaques in the arteries - i.e. atherosclerosis - eventually leading to blockage and strokes. Damage to proteins hardens the blood vessels, leading to heart failure and aneurysm. Chronic inflammation underlies arthritis and possibly osteoporosis. Senescence itself also leads to loss of cells, including muscle loss.

Finally, in very old age, the whole process can accelerate: ROS produced by senescent cells can cause damage in adjacent cells, inching them closer to senescence as well. The more cells senesce, the more damage they deal to healthy cells. Eventually the whole thing goes supercritical (though on a slow timescale), leading to an exponential acceleration of disease progression in old age.

A whirlwind tour of Ethereum finance

As a hacker and cryptocurrency liker, I have been hearing for a while about "DeFi" stuff going on in Ethereum without really knowing what it was. I own a bunch of ETH, so I finally decided that enough was enough and spent a few evenings figuring out what was going on. To my pleasant surprise, a lot of it was fascinating, and I thought I would share it with LW in the hopes that other people will be interested too and share their thoughts.

Throughout this post I will assume that the reader has a basic mental model of how Ethereum works. If you don't, you might find this [intro & reference](#) useful.

Why should I care about this?

For one thing, it's the coolest, most cypherpunk thing going. Remember how back in 2012, everyone knew that Bitcoin existed, but it was a pain in the ass to use and it kind of felt weird and risky? It feels exactly like that using all this stuff. It's loads of fun.

For another thing, the economic mechanism design stuff is really fun to think about, and in many cases nobody knows the right answer yet. It's a chance for random bystanders to hang out with problems on the edge of human understanding, because nobody cared about these problems before there was so much money floating around in them.

For a third thing, you can maybe make some money. Specifically, if you have spare time, a fair bit of cash, appetite for risk, conscientiousness, some programming and finance knowledge, and you are capable of and interested in understanding how these systems work, I think it's safe to say that you have a huge edge, and you should be able to find places to extract value.

General overview

In broad strokes, people are trying to reinvent all of the stuff from typical regulated finance in trustless, decentralized ways (thus "DeFi".) That includes:

- Making anything that has value into a transferable asset, typically on Ethereum, and typically an [ERC-20 token](#). A token is an interoperable currency that keeps track of people's balances and lets people transfer it.
- Making liquid exchanges where you can swap all of those tokens at market prices.
- Making schemes for moving those tokens over time, like borrowing, futures, etc.
- Making elaborate scams and arbitrages to obtain other people's tokens.

It's not completely clear to me what the main value proposition of all of this is. It's easy to generate things about it that seem somewhat valuable, but hard to say how each stacks up. Some possible value includes:

- Evading regulation, like securities laws, money laundering laws, sanctions, capital controls, laws against online gambling, etc. etc.
- Allocation of capital among projects that can raise money using cryptocurrency tokens (because somehow they have a scheme to tie the success of their project to the value of the token, making it a kind of virtual equity.)
- Having less middlemen than existing financial systems, making it more trustworthy and cheaper. (It is not currently more trustworthy or cheaper than mainstream American institutions, but it plausibly could be in a few years.)

Tokenization

The first step is to make everything into an ERC-20 token. This will let all the other products work with everything, because they will interoperate with ERC-20 tokens.

Stablecoins and pegs

It's common for someone to want to own an Ethereum version of some other asset that is not Ethereum, so that they can use it on Ethereum. The most typical example of this is US dollars. A token whose price is designed to be pegged to an external thing like this is called a [stablecoin](#).

There are a few techniques people use to accomplish this. The most popular one is to have a giant pile of US dollars somewhere under someone's control, and have that person act as a counterparty for anyone who wants to buy or sell 1 US dollar for 1 token. This is what [Tether](#) and [USDC](#) do.

A more complicated version is to have a giant pile of something that is not US dollars, but is worth something, and then be willing to trade 1 US dollar worth of whatever, for 1 token. This seems like what [Reserve](#) does. In a way it's also what [MakerDAO](#) does; I'll discuss this later when I get around to talking about loans.

Another asset people really like to own on Ethereum is bitcoin. This is accomplished by [WBTC](#) ("wrapped bitcoin") using a giant-pile-of-bitcoin technique, with a few useful twists; firstly, there's a cryptographic scheme by which you can trustlessly mint the WBTC tokens on Ethereum and send the BTC to the custodian on Bitcoin atomically. (This is [a general mechanism](#) for cross-blockchain transfers.) Secondly, you can go look at the Bitcoin balance of the custodian, so you know they actually have some collateral somewhere.

There are some other schemes for doing this that are trying to get around the requirement of having a giant pile of assets, by making it so that more of the coin is minted when the value of the coin goes above the peg, and the coin is burnt somehow when the value goes below the peg. [Basis](#) kind of tried to do this. It seems like nobody quite knows yet whether or not this is possible.

Governance tokens, i.e. tokenized equity

It's extremely common for new cryptocurrency projects to want to essentially give out equity in their project as a way to make money. In particular, a very common way to do it seems to be to mint "governance" tokens representing equity, give some of them to investors, and also give a bunch to early users, as a way of establishing a big userbase and attaining some network effects.

Frequently these tokens have a profit sharing component. Almost always, they let the owners of the tokens make some kinds of decisions relevant to the project -- parameter values, future redistribution of profits, development direction, project mergers, and so on, under the assumption that they will do so in a self-interested way as the stakeholders in the project.

I've found that governance tokens seem surprisingly valuable on the market, even when there is no explicit profit sharing component. For example, the [Uniswap governance token has a market cap of \\$6B](#). Like companies that don't issue dividends or buy back shares, it's slightly puzzling at first glance where so much value is coming from. Expectation of future profit sharing mechanisms? The ability to control the future direction of the project in other ways? Maybe people just like the stock. A lot of how people are making money right now on these platforms is by taking advantage of aggressive governance token rewards for using new and speculative projects.

Exchanges

Once everything is a token, you want to be able to trade the tokens. The old cryptocurrency model of having centralized, NYSE-style exchanges hasn't gone away; you can still trade lots of tokens on Coinbase and so on. However, a totally new variety of trustless, decentralized exchanges have popped up in the past few years. I will explain [Uniswap](#), which is just about the simplest possible one, and then talk about some variations on a theme.

Automated market making

Uniswap is a [constant function market maker](#). More specifically, it's a constant product market maker. This means it is an Ethereum contract that manages a giant pool of two tokens, and anyone can trade one token for the other against the pool as long as their trade preserves the product of the two token quantities in the pool.

The main reason you would do this is if you wanted one instead of the other. But even if you didn't, you would do it anyway, if the current rate being offered (i.e. the ratio of the two token quantities in the pool, times the constant) is different than the rate you can get on another liquid market. If it's better, you trade one way on the pool and the other on the market. If it's worse, vice versa. As a result of this arbitrage, Uniswap's rate will always magically be at the prevailing market rate.

Well, not quite. It will always magically be somewhere inside the prevailing market spread, plus or minus transaction fees, for the first epsilon quantity you're trading. After that, it will slip to a worse rate. The [slippage](#) will depend on how giant the pool of tokens is compared to your trade.

There are a number of appealing properties about this compared to a centralized exchange with an order book:

- All the machinery for making new pools is all automated and anyone can do it. My understanding is that you can invent two new ERC-20 tokens right now, mint a bunch of them, and make your very own Uniswap pool, without asking anyone's permission.
- The trade I described is way less computationally expensive than keeping track of an order book, so it costs way less gas. This is a huge concern right now in Ethereum land.
- The software is quite simple. That means there probably aren't many bugs.
- You only need to trust the software; you don't have to put your money into a questionable cryptocurrency exchange platform.

As a result of these properties, a lot of people are making similar kinds of automated market makers for different purposes. For example, [Curve](#) is a CFMM which uses a function that has less slippage when the two tokens have very similar value; it's used for trading stablecoins and tokens that are pegged against each other. [Balancer](#) is a CPMM that works with a set of more than two tokens and maintains the constant product invariant for all of them, so you can trade any of them for the rest of them.

Yield farming

All of these automated market makers operate on the same model of incentivizing people with money ("liquidity providers") to make a really giant pool of tokens for them, so that they can support lots of trade volume with minimal slippage. There are typically two incentives: one is that you get transaction fees on the trades proportional to your portion of the pool, the other is that you get governance tokens for the exchange platform.

These frequently add up to what looks like a [really good APR](#) on your money. The shallower the pool and the more the trading volume, the more of the transaction fees you are making. So a lot of people spend their time looking for AMM pools with attractive yields, buying whatever tokens that pool supports, and depositing them in the pool. Then later you can take your tokens back out plus whatever profits you made in the meantime. This is called "[yield farming](#)".

At a glance this sounds like a relatively safe way to make money, but there are a few non-obvious risks. One is your exposure to the tokens you bought and put in the pool. If the pool is so shallow, it's probably because one or the other of those tokens is new or has questionable value, so you might not be thrilled to own them. Another is systemic risk based on smart contract hacks, malicious governance, and so on. Another is [impermanent loss](#); it turns out that for many AMMs, liquidity providers will not be indifferent to price movements of the tokens. So it's a way to make money, but not necessarily a very safe way to make money.

Loans

There seems to be a lot of demand for loans of cryptocurrency. Like a lot of this technology, it's not totally clear to me where the demand is coming from. Some possibilities:

- Borrowing for the purposes of shorting this or that token.
- Borrowing as a way to invest on margin.

Because it's supposed to be trustless and decentralized, all loans are necessarily at least fully collateralized (except for flash loans, which I describe later.) I'll explain two prominent lending systems.

MakerDAO & the DAI stablecoin

[MakerDAO](#) is one of the oldest Ethereum projects. It has a mechanism that not only results in a way to borrow with collateral, but also results in a stablecoin just sort of popping out as a consequence, which is cool.

The way it works is, you can send some stuff to a Maker "vault" contract. The stuff can be any token that has value according to a big list that Maker manages collectively. The USD value (per some governance-approved oracle) of the stuff acts as your collateral. Once you have some stuff in there, you can mint a token called DAI proportional to the amount of your collateral (everything is somewhat overcollateralized, because most of the stuff has a lot of price variance) and do whatever you like with the DAI. Later, you can pay back the DAI to get your collateral back.

What if a loan becomes insufficiently collateralized, because the market value of the collateral went down? Maker solves this problem by making it profitable for other people to liquidate you in this case; other people can [bid DAI in an auction](#) to pay off your loan and take your collateral. Naturally, this should happen basically instantaneously.

There are some other mechanisms in play; as usual, there's a governance token MKR that shares profits from a fee you pay on your loans. There's also a kind of [emergency shutdown lever](#) that MKR holders can pull to liquidate everyone and redistribute the collateral to everyone who owns DAI, presumably in case Maker becomes self-aware.

Why is DAI stable? Well, if DAI got much more expensive than \$1, you can make more DAI, sell the DAI on the market, and then default. If DAI got much less expensive than \$1, people with outstanding loans could buy DAI off the market and pay off their loans with it. If DAI really tanked, DAI holders could invoke the emergency shutdown and reclaim the collateral. And when DAI starts drifting a bit, Maker governance tries to nudge it using a familiar central bank tool -- [the interest rate on Maker loans](#). All of these things help peg DAI to the dollar and they seem to mostly work.

Compound

Whereas Maker is basically a way to "lend to yourself", or take out leverage, [Compound](#) is an interpersonal lending system.

You start by putting a bunch of stuff into Compound, which will serve as collateral. Hopefully, other people have also put some stuff into Compound that you would like to borrow. If so, you can borrow it -- like Maker, you can borrow \$1 of stuff for every \$1.50 of collateral. Interest will accrue at some rate, and you can repay the loan and get your collateral back any time you choose. Meanwhile, the other people who put the stuff you borrowed there get the interest you paid.

What's the interest rate? Each token you can lend or borrow ([e.g.](#)) has an "interest rate model", chosen by platform governance (I don't know how they decide what it ought to be) which defines the interest rate as some function of the supply and demand for loans of that token. So if the demand for loans is high, but there isn't much of that token sitting around, lenders will get a high rate when they put that token into the pool.

Compound has a different approach to liquidating positions that have dipped below their collateral requirements due to market movements. If a loan is not sufficiently collateralized, any random person can call a function on the contract to steal some of the collateral and pay back some of the loan, at a better than market rate; so they will. It's a kind of "whoever notices it first" situation rather than Maker's slow auction. To me this seems simpler and effective -- I wonder why Maker didn't do it.

Malarkey

It turns out if you create a bunch of interoperable software that lets you transfer utility around, the software sometimes has unanticipated problems. It's very common for people to notice some way to [Dutch book](#) some combination of these services and pump money out until someone figures out how to do something about it.

Recently, someone created an amazing new invention which is like the [Low Orbit Ion Cannon](#) of arbitrage, called a [flash loan](#). A micro loan is a very small loan; a flash loan is a very fast loan. You can borrow the money -- with zero collateral required, and with a very small fee -- if and only if you can repay it at the end of the same Ethereum transaction. I find this quite remarkable.

[Aave](#) (how do you pronounce this?) is the leading flash loan service. To use it, you have to deploy a contract that calls the loan contract, does whatever you wanted the money for, and calls the loan contract again to repay it. I guess there are also some other use cases for flash loans. For example, if you have a collateralized loan on Compound, and you don't want to repay the loan, but you wish the collateral were something else, you could (as I understand it) take a flash loan of the other thing, deposit it, withdraw your old collateral, and sell the old collateral to repay your flash loan.

But the [really sweet use case](#) is incinerating slightly imperfect systems made of Ethereum contracts and taking out all the money. Check out some of these autopsies.

[Burninating the peasants](#) at bZx, a kind of margin trading platform:

The origin account of the transaction starts with nothing, then borrows and moves a pile of cash, causes two huge Uniswap orders (in both directions) in the course of the same transaction, and ends up with 65 ETH. That definitely looks fishy.

[Burninating the countryside](#) at Harvest, a decentralized hedge fund:

The attacker repeatedly exploited the effects of impermanent loss of USDC and USDT inside the Y pool on Curve.fi. They used the manipulated asset value to deposit funds into the Harvest's vaults and obtain vault shares for a beneficial price, and later exit the vault at a regular share price generating a profit...The value lost is about \$33.8 million, which corresponded to approximately 3.2% of the total value locked in the protocol at the time before the attack.

[Burninating the thatched-roof houses](#) of something called Value DeFi after they [tweeted](#) about their new "flash loan attack protection":

The attacker returned \$2 million to the protocol and pocketed \$6 million — and with it left one audacious message stating, “do you really know flashloan?”

Value Defi said it suffered a “complex attack that resulted in a net loss of \$6 million.”

-
1. FlashLoan 80k ETH from Aave
 2. FlashSwap 116M DAI from Uniswap
 3. Swap 80k for 31M USDT on Uniswap (effective 76.6k remaining 3.3k ETH)
 4. Deposit 25M DAI on ValueMultiVaultBank
 5. Swap 91M DAI to 90.2M USDC on Curve
 6. Swap 31M USDT to 17M USDC on Curve
 7. Withdraw 33M 3CRV from ValueMultiVaultBank
 8. Swap 17.3M USDC to 30.9 USDT on Curve
 9. Swap 90.2M USDC to 90.9 DAI on Curve
 10. Remove liquidity with 33M 3CRV for 33.1M DAI on Curve
 11. Swap 30.9 USDT for 76k ETH on Uniswap
 12. Payback FlashSwap 116M DAI to Uniswap
 13. Swap 283k DAI to 606.9 ETH on SushiSwap
 14. Payback FlashLoan 80.072k to Aave
 15. Transfer 2M DAI to Value Deployer
 16. Transfer 5.4M DAI to exploiter

You get the idea.

Should I really send my money to this computer program?

Good question. After a few evenings of looking at this I haven't been able to easily quantify the risk involved in using any of these tools to actually attempt to make money. If you go look at being a liquidity provider for stablecoin pairs on Curve -- which "should" be a very low risk investment when everything works -- you're getting a rate of a few dozen percent APR. But the potential, hard-to-quantify costs and risks abound:

- Is there a bug in Curve's contracts? Maybe. Maybe not, if nobody has found it yet? But who knows?
- Will you screw up in the process of moving your money onto the platform, or get owned by malware that nabs your keys?
- Will the underlying tokens, like DAI or USDC, lose their value for a hard-to-understand reason, even though they are not supposed to?
- What will the APR be like three months from now? You're getting rewarded in CRV governance tokens, and it's expensive in gas to sell them instantly -- are those going to hold their value? Who exactly is on the other side of this trade?
- Speaking of gas, given current prices, if you don't put a lot of money in, you're probably not looking at an attractive proposition.
- Doing US taxes on this is ridiculous. There's software that is supposed to help build your return, but it's hard to keep the tax implications in your head and take them into account.

So... a few dozen percent APR is a lot. And it's not *designed* to be a scam. To paraphrase Eliezer, "The contract does not hate you, nor does it love you, but you have deposited ERC-20 tokens which it can use for something else."

I advise the *emptor* to *caveat*. But I also advise you to check this stuff out, if you independently value exploring an exciting new land of nonsense.

Seven Years of Spaced Repetition Software in the Classroom

Description

This is a reflective essay and report on my experiences using [Spaced Repetition](#) Software (SRS) in an American high school classroom. It follows my [2015](#) and [2016](#) posts on the same topic.

Because I value concise summaries in non-fiction, I provide one immediately below. However, I also believe in the power of narrative, in carefully unfolding a story so as to maximize reader engagement and impact. As I have applied such narrative considerations in writing this post, I consider the following summary to be a spoiler.

I'll let you decide what to do with that information.

Summary (spoilers)

My earlier push for classroom SRS solutions was driven by a belief I came to see as fallacious: that forgetting is the undoing of learning. This epistemic shift drove me to abandon designs for a custom app that would have integrated whole-class and individual SRS functions.

While I still see value in classroom use of Spaced Repetition Software, especially in basic language acquisition, I have greatly reduced its use in my own classes.

In my third year of experiments (2016-17), I used a windfall of classroom computers to give students supervised time to independently study using an SRS app with individual profiles. I found longer-term average performance to be slightly worse than under the whole-class group study model, though students of high intelligence and motivation saw slight improvements.

Intro and response to Piotr Woźniak

I have recently received a number of requests to revisit the topic of classroom SRS after years of silence on the subject. Understandably, the term "postmortem" has come up more than once. Did I hit a dead end? Do I still use it?

Also, I was informed that SRS founding father Piotr Woźniak recently added [a page](#) to his SuperMemo wiki in which he quoted me at length and claimed that SRS doesn't belong in the classroom.

Well, I don't have much in the way of rebuttal, because Woźniak's main goal with the page seems to be to use my experience as ammunition against the perpetuation of school-as-we-know-it, which seems like a worthy crusade. He introduces my earlier classroom SRS posts by saying, "This teacher could write the same articles with the same conclusions. Only the terminology would differ." I'll take that as high praise.

If I were to quibble, it would be with the part shortly after this, where he says:

The entire analysis is made with an important assumption: "*school is good, school is inevitable, and school is here to stay, so we better learn to live with it*".

Inevitable? Maybe. Here to stay? Realistically, yes. But good? At best, I might describe our educational system as an "[inadequate equilibrium](#)". At worst? A pit so deep we still don't know what's at the bottom, except that it eats souls.

Other than that, let me reiterate my long-running agreement with Woźniak that SRS is best when used by a self-motivated individual, and that my classroom antics are an ugly hack around the fact that self-motivation is a rare element this deep in the mines.

Anyone who can show us a way out will have my attention. In the meantime, I'll do my best to keep a light on.

Prologue

At the end of my 2016 post, I teased a peek at a classroom SRS+ app I was preparing to build. It would have married whole-class and individual study functions with some other clever features to reduce teacher workload.

I had a 10k word document in hand: a mix of rationale, feature descriptions, and hypothetical “user stories”. I wasn’t looking for funding or a co-founder, just some technical suggestions and moral support. I would have been my own first user, and I had to keep my day job for that anyway.

But each time I read my draft, I had this growing, sickening sense that I was lying to myself and my potential customers, like a door-to-door missionary choking back a tide of latent atheism. And I should know, because the last time I had felt this kind of queasiness I was a door-to-door missionary choking back a tide of latent atheism.

I thought maybe this was just the kind of general self-doubt common to anyone undertaking something audacious, but I paused my work on it for another school year while I tried the obvious thing: providing students individual SRS app profiles and supervised class time in which to use them.

This is a two-part essay, and in Part 2, I’ll tell you how that went. But in Part 1, I’m going to make the case that Part 2 doesn’t matter very much.

Part 1: Everybody Poops

A great and terrible vision

As I wrapped up my Third Year experiment, I again tried to sort out my feelings about my visionary SRS app design, which I hadn't updated despite a year of fresh experience. Was it just self-doubt?

The fact that I could only code at a minimal hobbyist level didn't feel like the biggest hurdle. I think I could have picked up enough skill in that area. But even with a magical ability to translate my vision into code, I would have been up against a daunting base rate of failure for education startups. Also, I didn't consider myself a very typical teacher: What sounded brilliant and intuitive to me would probably seem pointless and nonsensical to 95% of my peers.

Still, I pulled out my Eye of Agamotto and checked out all of the futures where I developed the app. In almost all of these, nothing came of it. But in the few where my app saw high adoption, the result was... dystopia! Students turned against their teachers, and teachers against their students. Homework stretched to eternity. Millions of children cursed my name. The 'me' in these futures wore an ignominious goatee and a haunted stare.

Used judiciously for the right concepts, in the right courses, by the right teachers, I still think my imagined app could be a powerful tool. But I don't see any way to keep it from being abused. Well-intentioned teachers would put too much into it and demand too much from students. Any safeguards I put in to prevent this would just invite my app to be outcompeted by an imitator who removed these safeguards (which would seem arbitrary and restricting to most users).

I'm convinced of this because the me who wrote the original "A Year of Spaced Repetition..." post would have abused it. Let's see... He was averaging seven new cards a day? (That's 2-3 times what I would recommend today.) He uncapped the 20 new card/day limit? He knew even then that he was adding too many cards, but failed to cut back the following year? I'm not encouraged.

"But wait," you say. "You didn't think you were a typical teacher. Maybe a typical teacher could be trusted?"

No.

In defense of forgetting

The "problem" is that teachers instinctively introduce far more content than students can be expected to remember. This was obvious to me when I was averaging seven new cards a day, which still felt like a brutal triage of my total content.

Covering more material than can be retained isn't bad teaching, though. In fact, it's a good and necessary practice. Content — the more the merrier — is the training data the brain uses to form and refine mental models of the universe.^[1] These models tend to be long-lived, and allow the brain to re-learn the content more deeply and efficiently if it ever comes up again. They also allow it to absorb new-but-conceptually-adjacent contents more readily. In cognition, as in nutrition, you are what you eat — and good digestion naturally produces solid waste. The original training data is subject to lossy compression, with only a few random fragments left whole and unforgotten. (Tippecanoe, and Tyler Too! The mitochondria

is the powerhouse of the cell!) Such recollections are corn kernels bobbing top-side up in a turd floating down the river Lethe.

This is normal and fine. *Regular*, even.

But the educational establishment doesn't see it that way. The teacher I was seven years ago didn't see it that way. And I now realize that the teacher I was five and six years ago had queasy feelings because he was *starting* to see it that way. Following my gut, without fully understanding or even entirely registering what I was doing, I slowly turned around and started walking the other way, abandoning my app design and the unfinished "Third Year" report.

The orthodox view equates forgetting with failure. It's not "Everybody poops". It's "Poop is inadequate. How can we get more corn, less poop?" This belief is implicit whenever [someone](#) laments the "summer slide", or opines that students missing school during the Covid pandemic are "losing" months of learning — as if kids are spinning their progress meters backwards, just pooping away without anyone trying to stop them. Under this view, we keep kids in school partly to stop the leaks, and partly to stuff them with new knowledge faster than they can expunge old knowledge.

If this is how you see education, SRS is a tool to keep students from pooping. It offers the tantalizing possibility of learning without forgetting. Two steps forward, no steps back. Why *wouldn't* you push it as hard as possible?

Don't get me wrong. All else being equal, learning without forgetting would be great. But the most important effects of learning — lasting changes to our mental machinery — happen whether or not we forget the content. Once the lesson is over, dear teacher, your best shot at lasting growth has already left the harbor. So why are you still trying to hold back the tide? Why are you planning to punish your students for pooping on Tuesday, the day before your test, instead of Thursday, the day after it?[\[2\]](#)

In defense of remembering

This is not a "How I Learned to Stop Worrying and Love Forgetting" essay. I don't love forgetting. I will be the first to argue the merits of not forgetting right away. The longer we can keep ideas floating around in our heads, the greater their "cross-section", [as I put it in 2016](#), with more opportunities to make associative connections that cause useful long-lived updates to our mental models.

Unfortunately, I have not found SRS to be great at fostering the sorts of reflective mental states conducive to insight, except when studying on my own at a deliberately slow pace, as while on a walk. In such a use case, SRS no longer has quite the time-efficiency advantage that is its main selling point. The opportunity cost of using it goes up. In a whole-class SRS session, long reflective pauses between cards would invite frustration and misbehavior, and we wouldn't get through very many cards.

In defense of remembering, I will also argue that some skills are simply impossible without a continuous retention of specific dependencies. These skills tend to be technical. Heck, this might be the definition of a technical skill.

With a few mostly upper-level exceptions, though — math, physics, chemistry — most of what we teach in school is more conceptual than technical. We make you take history so you have a better model of how civilizations and governments work, not so you remember who shot Alexander Hamilton. We make you take English to improve your word-based input and output abilities, not so you remember the difference between *simile* and *metaphor*. At least, I hope we do.[\[3\]](#)

Besides, even in the technical classes, forgetting is the near-universal outcome, and the long-term benefits are mostly conceptual — for if you don't use these skills continuously for the rest of your life, you're almost certainly going to lose them. Maybe more than once.

I've forgotten algebra twice. I've forgotten how to write code at least three times. I can't do either one at the moment. But I'm still changed by having known them. I have an intuition for what sorts of problems ought to be mathematically solvable. I can think in terms of algorithms. And I could relearn either skill more easily than on the first or second occasions. Also, relearning has an anecdotal tendency to deepen understanding in a way that continuous retention may not, especially when approached from a different direction.

Still, as long as I'm defending retention, I think it's valid to ask whether we should force kids (and often, by extension, their parents) to relearn math every frickin' year. Consider: The conventional wisdom is that technical companies begrudgingly expect to have to (re)train most new workers in the very specific areas they need. They look to your resume and transcripts mostly for evidence that you have learned technical skills before and can presumably learn them again. I don't think they care if you've re-learned them three times already instead of six. So, if we're going to force kids to demonstrate intermediate math chops to graduate (a dubious demand), perhaps we could at least wait until the last practical moment, and then do it in bigger continuous lumps — like two-hour daily block classes starting in grade 9 or 10 — so they would have fewer opportunities to forget as they climb the dependency pyramid. Think of the tears we could save (or at least postpone).

The value proposition of classroom SRS

Anyway, classroom SRS has its strengths, but midwifing conceptual insights doesn't feel like one of them. I think it's also reasonable to assume that students forget almost everything from a classroom SRS deck as soon as they stop using it.

Adjusting for these two assumptions, the terrain where classroom SRS can beat out its opportunity costs dramatically shrinks. But I believe it still exists, at the intersection of high automaticity targets and medium-term objectives.

With [high automaticity targets](#), what you're trying to train is a reflexive response to a stimulus that is going to look a lot like the study card. Foreign language vocabulary is my poster child for this. You're not drilling the words to unearth insights. You're drilling for speed, so that they can keep up when a word pops up in a real-time conversation.

You're also trying to drill away the need for conscious awareness. You want that front-side combination of sounds or letters to cause the back-side set of sounds or letters to pop automatically into their heads. This is my intent when I drill my English students in word fragments (prefixes, roots, suffixes), which are really just bits of foreign language (Greek, Latin). If it's not automatic, then they'll gloss right over the possible meaning of "salubrious", even though they have learned that "salu" usually means "health".

By *medium-term objective*, I mean "I want my students to have automatic fluency with the content of these cards on Day X", where X is a date between one week and three months in the future. It shouldn't be sooner than that, in accordance with [Gwern's "5 and 5" rule](#): You probably need at least five days to get any real advantage from SRS. And it shouldn't be later than a few months, for two reasons: First, we're assuming the students will forget it all once they stop studying, which is all but guaranteed after the end of the course; there's little point in keeping those cards in rotation after Day X. Second, I probably don't want to start those cards until the last practical minute, which is unlikely to be more than three months ahead of time.

Why three months and not six? It's not a hard-and-fast rule, but from the experience of my first three years of classroom SRS, if you're trying to retain things for more than a few

months, the total number of cards is likely to become greater than you can productively study every day, and many cards will languish unseen. Plus, your roster can change, especially over a semester break. The set of students you have in six months might only have 70% overlap with the set you have now. Really, you should wait until the last practical minute.

But what constitutes a worthy “Day X”? It might be a test. But if it’s *your* test, you may not have been listening. *Your* test may just be arbitrarily punishing some kids for forgetting a little sooner than others. However, if it’s an external test, with high stakes for you and your students, then it could be a worthy Day X indeed. For me, Day X is the day of the big state test — the one used to compare students to students, teachers to teachers, and schools to schools.

When your students do well on an external test, though, please keep a healthy perspective. A high test score doesn’t mean they can do the hard things now and forever. It means they were able to earn a high test score on Day X. They will forget almost all of it afterwards. But you will have given them their best chance to signal to others that they can learn hard things, and that you can teach them hard things, and that your school has teachers who can teach hard things.

Day X doesn’t have to be a test. If you’re optimizing for brain change that persists after they forget all of your content, Day X could be an immersive event. Maybe your Spanish class is going to Madrid. You know they will have a deeper experience if you can bring their vocabulary to a peak of richness and automaticity on the eve of departure. Yes, they’ll still forget almost all the words later. But they might retain a glimpse of how the world looked when seen through another language.

Maybe your event is smaller. A virtual trip. An in-class conversation day where we pretend we’re at the beach (“¡En la Playa!”). Maybe their long-term takeaway will be an appreciation for how different languages use different grammars, which is not something most people even consider until they’ve studied a second language. Get their mental gears turning hard enough, and they might even see grammar as an arbitrary construct with tunable parameters and tradeoffs that influence what can be communicated easily. Maybe they’ll independently rediscover the Sapir-Whorf Hypothesis. But they’re not going to remember how to say ‘sand’. Nope. ‘Shark’, maybe (¡Tiburón!). But you can’t predict this, and it’s probably not worth the effort to try.

But maybe you’re not teaching a foreign language. No matter your subject, Day X could be any conceptually demanding lesson or unit that is difficult to even talk about without fluency in a given set of terms. These aren’t very common in 10th Grade English, though they come up more often in my Creative Writing class. In these cases, however, the dependent terms are conceptually rich enough that they don’t lend themselves very well to cards, and I find it’s better to just quickly re-teach them in front of the lessons that use them. “Remember how we said...”[\[4\]](#)

How I currently use classroom SRS

As you may have guessed, I’ve radically scaled back my usage of classroom SRS since those first three years. In fact, for the last four years, I’ve only used it during a two-to-three month span leading up to the state test. And for the last two of *those* years, I’ve only used it for word fragments. I’m very unlikely to abandon its use for word fragments, though, because the most important thing I teach my students by using SRS is [the existence of SRS](#). Word fragments are my favorite way to demonstrate how efficient study time can be. I add no more than about ten cards per week, which means that most days’ study takes less than two minutes. (This is good, because my own enthusiasm now begins to flag by the two minute mark.) I give very short quizzes on the fragments so they can do well on them and see how a

little study can have a big payoff. (Remember that most of my students don't ever study on their own.)

I'm still using Anki, with different profiles for each class. I run the review in a call-and-response style, where I show and say the card, and they know to simply shout out the answer. On a good day, it becomes a kind of chant. The number, speed, accuracy, and confidence of the responding voices tells me which button to press, and there's usually a bellwether student I can listen for as I make my decision. Because I'm striving for very high automaticity, I almost always press either 2 (the shortest affirmative next-study delay) or 1 (the negative start-it-from-scratch button).

My students mostly like the call-and-response flow, as archaic as it sounds, and I will refer you to [an older footnote](#) about that time I observed a traditional one-room Mennonite schoolhouse:

I once had the privilege of observing part of a lesson in a traditional [Mennonite](#) one-room schoolhouse. I don't speak a word of Low German, but it was clear the kids knew whatever it was they were drilling as they stood up and recited together. Most striking was the fact that they were *all* on the same page. There were no stragglers spacing out, slumped over, dozing off. The teacher could confidently build up to whatever came next without fear of leaving anyone behind.

For at least a minute or two every day, even worldly American kids can enjoy the routine. As I put it elsewhere in that Second Year report, "They enjoy the validation they get with each chance to confirm that they remember something. They enjoy going with the flow of a whole class doing the same thing. They enjoy the respite of learning on rails for a change, without any expectation that they take initiative or parse instructions."

It probably goes without saying, but this call-and-response format only works well with cards with a very short answer that can be recalled very quickly. This is why I now only use SRS for word fragments. If I taught a foreign language, or even a lower-grade reading class with more basic vocab words, I would be using it more. My wife taught high school Spanish for a number of years, experimented with SRS, and is on the record as saying [Duolingo](#) deserves to eat the world. Anyone she could get to use it independently didn't really need her class to do well on the final assessments.

After the state test, my students will forget almost all of their word fragments. That is the way of things. Ashes to ashes, circle of life, or, to get back to my controlling analogy, "All drains lead to the ocean, kid." What I'm hoping will remain is an updated appreciation for what a little regular study can do, and a vague recollection that there are these apps out there that are, you know, like smart flash cards, that make it *fast* to memorize stuff.

Against apathy, toward apprenticeship

I'm nearing the end of Part 1, which means I'm nearing the end of my labors on this post, since Part 2 was mostly written five years ago. As writing projects go, I have found this one extraordinarily difficult. Over the course of its creation, I have pooped five times. It wants to be a book (or at least a blog), as everything I say tries to come out as a chapter of explanation having little to do with SRS.[\[5\]](#)

Well, I'm now going to indulge in several paragraphs where I don't tie it back to SRS, so I can tell you the story of how I reinvented myself after my third year of spaced repetition software in the classroom. This included moving to a new school where I would have greater freedom to pursue my evolving views about learning. For what it's worth, this story at least starts with SRS.

You see, it was during those dangerously long classroom Anki sessions six and seven years ago that I honed my sensitivity to students' moods, to my own mood, and to how these feed off of each other. Sustaining a session without losing the room was like magnetically confining hot deuterium plasma — dicey, volatile, but occasionally, mysteriously, over unity. [6] I came to view anti-apathetic moods as a kind of energy that can be harnessed to do work *and to create new energy*.

Apathy, you may recall, is the true enemy. I've always known that. I [called her out](#) five years ago[7], but soon came to realize I had been fighting her on the wrong front.

I had been preoccupied by the fact that students who don't care won't activate enough of their brain to get any benefits from our daily review. To be fair, that *is* a problem, if I'm trying to prime them for success at a Day X event. But the more insidious issue is that a student in the thrall of Apathy won't be churning their mental gears on any of the content I may have tricked them into learning, which means they'll just forget it all without having made any lasting changes to their models. That's not just an Anki-time problem. That's an all-the-time problem. If they don't engage with anything, they don't keep anything.

I set off on a holy quest for anti-apathetic energy.

My errantry led me, for a time, to study stand-up comedy, not just because humor creates energy, but because a big part of that craft is an acting trick where you deliver incredibly polished lines in a way that sounds like you're coming up with them right there in that moment.[8] Perceived spontaneity is a powerful source of energy even more versatile than humor.

I don't know if I learned much about scripted spontaneity that I could articulate, but I felt like some of it rubbed off on me just by watching the experts closely over extended periods. And you know what? A lecture isn't so different from a bit. A lesson isn't so different from a set. A single changed word, a half-second delay, a subtle shift in facial expression can completely change the way the moment feels to the audience class. And like a comedian workshopping new material on the road, I could use the fact that I might teach the same lesson five times in one day to test variations, trying to provoke more engagement, better questions, bigger laughs.

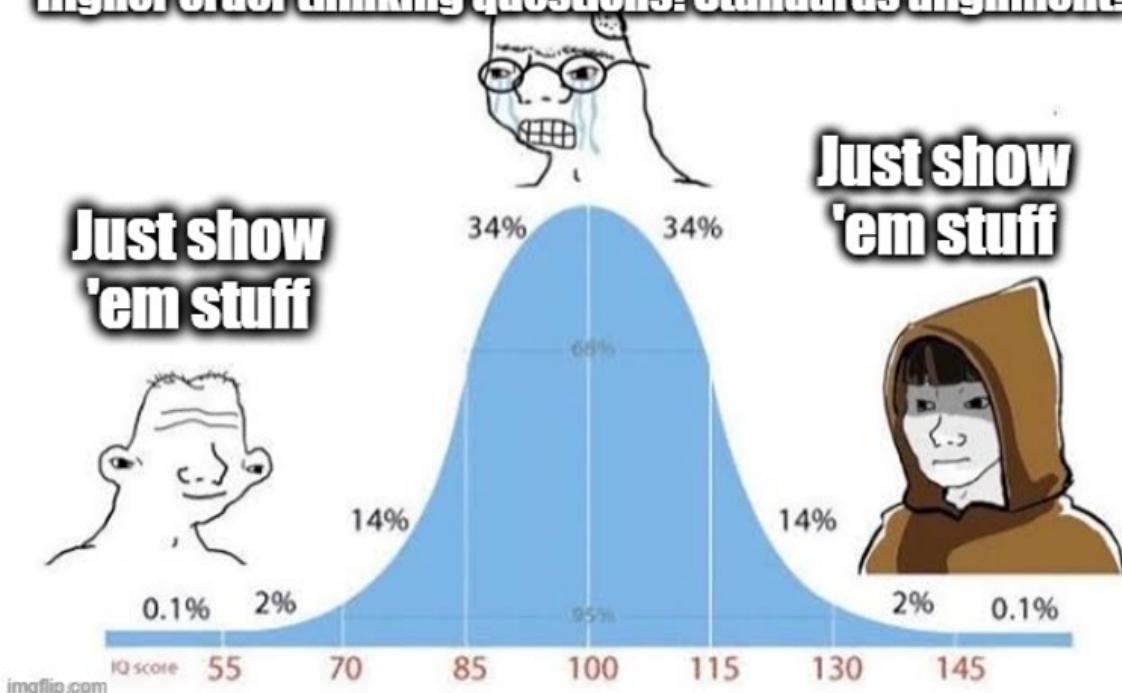
Equally important: I recognized that the process of refining the performance art was fun for me, and that my own engagement was the most powerful source of classroom energy. I could transmit it to my students, and maybe even get some energy back from them while I directed some of it into activity that would get their mental gears turning. Instead of burning out, I could burn brighter, and longer. On a good day, it became self-sustaining. On a great day, it could go supercritical, sending me home after my last class with my head spinning in a buzz of positive vibes and deep thoughts.[9]

During this same era, as part of my ongoing study of creative writing, I was binge-listening to interviews with television writers. One pattern that struck me was that it wasn't too uncommon for someone to just kind of find themselves working in that highly rarefied field simply because they had spent a lot of time around others who were already doing it. Without any organized instruction, they picked up on how it worked.

Did you catch it? That was twice that I had noticed how arcane expertise can rub off on people through prolonged proximity. That got me thinking about the [German Apprenticeship Model](#), and its medieval — nay, prehistoric — roots. It's how we used to learn everything, right? We followed mama out to the berry bushes, and papa out to the hunting grounds. The fact that it seemed to work for television writers told me that apprenticeship wasn't just for blue collar skills.

So, with the longer leash I enjoyed under my new bosses, I decided to move my instructional style closer to something resembling an apprenticeship where I mentored groups of 20-30 padawans in my arcane expertise.

**Nooo! Formative assessment! Summative assessment! Scaffolding! Differentiation!
Higher order thinking questions! Standards alignment!**



Yeah, I jumped on a trendy meme. Note my careful word choice: ‘show’, not ‘tell’. This, to me, is the defining action in mentor-apprentice relationships.

By switching schools, I lost my interactive whiteboard. So I replaced it with something even better: an extra computer on a make-shift stand-up desk (a narrow kitchen prep cart with fold-out boards.). A cheap second-hand monitor could face me while I mirrored that screen to the projector. Now I could do what I had seen coders do at instructional meet-ups: face the class while typing.

This meant I could *show* students what I do as a writer in real time, thinking out loud and watching their reactions as I typed. This could easily bore them, of course, but with strong energy-fu, old-school touch typing speed, and face-to-face interaction, I can pull it off more often than you might expect. On a good day, they find it fascinating. On one very special occasion each year, I do it for the full period, writing a 400+ word essay from scratch in 40 minutes with no prior knowledge of the prompt. Students have to hold their questions that day, and instead take observation notes, which become fodder for an extended debriefing discussion the next day.

The most important thing I’ve learned from those debriefings is that everyone can pick up something from a holistic demonstration like that, regardless of their skill level.[\[10\]](#) An advanced student might ask about my bracket substitution of a pronoun in a quote. An average student might say, “You used a lot of small and medium-sized body paragraphs instead of three big ones.” A sub-level student might say, “You didn’t like it if you used the same word too soon after you used it before.” And I always seem to get at least one surprising question about something I never would have thought to teach them, like, “How did you suck words into your cursor?” Then I’m like, “Oh, let me show you the difference between the Backspace and Delete keys...”

Did I make them memorize anything with that “lesson”? Nah. Did they make lasting updates to their mental models? Probably! Are you thinking of asking me, “But how do you test them on it?” Because if you are, then you *really* haven’t been paying attention!

There’s plenty more to be said about apprenticeship, but I think you get the idea, and this is still nominally an essay about classroom SRS.

If I had to summarize my self-reinvention in too many words, I would say that I’m now optimizing for “good days” at the high-energy intersection of “engaging for me”, “engaging for them”, and “conducive to lasting and worthwhile updates to their mental models”, with less regard for curricular scope and sequence.

In practice, this means... well, a lot of things. But it’s time I pinch off Part 1. That, “or get off the pot,” as they say.

Part 2: A Third Year of Spaced Repetition Software in the Classroom (2017)

[In this excavated report, text in brackets in commentary I'm adding in 2021. Anything out of the brackets is direct from my 2017 draft, or constructed from my notes to fit the perspective I had at the time.]

Synopsis and disclosure

I tried the obvious thing this year. Instead of game show-style whole-class front-of-the-room [Anki](#), I arranged for every student to be able to independently study material I created in [Cerego](#), both in and out of class.

Disclosure: Cerego provided me a free license for the year in exchange for some detailed feedback, which I gave them. This feedback was mostly about user interface issues and reports, the latter of which required some ugly scripting on my end to get numbers I found useful. As the Cerego team seemed to be rapidly iterating, I imagine they have made many changes and improvements to their app since 2017, though I have not used it since. Please keep this in mind as you read these years-old notes.]

Despite many small hang-ups, I was pleased with the Cerego's features and reliability. In exchange for a great deal of up-front effort, it gave me a unique window into student engagement and progress. Consequently, it proved to be an overwhelmingly potent tool for winning "the blame game", although I eventually came to feel uneasy about using this power.

Longer-term learning outcomes seemed, on average, to be slightly worse than with the whole-class Anki method. While highly motivated students benefited from being able to study more aggressively and efficiently than before -- and their objective scores were higher than ever -- their learning seemed less transferable to more authentic contexts. Students of lower motivation, while seeming to get little from either approach, got even less from this digital 1:1 method, and their slump accounts for the overall decline.

Setup

I taught a mix of regular (not honors) 9th and 10th English classes again, but over the summer of 2016 I was invited to move my classroom into an unusually-spacious converted computer lab in which 16 older desktop PCs were kindly left at my request. I had these arranged facing the sides of the room so I could see all screens easily. I allocated those PC seats on a semi-permanent basis as needed and requested. The balance of students sat at normal desks and used their phones for study.

This came with challenges. School WiFi was officially off-limits to students (though many always had the password anyway), and many students said they were at the whim of data caps they regularly pushed up against. Their phones, in most cases, were a generation or three behind state-of-the-art, with degraded batteries and exhausted storage capacity. A few students had difficulty even making room for the Cerego app that first week.

While our setup was marginal, between the PCs and phones, we only rarely ran into a situation where not everyone could be studying at the same time.

On the software side, it must be said that, for all its features, Cerego wasn't designed for my specific use case. The company's featured customers are business and colleges, who use the

product as part of packaged training programs and distance learning courses. Importantly, the app favors adding content into the learner's study rotation in blocks, *on the learner's own schedule*, rather than making it on the fly and trickling it immediately. It was also not designed to give a teacher "panopticon"-style real-time monitoring, nor to thwart adversarial users who want to look studious without studying.

Procedure

Before the start of each school day, I would consider the previous day's lesson content and add to the relevant Cerego study sets as appropriate. This process could be lumpy and not necessarily daily; some lessons invited a great deal of suitable content, and others none at all. Content additions were also far more common first semester than second semester, as I intentionally front-loaded material to maximize the time we would have to reinforce and apply it. During an average week where I added cards, we probably averaged about 50 additions. [!]

With a prominent timer at the front of the room, I allocated 10-12 minutes at the start of every 57 min class period as specially designated "Cerego Time". During Cerego Time, I would periodically patrol the room to ensure students were on task and to provide support.

Students were allowed to read a pleasure-reading book during this time instead, if they chose. This allowance was most obviously meant for anyone with extra time after catching up with their study, but I wasn't about to interfere with any teenager reading a book on their own volition. Not all regular readers (2-5 per class) were conscientious Cerego-ers.

Students were strongly encouraged to also use Cerego outside of class whenever the app recommended, if they wanted maximal retention for minimal time spent.

About once a week, usually without warning, I would give a ten question multiple choice quiz that could include questions directly taken from any content that had been in Cerego for at least a week, no matter how old. This was a multiple choice quiz done digitally in [Canvas](#). Before I put the grade into my book, I would add a 10% adjustment (not to exceed 100%), respecting the wisdom that aggressive study sees diminishing returns as one approaches a goal of 100% retention on large bodies of knowledge. My students were aware of this free 10% and my reasoning behind it.

To account for students just joining my class at the start of second semester, and for those who inevitably studied nothing for the seventeen calendar days between semesters — and even for those simply desperate for a fresh start — I had a lengthy grace period of sorts in January and February. Older stuff was temporarily not included in the "quizzable" question pool. I posted dates for when I would consider each old set fair game again; every week or two, a set would find itself back in the pool according to this schedule, and stay there for the rest of the year.

I did not use Cerego stats directly for any kind of grade, instead using my Canvas quizzes for this. My reasons:

- I wasn't sure every student would consistently be able to use the app, and didn't want to deal with the push-back from students and parents claiming (honestly or otherwise) insurmountable tech obstacles to using Cerego outside of class.
- Due to limitations in Cerego's reporting, I wasn't sure how to regularly compute a fair grade based on Cerego stats.
- I wasn't sure how far I would be able to trust that a student's stats weren't being run up by a smarter friend using the app on their behalf.
- I didn't want to discourage students from using Cerego Time to instead read their pleasure books (a habit of immense, scientifically-backed value that I do everything I can to promote).

- I didn't want to give the impression that Cerego is necessarily the best or only way to study, but instead to make it clear that knowing the content was *their* responsibility, however they chose to do it; my providing them with Cerego cards and time to study them was simply a function of my being a Really Nice Guy.

Points of friction

This section is not a critique of Cerego specifically, but rather a reminder that classroom technology is not inherently good. The mythical 1:1 student tech ratio doesn't suddenly make impossible dreams reality, and in fact comes with ongoing costs that must be weighed against the benefits. Here were some points of friction I encountered:

- Forgotten login information for the school PCs or Cerego.
- Slow startup, login, and load times on outdated equipment. [Fun fact: I've found that as my current school cuts down on the need for different logins through [Clever](#), they create a separate problem of longer and more fragile authentication chains — handshaking from one site to another — that can fail on slow machines or under spotty WiFi.]
- Old or abused keyboards and mice that intermittently fail.
- The occasional bigger problem, like a blown power supply.
- For phone users: discharged, confiscated, lost, or broken devices.
- Distractions and inappropriate behaviors that wouldn't be possible if students didn't have their own screen to command.

All of the above adds up to a kind of tax on your time and energy, even when you have enough respect from your students to minimize deliberate abuse. (I had maybe 2-3 bad eggs during the year committing occasional acts of minor sabotage.) Moreover, every possible point of friction becomes amplified by a student who doesn't feel getting to the objective, like a child who finds an hour's worth of [yak shaving](#) to do whenever bedtime rolls around.

Problems with multiple-choice study cards

Unlike Anki and other personal-use SRS, where the user self-assesses performance and collaborates with the app to schedule the next review, apps like Cerego are built to measure retention objectively. This changes how study cards have to be constructed. Although options [even in 2017] are varied, the most practical and straightforward method is usually a "front" side card with a question or term and a "back" side of multiple-choice responses.

Some problems with multiple-choice format:

- Responding to a multiple-choice question (or any kind of question) takes more time than pressing a self-assessment button.
- In general, it's more work to create study cards that can be assessed by the app. This is true even in the ideal case, which for Cerego is when you can assign a set of cards where the correct answer in one card can automatically become a multiple-choice distractor (wrong answer) for other cards in the set. But many cases are not ideal, and the only plausible distractors will be ones you add manually.
- Students can get confused when distractors contaminate tenuous mental associations. This is a well-studied effect with testing in general, and I had one student (motivated, but lower IQ) who I feel was positively ruined by it this year.
- Students mostly don't try to recall the answer before looking at multiple-choice options, instead defaulting to the following heuristic: "Look for an answer that feels right -- if none do, press 'None of the above'". This is a problem, because the act of trying to recall the specific thing is known to be the critical step that reinforces the memory; in contrast, merely recognizing familiar facts (as when "going over notes") is known to give students false confidence.

I gave my Cerego contacts some ideas I had for minimizing some of the downsides of multiple-choice. Because my students were largely deaf to my pleading that the “front” card screen — the one containing only the question — is where the learning actually happens, there could be a mandatory (or at least default, opt-out) short delay on that screen, especially when the app detects inhumanly rapid clicking.

Cerego actually asks “Do you know this?” on that screen, giving them a chance to self-assess in the negative without going to the multiple choices, but the vast majority of students never saw this screen as anything but a speed bump to click through.

My thought was that Cerego could occasionally not show the multiple choice options right away when they click “I Know It”, but instead call their bluff, asking, “Oh? How confident are you?” and prompting them to select a confidence level on a slider bar before showing the choices. Not only might this end the bad habit, it could also provide an opportunity to help them with their [credence calibration](#), a useful skill that might make them better thinkers and learners. I also suggested Cerego might be able to use this data to learn more about a learner and better judge their mastery level through sexy Bayesian wizardry.

[My aborted app design would have taken that concept to its logical conclusion: letting trusted users fully self-assess most of the time, but occasionally performing “reality checks” where it made the user respond in a way it could verify. It could then use straightforward Bayesian updates from these checks to decide how often to do them for each user.]

New failure modes

New format, new failure modes:

- **Performative clicking.** I would commonly have students who didn’t want the discomfort of getting called to task, but also didn’t want to actually do the task, so they would put up a show of productivity, continually clicking random answers over and over again without reading. Others would loiter in the stats screens, play with the cursor, check their grades... anything that wouldn’t require actual thinking.
- **Exploits.** Some students realized that mindless clicking moved Cerego’s progress bar on their study session forward. In some cases, it even raised their score. One enterprising young man demonstrated this for me, proudly resting a textbook over the Enter key, then kicking back as he “studied” his sets in record time. It was hard to be mad at him, as I could see myself doing the same at his age. Indeed, I was impressed. But he was in no way discouraged by my reminder that I didn’t use Cerego reports for grades, and that his trick wouldn’t leave him any better prepared for the quizzes that counted. (His mind was a steel trap, though; he did just fine.)
- **Hunkering.** Cerego is set up such that students don’t have new cards added to their rotations until they make an active choice to press a button that does this. Thus, many students would endlessly study only the first twenty cards from the start of the year, never pushing themselves with anything new. In their defense, one of my feedback notes to Cerego was that the UI [in 2017, remember] didn’t make it very clear that they had new material awaiting activation. But even after interventions where I walked them through the process, many of these fox-holed students would fail to activate newer cards on their own initiative.
- **Idleness and moping.** Apathy often manifests as lethargy combined with half-hearted complaints, voiced only when confronted, that it’s “too hard” or that “I don’t understand it”. Even though neither of those complaints made much sense when studying limited subsets of word-definition vocab pairs (the most common card set), I still heard both of them regularly from the hibernating bears I dared to poke. (Metaphorically. Never touch students.)

This was further evidence of something I already believed: that these complaints, in these contexts, are a means of disincentivizing teachers from bothering them, as

opposed to cries for help. After all, if such a student stands by their claim of not understanding it, what is a responsible teacher supposed to do except to stand there and reteach them the whole thing, or schedule one-on-one tutoring, holding their hand with every “I don’t get it” until the work is done for them? If the student had really wanted to understand and do the work, they would have raised their hand as soon as they encountered difficulty instead of trying to be inconspicuous.

[I’ve always been more sympathetic to apathetic students than I probably sound here. Public education demands more directed attention from teenagers than most of them can realistically muster for 35 hours a week.]

Dominating the blame game

Teachers are regularly asked by their bosses how they are “differentiating” instruction, adjusting lessons for students across a class’s range of skill levels, learning disabilities, and language deficiencies. They are also asked by parents what their children can do to improve their grade.

Cerego gave me a ready answer to both questions: “Well, in my class we use a free study app that I load with all of the terms, vocab and such that could be on my quizzes. It’s like smart flash cards that let you know when you need to study to avoid forgetting things. They adjust to give you more practice with the things you struggle with. Not only do I provide time to use it during class — even providing a computer if they need it — but it works on any internet device. Students can use it as often they like to be as prepared as they want to be.” Nobody ever complained about this answer, and some were quite impressed with it — more than I was, to be honest.

I also had powerful ammunition in the all-too-common scenario where, at a meeting with all of the child’s teachers, a parent blames poor grades on the teachers’ not adjusting to their child’s very special needs, instead of on their child’s ridiculously obvious laziness.

We can’t, of course, just come out and call it like we see it. But we *can* show parents our data and let them connect the dots. So, in these cases, I would just repeat my “Well, in my class we use a free study app...” spiel, emphasizing the “as prepared as they want to be” part. I would then add, “According to the app, your child has spent [x] minutes studying over the last week, which is about [y]% of the time my average ‘A’ student spent in that same period, and, come to think of it,” I would say, scratching my head for effect, “far less than the time I provide *in class* for it.”

Cue evil gaze from parent to child, squirming discomfort from child, envious awe from my fellow teachers.

It’s true! Here is a snapshot of one type of output I collected from my report-processing scripts for one of my students. You’re looking at one block of a larger data sheet I brought to parent meetings and included in periodic emails sent home. This one was for a fairly average student who put in the minimum expected time but didn’t push themselves very hard. A slacker’s would be more brutal.

Cerego Usage (study app)		
Not graded directly, but all quizzable material can be reviewed here independently, and, most days, during first 10 mins of class.		
Mastery:	49%	(as a percentage of the cumulative 'set level' attained by the median 'A' student currently [1.8/3.7])
Initiative:	62%	(as a percentage of the cards put in rotation vs. the median 'A' student currently [382/620])
Log time:	97%	(as a percentage of the total hours logged by median 'A' student so far [19.9/20.5]. In min/calender day that's [4.4/4.5].)
Last 7 days:	41.9 min	(with 13 new memories started. 4/21/2017-4/28/2017.)

Like I said, absolute dominance.

But like a lot of games, beating the “blame game” just made me tired of playing it, and ready to move on to something else. The enemy is not the apathetic student. The enemy is Apathy herself. I want to teach the lazy student, not destroy them with my Orwellian gaze.

Results and discussion

Table

In the following table, n=129, the sum of the 9th and 10th grade students that finished second semester with me. The procedures were identical in both grades, and I didn't find much reason to divide them, preferring the larger total sample. I then divided the combined sample into quintiles as shown:

Percentile (by Sem 2 Grade)	Avg. Sem 2 Grade	Avg. Hours Studied	Avg. Set Level Reached	Avg. % of Cards Started
80-100	95	23.4	3.7	93
60-79	86	21.7	2.1	58
40-59	80	21.9	2.1	57
20-39	72	20.3	1.3	36
0-19	57	18.7	0.9	27

The "Sem 2 Grade" is their course grade from just the second semester, but the other stats are all cumulative for the year. (No, I don't have any state test data for this group, and I never will. Having switched employers, I am not privy to the results, which arrive in late summer or early fall.)

"Set Level" is Cerego's signature rating of overall progress and retention, on a 4-point scale.

"% of Cards Started" is the fraction of the total cards I had prepared that the students had added into their rotations. (Remember that Cerego did not do this automatically). For 9th graders, there were 648 cards. For 10th grade, there were 749.

Study time analysis

As a sanity check, I crudely estimate that we had study time on 160 of our 180 school days, spending an average of 11 minutes each time. That would add up to 29.3 hours of total in-class study time. That the actual averages are lower does not surprise me, due to a combination of absences, roster changes, and start-up times. What we can conclusively say is that there was not a massive amount of outside-of-class study going on.

Of course, not all of those logged study minutes were productive study time. It wasn't always clear to me when Cerego counted a minute towards study vs. idle, or whether it detected idleness at all on the mobile app. Indeed, there were several cases where a student's mobile app seemed to have logged continual study overnight, and even, in one case, for multiple continuous days. The above chart has not been adjusted for known or unknown anomalies of this kind.

Regardless, as you can see, while time spent studying was correlated with performance, there was barely a 25% difference in study time separating the top and bottom grade quintiles. Even this is less exciting than it looks, as the lowest scorers were also more likely to be absent, missing their in-class study time. I have made no effort to adjust for this.

One thing you can't see in that chart is the high variance that existed within the top quintile. In this group, time spent studying varied from 33 hours to 12 — and 12 was the top student! Anecdotally, I perceived two distinct subgroups of high performers: highly motivated learners who had a natural disadvantage, like being a foreign exchange student speaking a second language, and high IQ avid reader types. The former put in far more hours than the latter. In fact, that second group put in less time than the average bottom quintile student.

Only a very small number of highly motivated students showed signs of studying over weekends and breaks.

SRS signal, or just conscientiousness?

While you can see a much stronger signal in the "Set Level" and "% of Cards Started" columns, it's hard to know how much this is just measuring conscientiousness. Good students are going to do what they're asked to do, and get the good grade no matter what, but this doesn't mean that what they're asked to do is always necessary to get the good grade — or that the grade reflects anything worthwhile in the first place.

People persons

At least a few of the students I could never get to study Cerego were very on-the-ball whenever we did any kind of verbal review.

[I've seen a lot of this pattern during the pandemic. Students who seemed like inert lumps online, with very low grades, have in many cases returned to the classroom and revealed themselves to be dynamic and invested. An engaging human at the front of the room really is the "value add" of in-person instruction. This is something I encourage my peers to keep in mind whenever deciding between autonomous work and teacher-student interaction.]

High automaticity in high achievers

When it came to automaticity, outlier results were more impressive than ever. The very small number of students at the overlap of highly motivated, highly intelligent, and highly competitive absolutely crushed it in the review game we regularly played at my interactive whiteboard, beating *me* on several occasions, which almost never happened previously.

Weak transference?

However, transference to other contexts was less evident. In my first report, I had remarked on [anecdotal impressions](#) of higher-quality discussion and essay responses from those who had embraced our Anki review, suggesting that they had truly enlarged their lexicon to be able to talk about more complex ideas. I saw less of that this year. I don't know what that means. It could just be that this mix of students was less open with their thoughts. But I can also see how they may have seen the Cerego universe as [distinct](#) from the universe of essay and discussion. Whole-class Anki might be more resistant to this bifurcation by making us say the words out loud to each other, normalizing their use.

Drama benchmark analysis

To compare methodologies as directly as possible, for a third year running I handled my Drama unit the way I accidentally had during my first year of classroom SRS: some terms taught before the pre-test, most taught after the pre-test, an identical post-test much later, and no review of any of it except through the SRS.

The overall results in the Drama unit were slightly worse this year. This was surprising. This cohort started lower on the pre-test, which was consistent with my impression of them, but I predicted that we would at least match or exceed last year's gains, as we had more room to improve. We did not. Retention of some reliable bellwether terms actually dropped prior to the post-test. In picking through individual scores, my impression was that whole-class Anki and independent in-class Cerego were statistically equivalent for motivated learners, but whole-class Anki won easily with less motivated learners. As always, there were plenty of truly unmotivated students who got nothing from either method.

I tried to tease this out even further. This was pretty unscientific, but I took the pre and post-test scores of twenty students from last year, and aligned them individually to students from this year with similar pre-test scores and, in my view, similar work ethics. Highly motivated students starting very low may have done slightly better with Cerego than with Anki, but poorly motivated students starting low did somewhat better with Anki.

I'm sure a lot of this came down to how Cerego makes new card sets "opt-in". Students of lower motivation were less likely to encounter the Drama terms in their study rotation at all!

Phone vs. Computer seemed to make a difference here, too. Stuck with a very visible PC, some low performers would occasionally have good days and get in a groove. The ones glued to their phones found anything to do except Cerego.

Conclusions (2017)

If I see students as being ultimately responsible for their own learning, independent Cerego is the fairer approach that will help students get what they "deserve". If I see things more pragmatically and utilitarian (as I do), the numbers favor the whole-class Anki approach. And yet...

If I were staying at that school, with my classroom computers, I would have tried to get the best of both worlds. It was my plan to use Cerego again — having already done most of the legwork — and try to make it friendlier, with more teacher interaction, supplementing with some whole-class Anki. I would have pushed Cerego's developers to make some of my most wanted changes, and I would have pushed myself to cut back on the number of cards I used.

But it's moot, now. I won't have computers at my new school. And part of the reason I left was because I didn't like the feel of the groove I was settling into.

Whole-class Anki review wins for simplicity and camaraderie. Cerego wins for surveillance and power. Which would you want to see stamping on a teenage face forever?

Trick question! It's not nice to stamp on faces. I feel like I've been pushing SRS too far past the point of diminishing returns, and I don't know why it has become an annual tradition for me to vow to cut back next year and then fail to do so. I should probably break that cycle. Apathy is the enemy, and she remains unbowed. I've been looking for a technological fix, but I think the solution is, at best, only partly technological.

[My notes here spiraled off into very technological solutions (sigh) to add to my dream SRS+ app, which I had already postponed again but still wasn't ready to abandon. I suppose I can give myself a little credit for brainstorming features to encourage human interaction and conceptual connections. Eventually, my notes came back to some thoughts about what makes a class thrive, which I have translated into coherent sentences below.]

From a scalability standpoint, it's nice that something like Cerego doesn't depend on a teacher's charm the way my whole-class Anki approach does. Teachers could do a lot worse than a standardized pack of quality Cerego sets that reinforce matching cookie-cutter lessons. But couldn't teachers also do better? I think I could do better. Cerego and Canvas quizzes create distance between me and my students. But I want to bring us closer and dial up the enthusiasm.

I don't think gamification is the answer. I've been noticing that the appeal of games is pretty niche, failing to capture many from the apathetic middle, and then for the wrong reasons, with the wrong incentives.

So what would work?

In education research, it always looks like *everything* works at least a little bit. This is probably a combination of publication bias and the fact that teachers sometimes get excited to try something new. Excitement is infectious. This gets students more engaged, which then improves outcomes. My early success with classroom SRS — and subsequent disappointments — would certainly fit that pattern.

Maybe I should make a point of trying new things each year for the explicit purpose of exploiting the excitement factor? How would I explain that to my bosses? "Well, I deliberately diverged from the curriculum and accepted best practice because I grew weary of them."

[Yes, actually. My new bosses are great that way.]

Thesis, Antithesis, Synthesis (2021)

As a student of storytelling, I can't help but find an arc to my fourteen years of teaching up to this point.

When I first started out, I didn't know what I was doing but kept Apathy at bay through sheer passion. I worked harder than anyone. I couldn't wait to try my stuff out, and students responded to all but my cringiest overtures.

When this inevitably exhausted me, I had a hard slump. Lessons that used to work fell flat. I still didn't know what I was doing, and now lacked sufficient passion to brute force success. So I retreated into systems and structure, building word banks, prompt banks, quiz banks; rubrics, charts, and randomizers; running reports; slinging code. A suit of high-tech power armor to augment my feeble form. A different kind of brute force.

My systems gave me stability and staying power, and, eventually, the confidence to explore. My three years of heavy SRS experimentation were the culmination of this phase. I

stretched. I grew. But I still felt plateaued and frustrated, perhaps having taken systems as far as they could go.

Apathy still mocked me from her emoji throne.

I step out of the armor and find I no longer need it. One by one, my systems clatter to the ground. I know who I am. I know where my power comes from. And I know my enemy.

She will lose, because she is overconfident. She won't prepare, because she is indifferent. And she won't hear my warning, because I issue it now in the one place I know she'll never reach: the bottom of a 10,000 word essay.

I'm coming for you.

Dark Matters

This post will be about the main points of evidence for the existence of dark matter. To evaluate whether a competing theory to dark matter is plausible, it's important to know what the actual arguments in favor of dark matter are in more detail than just "dark matter is the stuff you have to add to get galactic rotation curves to work out". A competitor has to address the *strongest* arguments in favor of the existence of dark matter, not just the weaker fare like galactic rotation curves.

So, when reading some hot new arxiv paper about dark matter or the lack thereof, it is fairly useful to know the top five lines of evidential support for dark matter (in my own personal estimation, others may differ). This lets you at least check whether the result is directly addressing the major cruxes that the case for dark matter rests upon, or just picking off one particular piece of evidence and sweeping the rest under the rug, even if you lack the full technical ability to evaluate the claimed result.

This post will be saving the best for last, so if you're not going to read the whole thing, skip down to sections 4 and 5.

Also, what exactly is meant when the term "dark matter" is used in this post? *Anything with mass (so it's affected by gravity and gravitationally influences other things) which does not interact via the electromagnetic force.* Electrons, protons, nuclei, and atoms emphatically do not count. Black holes, neutrinos, WIMPS (weakly interacting massive particles), and axions would count under this definition. The last two are theoretical, the first two are very much established. Of course, it would be a massive cop-out to go "neutrinos exist, therefore dark matter does", so "dark matter" will be used with a followup connotation of "and whatever the heck is (we don't know yet), there *must* be 5x more of it in the universe than matter made of atoms or atom parts, no way around that whatsoever"

Point 1: Galactic Rotation Curves

The story begins with galaxy rotation curves, which were the original motivation for postulating dark matter in the first place. Given a point gravitational mass, it's pretty simple to calculate the velocity of something orbiting around it, depending only on how far away the object is orbiting and how much mass is in the central point. Stuff orbiting further out from a point mass will be orbiting at a lower velocity.

With a bit more work, given a disc of mass, you can calculate the velocity of something orbiting around or within it. For this, the graph of orbital velocity vs distance from the center of the disc first rises, then falls. Orbital velocities are low in the center because stuff orbiting near the center of the disc isn't orbiting around very much mass, and orbital velocities are low at the outside of the disc, because you get closer to being able to approximate things by the situation "your distant object is orbiting around a central point mass", which, as previously discussed, already exhibits the "stars on further-out orbits move more slowly" behavior.

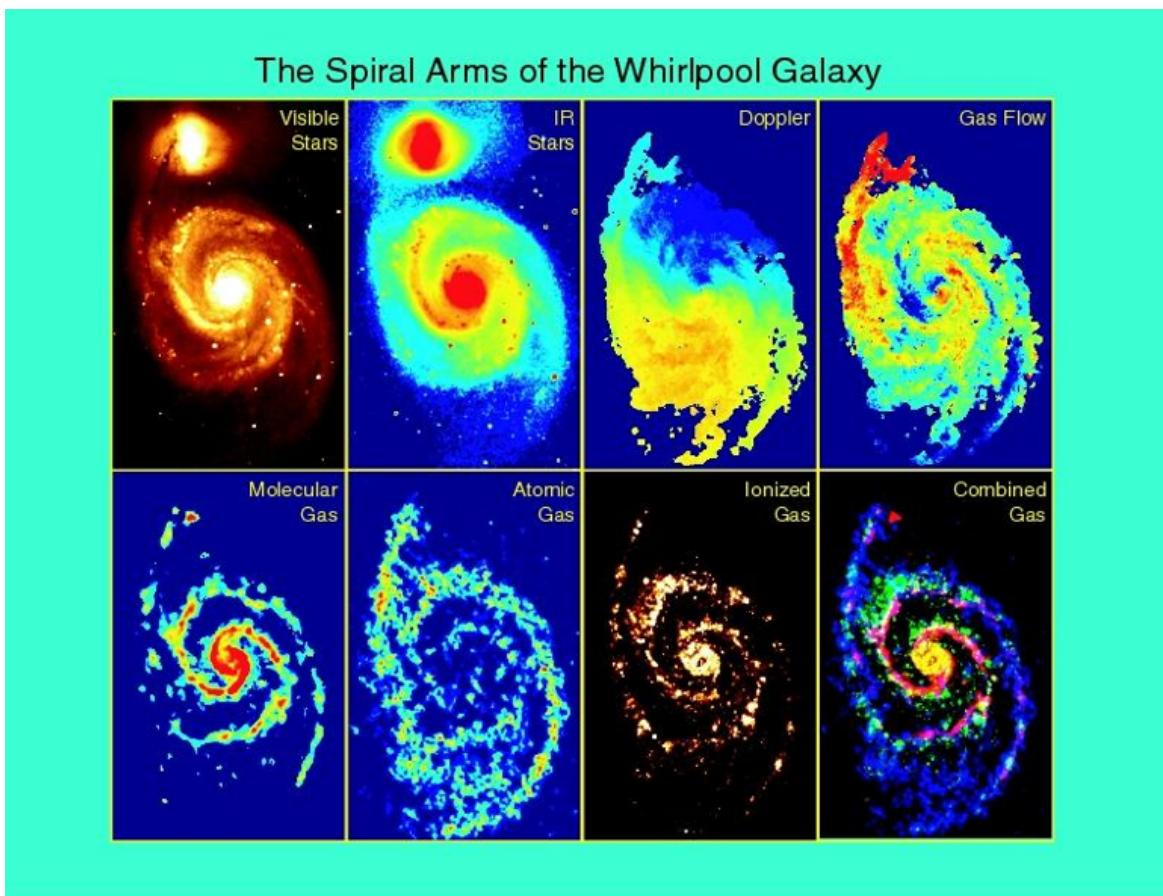
Computing this in practice requires knowledge of two things, however. First, you need to know how fast the stars in the galaxy are orbiting around the center. Second, you need to know the radial distribution of mass in the disc or ellipse.

It's pretty easy to tell how fast stars in a galaxy are orbiting around the center, for suitably chosen galaxies. Stars have emission and absorption lines at very specific frequencies measured to very high accuracy, which only depend on details of atomic physics that don't change in different galaxies. So, as an example, you could pick an edge-on spiral galaxy, and

look at the position of the absorption lines in the center of the galaxy. Then, you can look at the two edges of the galaxy, and in one edge, the emission/absorption line will be shifted to higher frequencies (the side of the galaxy that is rotating towards you), and in the other edge which is rotating away from you, the line will be shifted to lower frequencies, and with this you can accurately measure the rotational velocity of stars around the center of the galaxy.

But what about figuring out the radial distribution of mass? Well, for stars, by looking at a bunch of binary stars in the Milky Way, and stellar evolution models, we have an accurate idea how the color of a star corresponds to its mass and luminosity, so if you know how far away a galaxy is, and it has a spectrum corresponding to a bunch of orange K-type stars, you can use how bright the galaxy is, how far away it is, and the luminosity-to-mass relationship for K-type stars to figure out about how much stellar mass is there. It's more complicated than that because there's many different types of stars, but it can be done.

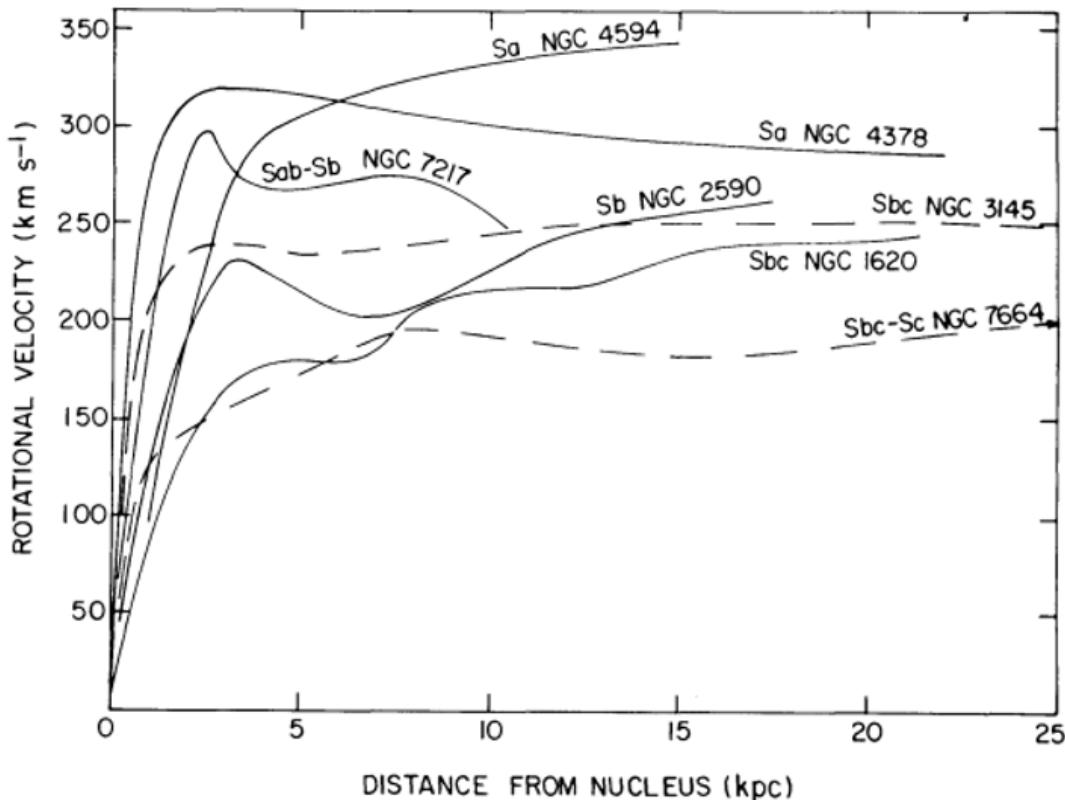
Of course, stars aren't all the mass we know of. In fact, they often make up a rather small portion of the visible mass in a galaxy. There are also clouds of hot ionized hydrogen, which emit high-frequency light, UV and up. And clouds of monoatomic hydrogen can be picked out in the radio spectrum from their characteristic 21-cm radio waves. Similarly, clouds of diatomic hydrogen, H₂, have their own characteristic spectral lines. Just check the luminosity of those vs the distance of the galaxy, and you can figure out about how much gas of the various types there is.



Of course, this might not account for *all* the mass in a galaxy, there might be some stuff that was missed. But we can try adding up all the mass that we see and where we see it, and

check that against the orbital velocity of the stars to see if we found everything or missed some stuff.

And, lo and behold, if you check the speed at which stars orbit in galaxies, it does *not* exhibit the expected behavior. Yes, orbital velocities are low in the center, but towards the outskirts, where you'd expect most of the mass to be more towards the middle of the galaxy so the stars should be slowing down in their orbits, they're either orbiting just as fast or even speeding up slightly. We can check even further out by looking at the 21 cm line of cold monoatomic hydrogen, and it's *still* rotating fast. Which would all be accounted for by the galaxy being embedded in a big cloud of mass.



On its own, this isn't terribly convincing. The estimates of the rotation curves rely on us managing to account for all the matter we saw. So maybe there are just big clouds of hydrogen that we haven't managed to find yet. Yes, there's a bunch of missing mass, but jumping from that to concluding that it isn't made of atoms is quite a stretch. Or maybe a full GR (General Relativity) treatment of the galaxy would account for things? Or maybe this is pointing towards a change in the laws of gravity itself on large scales? More on these options later.

For now, I'll observe something quite interesting. While most galaxies have rotation curves like this which indicate that there's about 6x more mass there than can be seen (*on average*), this number certainly isn't uniform across the universe. It's a good average for galaxy clusters, but the ratio experiences a lot more variation as the galaxy gets smaller, with the most extreme cases being found in dwarf galaxies. There are [some dwarf galaxies](#) where the visible matter pretty much accounts for the galactic rotation curve, and [some dwarf galaxies](#) where the rotation curves seem to indicate that there's >100x more matter there than was observed. So, whatever explanation you invoke to account for galactic

rotation curves, it should hopefully account for the existence of the occasional dwarf galaxy where what you see is what you get.

I count this as one of the weaker pieces of evidence in favor of dark matter, partially because it was the piece of evidence that lead to dark matter being postulated in the first place. It does no good to go "look at how dark matter perfectly accounts for galactic rotation curves" when this evidence was what lead to dark matter being promoted to a live hypothesis in the first place.

The other reason why it's one of the weaker pieces of evidence is that dark matter isn't a perfect fit, just a pretty good one. Pretty much every competitor theory to dark matter starts out with trying to replicate the galactic rotation curves, and so they also do a fairly good job at it.

Simulations of galactic evolution are quite hard, and the existence of an extended dark matter halo *does* account for the rotation curves quite well (as it was supposed to do), but there's two predictions from these simulations which reality doesn't quite seem to bear out.

The first is the cuspy halo problem. The simulations seem to indicate that there should be a considerable overabundance of dark matter in the core of a galaxy, where empirically it looks more like a uniform distribution given how the stars are moving in there. The best explanation I've seen is [this paper](#) which indicates it's a computational artifact of the algorithms used in the simulations to get a massive N-body physics problem down from $O(n^2)$ interactions per timestep to $O(n \log n)$ computations per timestep, though I lack the technical expertise necessary to fully evaluate said paper.

The second is the dwarf galaxy problem. The simulations seem to indicate that there should be about 50x more dwarf galaxies around a galaxy than actually seem to exist. From searching harder, we *have* found more ultra-dim dwarf galaxies around the Milky Way than were thought to exist at the time, and these new ones seem to have very high ratios of dark matter in them by how the stars are moving. So, one of the proposed explanations is that maybe there *are* in fact a whole bunch of dwarf-galaxy-sized lumps of dark matter around our galaxy that have so little gas and stars associated with them that we don't know they're there.

A second possible explanation to the dwarf galaxy problem is warm dark matter. Roughly, if we assume dark matter is made up of a bunch of particles, we've got three possibilities. The first is that they're going near light-speed, as neutrinos do. This is hot dark matter. Or they could be traveling at fairly low velocities, like <10 percent of lightspeed. This is cold dark matter. Or they could be something in the middle, ie, warm dark matter. If all your particles are going super-fast and only interact via gravitation, they would form an enormous spread-out fluffy mass, while if all your particles aren't going fast at all, gravitation can draw them into much smaller, clumpier, and denser structures.

Hot dark matter is ruled out (well, technically, neutrinos are a form of hot dark matter, but they don't account for anywhere near the 5x mass excess), as the size of its clumps would be considerably larger than a galaxy, and simulations of galaxy formation with the dark matter being hot don't produce anything like what we see today. So that rules out neutrinos as a possibility, because if they made up most of the dark matter, you wouldn't get galaxies looking like they currently do.

Cold dark matter, being much more clumpy, produces that profusion of dwarf galaxies which may or may not exist. So, if the dark matter is warm, it would smooth things out enough to still correctly account for what occurs at the scale of a big galaxy, but not a tiny little dwarf galaxy, as the stuff is moving too fast to form structures of that small size. From talking to someone in the field, apparently cold dark matter is favored, though I'm not entirely sure why. It has a lot more particle candidates from theoretical physics than warm dark matter does, but there's probably some other reason I missed.

So, for dark matter vs competitors, galactic rotation curves don't settle it, as all the competitors for dark matter try to explain those rotation curves as well. There's also some unsolved issues where dark matter accounts for galactic rotation curves, but the simulations don't quite match up with what we see, what with their cuspy halos and profusion of dwarf galaxies. But then we get into meatier fare.

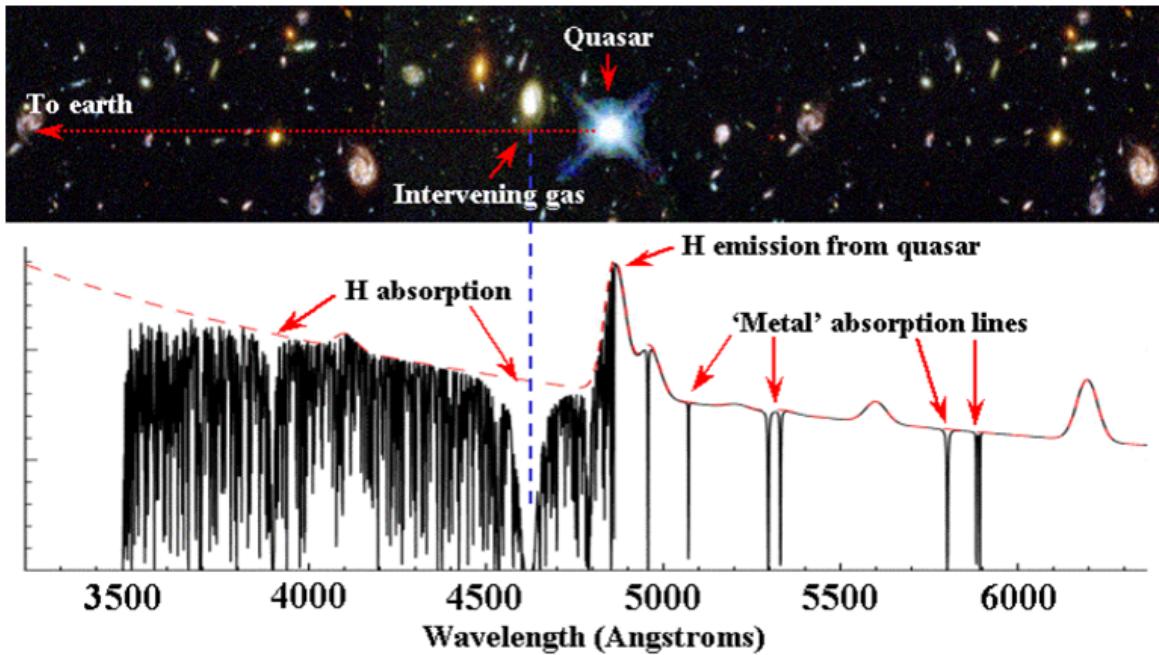
Point 2: Large-scale Universe Structure

We can run simulations of what the large-scale structure of the universe would look like, starting from the conditions a few thousand years after the Big Bang. Just throw in different amounts of regular matter (hydrogen-helium gas mix), light, dark energy, and matter which doesn't interact with electromagnetism, introduce some overdensities and underdensities of the magnitude and spacing we see in the Cosmic Microwave Background, and run it with General Relativity (this is more tractable than giving galactic formation the full GR treatment). Then just compare against the observed large-scale structure of the universe, to rule out combinations of these components. We should get a match with the pattern of voids and filaments and their densities and spacing we observe in the universe.

Well, first, how do we know the large-scale structure of the universe? Lots of large-scale automated deep-sky surveys, with redshift as a fairly accurate distance proxy, calibrated against type Ia supernova.

There's also another neat trick with hydrogen gas. If you have some super-bright source of high-energy radiation, like a quasar (the brightest things in the universe), and the light passes through a cloud of neutral hydrogen gas, the hydrogen gas will absorb a particular frequency of UV light. Also, the expansion of the universe doesn't just redshift light because the galaxy is traveling away from us, it also redshifts light because the light literally stretches out due to the expansion of space over the course of its journey across the universe.

So, if you've got a quasar, and the light passes through a cloud of hydrogen gas, that particular UV frequency of light will get absorbed, making a spectral line. Then, if the light keeps going and redshifting and runs into another cloud of hydrogen gas in its travels, the old UV spectral line will have been redshifted down, and light which used to be above that particular spectral line will get redshifted to line up with that frequency of light. So now, looking at the quasar spectrum, there are two absorption lines, the lower one for when the light first went through a cloud of hydrogen, and the higher one for when the light more recently went through a hydrogen cloud. Although, usually, there's a lot more than two lines. By looking at how deep the absorption lines are in the spectrum, and where they are, you can build up a picture of how much cold hydrogen gas is at what distances along the line-of-sight between us and the quasar.



So, how do the simulations do at reproducing what we see in the universe around us? Well, first-up, this is an opportunity to falsify the usual picture of dark matter. If a 5-to-1 ratio of dark matter to hydrogen/helium gas *doesn't* replicate the large-scale structure of the universe, the theory dies right here.

Since I'm citing this as a piece of evidence, it obviously must have passed. In fact, this was known as a problem early on, that for the standard accounting of known mass in the universe, it wasn't enough to permit large clusters and walls of galaxies to form. The ordinary matter would be too spread out and you'd just get little clumps of galaxies instead of the larger superclusters and walls. That was another reason dark matter was an attractive hypothesis, because if there's a lot more mass in the universe, you could get enough mass for large-scale structure formation. The dark matter would gravitate into clumps, which would provide a big gravitational well for the ordinary matter to accrete at the bottom of and form galaxies. At the time, this wasn't super-detailed, the more detailed computer simulations and large-scale surveys of the cosmos (which, again, could have falsified dark matter being in 5x abundance) came afterwards.

This is *not* accounted for by just gravitoelectromagnetic effects. Modified gravity theories also have a hard time explaining this one. It still leaves open the possibility that maybe there's just a whole boatload of extra hydrogen and helium out there that we just haven't spotted yet for whatever reason.

So that's the second line of evidence, large-scale simulations of the universe most accurately replicate its structure with about 6x more matter than is currently accounted for.

Point 3: Galactic Cluster Lensing

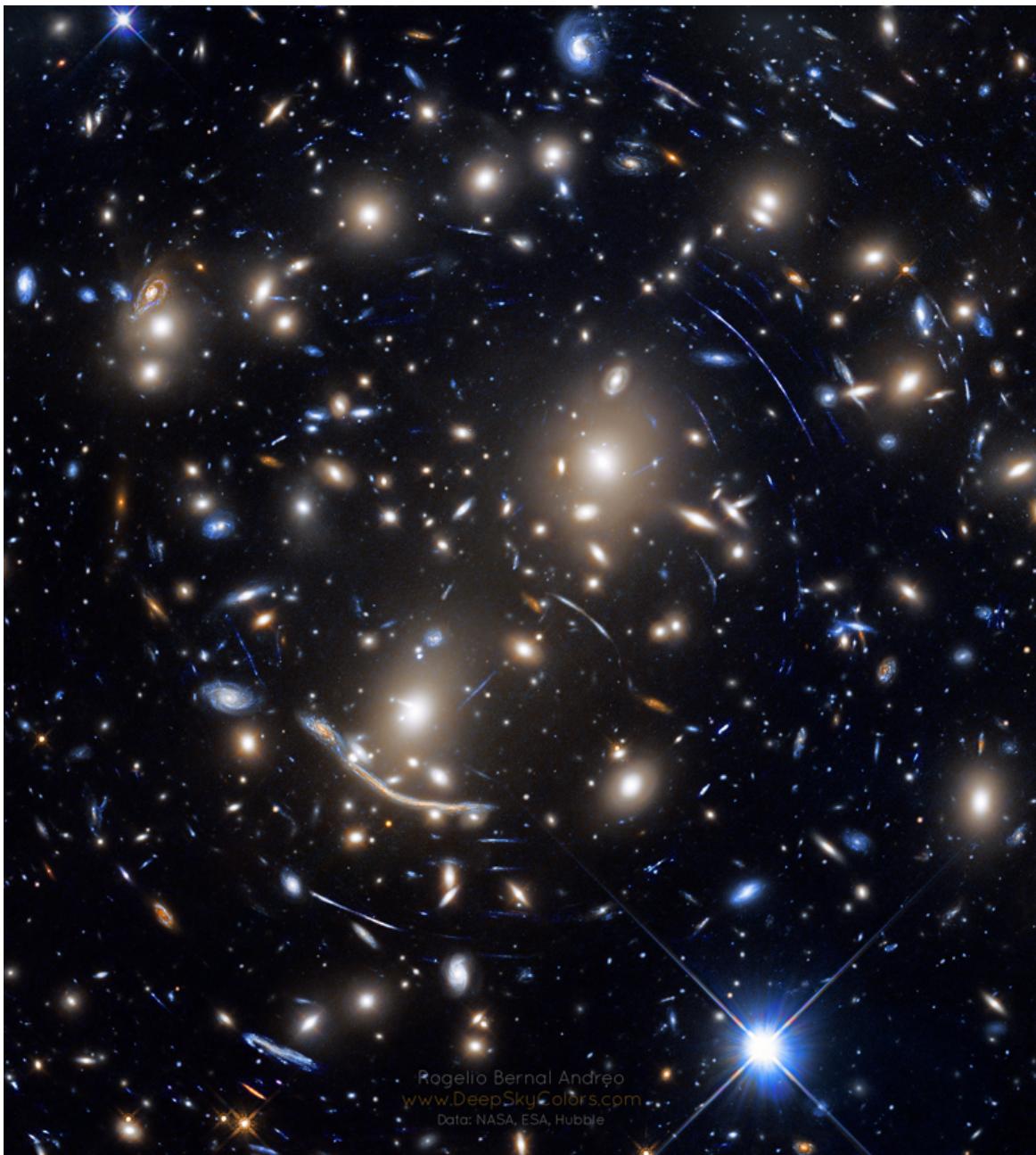
And now we get into the third thing, to further drive the nail into gravity-based competitors.

One of Einstein's famous original predictions was of the magnitude of displacement of the image of a star from its usual position during a solar eclipse, due to gravitational lensing from the sun. GR passed with flying colors. So, another conceivable test that you can carry out for dark matter (once you gain a Hubble Telescope to play around with) is to find a bunch of galaxy clusters which are big enough to have gravitational lensing effects, and see how

much of a dent they make in the images of the background galaxies. If it indicates that there's 6x more matter than you see, then that's another test passed.

Much of the visible mass in a galactic cluster is in the form of hot plasma which emits X-rays and UV light, from the galaxies moving around and running into each other at high speed which heats up their surrounding gas. This can be weighed by checking the plasma luminosity, and using the spectra to infer temperature of the gas and how brightly it *should* be glowing, and then if you know far away the galactic cluster is, you can solve for the amount of plasma present.

A bunch of studies on this are carried out using weak lensing, which looks at subtle distortions of the background galaxies, but there are galactic clusters with far more spectacular instances of gravitational lensing, as pictured below.



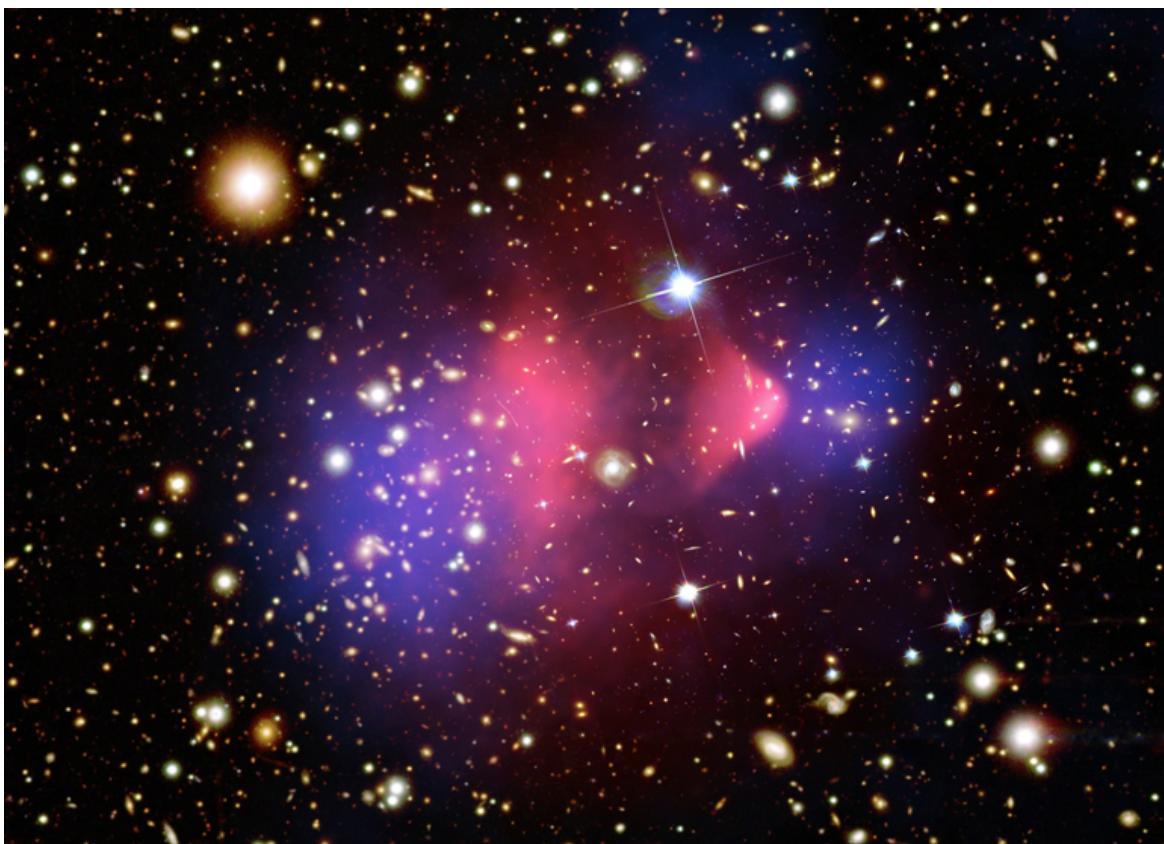
And, what do you know, the gravitational lensing of galaxy clusters all indicate a mass about 6x more than the hot gas and stars and every other detectable source of mass would indicate. Again, this is another thing that modified gravity theories really struggle to deal with. They ace the galactic rotation curves, but don't do so well with galaxy cluster dynamics.

Point 4: The Bullet Cluster

And while we're on the topic of galaxy clusters, there's also the Bullet Cluster. Which wasn't exactly an advance prediction, but it's another thing that's really hard for competitors to explain, while the standard picture of dark matter makes perfect sense of it.

Pretty much, the Bullet cluster is a pair of galaxy clusters that rammed into each other. Stars are very widely spaced. If the Milky Way and Andromeda collided, there would only be about one stellar collision in the whole interaction, IIRC. So the stars should mostly just go through each other. If dark matter mostly doesn't interact with itself, the dark matter blobs should just go through each other. The same cannot be said for the hot ionized plasma which makes up much of the mass in a galactic cluster, however. The particles in it are charged, and would, with high probability, get slowed down and heated up from ramming into a galactic-cluster-sized blob of hot plasma heading the other way.

And so, we can view the cluster in X-rays to see where most of the visible mass, including stars, is. And we can check the subtle gravitational lensing of background galaxies to see where the actual mass is. And now... check out the following image. Pink is the X-ray emitting gas, purple is where the mass is according to background gravitational lensing.



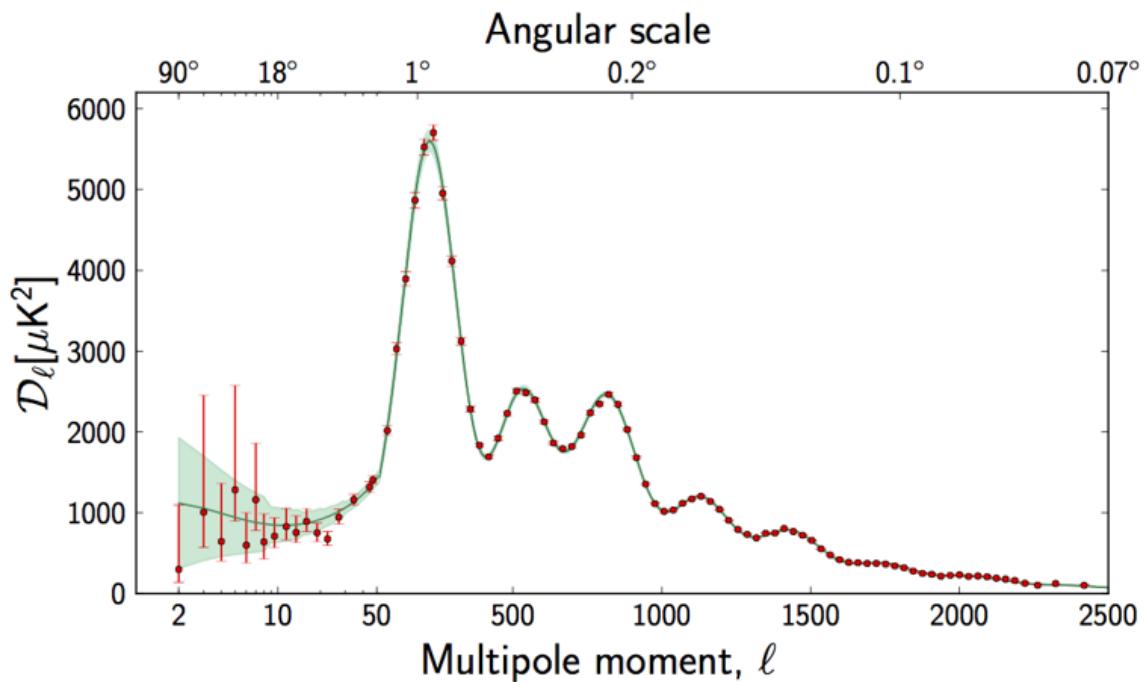
Explain *this* with modified gravity.

This also weighs against the "maybe it's just a bunch of ordinary hydrogen and helium in some gas form we can't see" hypothesis, because it's rather hard to shoot two galaxy-cluster-sized blobs of gas of each other and have them pass *straight through* each other.

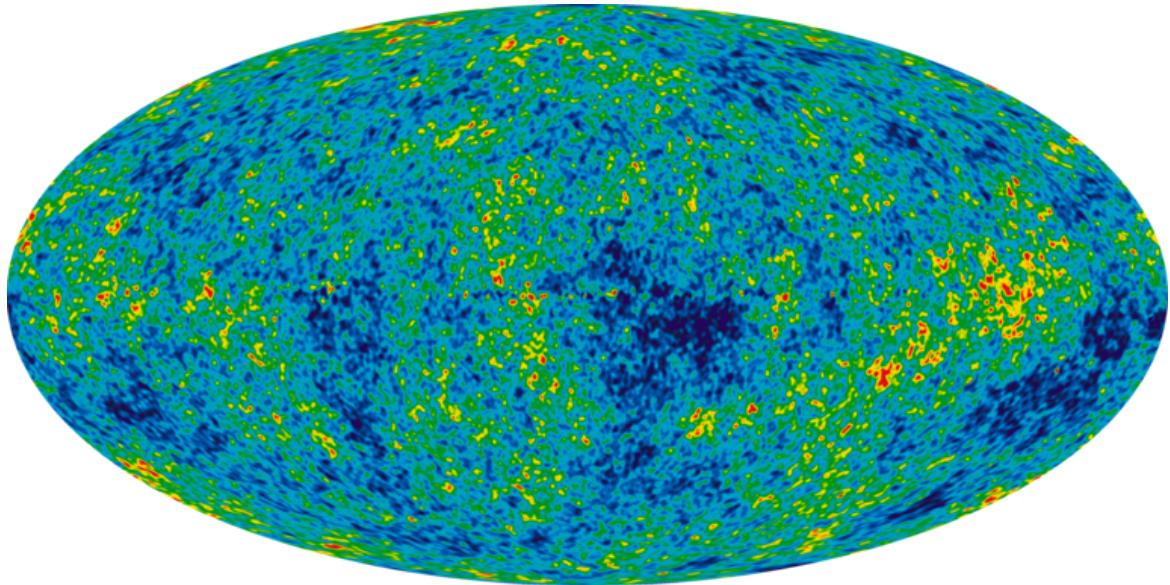
Point 5: The CMB Power Spectrum

And now we come to my personal favorite piece of evidence. The Bullet cluster is usually taken as the most spectacular line of evidence for dark matter, but the CMB power spectrum is what lets us conclusively rule out that dark matter is anything born of atoms.

This was very much an advance prediction. The detailed measurements of the pattern of ripples in the cosmic microwave background (CMB) radiation, in enough detail to get the following graph, were not around at the time dark matter was postulated. It came in around 2000 and later, mostly from the WMAP probe, and was then refined by the Planck probe.



Roughly, this graph is telling you the magnitude of the fluctuations on different scales of the cosmic microwave background. On the far left of this graph, it's plotting the amplitude of CMB variations over larger regions of space, where the cosmic microwave background looks pretty uniform. On the far right of this graph, it's plotting the amplitude of CMB variations on the smallest scales. It says that most of the fluctuations in the CMB are accounted for on the scale of about 1 degree. This spectrum encodes information about what was going in the early universe, the characteristic length scale on which the universe had ripples during its formation.



The positioning of the first peak tells you about the curvature of the universe. It's the dominant length scale on which the universe has clumps or voids. This can be measured now, from large-scale universe structure. If the first peak was shifted to the left or right, it would correspond to the major fluctuation scale of the universe looking bigger/smaller at early times than it does now, which is characteristic of large-scale universe curvature. The universe, as near as we can tell, looks flat (ie, like \mathbb{R}^3 , euclidean 3-space) on large scales.

The curvature of the universe depends on what it has in it. A universe with a whole bunch of matter/energy in it would be positively curved. The energy in empty space which would account for the cosmological constant (whatever the hell it comes from), accounts for 69% (plus or minus 0.6%) of the total mass-energy budget you need for a flat universe. And, you guessed it, for the remaining known mass, you don't get anywhere near the 31% necessary.

But this isn't the only way that you can check up on dark matter with the CMB power spectrum! In the early universe, everything was hot and dense and there was a lot of radiation/light around. So, matter which interacts with radiation would heat up and expand back again if it started compressing, while matter which doesn't interact with light but *does* interact with gravity wouldn't have that effect. Having different amounts of ordinary atomic matter vs dark matter early on in the universe produces different characteristic patterns in the spectrum, with ordinary atomic matter tending to enhance the even-numbered peaks, and dark matter tending to enhance the odd-numbered ones.

And, lo and behold, the CMB power spectrum (particularly the second and third peaks) perfectly fits with ~5% of the universe's mass-energy being accounted for by matter, and ~26% of the universe's mass-energy being accounted for by dark matter, summing up to 31%.

This is how we know dark matter can't be anything made of atoms. Early on in the universe, it was more of a very hot homogenous soup, so everything made of atoms in the current day was in the form of a hot dense soup of gas way back when. So this is directly measuring the proportion of "matter made of atoms" to "??? stuff that has gravity but doesn't interact with light". And it fits!

From an advance-prediction standpoint, we get something else interesting. The CMB ripples seemed to indicate that there was *more* baryonic (hydrogen and helium) matter in the universe than had previously been accounted for around galaxies, which was promptly

dubbed the "Missing Baryon problem". It was a bit tricky to account for, because much of it was in the form of incredibly wispy warm hydrogen in great streamers between the galactic clusters and in the voids, but by 2020, the last missing 30% of the CMB-indicated hydrogen and helium was accounted for and the missing baryon problem was solved.

Conclusion:

So, the five main lines of support for dark matter are: galactic rotation curves, accurately replicating the large-scale structure of the universe in simulations, the extra mass still being present when you look at galaxy cluster lensing, the bullet cluster having the mass and the visible gas in different locations, and finally, perfectly accounting for the pattern of ripples in the cosmic microwave background in two different ways.

So, this should hopefully provide some background on why dark matter is widely accepted to be a Thing, and furnish some useful heuristics on reading a paper on competitors to dark matter for explaining things. If they only talk about galactic rotation curves and don't try to explain the CMB spectrum and Bullet Cluster, you can safely throw the paper in the trash, as these are the two major hurdles that the dark matter competitors fail to clear. Conversely, if there's a paper proposing an alternative to dark matter that manages to explain the CMB power spectrum and the Bullet Cluster, it's now safe to take it more seriously. It's the astrophysics equivalent of Scott Aaronson's heuristics on whether a P vs NP proof is worth taking seriously.

But what even is dark matter, anyways? Well... there are a lot of possibilities, and it would be premature to conclude that it's any particular one of them, or even that we won't get hit with something out of left field. The basic constraints we know are: It's affected by gravity and has gravity. It doesn't interact through electromagnetism, otherwise it wouldn't be consistent with the CMB spectra. If it does interact through the weak or strong force, the interaction will be really weak, because we've built a bunch of big particle detectors to check for dark matter with appreciable weak or strong force interactions and haven't come up with shit. Dark matter shouldn't be hot/travel at a sizeable chunk of the speed of light, because otherwise the large-scale structure of the universe would be too uniform and galaxies wouldn't have the structure they do, which rules out neutrinos. And you should be able to fire two big lumps of it at each other and have them pass straight through each other, because the Bullet Cluster is a thing.

Hopefully this post was informative.

Strong Evidence is Common

This is a linkpost for <https://markxu.com/strong-evidence>

Portions of this are taken directly from [Three Things I've Learned About Bayes' Rule](#).

One time, someone asked me what my name was. I said, “Mark Xu.” Afterward, they probably believed my name was “Mark Xu.” I’m guessing they would have happily accepted a bet at 20:1 odds that my driver’s license would say “Mark Xu” on it.

The prior odds that someone’s name is “Mark Xu” are generously 1:1,000,000. Posterior odds of 20:1 implies that the odds ratio of me saying “Mark Xu” is 20,000,000:1, or roughly 24 bits of evidence. That’s **a lot** of evidence.

Seeing a Wikipedia page say “X is the capital of Y” is tremendous evidence that X is the capital of Y. Someone telling you “I can juggle” is massive evidence that they can juggle. Putting an expression into Mathematica and getting Z is enormous evidence that the expression evaluates to Z. Vast odds ratios lurk behind many encounters.

One implication of the Efficient Market Hypothesis (EMH) is that it is difficult to make money on the stock market. Generously, maybe only the top 1% of traders will be profitable. How difficult is it to get into the top 1% of traders? To be 50% sure you’re in the top 1%, you only need 200:1 evidence. This seemingly large odds ratio might be easy to get.

On average, people are overconfident, but [12% aren't](#). It only takes 50:1 evidence to conclude you are much less overconfident than average. An hour or so of [calibration training](#) and the resulting calibration plots might be enough.

Running through Bayes’ Rule explicitly might produce a bias towards middling values. Extraordinary claims require extraordinary evidence, but extraordinary evidence might be more common than you think.

Fun with +12 OOMs of Compute

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

Or: Big Timelines Crux Operationalized

What fun things could one build with +12 orders of magnitude of compute? By ‘fun’ I mean ‘powerful.’ This hypothetical is highly relevant to AI timelines, for reasons I’ll explain later.

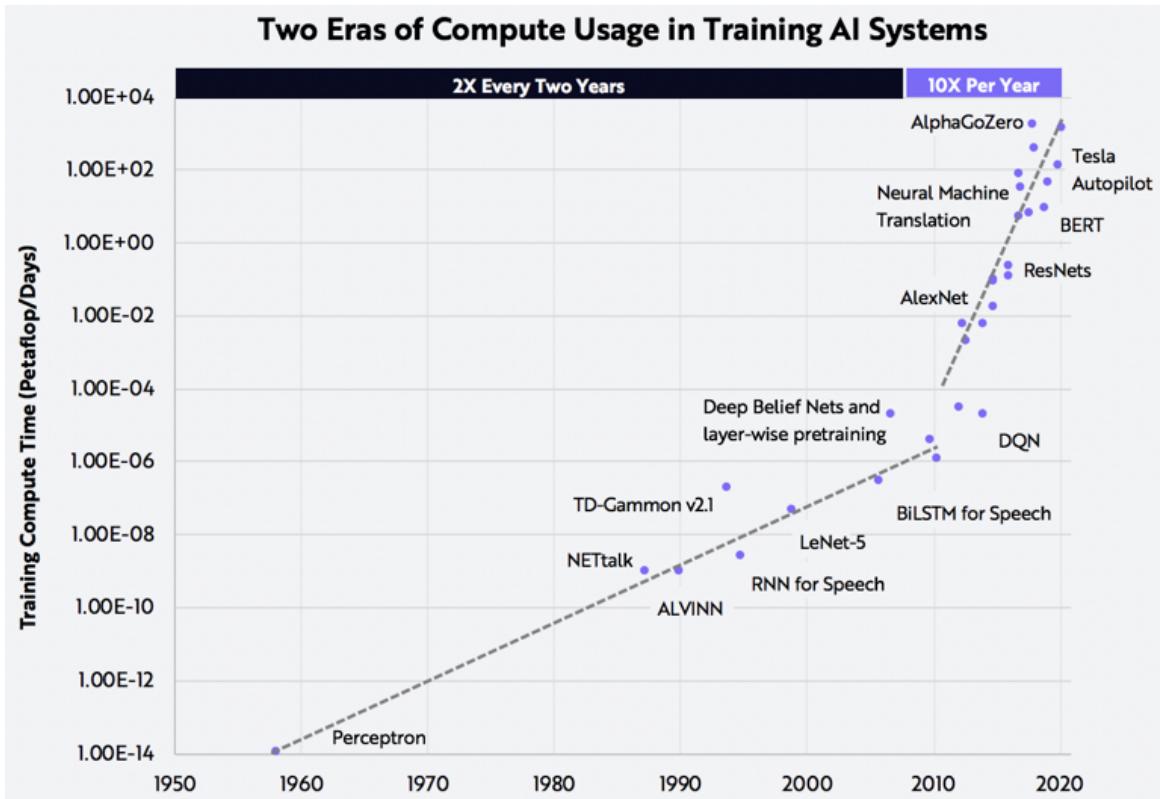
Summary (Spoilers):

I describe a hypothetical scenario that concretizes the question “*what could be built with 2020’s algorithms/ideas/etc. but a trillion times more compute?*” Then I give some answers to that question. Then I ask: How likely is it that some sort of TAI would happen in this scenario? This second question is a useful operationalization of the (IMO) most important, most-commonly-discussed timelines [crux](#): “Can we get TAI just by throwing more compute at the problem?” I consider this operationalization to be the main contribution of this post; it directly plugs into Ajeya’s timelines model and is quantitatively more cruxy than anything else I know of. The secondary contribution of this post is my set of answers to the first question: They serve as intuition pumps for my answer to the second, which strongly supports my views on timelines.

The hypothetical

In 2016 the Compute Fairy visits Earth and bestows a blessing: Computers are magically 12 orders of magnitude faster! Over the next five years, what happens? The Deep Learning AI Boom still happens, only much crazier: Instead of making AlphaStar for 10^{23} floating point operations, DeepMind makes something for 10^{35} . Instead of making GPT-3 for 10^{23} FLOPs, OpenAI makes something for 10^{35} . Instead of industry and academia making a cornucopia of things for 10^{20} FLOPs or so, they make a cornucopia of things for 10^{32} FLOPs or so. When random grad students and hackers spin up neural nets on their laptops, they have a trillion times more compute to work with. [EDIT: Also assume magic +12 OOMs of memory, bandwidth, etc. All the ingredients of compute.]

For context on how big a deal +12 OOMs is, consider the graph below, from [ARK](#). It’s measuring petaflop-days, which are about 10^{20} FLOP each. So 10^{35} FLOP is $1e+15$ on this graph. GPT-3 and AlphaStar are not on this graph, but if they were they would be in the very top-right corner.



Question One: In this hypothetical, what sorts of things could AI projects build?

I encourage you to stop reading, set a five-minute timer, and think about fun things that could be built in this scenario. I'd love it if you wrote up your answers in the comments!

My tentative answers:

Below are my answers, listed in rough order of how ‘fun’ they seem to me. I’m not an AI scientist so I expect my answers to overestimate what could be done in some ways, and underestimate in other ways. Imagine that each entry is the best version of itself, since it is built by experts (who have experience with smaller-scale versions) rather than by me.

OmegaStar:

In our timeline, it cost about 10^{23} FLOP to train [AlphaStar](#). ([OpenAI Five](#), which is in some ways more impressive, took less!) Let’s make OmegaStar like AlphaStar only +7 OOMs bigger: the size of a [human brain](#).^[1] [EDIT: You may be surprised to learn, as I was, that AlphaStar has about 10% as many parameters as a honeybee has synapses! Playing against it is like playing against a tiny game-playing insect.]

[Larger models seem to take less data to reach the same level of performance](#), so it would probably take at most 10^{30} FLOP to reach the same level of Starcraft performance as AlphaStar, and indeed we should expect it to be qualitatively better.^[2] So let’s do that, but also train it on lots of other games too.^[3] There are [30,000 games in the Steam Library](#). We

train OmegaStar long enough that it has as much time on each game as AlphaStar had on Starcraft. With a brain so big, maybe it'll start to do some transfer learning, acquiring generalizable skills that work across many of the games instead of learning a separate policy for each game.

OK, that uses up 10^{34} FLOP—a mere 10% of our budget. With the remainder, let's add some more stuff to its training regime. For example, maybe we also make it read the entire internet and play the “Predict the next word you are about to read!” game. Also the “Predict the covered-up word” and “predict the covered-up piece of an image” and “predict later bits of the video” games.

OK, that probably still wouldn't be enough to use up our compute budget. A Transformer that was the size of the human brain would only need 10^{30} FLOP to get to human level at the predict-the-next-word game [according to Gwern](#), and while OmegaStar isn't a transformer, we have 10^{34} FLOP available.[\[4\]](#) (What a curious coincidence, that human-level performance is reached right when the AI is human-brain-sized! [Not according to Shorty.](#))

Let's also hook up OmegaStar to an online chatbot interface, so that billions of people can talk to it and play games with it. We can have it play the game “Maximize user engagement!”

...we probably still haven't used up our whole budget, but I'm out of ideas for now.

Amp(GPT-7):

Let's start by training GPT-7, a transformer with 10^{17} parameters and 10^{17} data points, on the entire world's library of video, audio, and text. This is almost 6 OOMs more params and almost 6 OOMs more training time than GPT-3. Note that a mere +4 OOMs of params and training time is predicted to reach near-optimal performance at text prediction and [all the tasks](#) thrown at GPT-3 in the [original paper](#); so this GPT-7 would be superhuman at all those things, and also at the analogous video and audio and mixed-modality tasks.[\[5\]](#) Quantitatively, the gap between GPT-7 and GPT-3 is about *twice as large* as the gap between GPT-3 and GPT-1, (about 25% the loss GPT-3 had, which was about 50% the loss GPT-1 had) so try to imagine a qualitative improvement twice as big also. And that's not to mention the possible benefits of multimodal data representations.[\[6\]](#)

We aren't finished! This only uses up 10^{34} of our compute. Next, we let the public use [prompt programming](#) to make a giant library of GPT-7 functions, like the stuff demoed [here](#) and like the stuff being built [here](#), only much better because it's GPT-7 instead of GPT-3. Some examples:

- Decompose a vague question into concrete subquestions
- Generate a plan to achieve a goal given a context
- Given a list of options, pick the one that seems most plausible / likely to work / likely to be the sort of thing Jesus would say / [insert your own evaluation criteria here]
- Given some text, give a score from 0 to 10 for how accurate / offensive / likely-to-be-written-by-a-dissident / [insert your own evaluation criteria here] the text is.

And of course the library also contains functions like “google search” and “Given webpage, click on X” (remember, GPT-7 is multimodal, it can input and output video, parsing webpages is easy). It also has functions like “Spin off a new version of GPT-7 and fine-tune it on the following data.” Then we fine-tune GPT-7 on the library so that it knows how to use those functions, and even write new ones. (Even GPT-3 [can do basic programming](#), remember. GPT-7 is much better.)

We still aren't finished! Next, we embed GPT-7 in an amplification scheme — a “[chinese-room bureaucracy](#)” of calls to GPT-7. The basic idea is to have functions that break down tasks into sub-tasks, functions that do those sub-tasks, and functions that combine the results of the sub-tasks into a result for the task. For example, a fact-checking function might start by dividing up the text into paragraphs, and then extract factual claims from each paragraph, and then generate google queries designed to fact-check each claim, and then compare the search results with the claim to see whether it is contradicted or confirmed, etc. And an article-writing function might call the fact-checking function as one of the intermediary steps. By combining more and more functions into larger and larger bureaucracies, more and more sophisticated behaviors can be achieved. And by fine-tuning GPT-7 on examples of this sort of thing, we can get it to understand how it works, so that we can write GPT-7 functions in which GPT-7 chooses which other functions to call. Heck, we could even have GPT-7 try writing its own functions! [\[7\]](#)

The ultimate chinese-room bureaucracy would be an agent in its own right, running a continual [OODA loop](#) of taking in new data, distilling it into notes-to-future-self and new-data-to-fine-tune-on, making plans and sub-plans, and executing them. Perhaps it has a text file describing its goal/values that it passes along as a note-to-self — a “bureaucracy mission statement.”

Are we done yet? No! Since it “only” has 10^{17} parameters, and uses about [six FLOP per parameter per token](#), we have almost 18 orders of magnitude of compute left to work with. [\[8\]](#) So let’s give our GPT-7 uber-bureaucracy an internet connection and run it for 100,000,000 function-calls (if we think of each call as a subjective second, that’s about 3 subjective years). Actually, let’s generate 50,000 different uber-bureaucracies and run them all for that long. And then let’s evaluate their performance and reproduce the ones that did best, and repeat. We could do 50,000 generations of this sort of artificial evolution, for a total of about 10^{35} FLOP.[\[9\]](#)

Note that we could do all this amplification-and-evolution stuff with OmegaStar in place of GPT-7.

Crystal Nights:

(The name comes from an [excellent short story](#).)

Maybe we think we are missing something fundamental, some unknown unknown, some [special sauce](#) that is necessary for true intelligence that humans have and our current artificial neural net designs won’t have even if scaled up +12 OOMs. OK, so let’s search for it. We set out to recapitulate evolution.

We make a planet-sized virtual world with detailed and realistic physics and graphics. OK, not *perfectly* realistic, but much better than any video game currently on the market! Then, we seed it with a bunch of primitive life-forms, with a massive variety of initial mental and physical architectures. Perhaps they have a sort of virtual genome, a library of code used to construct their bodies and minds, with modular pieces that get exchanged via sexual reproduction (for those who are into that sort of thing). Then we let it run, for a billion in-game years if necessary!

Alas, [Ajeya estimates](#) it would take about 10^{41} FLOP to do this, whereas we only have 10^{35} .[\[10\]](#) So we probably need to be a million times more compute-efficient than evolution. But maybe that’s doable. Evolution is pretty dumb, after all.

1. Instead of starting from scratch, we can start off with “advanced” creatures, e.g. sexually-reproducing large-brained land creatures. It’s unclear how much this would save but plausibly could be at least one or two orders of magnitude, since Ajeya’s

estimate assumes the average creature has a brain about the size of a nematode worm's brain.[\[11\]](#)

2. We can grant "magic traits" to the species that encourage intelligence and culture; for example, perhaps they can respawn a number of times after dying, or transfer bits of their trained-neural-net brains to their offspring. At the very least, we should make it metabolically cheap to have big brains; no birth-canal or skull should restrict the number of neurons a species can have! Also maybe it should be easy for species to have neurons that don't get cancer or break randomly.
3. We can force things that are bad for the individual but good for the species, e.g. identify that the antler size arms race is silly and nip it in the bud before it gets going. In general, more experimentation/higher mutation rate is probably better for the species than for the individual, and so we could speed up evolution by increasing the mutation rate. We can also identify when a species is trapped in a local optima and take action to get the ball rolling again, whereas evolution would just wait until some climactic event or something shakes things up.
4. We can optimise for intelligence instead of ability to reproduce, by crafting environments in which intelligence is much more useful than it was at any time in Earth's history. (For example, the environment can be littered with monoliths that dispense food upon completion of various reasoning puzzles. Perhaps some of these monoliths can teach English too, that'll probably come in handy later!) Think about how much faster dog breeding is compared to wolves evolving in the wild. Breeding for intelligence should be correspondingly faster than waiting for it to evolve.
5. There are probably additional things I haven't thought of that would totally be thought of, if we had a team of experts building this evolutionary simulation with 2020's knowledge. I'm a philosopher, not an evolutionary biologist!

Skunkworks:

What about [STEM AI](#)? Let's do some STEM. You may have seen this now-classic image:

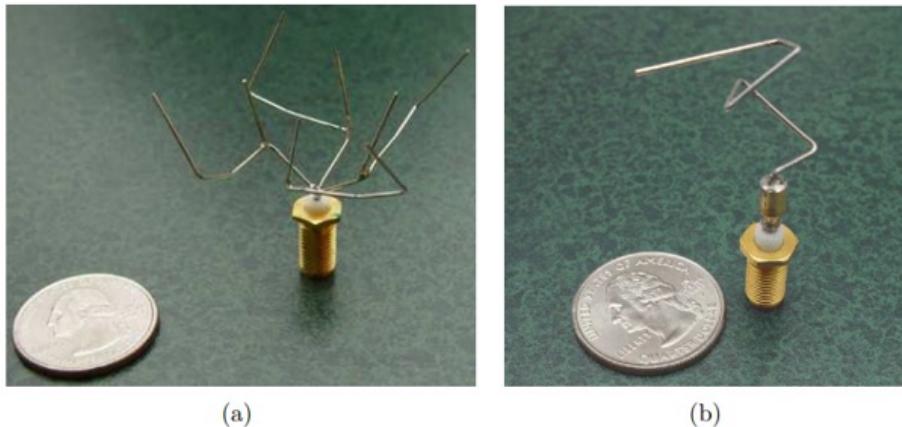


Figure 2. Photographs of prototype evolved antennas: (a) the best evolved antenna for the initial gain pattern requirement, ST5-3-10; (b) the best evolved antenna for the revised specifications, ST5-33-142-7.

These antennas were designed by an evolutionary search algorithm. Generate a design, simulate it to evaluate predicted performance, tweak & repeat. They flew on a NASA spacecraft fifteen years ago, and were massively more efficient and high-performing than the contractor-designed antennas they replaced. Took less human effort to make, too.[\[12\]](#)

This sort of thing gets a lot more powerful with +12 OOMs. Engineers often use simulations to test designs more cheaply than by building an actual prototype. SpaceX, for example, [did this](#) for their Raptor rocket engine. Now imagine that their simulations are significantly more detailed, spending 1,000,000x more compute, and also that they have an evolutionary

search component that auto-generates 1,000 variations of each design and iterates for 1,000 generations to find the optimal version of each design for the problem (or even invents new designs from scratch.) And perhaps all of this automated design and tweaking (and even the in-simulation testing) is done more intelligently by a copy of OmegaStar trained on this “game.”

Why would this be a big deal? I’m not sure it would be. But take a look at this list of [strategically relevant technologies and events](#) and think about whether Skunkworks being widely available would quickly lead to some of them. For example, given how successful [AlphaFold 2](#) has been, maybe Skunkworks could be useful for designing nanomachines. It could certainly make it a lot easier for various minor nations and non-state entities to build weapons of mass destruction, perhaps resulting in a [vulnerable world](#).

Neuromorph:

According to [page 69 of this report](#), the Hodgkin-Huxley model of the neuron is the most detailed and realistic (and therefore the most computationally expensive) as of 2008. [EDIT: Joe Carlsmith, author of [a more recent report](#), tells me there are more detailed+realistic models available now] It costs 1,200,000 FLOP per second per neuron to run. So a [human brain](#) (along with relevant parts of the body, in a realistic-physics virtual environment, etc.) could be simulated for about 10^{17} FLOP per second.

Now, presumably (a) we don’t have good enough brain scanners as of 2020 to actually reconstruct any particular person’s brain, and (b) even if we did, the Hodgkin-Huxley model might not be detailed enough to fully capture that person’s personality and cognition.[\[13\]](#)

But maybe we can do something ‘fun’ nonetheless: We scan someone’s brain and then create a simulated brain that looks like the scan as much as possible, and then fills in the details in a random but biologically plausible way. Then we run the simulated brain and see what happens. Probably gibberish, but we run it for a simulated year to see whether it gets its act together and learns any interesting behaviors. After all, human children start off with randomly connected neurons too, but they learn.[\[14\]](#)

All of this costs a mere 10^{25} FLOP. So we do it repeatedly, using stochastic gradient descent to search through the space of possible variations on this basic setup, tweaking parameters of the simulation, the dynamical rules used to evolve neurons, the initial conditions, etc. We can do 100,000 generations of 100,000 brains-running-for-a-year this way. Maybe we’ll eventually find something intelligent, even if it lacks the memories and personality of the original scanned human.

Question Two: In this hypothetical, what’s the probability that TAI appears by end of 2020?

The first question was my way of operationalizing “*what could be built with 2020’s algorithms/ideas/etc. but a trillion times more compute?*”

This second question is my way of operationalizing “*what’s the probability that the amount of computation it would take to train a transformative model using 2020’s algorithms/ideas/etc. is 10^{35} FLOP or less?*”

(Please ignore thoughts like “But maybe all this extra compute will make people take AI safety more seriously” and “But they wouldn’t have incentives to develop modern parallelization algorithms if they had computers so fast” and “but maybe the presence of the

Compute Fairy will make them believe the simulation hypothesis?" since they run counter to the spirit of the thought experiment.)

Remember, the definition of [Transformative AI](#) is "AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution."

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Did you read those answers to Question One, visualize them and other similarly crazy things that would be going on in this hypothetical scenario, and think “Eh, IDK if that would be enough, I’m 50-50 on this. Seems plausible TAI will be achieved in this scenario but seems equally plausible it wouldn’t be.”

No! ... Well, maybe you do, but speaking for myself, I don’t have that reaction.

When I visualize this scenario, I’m like “Holyshit *all five* of these distinct research programs seem like they would probably produce something transformative within five years and perhaps even *immediately*, and there are probably more research programs I haven’t thought of!”

My answer is 90%. The reason it isn’t higher is that I’m trying to be epistemically humble and cautious, account for unknown unknowns, defer to the judgment of others, etc. If I just went with my inside view, the number would be 99%. This is because I can’t articulate any not-totally-implausible possibility in which OmegaStar, Amp(GPT-7), Crystal Nights, Skunkworks, and Neuromorph and more *don’t* lead to transformative AI within five years. All I can think of is things like “Maybe transformative AI requires some super-special mental structure which can only be found by massive blind search, so massive that the Crystal Nights program can’t find it...” I’m very interested to hear what people whose *inside-view* answer to Question Two is <90% have in mind for the remaining 10%+. I expect I’m just not modelling their views well and that after hearing more I’ll be able to imagine some not-totally-implausible no-TAI possibilities. My inside view is obviously overconfident. Hence my answer of 90%.

Poll: What is your *inside-view* answer to Question Two, i.e. your answer *without* taking into account meta-level concerns like peer disagreement, unknown unknowns, biases, etc.

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Bonus: I've [argued elsewhere](#) that what we really care about, when thinking about AI timelines, is AI-induced points of no return. I think this is likely to be [within a few years](#) of TAI, and my answer to this question is basically the same as my answer to the TAI version, but just in case:

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

OK, here's why all this matters

Ajeya Cotra's excellent timelines forecasting model is built around a probability distribution over "the amount of computation it would take to train a transformative model if we had to do it using only current knowledge."[\[15\] \(pt1p25\)](#) Most of the work goes into constructing that probability distribution; once that's done, she models how compute costs decrease, willingness-to-spend increases, and new ideas/insights/algorithms are added over time, to get her final forecast.

One of the great things about the model is that it's interactive; you can input your own probability distribution and see what the implications are for timelines. This is good because

there's a lot of room for [subjective judgment and intuition](#) when it comes to making the probability distribution.

What I've done in this post is present an intuition pump, a thought experiment that might elicit in the reader (as it does in me) the sense that *the probability distribution should have the bulk of its mass by the 10^{35} mark*.

Ajeya's best-guess distribution has the 10^{35} mark as its median, roughly. As far as I can tell, this corresponds to answering "50%" to Question Two.[\[16\]](#)

If that's also your reaction, fair enough. But insofar as your reaction is closer to mine, you should have shorter timelines than Ajeya did when she wrote the report.

There are lots of minor nitpicks I have with Ajeya's report, but I'm not talking about them; instead, I wrote this, which is a lot more subjective and hand-wavy. I made this choice because the minor nitpicks don't ultimately influence the answer very much, whereas this more subjective disagreement is a pretty big [crux](#).[\[17\]](#) Suppose your answer to Question 2 is 80%. Well, that means your distribution should have 80% by the 10^{35} mark compared to Ajeya's 50%, and that means that your median should be roughly 10 years earlier than hers, all else equal: 2040-ish rather than 2050-ish.[\[18\]](#)

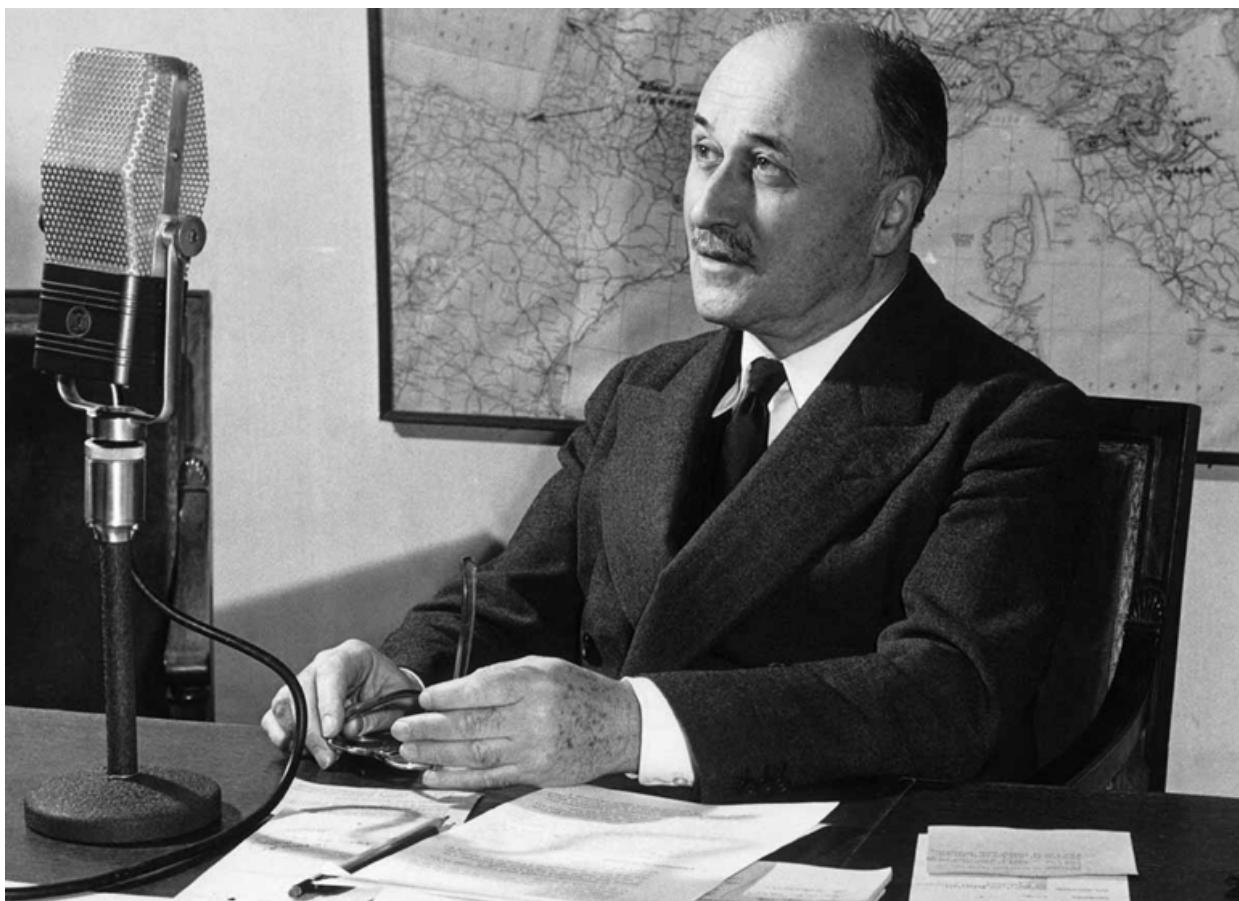
I hope this post helps focus the general discussion about timelines. As far as I can tell, the biggest crux for most people is something like "Can we get TAI just by throwing more compute at the problem?" Now, obviously we *can* get TAI just by throwing more compute at the problem, there are theorems about how neural nets are universal function approximators etc., and we can always do architecture search to find the right architectures. So the crux is really about whether we can get TAI just by throwing a *large but not too large* amount of compute at the problem... and I propose we operationalize "large but not too large" as " 10^{35} FLOP or less."[\[19\]](#) I'd like to hear people with long timelines explain why OmegaStar, Amp(GPT-7), Crystal Nights, SkunkWorks, and Neuromorph wouldn't be transformative (or more generally, wouldn't cause [an AI-induced PONR](#)). I'd rest easier at night if I had some hope along those lines.

This is part of my larger investigation into timelines commissioned by [CLR](#). Many thanks to Tegan McCaslin, Lukas Finnveden, Anthony DiGiovanni, Connor Leahy, and Carl Shulman for comments on drafts. Kudos to Connor for pointing out the Skunkworks and Neuromorph ideas. Thanks to the LW team (esp. Raemon) for helping me with the formatting.

Jean Monnet: The Guerilla Bureaucrat

I have written about coordination problems from various points of view in the past ([biology](#), [economics](#), [sociology](#), [political science](#)) but this time I am about to focus not on the theory, but on the practice.

Jean Monnet was one of the founding fathers of the European Union. One may even say that he was the architect of the European Union. However, as founding fathers go, he was rather unusual. His background was unusual: He was neither a political leader, nor a lawyer, a philosopher or a military commander. He was a son of a brandy merchant from the small town of Cognac near Bordeaux and himself a merchant by trade. He dropped out of school at sixteen and never got any extensive formal education.



But also his approach was unusual: He never held an elected position, he has never put himself to the forefront, he almost never made big speeches and is not known for memorable quotations. Rather, he was always in the background, busy with the boring technical work, hanging around politicians, showing them his famous balance sheets and trying to convince them to do the sensible, if unexpected, thing.

He was, in fact, so undistinguished that, when Fortune magazine run a story about him, they have given up on inventing a proper title for him and introduced him simply as "Monsieur Jean Monnet of Cognac". But whoever he was in his life - a trader, a banker,

a civil servant - the only description that truly fits is that he was a solver of coordination problems.

The Monnet Method

This article will explore what Mario Draghi (former president of European Central Bank, and now, quite unexpectedly, the Italian prime minister) calls "[the Monnet method](#)", a bunch of principles that guided the effort to unite the continent divided by centuries of incessant wars and feuds.

But while Draghi is focusing on the lessons that may be relevant in the current state of the European Union, my interest is a bit broader: How does one solve coordination problems in general? And how does to do it as successfully as Jean Monnet once did?

In this article we are going to examine that question. Yet, before we begin, a warning is due. Monnet himself, in his memoirs, refuses to write down his method:

I might have written a series of practical maxims; but I distrust general ideas, and I never let them lead me far away from practical things. I have described the dramatic events I have lived through and the lessons I have learned from them, in the hope of preventing their happening again. My purpose is very practical. Some may call it a philosophy, if they prefer: but the essential point is to make it useful beyond the experience of one individual.

And that, I think, is not Monnet being modest. It's the core of his approach. The only way to break out of inadequate equilibria, to solve the coordination problems, is to take advantage of the unexpected. Everything that is expected, after all, just feeds into the equilibrium and makes it persist. And to take advantage of the unexpected, one should not bind himself to a specific, predictable method.

Bypassing the Hierarchy

One recurring theme in Jean Monnet's life was working outside of the existing institutions. The common sense would have it that to change how Europe works, he should have found a humble job at French ministry of foreign affairs and work his way up the hierarchy until he had enough say to push his ideas forward. Instead, it's 1914, the beginning of the Great War. Monnet is 26 years old and has no prior political experience:

One of our friends at Cognac was a lawyer, Maitre Fernand Benon, who happened to know René Viviani, the Prime Minister, quite well [...] and he agreed to introduce me to [him].

[...]

Viviani said to me: "Sir, I gather that you have some interesting proposals. Tell me."

"The problem lies in using to the full the decisive contribution of Britain's economic Power. At the moment, we don't know how to do it. And so long as we fail to allocate responsibilities according to the ability of each side, the Alliance will remain a mere juxtaposition of two separate Powers. At present, despite all the good intentions of those responsible, there are absurd instances of waste and unnecessary duplication."

Viviani interrupted. "Can you give me some examples?"

"The merchant fleets have not been fully requisitioned. There are good reasons for that, I know. But is there any reason why they should compete with each other, why they shouldn't charge the same freight rate and why their cargoes shouldn't be co-ordinated so that at least priority supplies get through quickly? You're worried at the moment because the price of oats has gone up. But it's not the price that's gone up - it's the cost of shipping them."

"What do you propose?"

"We need to set up joint bodies to estimate the combined resources of the Allies, share them out, and share out the costs."

"But we already have machinery for inter-Allied co-operation, and I'm told that it works well."

"That's nothing more than a communications system. It doesn't take decisions or make choices. We're beginning to suffer from shortages, and we must devote our resources to the most rational ends - all our resources - all our joint resources. It's this, I believe, that's still not understood. Allied solidarity must be total. In other words, neither side must be free to use its men, its supplies, or its shipping in ways that haven't been agreed by both."

"I see what you mean: but you must realize that we are talking about two Governments and two sovereign Parliaments. Can you imagine these joint decisions being taken simultaneously?"

"I know the British well enough to be sure that we can reach a real agreement with them if we appeal to their loyalty and if we play fair. They know what a terrible burden the French armies are bearing for the common cause. They will agree to make the biggest contribution in the fields where they are supreme - in production and shipping."

"I think so too, but it's hard to broach the subject at a time when we're asking them to send more troops. You seem to have some idea of how to go about it. Try. I'll tell Millerand [French minister of war] to expect you. Explain to him what you've just told me."

I've quoted the story in full because it captures the essence all the later conversations Monnet had with politicians. Both in its substance - that is, coordination between countries - and its unexpected, bold, almost cheeky style. A random nobody arrives from the blue and makes grandiosely far-fetched proposals, which, nonetheless, often get accepted by the people in power.

Here's what Monnet has to say on the topic himself:

Although it takes a long time to reach the men at the top, it takes very little to explain to them how to escape from the difficulties of the present. This is something they are glad to hear when the critical moment comes. Then, when ideas are lacking, they accept yours with gratitude - provided they can present them as their own. These men, after all, take the risks; they need the kudos.

Or, hitting closer to the problem of inadequate equilibria, he explains:

However far-sighted they may be, Governments always find it difficult, and very often impossible, to change the existing state of affairs which it is their duty to administer. In their hearts they may wish to do so; but they have to account for their actions to Parliament, and they are held back by their officials, who want to keep everything just so.

In short: When it comes to coordination problems, always speak to the most powerful person around. For it is they who, if anyone, have enough power to break the existing institutions and thus escape the existing deadlock.

The Better Nature of Men

As a side point to the previous section - and although Monnet doesn't explicitly say so - speaking to a person, as opposed to dealing with an institutional process, brings in considerations that don't exist within the institution.

Here's how Monnet finishes the story above:

Leaving Viviani's office, I found Fernand Benon waiting in the hall of the Faculty of Letters. He told me that Viviani had just that morning heard of the death of his two sons in the Battle of the Marne.

He puts it as dryly as that. He doesn't elaborate. But the reader is left to wonder whether such a personal tragedy have made Viviani more prone to act in unorthodox ways. Whether he was more prone to disregard the business as usual and focus on efficiency, even if it meant supporting a young man with ludicrously far-fetched proposals.

Similar point can be made about creation of European Coal and Steel Community, the predecessor to the European Union. The people involved may have come from different countries, often traditional enemies, locked into inadequate equilibria, promoting their own interests at the expense of the whole. But, on the other hand, each of them has lost friends and relatives in the war and often it took as little as to look out of the window to see the ruins and the destruction caused by the malfunctioning system.

A casual visitor to the community offices in 1955 notes:

The men who worked on the Treaty of Paris were men who had fought two wars, lived through two wars, and were determined to forge a future that would turn its back on that past. [...] In the General Secretariat there was Kohnstamm, who was Dutch and had been in the Resistance, and opposite him, there was a German called Winrich Behr. Winrich Behr had been on the staff of Field Marshal Paulus during the battle of Stalingrad and had had the task of bringing to the Führer the news of the Stalingrad army's surrender; he was also on Rommel's staff in the Afrika Korps during the North African campaign. [...] So there was a German, a Frenchman, Rollmann, who was a Luxembourger, there were Belgians, and so on. [...] these men, who did not know one another and had even served in opposing armies, came together and a friendship sprang up among them. That is what I felt, and as a Swiss, it made a great impression on me. My conclusion was that these men who had fought were now closer to one another than we, who remained or found ourselves outside.

Monnet expresses the same sentiment in a less poetic way:

Werner Klaer and Roger Hutter, one German and the other French, sat opposite each other in the same office. They recognized and confessed the ways in which, in their national rail systems, they had both manipulated freight-rates to distort free competition. Together, in the closest collaboration, they now spent months undoing a skein of national discrimination.

Crises are Opportunities

At this most fateful moment in the history of the modern world the Governments of the United Kingdom and the French Republic make this declaration of indissoluble union and unyielding resolution in their common defence of justice and freedom against subjection to a system which reduces mankind to a life of robots and slaves.

The two Governments declare that France and Great Britain shall no longer be two nations, but one Franco-British Union.

The constitution of the Union will provide for joint organs of defence, foreign, financial and economic policies.

Every citizen of France will enjoy immediately citizenship of Great Britain; every British subject will become a citizen of France.

Both countries will share responsibility for the repair of the devastation of war, wherever it occurs in their territories, and the resources of both shall be equally, and as one, applied to that purpose.

During the war there shall be a single war Cabinet, and all the forces of Britain and France, whether on land, sea, or in the air, will be placed under its direction. It will govern from wherever best it can. The two Parliaments will be formally associated. The nations of the British Empire are already forming new armies. France will keep her available forces in the field, on the sea, and in the air. The Union appeals to the United States to fortify the economic resources of the Allies, and to bring her powerful material aid to the common cause.

The Union will concentrate its whole energy against the power of the enemy, no matter where the battle may be.

And thus we shall conquer.

After this message was dictated over telephone on June 16th, 1940 there was a short silence on the French side.

Had the text been approved by Churchill himself?

Churchill picked up the telephone and said: "Hold on! De Gaulle's leaving now: He'll bring you the text... And now, we must meet quickly. Tomorrow morning at Concarneau."

Full irreversible political union of Britain and France would have been inconceivable at any other moment. However, given the grave military circumstances, Monnet was able to persuade both Churchill and De Gaulle to support the proposal.

To give some context, Churchill was generally in favour of European integration, but he imagined it as a continental matter, with Britain standing benevolently on the side. However, in 1940 the matters looked grim indeed. If France had signed an armistice - which was almost certain to happen - Britain would be left fighting Germany all by itself. (At the time neither the US, nor the USSR have been involved.) So, Churchill swallowed his disgust and put his weight behind the proposal.

De Gaulle, on the other hand, was a nationalist. In the postwar era he single-handedly hindered the progress of European integration more than anyone else. He rejected two British applications to join the block, caused the "empty chair" crisis and so on. Yet, facing the immediate prospect of France surrendering to Germany, he was willing to support the full unification of the two countries.

Unfortunately, the proposed meeting at Concarneau never happened and France has signed the armistice with Germany on June 22nd.

This anecdote brings in a new point. Speaking to the people in power may not, by itself, solve a coordination problem. Often, a crisis is needed to make them more willing to break the mold and act in unorthodox ways.

In practice, this means that the business of breaking the inadequate equilibria often boils down to waiting for the crisis, building social networks in the meantime and preparing solutions that could be put on table once the crisis hits.

Monnet:

I can wait a long time for the right moment. In Cognac, they are good at waiting. It is the only way to make good brandy.

Of course, the method is not guaranteed to work. It's a gamble. You can only win by trying over and over again.

The proposal for Franco-British union, as already said, has failed. So did the European Defense Community, an attempt in early fifties to establish a common European army.

The crisis that triggered the effort hit in 1950 with the outbreak of Korean War. The feeling at the time was that the same may happen in Europe. Recall that half of Europe, most significantly the east part of Germany, was occupied by Soviet forces and that the relations were everything but friendly. (The truly medieval siege of Berlin has ended just a year ago.)

On the one hand, there was a strong pressure to rearm West Germany, so that it was not an easy prey to Soviets. But at the same time and quite understandably, the prospect of the resurrected German army caused quite a lot of uneasiness in France and Benelux countries.

EDC has been an attempt to solve the problem by creating a common European army. In Monnet's words:

First soldier placed under arms in Germany should be a European soldier.

However, by the time when the treaty was voted on it the French parliament, the window of opportunity has already closed. Korean war was over and the treaty was rejected.

Given that the proposal for Franco-British union failed, it may not be the best illustration for the principle. However, Monnet didn't sit idly during the war. He tried to solve the problem by focusing on the US instead of France. This is what Maynard Keynes has to say on the topic:

When the United States of America entered the war, Roosevelt was presented with an aircraft production programme which all the American experts thought would require a miracle. Jean Monnet was bold enough to find it inadequate... The President came to agree with this point of view. [...] This crucial decision may well have shortened the war by a whole year.

Little steps

There were, in the inter-war period, European federalists, people like count Richard von Coudenhove-Kalergi, trying to push for immediate establishment of the United States of Europe.

Coudenhove-Kalergi was an interesting character. A child of an Austro-Hungarian diplomat and a Japanese mother (their [wedding photo](#) is too awesome to not to link to) he was the founder of the Pan-European movement. His adventures during the war also served as a basis for Victor Laszlo, a character in the movie Casablanca.

I don't claim to fully understand what the Pan-European movement was about. From the brief look it looks like they were aiming at some kind of improved version of former Austria-Hungary. The fact that Coudenhove-Kalergi was succeeded as the president of the movement by Otto von Habsburg definitely points in that direction.

That being said, Coudenhove-Kalergi did manage to get support from some politicians (French prime minister Aristide Briand) and intellectuals (Einstein, Freud) and so the movement wasn't totally irrelevant.

But: The Pan-European project would have required immediate giving up of most of the national sovereignty of the concerned nations. And, as became apparent during the later unification of the continent (and also, more recently, during Brexit) the nations would engage in all kinds of disruptive behaviour before letting go even a smallest piece of their sovereignty. In short, the Pan-European project was a political non-starter.

On the other hand, there were attempts to establish peace in Europe without nations giving up their sovereignty. This is the line of thought represented by the League of Nations and later by the United Nations as well as by the Council of Europe. (Not to be confused with European Council or Council of the European Union, which are EU institutions!)

And while Monnet had nothing to do with the federalists, he was personally involved in the League of Nations. He was the deputy secretary-general of the organization while it was still in its beginnings, when the secretariat has done the most work and consisted maybe of twenty people.

And the importance of these organizations should not be downplayed. League of Nations managed to solve tricky problems like the Problem of Silesia, the problem of Danzig or to prevent a full economic collapse of the newly established Austria. These organizations also provide the institutional backing for the modern international law (e.g. International Court of Justice in the Hague). And having a common international discussion forum, such as UN, even if it had no real power, is still worth it.

But Monnet has also seen the problems first hand. As he explains, where the League of Nations succeeded (e.g. [Silesia](#)) it was only because the allies didn't want to rock the boat so early after the war and so they handed the problems they didn't agree on to the League, along with the power to solve them.

Bringing Governments together, getting national officials to co-operate, is well-intentioned enough; but the method breaks down as soon as national interests conflict, unless there is an independent political body that can take a common view of the problem and arrive at a common decision. I became convinced of this twenty years later. What success we had in Geneva is more simply explained. The important agreements that were reached there became possible in so far as the Great Powers, in particular France and Britain, thought it to be in their interest to avoid a dispute. When this only was the case, we were free to seek solutions.

But this only works for a while. Once the memories of the war wane away, there are no more incentives to hand problems to the common institutions and national sovereignty reigns supreme once again. This is, more or less, the state of affairs we can see in the UN security council in the present. Any proposal by the US or the UK get vetoed by Russia and China. Any proposals by Russia or China are shot down by the US and the UK. In the end, nothing gets done.

During this time I was busy between London and Paris, winding up the units of which I had been in charge, and I had no hand in the drafting of the Hague Covenant. Those who did draft it were careful to avoid setting up a genuine authority independent of the member States, or even a first nucleus of autonomous international power. The whole of the League depended on the Council, which alone was empowered to take decisions, and even then by unanimous vote. The Assembly could issue only opinions, resolutions, and recommendations. The role of the Secretariat was to assist the Council in its work. Quite obviously such an organization was incapable of expressing and imposing a common will. That, at least, is the conclusion I came to later. But at the time I did not see the pooling of sovereignty as a way of solving international problems. Nobody did, even if their words seemed to imply an appeal to some authority that would be above nations.

And:

One scene among others sticks in my memory: it was a meeting of the Council to discuss the world distribution of raw materials. The Italian representative, Marchese Imperiale, was pressing for a certain decision to be taken. As usual, the British representative, Lord Balfour, looked as if he were asleep. Then his turn came, he got up and said simply: 'His Majesty's Government is against.' Then he returned to his doze. The question was settled.

We can already see the problem that comes up in many, if not all, coordination problems. Either the parties in question get full power to decide for themselves, that is, power to veto any common decision (League of Nations), or their ability to decide for themselves is constrained, which they would never agree on in the first place (Pan-Europa). The former option means that they will never agree, the latter option means that they will never even get to the negotiation table.

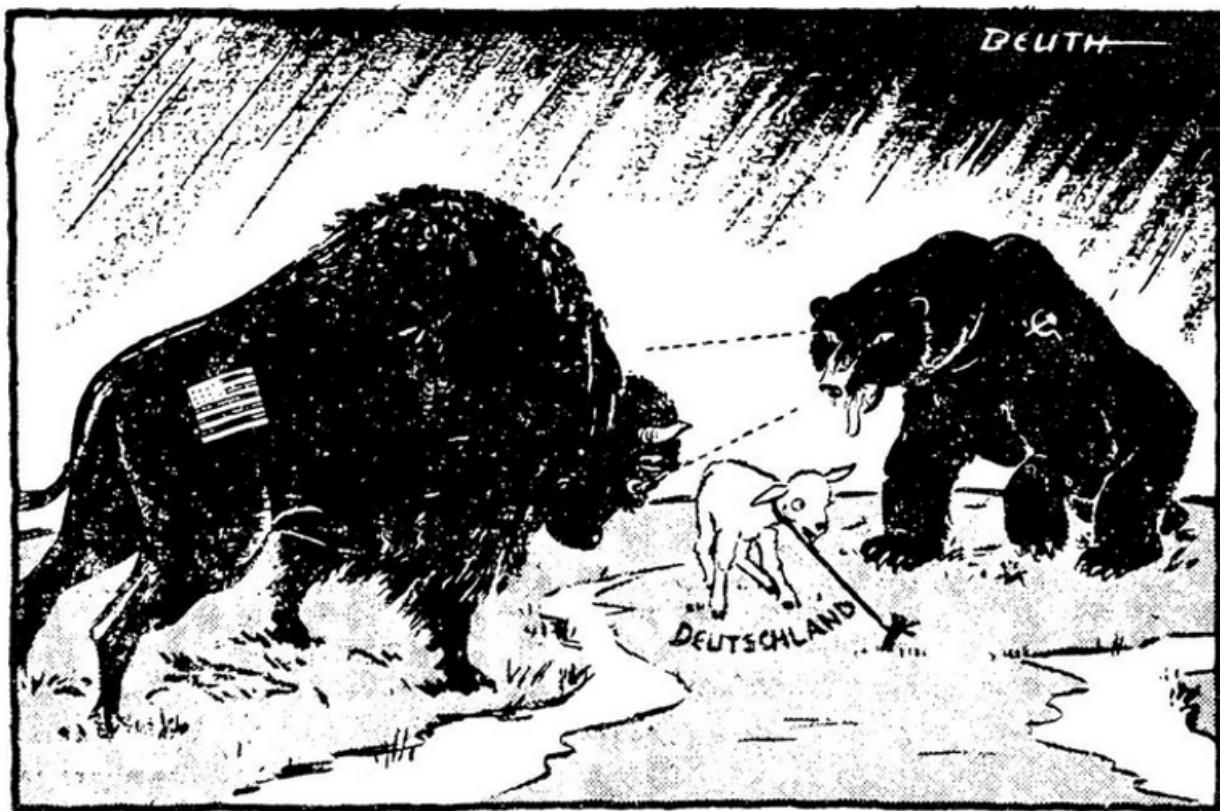
In the early 50's in Europe, the problem was solved by the method of "small steps". The states were not asked to relinquish all their sovereignty in one go. Rather, a very specific area of interest was singled out (coal and steel industry) and the delegation of sovereignty was limited to that area.

But this is a hard trick to pull off. One needs all the following at once:

- The area of interest must be limited enough not to scare individual actors away.
- There must be a crisis serious enough to override all the remaining fear of the sovereignty loss.
- The proposal must be done at the right time, by the right actor.
- The chosen area must be impactful enough to be worth the effort.

The idea that Europe can be somehow made more peaceful by increased economic cooperation was floating around for a long time. What required political genius was to meet all those preconditions at once.

In that particular case, the crisis was caused by French fear of revived Germany on one side and German desperation of being caught unarmed and occupied in the center of conflict of the great powers. Maybe a contemporary cartoon explains the mood better than I can do.



The solution was to internationalize the coal and steel industry, which in effect, meant giving French fair access to the coal from the Rhine-Ruhr region. Coal and steel at the time were the most important resources needed to wage a war, similar to the oil today. The common market in coal and steel not only meant that one country can't easily get a large military advantage simply by owning a specific coal-producing region, but also made the market much more transparent, allowing the participants to closely watch each other.

For Germany, on the other hand, the solution meant that it was, for the first time since the end of the war, invited to an international organization as an equal among equals. By making it less dangerous, it lowered the pressure to keep occupation regime in

place and paved a way back to the normal. And returning to the normal was the priority number one. In the hunger winter on 1946/47, the average calorie intake per day per person in Germany is believed to have been around 1000. People were starving. Archbishop of Cologne gave his blessing to those who stole to feed and warm their families. But at the same time, German industry was being disassembled. Things had got somehow better by 1950, but getting out of the deadlock and restarting the German economy was still of utmost importance.

At the same time, economic cooperation is impactful, in the sense that once you start doing it it naturally expands. Having a common market for coal and steel may be great, but if the freight costs are unfair, then you are back to your original problem. To solve it, you need common transportation policy. And indeed, in the subsequent decades the economic cooperation expanded until we've got the full common market of today.

Who made the proposal was also important. Just few months before Schumann declaration (the proposal by French to form the steel and coal community) similar idea was floated by German prime minister Konrad Adenauer. But he got laughed down in the political circles as well as in the press. Germany was not in the position to make proposals at the time.

But, in the end, even all of that would not suffice if the procedure of the negotiations were not what it was.

Two-layered approach

The problem there, you see, is that if the participants don't agree to the solution of the crucial coordination problem in advance it will become a bargaining token in the subsequent negotiations on the technical issues. That way, it will get gradually watered down if not completely removed.

Consider a modified version of the [prisoner's dilemma](#). This time, the prisoners are allowed to communicate, but they also have to solve an additional technical problem, say, how to split the loot. They may start with agreeing on not betraying each other to the prosecutors, but later one of them may say: "I've done most of the work. I want 70% of the loot, otherwise I am going to rat on you." It's easy to see how the problem would escalate and end up in the prisoners betraying each other.

Similar dynamics could be observed in the League of Nations:

At every meeting, people talked about the general interest, but it was always forgotten along the way: everyone was obsessed by the effect that any solution would have on him - on his country. The result was that no one really tried to solve the actual problems [...]. This was inevitable in a body subject to the unanimity rule.

The European coal and steel effort avoided this problem by making the agreement on the coordination problem (that is, delegating the national sovereignty) a condition to participate in the negotiations. Both France and Germany were willing to do so and the project started as a Franco-German endeavour. However, other countries were invited to join.

French memorandum sent to London, Rome and Benelux countries:

The Governments of ... are resolved to carry out a common action aiming at peace, European solidarity, and economic and social progress by pooling their coal and steel production and by the institution of a new High Authority whose decisions will bind ... and the countries which may adhere to it in the future.

And the British reply:

His Majesty's Government have received the French Government's Memorandum [...] It should [...] be realized that if the French Government intend to insist on a commitment to pool resources and set up an authority with certain sovereign powers as a prior condition to joining in the talks, His Majesty's Government would reluctantly be unable to accept such a condition. His Majesty's Government would greatly regret such an outcome.

There you go. National sovereignty is now called out by name.

In the end, Italy and Benelux countries accepted the offer. Britain did not. The coordination problem was solved, albeit at the cost of sacrificing Britain's membership in the new project.

Tribalism

The previous discussion begs a question. The founding fathers of the EU took for granted that people assigned to supranational European organizations would work for the good of Europe as a whole rather than for the benefit of their native countries. But that's far from obvious. Would a person abandon their tribe and join a super-tribe just because their job descriptions tells them to do so? If so, then tribalism is less of a problem than we thought.

And looking at concrete examples, we observe that it can go both ways. American congresspeople, for example, clearly work for benefit of their party, not for the benefit of the whole. On the other hand [Swiss Federal Councilors](#), despite being from different parties, work for the benefit of the entire Switzerland.

My first guess would be that allegiance to a tribe follows the accountability. If US congresspeople are primarily accountable to their parties (in the sense of being nominated by the parties) they will split into tribes along party lines. If Swiss Federal Councilors are accountable to the parliament (by being elected by the majority of parliamentarians and thus needing support from multiple parties) then they'll work for the common cause.

However, European Commission seems to defy that rule. The members are nominated by the national governments, yet, they seem not to give unfair advantage to their native countries.

The alternative explanation would be that the common sense outcome of the [Robbers Cave experiment](#) applies: If people are put in a single room, working to solve common problems, they will eventually form a coherent tribe.

But again, the US case seems to contradict that conclusion. There is certainly more to think about here.

Instead of Conclusion

While this article has been mostly about breaking old inadequate institutions, I would like to finish with a quote paying homage to institutions as such.

Monnet, at the first meeting of the European Council, remarks:

The union of Europe cannot be based on goodwill alone. Rules are needed. The tragic events we have lived through and are still witnessing may have made us wiser. But men pass away; others will take our place. We cannot bequeath them our personal experience. That will die with us. But we can leave them institutions. The life of institutions is longer than that of men: if they are well built, they can accumulate and hand on the wisdom of succeeding generations.

And if I am allowed to expand on that thought, the institutions of the European Union accumulate not only the wisdom imparted on us during the world wars, but, by the virtue of its gradual expansion, also the wisdom of decades of living under Franco or Salazar, the lessons learned during the Troubles in Northern Ireland, the wisdom of balancing the contradictory influences from the West and the East in neutral Austria, the wisdom of living under communist regimes everywhere from Estonia to Bulgaria and, with accession of Croatia (and soon, hopefully, other Balkan countries) the lessons painfully learned in the wars of former Yugoslavia.

Most of the quotes in this article were taken from [Jean Monnet's memoirs](#). Huge trove of resources about the history of European integration process can be found [here](#) (warning: lot of stuff is available only in French or German).

The case for aligning narrowly superhuman models

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

*I wrote this post to get people's takes on a type of work that seems exciting to me personally; I'm not speaking for Open Phil as a whole. Institutionally, we are very uncertain whether to prioritize this (and if we do where it should be housed and how our giving should be structured). **We are not seeking grant applications on this topic right now.***

Thanks to Daniel Dewey, Eliezer Yudkowsky, Evan Hubinger, Holden Karnofsky, Jared Kaplan, Mike Levine, Nick Beckstead, Owen Cotton-Barratt, Paul Christiano, Rob Bensinger, and Rohin Shah for comments on earlier drafts.

A genre of technical AI risk reduction work that seems exciting to me is trying to align existing models that already are, or have the potential to be, “superhuman”^[1] at some particular task (which I’ll call **narrowly superhuman models**).^[2] I don’t just mean “train these models to be more robust, reliable, interpretable, etc” (though that seems good too); I mean “figure out how to harness their full abilities so they can be as useful as possible to humans” (focusing on “fuzzy” domains where it’s intuitively non-obvious how to make that happen).

Here’s an example of what I’m thinking of: intuitively speaking, it feels like GPT-3 is “smart enough to” (say) give advice about what to do if I’m sick that’s better than advice I’d get from asking humans on Reddit or Facebook, because it’s digested a vast store of knowledge about illness symptoms and remedies. Moreover, *certain ways of prompting it* provide suggestive evidence that it could use this knowledge to give helpful advice. With respect to the Reddit or Facebook users I might otherwise ask, it seems like GPT-3 has the potential to be narrowly superhuman in the domain of health advice.

But GPT-3 doesn’t seem to “want” to give me the best possible health advice -- instead it “wants” to play a strange improv game riffing off the prompt I give it, pretending it’s a random internet user. So if I want to use GPT-3 to get advice about my health, there is a gap between what it’s capable of (which could even exceed humans) and what I can get it to actually provide me. I’m interested in the challenge of:

How can we get GPT-3 to give “the best health advice it can give” when humans^[3] in some sense “understand less” about what to do when you’re sick than GPT-3 does? And in that regime, how can we even tell whether it’s actually “doing the best it can”?

I think there are other similar challenges we could define for existing models, especially large language models.

I'm excited about tackling this particular type of near-term challenge because it feels like a microcosm of the long-term AI alignment problem in a real, non-superficial sense. In the end, we probably want to find ways to meaningfully supervise (or justifiably trust) models that are more capable than ~all humans in ~all domains.^[4] So it seems like a promising form of practice to figure out how to get particular humans to oversee models that are more capable than them in specific ways, if this is done with an eye to developing scalable and domain-general techniques.

I'll call this type of project **aligning narrowly superhuman models**. In the rest of this post, I:

- Give a more detailed description of what aligning narrowly superhuman models could look like, what does and doesn't "count", and what future projects I think could be done in this space ([more](#)).
- Explain why I think aligning narrowly superhuman models could meaningfully reduce long-term existential risk from misaligned AI ([more](#)).
- Lay out the potential advantages that I think this work has over other types of AI alignment research: (a) conceptual thinking, (b) demos in small-scale artificial settings, and (c) mainstream ML safety such as interpretability and robustness ([more](#)).
- Answer some objections and questions about this research direction, e.g. concerns that it's not very neglected, feels suspiciously similar to commercialization, might cause harm by exacerbating AI race dynamics, or is dominated by another type of work ([more](#)).
- Briefly discuss where I think some AI alignment researchers currently stand on this work ([more](#)).
- Summarize takeaways and possible next steps for readers ([more](#)).

There aren't a large number of roles where someone could do this right now, but if aligning narrowly superhuman models is a good idea, *and* we can build a community consensus around it being a good idea, I think we have a good shot at creating a number of roles in this space over the coming years (allowing a larger number of people to productively contribute to AI x-risk reduction than would be possible otherwise). To discover whether that's possible, **I'd appreciate it if people could react with pushback and/or endorsement**, depending on where you're at.

What aligning narrowly superhuman models could look like

I'm a lot less confident about a particular agenda or set of project ideas than I am about the high-level intuition that it seems like we could somehow exploit the fact that today's models are superhuman in some domains to create (and then analyze and solve) scaled-down versions of the "aligning superintelligent models" problem. I think even the basic framing of the problem has a lot of room to evolve and improve; I'm trying to point people toward something that seems interestingly analogous to the long-run alignment problem rather than nail down a crisp problem statement. With that said, in this section I'll lay out one vision of what work in this area could look like to provide something concrete to react to.

First of all, it's important to note that not all narrowly superhuman models are going to be equally interesting as alignment case studies. AlphaGoZero (AGZ) is narrowly

superhuman in an extremely strong sense: it not only makes Go moves better than the moves made by top human players, but also probably makes moves that top players couldn't even reliably *recognize* as good. But there isn't really an outer alignment problem for Go: a precise, algorithmically-generated training signal (the win/loss signal) is capable of eliciting the "full Go-playing potential" of AGZ given enough training (although at a certain scale [inner alignment issues](#) may crop up). I think we should be focusing on cases where both inner and outer alignment are live issues.

The case studies which seem interesting are models which have the potential to be superhuman at a task (like "giving health advice") for which we have no simple algorithmic-generated or hard-coded training signal that's adequate (which I'll call "**fuzzy tasks**"). The natural thing to do is to try to train the model on a fuzzy task using human demonstrations or human feedback -- but if (like AGZ) the model actually has the capacity to improve on what humans can demonstrate or even reliably recognize, it's not immediately obvious how to elicit its "full potential."

Here's an attempt at one potential "project-generation formula", where I try to spell out connections to what I see as the main traditional sub-problems within academic AI alignment research:

Choose a helpful "fuzzy" task (e.g. summarization, question-answering, advice-giving, story-writing) for which we have suggestive evidence that makes us suspect a state-of-the-art model has the capacity to significantly outperform some reference set of humans (e.g. Mechanical Turk workers) given the right training signal. Then,

1. **Reward learning:** Find a training procedure that allows those reference humans to train the model to do the fuzzy task better than they could do it (and ideally, better than they could even recognize or verify unaided). This procedure shouldn't rely on the researchers' own understanding of the particular domain in a way that wouldn't generalize across domains.
2. **Scalability and competitiveness:** Argue or empirically demonstrate that the human oversight work wouldn't have to scale up much if the model were 10x or 100x bigger, or each instance of the task took 10x or 100x longer to demonstrate or evaluate.
3. **Interpretability and robustness:** Once you've done this, try to understand its behavior and stamp out whatever pathologies (e.g. lying, going off the rails) may have cropped up.^[5]

This is just one type of project you could do in this space. The larger motivating question here is something like, "It looks like at least some existing models, in at least some domains, 'have the ability' to exceed at least some humans in a fuzzy domain, but it's not obvious how to 'draw it out' and how to tell if they are 'doing the best they can to help.' What do we do about that?"

I don't think the project-generation formula I laid out above will turn out to be the best/most productive formulation of the work in the end; I'm just trying to get the ball rolling with something that seems concrete and tractable right now. As one example, the project-generation formula above is putting reward learning / "outer alignment" front and center, and I could imagine other fruitful types of projects that put "inner alignment" issues front and center.

Existing work in this area

This kind of work only became possible to do extremely recently, and mostly only in industry AI labs; I'm not aware of a paper that follows all three steps above completely. But "[Learning to summarize from human feedback](#)" (Stiennon et al., 2020) accomplishes the easier version of 1 and a bit of 2 and 3. The authors chose the fuzzy task of summarizing Reddit posts; there was an existing corpus of human demonstrations (summaries of posts written by the posters themselves, beginning with "TL;DR"):

1. **Reward learning:** Ultimately, the quality of summaries generated by a large language model fine-tuned with RL from human feedback exceeded the quality of the Reddit summaries (i.e. it exceeded what some set of reference humans generated). But it didn't really exceed what the human workers could *evaluate* -- except in the fairly straightforward (but IMO meaningful) sense that the authors figured out quality control procedures, human rating aggregation algorithms, easier framings of the question, training and feedback for workers, etc that allowed them to get better performance than they would have gotten using the most naive implementation of "train on human ratings."
2. **Scalability:** I don't think the paper makes explicit arguments about scalability, but the method is very domain-general and could plausibly work for significantly harder tasks, especially combined with decomposition (and I'd like to see that systematically attempted).
3. **Interpretability and robustness:** The paper doesn't dig deep into interpretability, reliability, and pathological behavior, but it does demonstrate that optimizing the reward model (learned from human judgments) "too hard" leads to weird pathological summaries that are repetitive, offensive, etc., and addresses this by applying a penalty for diverging too far from the human demonstration distribution.

What kinds of projects do and don't "count"

In the high-level description of this research area, I've aimed to be as broad as possible while picking out the thing that seems interestingly different from [other research in alignment right now](#) (i.e. the focus on narrowly superhuman models). But given such a broad description, it can be confusing what does and doesn't count as satisfying it. Would self-driving cars count? Would [MuseNet](#) count? Would just training GPT-4 count?

Firstly, I don't think whether a project "counts" is binary -- in some sense, all I'm saying is "Find a model today such that it seems as non-obvious as possible how to align it, then try to align it." The more obvious the training signal is, the less a project "counts." But here are some heuristics to help pick out the work that currently feels most central and helpful to me:

- **You should probably be fine-tuning an existing large model:** I don't think we should be guessing what size models could have the potential to be narrowly superhuman in some domain; I think an alignment project should probably be inspired by noticing that an existing model seems to have some "knowledge" or "skill" that's it not adequately harnessing because it doesn't "want to", as in the example with GPT-3 and health advice above.^[6] I would guess the base model you start with should be >>1B parameters, and the larger the better -- this is

because the larger the model is, the more likely it is to have the capacity to be superhuman in an interesting, challenging domain. Less confidently, I would guess that you probably want to be fine-tuning a generative model like GPT-3 or MuseNet (as opposed to a supervised learning model like an image classifier or an RL model like AlphaGoZero or AlphaStar), because those models seem closest to being able to do “interesting real-world tasks” better than some humans can.

- **If you’re making the model larger, it doesn’t count:** I see the point of this work as “realizing the potential of existing state-of-the-art models in fuzzy domains”, rather than pushing forward the state-of-the-art in models’ raw potential. Note that this doesn’t mean I think scaling up models is always bad -- I definitely see risks there, but also potential benefits depending on who does it and how (e.g. new large models can also create new opportunities to do empirical alignment research like this). I think the question of the sign of scaling work is pretty complicated and situation-dependent. I just want to clearly distinguish between the projects of “aligning narrowly superhuman models” and “scaling models up to make them (more) superhuman”, and make it clear that someone could participate in one without participating in the other. So, for example, training GPT-4 would not count as aligning a narrowly superhuman model.^[7]
- **If you’re not dealing with humans, it probably doesn’t count:** I think that if you can get the model to achieve superhuman performance at some task without collecting any human feedback or human demonstrations, the task is probably not “fuzzy” enough. It shouldn’t be easy for humans to just write down an algorithm specifying what they want, and there shouldn’t be an existing dataset that just demonstrates what they want. In practice, I also don’t think human demonstrations alone will cut it (unless they are cleverly combined with an amplification-like scheme or somehow augmented or assisted); RL from human feedback will probably be necessary. My guess is that self-driving cars mostly fail on these grounds -- in a lot of self-driving car companies, only the recognition of objects in a scene is done with large neural nets, and those are trained almost entirely from labeled datasets.^[8] To the extent that large models are used for the actual driving policy (which they usually aren’t), relatively simple/algorithmic training signals like “how far is the car from other cars”, “how centered is it in the lane”, “how smooth is its acceleration”, etc seem probably adequate to elicit human-level or superhuman driving ability without bringing in feedback from human judgments.
- **If you didn’t make the model genuinely useful, it probably doesn’t count:** I think we should generally be choosing complex, multi-dimensional real-world tasks where there is a lot of room to improve on typical humans’ actions and/or judgments -- giving advice, summarizing research, coding, writing emails, translation, telling stories, etc. In the end, these models should feel impressive and valuable -- they generally wouldn’t constitute a commercial product on their own because commercial products are rarely “clean” or “pure ML”, but should ideally have the potential to become a product with some design and engineering work. If the selected task was not valuable or at least inherently interesting, I would guess that the alignment problem wasn’t hard enough and much of the [benefits of “practicing on something similar to the real deal”](#) would be reduced. Note however that **“genuinely useful” doesn’t mean optimized for usefulness alone** -- I expect this research will not look like the shortest path to creating a valuable product (e.g. by construction the [approach I propose below](#) makes it much harder than it has to be if you just want to train a model to be useful somehow). See [this objection and response](#) for more detail.

I think some projects that don't fit all these criteria will also constitute useful progress on aligning narrowly superhuman models, but they don't feel like central examples of what I'm trying to point at.

Potential near-future projects: “sandwiching”

I think a basic formula that could take this work a step beyond Steinnon et al, 2020 is a) “sandwich” the model in between one set of humans which is less capable than it and another set of humans which is more capable than it at the fuzzy task in question, and b) figure out how to help the less-capable set of humans reproduce the judgments of the more-capable set of humans. For example,

- First fine-tune a coding model to write short functions solving simple puzzles using demonstrations and feedback collected from expert software engineers. Then try to match this performance using some process that can be implemented by people who don't know how to code and/or couldn't solve the puzzles themselves.
- First fine-tune a model to answer long-form questions in a domain (e.g. economics or physics) using demonstrations and feedback collected from experts in the domain. Then try to match this performance using some process that can be implemented by people who know very little about the domain.
- First fine-tune a model to translate between English and French using demonstrations and feedback collected from people who are fluent in both languages. Then try to match this performance using some process that can be implemented by people who are fluent in one language and barely know the other (or don't know it at all and only have a dictionary). Something similar was done in [Lample et al., 2018](#), although they didn't use human feedback.

In all of these cases, my guess is that the way to get the less-capable group of humans to provide training signals of a similar quality to the more-capable group will involve some combination of:

- Training models to help the humans form better judgments (for example, training models to explain the meaning of technical terms or to fetch and summarize relevant papers for humans).
- Breaking down the problem and splitting it up among many humans (as in [Humans Consulting HCH](#)).
- Getting models to explain why they're doing what they're doing in simpler terms that connect to things the human overseers understand (this feels like it could fit under debate or interpretability).
- Figuring out how to train the human workers, and how to separate their good judgments from noise / mistakes.

It may not yet be possible to do these more ambitious projects (for example, because models may not be powerful enough yet to train them to meaningfully help human evaluators, engage in debates, meaningfully exceed what humans can recognize / verify, etc). In that case, I think it would still be fairly valuable to keep doing human feedback projects like Steinnon et al., 2020 and stay on the lookout for opportunities to push models past human evaluations; state-of-the-art models are rapidly increasing in size and it may become possible within a couple of years even if it's not quite possible now.

Importantly, I think people could make meaningful progress on aligning narrowly superhuman models using existing models without scaling them up any further, even if they are only superhuman with respect to human demonstrations for now -- there's a lot we don't know even just about how to do RL from human feedback optimally. And in the near future I expect it will be possible to use the larger models which will likely be trained to do even more interesting projects, which have the potential to exceed human evaluations in some domains.

(For more speculative thoughts on how we might go beyond "sandwiching", see [the appendix](#).)

How this work could reduce long-term AI x-risk

On the outside view, I think we should be quite excited about opportunities to get experience with the sort of thing we want to eventually be good at (aligning models that are smarter than humans). In general, it seems to me like building and iterating on prototypes is a huge part of how R&D progress is made in engineering fields, and it would be exciting if AI alignment could move in that direction.

If there are a large number of well-motivated researchers pushing forward on making narrowly superhuman models as helpful as possible, we improve the odds that we first encounter serious problems [like the treacherous turn](#) in a context where a) models are not smart enough to cause actually catastrophic harm yet, and b) researchers have the time and inclination to really study them and figure out how to solve them well rather than being in a mode of scrambling to put out fires and watching their backs for competitors. Holistically, this seems like a much safer situation to be in than one where the world has essentially procrastinated on figuring out how to align systems to fuzzy goals, doing only the minimum necessary to produce commercial products.

This basic outside view consideration is a big part of why I'm excited about the research area, but I also have some more specific thoughts about how it could help. Here are three somewhat more specific paths for working on aligning narrowly superhuman models today to meaningfully reduce long-term x-risk from advanced AI:

- **Practical know-how and infrastructure:** It seems likely that a successful long-run approach to (machine learning-based) alignment will involve somehow learning from human demonstrations and/or feedback as a key component, and also pretty likely that it will involve somehow using ML tools to help go beyond raw human judgment. I'd guess that a number of low level details about *how* ideas like "RL from human feedback" and "ML aiding human judgments" are implemented will make a difference to how successful the approach is: things like which human judges are selected, how well they are trained and how much practice they have, what exact types of questions are used to elicit the judgments, what judgment aggregation and quality assurance procedures are used, whether there are good off-the-shelf ML solutions for enhancing human judgments in certain ways, whether there are easy-to-use platforms that let researchers gather good human feedback at the push of a button, etc. Aligning narrowly superhuman models today could help build up tools, infrastructure, best practices, and tricks of the trade. I expect most of this will eventually be developed anyway, but speeding it up and improving its quality could still be

quite valuable, especially in short timelines worlds where there's a lot less time for things to take their natural course.

- **Better AI situation in the run-up to superintelligence:** If at each stage of ML capabilities progress we have made sure to realize models' full potential to be helpful to us in fuzzy domains, we will be going into the next stage with maximally-capable assistants to help us navigate a potentially increasingly crazy world. We'll be more likely to get trustworthy forecasts, policy advice, research assistance, and so on from our AI assistants. Medium-term AI challenges like supercharged fake news / clickbait or AI embezzlement seem like they would be less severe. People who are pursuing more easily-measurable goals like clicks or money seem like they would have less of an advantage over people pursuing hard-to-measure goals like scientific research (including AI alignment research itself). All this seems like it would make the world safer on the eve of transformative AI or AGI, and give humans more powerful and reliable tools for dealing with the TAI / AGI transition.^[9]
- **Chance of discovering or verifying long-term solution(s):** I'm not sure whether a "one shot" solution to alignment (that is, a single relatively "clean" algorithm which will work at all scales including for highly superintelligent models) is possible. But if it is, it seems like starting to do a lot of work on aligning narrowly superhuman models probably allows us to discover the right solution sooner than we otherwise would have. For one thing, people doing this work could test proposals (such as [Iterated Distillation and Amplification](#)) coming from more conceptual researchers, verifying or falsifying elements and proposing modifications informed by empirical understanding. It also seems plausible that a solution will emerge directly from this line of work rather than the conceptual work -- the latter is mostly focused on finding a one-shot solution *that will work under ~pessimistic empirical assumptions*,^[10] but it seems very plausible that a) it's impossible to find a one-shot solution that works under worst-case empirical assumptions, but b) it's possible to find one that works given the actual ways that models tend to learn or generalize. More broadly, "**doing empirical science on the alignment problem**" -- i.e. systematically studying what the main problem(s) are, how hard they are, what approaches are viable and how they scale, etc -- could help us discover a number of different avenues for reducing long-run AI x-risk that we aren't currently thinking of, one-shot technical solutions or otherwise.

I think both the broad outside view and these specific object-level benefits make a pretty compelling case that this research would be valuable on the object level. Additionally, from a "meta-EA" / "community building" perspective, I think pioneering this work could boost the careers and influence of people concerned with x-risk because it has the potential to produce conventionally-impressive results and demos. My main focus is the case that this work is valuable on the merits and I wouldn't support it purely as a career-boosting tool for aligned people, but I think this is a real and significant consideration that can tip the scales.

Advantages over other genres of alignment research

First, I'll lay out what seem like the three common genres of alignment research:

- **Conceptual research:** This is pen-and-paper thinking that often looks like a combination of math and philosophy, which is usually aiming to make progress toward a “one shot” solution (and also often involves a lot of [disentangling](#) and framing what the problem even is). The most prominent examples are [MIRI's work](#) and [Paul Christiano's work](#); a number of other posts on the Alignment Forum also fit in this category.
- **Gridworlds and games:** This work aims to demonstrate alignment problems such as wireheading or other reward hacking in a relatively small-scale artificial setting such as a simple game, and usually to solve the demonstrated problem(s) in the small-scale setting in a way that could shed light on how to solve larger-scale alignment problems. Two examples are [REALab \(Kumar et al., 2020\)](#) and [Inverse Reward Design \(Hadfield-Mennell et al., 2017\)](#).
- **Mainstream ML safety:** This is alignment-relevant work that existing ML researchers were independently working on; most of it fits under “reliability+robustness” or “interpretability.” This work is usually done on fairly large (though not always state-of-the-art) neural networks, but doesn’t usually pay special attention to the case where models are more capable or knowledgeable than humans. Some examples are the [OpenAI microscope](#) (interpretability), [Dathathri et al., 2020](#) (robustness and reliability), and the [Unrestricted Adversarial Examples Challenge](#) (robustness and reliability).

I’m broadly supportive of all three of these other lines of work, but I’m excited about the potential for the new approach described in this post to “practice the thing we eventually want to be good at.” I think on the outside view we should expect that doing whatever we can find that comes closest to practicing what we eventually want to do will be good in a number of ways (e.g. feeling and looking more “real”, encouraging good habits of thought and imposing helpful discipline, etc).

More specifically, here are some advantages that it feels like “aligning narrowly superhuman models” line of work has over each of the other three genres:

- Compared to conceptual research, I’d guess aligning narrowly superhuman models will feel meatier and more tractable to a number of people. It also seems like it would be easier for funders and peers to evaluate whether particular papers constitute progress, which would probably help create a healthier and more focused field where people are broadly more on the same page and junior researchers can get stronger mentorship. Related to both of these, I think it provides an easier opportunity for people who care about long-run x-risk to produce results that are persuasive and impressive to the broader ML community, as I mentioned above.
- Compared to gridworlds and games, I think this work stands a greater chance of scaling up to more capable systems -- I think it would probably provide some good discipline to do alignment work at a scale that’s large enough that it’s already kind of unwieldy, where models are already more capable than their overseers in some real-world-relevant ways, and researchers are forced to confront messy details and hard-to-foresee structural issues. When it’s possible to demonstrate an issue at scale, I think that’s usually a pretty clear win.
- Compared to mainstream ML safety, aligning narrowly superhuman models has some of the “discipline” advantages mentioned above of focusing on situations where models are more capable than humans. Additionally, lots of researchers work on interpretability and robustness for lots of different reasons, meaning the specific research priorities and “tastes” of the broader interpretability and robustness fields won’t be particularly optimized for reducing long-run x-risk. This can make it harder for newer researchers motivated primarily by x-risk to

zoom in on the most x-risk-relevant subproblems and get adequate mentorship on that; aligning narrowly superhuman models has the potential to be more x-risk-oriented from the start.

Finally and maybe most importantly, I think aligning narrowly superhuman models has **high long-run field growth potential** compared to these other genres of work. Just focusing on GPT-3, there are already a *lot* of different fuzzy goals we could try to align it to, and the number of opportunities will only grow as the ML industry grows and the number and size of the largest models grow. This work seems like it could absorb a constant fraction (e.g. 1% or 5%) of all the ML activity -- the more models are trained and the more capable they are, the more opportunity there is to align narrowly superhuman models to ever more tasks.

I think we have a shot at eventually supplying a lot of people to work on it too. In the long run, I think more EAs could be in a position to contribute to this type of work than to either conceptual research or mainstream ML safety.^[11] Conceptual research is often foggy and extremely difficult to make progress on without a particular kind of inspiration and/or hard-to-define “taste”; mainstream ML safety is often quite technical and mathematically dense (and ensuring the work stays relevant to long-run x-risk may be difficult).

A lot of work involved in aligning narrowly superhuman models, on the other hand, seems like it’s probably some combination of: a) software engineering and ML engineering, b) dealing with human contractors, and c) common sense problem-solving. Lead researchers may need to bring taste and research judgment to ensure that the work is well-targeted, but a number of people could work under one lead researcher doing tractable day-to-day work with reasonably good feedback loops. If there were institutional homes available to onboard people onto this work, I think a strong generalist EA with a software engineering background could plausibly retrain in ML engineering over 6-12 months and start contributing to projects in the space.

Right now there are only a few organizations that offer roles doing this work and that seems like a big bottleneck, but it could make sense to prioritize creating more institutional homes and/or rapidly expanding the ones that exist.

Objections and responses

In this section I’ve tried to anticipate some potential objections, and give my responses; I’d suggest skipping around and reading only the ones that interest you. I don’t think that I have knock-down answers to all of these objections, but I do remain holistically excited about this idea after reflecting on them some.

How would this address treachery by a superintelligence?

Elaboration of objection: *It seems like there is a “hard core” of the alignment problem that only crops up when models are very smart in a very general way, not just e.g. better than MTurkers at giving medical advice. The specific scariest problem seems to be the “treacherous turn”: the possibility that the model will appear to be helpful during training time even though it’s actually power-seeking because it’s aware that it’s being trained and has to act helpful to survive, and later cause*

catastrophic harm once it knows it's out of the training setup. It doesn't seem like the "aligning narrowly superhuman models" style of work will figure out a way to address the treacherous turn until it's likely too late.

I'm very uncertain how relevant the near-term work will turn out to be for more exotic problems like the treacherous turn, and I want to think more about ways to nudge it to be more relevant.^[12] I would be very excited to find empirical research projects on large models that specifically shed light on the treacherous turn possibility, and I agree it's a weakness of my [set of potential projects](#) that they aren't specifically optimized for unearthing and correcting treachery.

With that said, I don't think there are currently genres of work that feel similarly tractable and scalable that *do* tackle the treacherous turn head on -- of the [main genres of alignment work](#), I'd argue that only a subset of the conceptual work is aiming to directly generate a long-term solution to treachery, and I think the jury is very much out on whether it will be fruitful; gridworlds and games and mainstream ML safety largely don't seem to try for a long-term treacherous turn solution. So I think the *relative* hit that my proposal takes due to this consideration is fairly limited.^[13]

Even if they don't start off tackling the treacherous turn, I'd guess that researchers would have a decent shot at learning useful things about treachery down the line if they were pursuing this work. Basically, I think it's pretty likely that full-blown treachery will be preceded by mini-treachery, and with better understanding of how neural networks tend to learn and generalize, researchers may be able to specifically seek out domains where mini-treachery is especially likely to occur to better study it. Even if techniques used by empirical researchers don't work out of the box for the treacherous turn, empirical work eliciting and studying mini-treachery could still inform what kind of theoretical or conceptual work needs to be done to address it, in a way that seems more promising to me than eliciting micro-treachery in gridworlds and games.

Moreover, even though the treacherous turn seems like the scariest single source of risk, I don't think it totally dominates the overall expected AI risk -- a significant fraction of the risk still seems to come from more "mundane" outer alignment failures and various unforced errors, which this empirical work seems better-placed to address. Of the [three broad ways I listed that this work could reduce x-risk](#), the critique that it doesn't seem to address the treacherous turn very well applies most to the "Chance of discovering or verifying long-term solution(s)" category; even if it fails to address the treacherous turn, it still seems that "Practical know-how and infrastructure" and "Better AI situation in the run-up to superintelligence" matter.

Doesn't this feel suspiciously close to just profit-maximizing?

Elaboration of objection: *It sort of sounds like you're just telling EAs to make AI really useful to humans (and indeed push models to be superhuman if they can be); it feels like this would also be what someone who is into pure profit-maximization would be excited about, and that makes me suspicious about the reasoning here and nervous about calling it an alignment activity. Even if you're right that it helps with alignment, we might see a lot of people flock to it for the wrong reasons.*

I agree that there is overlap with commercial incentives, but I think there are three high-level ways that this type of work would be different from what you'd do if you were profit-maximizing:

- **Not making models bigger:** This work doesn't involve making models bigger; it involves making models of a given fixed size more helpful. In a commercial setting, often a cost-effective way of improving results would be to simply scale the model up.
- **Seeking difficult rather than easy problems:** The problem selection is different -- other things being equal, in a commercial setting you want to select the *easiest* possible tasks; in this type of work, people would select *interestingly difficult* tasks. For example, commercial incentives would push someone to focus on precisely those tasks where simply meeting (rather than exceeding) the human imitation benchmark is sufficient for being profitable. Profit-motivated people would also likely seek tasks where algorithmically generated or hard-coded reward signals would go a long way (for example, in robotics you might be able to get away with providing algorithmically generated feedback about whether the robot's actuators ended up in the right place). The [sandwiching approach](#) I propose above is by construction making things much harder than they need to be from a pure commercial standpoint: it involves refusing to use the "best human overseers for the job" in favor of trying to figure out how to help less-capable overseers provide an adequate training signal.
- **Seeking domain-general and scalable techniques:** There is a focus on scalability and generality of techniques that goes well beyond what would be commercially optimal. In commercial settings, I expect that people will make heavy use of hard-coded behaviors and "hacks" which fully exploit domain knowledge (as is the case with self-driving cars). Additionally, there is often a "right size model for the job" in commercial settings (image models only need to be so big to adequately power self-driving car perception), and there will often not be much incentive to find techniques that also work well for a model 100x bigger. A "clean", domain-general, and scalable technique is rarely what will make the most profit at the current moment.

More broadly, I think successful versions of this type of alignment work should get someone who deeply understands ML and its limitations to say something like, "Wow, it's cool that you got the model to do that." My sense is that most commercial projects wouldn't really elicit this reaction, and would look more like applying a lot of hard work to realize an outcome that wasn't very much in doubt.

Given these differences, I think there's a good shot at distinguishing this type of work from pure profit-seeking and cultivating a community where a) most people doing this work are doing it for altruistic reasons, and b) this is reasonably legible to onlookers, funders, potential junior researchers, etc.

Isn't this not neglected because lots of people want useful AI?

Elaboration of objection: Even if this is useful for alignment, and even adjusting for the fact that companies aren't focusing on the version that's specifically alignment-optimized, won't a ton of this work get done in AI labs and startups? Doesn't that mean that the EA community is less likely to make an impact on the margin than in other, less-commercially-incentivized types of alignment work?

I do think there's probably some work happening broadly along these lines from a commercial motivation, and there will probably be significantly more in the future. But I pretty strongly suspect that there are very few, if any, projects like the ones [proposed above](#) currently being done in a commercial setting, and what work is being done is less well-targeted at reducing long-run x-risk than it could be.

The vast majority of commercial work going into AI by dollars is a) hyper application-specific and hard-coding intensive such as self-driving cars, or b) focused on scaling big generic models. I don't actually think the resources going into any sort of project focused on human demonstrations and feedback is very large right now; I'd guess it's within an order of magnitude of the resources going into other alignment work (e.g. \$100s of millions per year at the high-end, where other alignment research absorbs \$10s of millions per year). And for the reasons outlined [above](#), not a lot of this will be focused on exceeding humans using scalable, domain-general techniques.

As an example to illustrate the relative neglectedness of this work, it was Paul Christiano (motivated by long-term alignment risk concerns) who led the the [Stiennon et al., 2020](#) work, and I think it's reasonably likely that if he hadn't done so there wouldn't have been a human feedback paper of similar scale and quality for another year or so. I'd guess the EA community collectively has the opportunity to substantially increase how much of this work is done before transformative AI with a strong push, especially because the "going beyond human feedback" step seems less commercially incentivized than the Stiennon et al. work.

Some additional thoughts on neglectedness:

- I think that it matters who is doing this work and why, not just that the work gets done somehow. It seems significantly better to have someone working on these problems who is self-awarely doing it to help with long-run x-risk reduction, and who is plugged into the broader alignment community, than someone who just happens to be doing work that might be relevant to alignment. It's valuable to be collaborating with and getting feedback from more theoretical alignment researchers, and to be mentally on the lookout for ways to make the work more analogous to the long-run challenge; a generic ML engineer working on human feedback to improve the newsfeed at Facebook would be much less likely to continue to keep focusing on long-run-relevant questions for their whole career. [\[14\]](#) (And one of the value propositions here is that the long-termists / AI alignment people, as a community, should be gathering this experience, so experience that's less accessible to the community is less valuable.)
- I think that for most people, [\[15\]](#) the value (roughly speaking, the importance multiplied by the tractability) of doing marginal work in an area as a function of its crowdedness is often an upside-down U-shape rather than strictly decreasing. When there's practically no one in an area, there's no one who can mentor you when you're getting started, no one who you can hire when you're experienced, and there's no built-in audience who can be swayed by your demonstrations or arguments and can act on that. My personal intuition is that for empirical alignment work, we're near the increasing returns part of this curve (though this situation can change rapidly). There's an existing group of people who have an incentive to work on something in this space and may ramp up soon, but I think EAs have a chance to set the tone and agenda for what exactly the work they do looks like, and what standards it should be held to. I could imagine a pretty broad range of outcomes for how much ML engineers working on productizing hold themselves to the standard of finding domain-general and scalable solutions, and I could imagine EAs having an impact on that culture.

Will this cause harm by increasing investment in scaling AI?

Elaboration of objection: Even if the people doing this research don't personally scale up models and focus on generalizable and scalable solutions to making models helpful, they will be demonstrating that the models have powerful and useful capabilities that people might not have appreciated before, and could inspire people to pour more investment into simply scaling up AI or making AI useful in much less principled ways, which could cause harm that exceeds the benefits of the research.

This is a very contentious question and people have a wide range of intuitions on it. I tend to be less bothered by this type of concern than a lot of other people in the community across the board. At a high-level, my take is that:

- We're in the middle of an AI investment boom that I expect to be sustained for several more years.
- The amount of effort going into AI as a whole (\$10s of billions per year) is currently ~2 orders of magnitude larger than the amount of effort going into the kind of empirical alignment I'm proposing here, and at least in the short-term (given excitement about scaling), I expect it to grow faster than investment into the alignment work.
- This means an additional dollar of effort going into the empirical language models alignment work would need to generate ~\$100 or more of investment into accelerating AI to have a proportionally large impact on accelerating AI as a whole, in a climate where investors are already excited and AI labs are already trying hard to make them more excited. This isn't out of the question, but doesn't seem likely to me, especially given that EAs would likely be partially displacing people who would do similar work from a pure profit motivation, and that we could try to consciously shape messaging to further reduce the expected impact on AI hype. (In general, it's hard to get a factor of 100 leverage on your spending even if you're optimizing for it.)
- It also seems plausible that there are positive side effects on others' investment, such as directing marginal money away from making models larger and toward fine-tuning models to be helpful.
- Finally, I am not personally fully convinced that speeding up AI as a whole would be net negative (it seems like timing interacts in extremely complicated ways with who is in power and what the global situation is like around the time of transformative AI), which claws back some of the expected damage from acceleration.

With that said, I do think that exciting demos are a lot more likely to spur investment than written arguments, and this kind of research could generate exciting demos. Overall, the case for caution feels stronger to me than the case for caution about discussing arguments about timelines and takeoff speeds, and this consideration probably net claws back some enthusiasm I have for the proposal (largely out of deference to others).

Why not just stick with getting models not to do bad things?

Elaboration of objection: Even if this is useful for alignment, worth doing on the margin, and not net-harmful, it seems like it would be dominated by doing practical/near-term work that's more clearly and legibly connected to safety and harm-reduction, like "getting models to never lie" or "getting models to never use racist slurs" or "getting models to never confidently misclassify something." That work seems more neglected and more relevant.

Some people might feel like "avoiding bad behaviors" is clearly the subset of near-term empirical alignment work which is most relevant to long-run alignment and neglected by profit-seeking actors -- after all, in the long run we're trying to avoid a big catastrophe from misaligned AI, so in the short run we should try to avoid smaller catastrophes.

I disagree with this: I think both "getting models to be helpful and surpass human trainers" and "getting models to never do certain bad things" are valuable lines of empirical alignment work, and I'd like to see more of both. But I don't think reliability and robustness has a special place in terms of relevance to long-run x-risk reduction, and if anything it seems somewhat less exciting on the margin. This is because:

- Most versions of "make a model more reliable" don't really get at scalability to tasks/domains that are more challenging for humans to supervise, and it seems especially valuable to specifically target that. It seems very plausible to me that the most interesting challenges that are most analogous to the long-run challenge will only come up when we're trying to get excellent or superhuman performance out of a model, rather than when we're trying to avoid certain specific bad things.
- I don't actually think that reliability work is more neglected than the work of getting models to be helpful in domains that are difficult for humans. There is a significantly larger academic field around reliability and robustness than around alignment, and the reliability/robustness problem is often harder to avoid or sidestep as a company: you can choose domains where human expertise is strong or automated reward signals exist, but you will still need to get your product to meet a fairly high bar of reliability before it is commercially viable.
- Robustness and reliability falls under multiple different "social good" brands. People concerned with "Fairness, Accountability, and Transparency" (FAT) tend to be very interested in the reliability and robustness space, as well as people concerned with e.g. autonomous weapons. Even though there is a worry that [the "make models helpful" work is too easy to confuse with commercialization](#), my weak best guess is that it would actually be *harder* to tell which people working in the robustness space are optimizing for reducing long-term x-risk from AI (vs for profit or other altruistic goals), and I'd guess it would be tougher to build a distinctive culture / brand around working on the sub-problems most relevant to long-term risk.

Why not focus on testing a candidate long-term solution?

Elaboration of objection: This proposal seems like it would lead to a lot of wasted work that isn't sufficiently optimized for verifying or falsifying a long-term solution to alignment. It would be better if the potential projects were more specifically tied in to testing an existing candidate long-term solution, e.g. [Paul Christiano's agenda](#).

I'll focus on Paul's agenda in my response, because the specific people I've talked to who have this objection mostly focus on it, but I think my basic response will apply to all the conceptual alignment agendas.

Some of the projects under the umbrella of "aligning narrowly superhuman models" seem like they could instead be reframed around specific goals related to Paul's agenda, like "prototyping and testing [capability amplification](#)", "prototyping and testing [imitative generalization](#)", "figuring out how [ascription universality](#) works", and so on. I do think one of the value propositions of this work is shedding light on these sorts of concepts, but I think it's probably not helpful to frame the whole endeavor around that:

- Verifying proposed long-term solutions is [only one way that the work could reduce AI x-risk](#), and I don't think it's overwhelmingly dominant, [16] especially not if restricted to the set of long-run solutions proposed *so far*. I want people who are committed to reducing long-run AI x-risk but don't believe in any of the existing conceptual research to be doing this work, too.
- Not a lot of people currently understand the agenda well enough that they could generate good research projects from the prompt of "prototype and test [concept from a Paul blog post]." Similarly, I don't think funders and peer reviewers understand the agenda well enough to tell if a research project with that goal was helpful.
- Paul's agenda is in very active development, and I think there's a reasonable chance the whole plan ends up looking pretty different within a year or two. Given this and the above point, I think empirical work testing specific Paul ideas is best done in close collaboration with him, and I'd guess even someone who believes in Paul's agenda would often be better off just targeting the slightly looser problem description absent a lot of access to him. This makes me think research under the frame of "test Paul's agenda" is a lot less scalable than research under the frame of "align narrowly superhuman models."

There could be some simple organizing goal or "tagline" for empirical alignment research that is *neither* "test [concept from a Paul blog post]" *nor* "align narrowly superhuman models" which would inspire better-targeted research from the perspective of someone who's bullish on Paul's work, but the ones I've thought about haven't been convincing, [17] and I'd guess it'll be hard to find a good organizing tagline until the theory work gets to a more stable state.

Current state of opinion on this work

One of my goals in writing this blog post is to help build some community consensus around the "aligning narrowly superhuman models" proposal if it's in fact a good idea. To that end, I'll lay out my current understanding of where various AI alignment researchers stand on this work:

- Paul Christiano spent a few years at OpenAI working on this kind of thing (as I mentioned above he was the team lead on the Stiennon et al., 2020 paper) and generally thinks it's important -- he feels the conceptual work he's currently doing beats it as a use of his own time, but believes that this kind of work is among the best *highly scalable* types of alignment research.
- Alignment researchers I've spoken to that primarily do research on large neural networks (unlike Paul, who does a mixture of this and conceptual thinking) tend

to be more enthusiastically positive on this and more likely to consider it the best kind of work they personally could do. They also tend to be more positive on even more “no holds barred” versions of this idea -- i.e., just trying to make helpful models without focusing in particular on ideas like “sandwiching.”

- My understanding of Eliezer Yudkowsky’s position is one of “cautious relative optimism” about something in this general space compared to other non-MIRI alignment work, though he would frame the core concern differently, with more emphasis on understandability of models’ answers and decisions (e.g. “GPT-3 has somewhere buried inside it knowledge of what to do when you’re sick; how do you extract all of that and how can you tell when you’ve succeeded?”). He was [reasonably positive](#) on Stiennon et al., 2020 when it came out, and would be happy to see more work like that. Evan Hubinger’s position seems broadly similar (he is specifically interested in [ascription universality](#)). I’m not sure where others at MIRI would land on this work.
- My sense is that people who do conceptual thinking work other than Paul and MIRI tend to have a position similar to or somewhat more optimistic than Eliezer’s or Evan’s. E.g. I think Rohin Shah feels that aligning narrowly superhuman models is a reasonably good baseline for what research to do (and is developing a [benchmark](#) related to this), but he has privileged insight that beats that baseline. My rough sense is that other researchers doing conceptual thinking are on average somewhat less excited about aligning narrowly superhuman models than Paul is, and a lot less excited than the pure ML alignment researchers, but I’m not sure.

I also think a number of AI alignment researchers (and EAs working in AI risk more broadly) simply haven’t thought a lot about this kind of work because it hasn’t really been possible until the last couple of years. Until 2019 or so, there weren’t really any models accessible to researchers which could exceed human performance in fuzzy domains, and research agendas in AI alignment were largely formed before this was an option.

Takeaways and possible next steps

I’ve laid out the hypothesis that aligning narrowly superhuman models would concretely reduce x-risk and has high long-run field growth potential (i.e., lots of people who don’t have particularly esoteric skills could eventually help with it). I think if the EA and AI alignment community is in broad agreement about this, there’s potential to make a lot happen.

In terms of immediate actionable takeaways:

- **If you disagree with this argument, say so** -- especially if you [think it would be harmful](#) or would be dominated by a different line of work that shares [similar practical advantages](#) of tangibility, good feedback loops, and potential-for-scale.
- If you have more or better [project ideas](#) in mind, say so -- especially if you have ideas about [how to target “treacherous turn” dynamics](#) more specifically or how to reframe the statement of the problem to make it more productive, well-targeted, etc.
- If you a) already agree with me, and b) are already in a good position to fairly immediately make this work happen (e.g. you are a PI at a university lab that is able to fine-tune open-source models like Google’s T5, or you are a senior ML researcher at a tech company with the freedom to do your own projects), then

consider doing a project in this space. For example, you could try to solve tasks in this [Minecraft human feedback benchmark](#) being developed by some researchers at [CHAI](#) when it's released. Getting more demos of what it looks like to do this research will help make it easier to think about how valuable it would be and build consensus around it if it is. Most people will *not* be in this position. As I said at the top, **Open Phil is not soliciting grant applications right now** from people who want to try it out -- this blog post is my personal viewpoint, and institutionally we're still figuring out how much we want to prioritize this (discussion and arguments surrounding this post will feed into that).

- If you agree with this case and might be in a position to work on aligning narrowly superhuman models a few years down the line (e.g. if you are a software engineer or a university student with a technical background), consider keeping this in the back of your mind and checking in about future opportunities. If you are ready to try to switch into this work sooner, there may be jobs available doing this sort of work at various AI labs including [OpenAI](#) and [DeepMind](#),^[18] though I'd guess roles right now are fairly limited.

Looking forward to hearing people's thoughts!

Appendix: beyond sandwiching?

Right now, models like GPT-3 are not "superhuman" at fuzzy tasks in the sense that AlphaGoZero is "superhuman" at playing Go. AGZ plays Go better than *any* human, while GPT-3 is only capable of giving better advice or writing better stories than *some* humans, which is what makes the "[sandwiching tactic](#)" an option. What happens when language models and other models get narrowly superhuman in a strong sense -- better than all humans in some fuzzy domain, e.g. stock-picking? How would we verify that we got the model to be "doing the best it can do to help" when there's no reference model trained on a ground truth signal to compare its performance to?

I'm definitely very unsure what this would look like, but an important starting assumption I have is that whatever techniques worked well to get less-capable humans to reproduce the judgments of more-capable humans in a "sandwich" setting stand a good chance of just continuing to work. If we were careful not to actually *use* the expertise of the more-capable set of humans in whatever systems/tools we used to assist/augment the less-capable set, and a similar set of systems/tools seemed to work across multiple domains and for humans at multiple different capability levels, there's no particular reason to believe they would not continue working once models go from slightly less capable than the best humans to slightly more capable than them at some task. So I think it's possible we could do most of the R&D in the regime where sandwiching works.

With that said, here are some thoughts about how we could try to probe whether our alignment techniques were actually successful at eliciting a model's full potential in a regime the model is more capable than the best humans:

- We could attempt the sandwiching technique with augmented or assisted humans at both the top and the bottom. For example, if a model is capable of better stock-picking than the best human hedge fund analysts, we could compare its performance when trained using feedback from top hedge fund analysts vs feedback from normal people -- where *both* the hedge fund analysts

and the normal people are assisted/augmented using whatever decomposition techniques, question elicitation techniques, ML assistants, etc. are the state of the art. Since the expertise of the individual humans within the system ideally shouldn't bear on how well they are able to provide supervision, the goal would be to get the "augmented normal people" to produce a model that behaves similarly to the model produced by the "augmented experts."

- We could try to exploit the fact that it's usually easier to recognize a good outcome (especially after some time has passed) than to generate good behaviors. For example, we could compare the performance of a stock-price-predicting model trained using feedback from augmented hedge fund analysts to the performance of a model trained with the benefit of hindsight about what stocks actually went up over a one-year period. Ultimately, I think it's scary if we have to rely entirely on medium- and long-run outcomes to train our superhuman models, [19] but in an R&D context it could make sense to use a model trained on medium-run outcomes to check whether our techniques for augmenting human judgment are managing to elicit the full potential of the model.
- More speculatively, we could build up interpretability tools to probe the model and try to understand why it's doing what it's doing to build up confidence that it's "trying its best" and making reasonable decisions. We could try to work on something like this in the near-term with AlphaGoZero itself -- try to develop interpretability tools that will let pro Go players learn new insights about how to play Go better by probing AGZ. If we develop tools like this and verify them on models like AGZ, top hedge fund analysts could later use them to probe a stock-picking model and develop a better understanding of whether it's truly "trying its best to pick the right stocks."

-
1. At least better than *some* salient large group of humans in a particular context, like "Mechanical Turk workers", "stackoverflow users", etc. Right now, models are only superhuman with respect to *all* humans in particular crisp domains like games. E.g. AlphaGoZero is better at Go than *any* human; GPT-3 probably has the potential to give better advice than *some* humans. [←](#)
 2. This idea isn't original to me -- a number of others (especially some people working on long-term AI alignment at OpenAI and DeepMind) have thought along similar lines. My own thinking about this has been informed a lot by discussions with Paul Christiano and Holden Karnofsky. [←](#)
 3. e.g., Mechanical Turk workers who are hired to give feedback to the model [←](#)
 4. Though if we could pull off a path where we build an AI system that is superhuman in certain engineering capabilities but not yet human-level in modeling and manipulating people, and use that system to cut down on x-risk from other AI projects without having to figure out how to supervise arbitrary superhuman models, that could be really good. [←](#)
 5. Note that I don't think this is the only way to study interpretability and robustness, or even necessarily the best way. In this project-generation formula, the domain and task were optimized to make *reward learning* an especially interesting and important challenge, rather than to make interpretability or robustness especially challenging, interesting, or important. I think it's good to be complete and to try to ensure interpretability and robustness in these domains, but we should probably also do other lines of research which choose

domains / tasks that are specifically optimized for interpretability or robustness, rather than reward learning, to be especially challenging and important. ↵

6. Pragmatically speaking, fine-tuning a large model rather than training from scratch is also orders of magnitude cheaper, and so a lot more accessible to most researchers. ↵
7. Another way of seeing why it wouldn't count is that "predict the next token" is an extremely non-fuzzy training signal. ↵
8. Human contractors make these labels, but they are not providing feedback. ↵
9. More speculatively, if we're realizing models' full potential as we go along, there's less chance of ending up with what I'll call an "unforced sudden takeoff": a situation where on some important set of fuzzy tasks models jump suddenly from being not-that-useful to extraordinarily useful, but this was due to not bothering to figure out how to make models useful for fuzzy tasks rather than any inherent underlying fact about models. I'm not sure how plausible an unforced sudden takeoff is though, and I'm inclined (because of efficient market intuitions) to think the strong version of it is not that likely. H/t Owen Cotton-Barratt for this thought. ↵
10. E.g., that whenever there are two or more generalizations equally consistent with the training data so far, models will never generalize in the way that seems more natural or right to humans. ↵
11. I think eventually gridworlds and games will probably fade away as it becomes more practical to work with larger models instead, and dynamics like the treacherous turn start to show up in messier real-world settings. ↵
12. One idea a couple of others have suggested here and which I'm generally interested in is "transparency in (narrowly superhuman) language models": finding ways to understand "what models are thinking and why," especially when they know more about something than humans do. I like this idea but am very unsure about what execution could look like. E.g., would it look like [Chris Olah's work](#), which essentially "does neuroscience" on neural networks? Would it look like training models to answer our questions about what they're thinking? Something else? ↵
13. Though you could think that in an absolute sense it and all the other approaches that aren't tackling treachery head-on are doomed. ↵
14. I would also prefer other things being equal that EAs focused on long-run x-risk get the recognition for this work rather than others, but as I said above I consider this secondary and think that this agenda is good on the merits, not just as career capital for EAs. ↵
15. There are some innovators for whom the value of being in an area is strictly decreasing in its crowdedness, because their main value-add is to "start something from nothing." But I don't think that applies to most contributors, even those who have an extremely large impact eventually (which might even be larger than the innovators' impact in some cases). ↵
16. Some people have argued that the "verifying long-run solutions" path is dominant because the other stuff is likely to happen anyway, but I'm not

convinced. I think all three paths to impact that I laid out are likely to happen one way or another, and there's room to speed up or improve all of them. I do think there could be some boost to the "verifying long-run solutions" path, but all in all I feel like it'll be $\frac{1}{3}$ to $\frac{3}{4}$ of the value, not >90% of the value. ↵

17. The most plausible competing pitch in my mind is "get language models to answer questions honestly", which seems like it could get at the "ascription universality" / "knowing everything the model knows" concept (h/t Evan H, Owen C-B, Owain E). That would narrow the focus to language models and question-answering, and rule out projects like "get non-coders to train a coding model." I think the "get language models to answer questions honestly" frame is reasonable and I want to see work done under that banner too, but I'm not convinced it's superior. It considerably narrows the scope of what's "in", cutting down on long-run field growth potential, and I think a lot of the projects that are "out" (like the coding project) could be helpful and informative. I also worry that the tagline of "honesty" will encourage people to focus on "avoiding harmful lies that are nonetheless pretty easy for humans to detect", rather than focusing on regimes where models exceed human performance (see [this objection](#) for more discussion of that). ↵
18. It's possible other places, like Google Brain or some other FAANG lab, would also have roles available doing this type of work -- I am just more unsure because there is less of a long-termist alignment researcher presence in those places. ↵
19. Eventually, when models are more strongly superhuman, I think it will get too hard to even tell whether *outcomes* were acceptable, because AI systems could e.g. compromise the cameras and sensors we use to measure outcomes. So relying on outcomes earlier on feels like "kicking the can down the road" rather than "practicing what we eventually want to be good at." "Don't kick the can down the road, instead practice what we eventually want to be good at" is the overall ethos/attitude I'm going for with this proposal. ↵

What's So Bad About Ad-Hoc Mathematical Definitions?

Suppose it's the early twentieth century, and we're trying to quantify the concept of "information". Specifically, we want to measure "how much information" one variable contains about another - for instance, how much information a noisy measurement of the temperature of an engine contains about the actual engine temperature.

Along comes [Karl Pearson](#), and suggests using his "correlation coefficient" (specifically the square of the correlation coefficient, $\rho(X, Y)^2$). As a measure of information, this has some sensible properties:

- If there's no information, then $\rho(X, Y)^2$ is zero.
- If $\rho(X, Y)^2$ is one, then there's perfect information - one variable tells us everything there is to know about the other.
- It's symmetric: the amount of information which X tells us about Y equals the amount of information which Y tells us about X.

As an added bonus, it's mathematically simple to calculate, estimate, and manipulate. Sure, it's not very "principled", but it seems like a good-enough measure to work with.



Karl Pearson. He'd make a solid movie villain; I get sort of a Tywin Lannister vibe.

Now an engineer from Bell Telephone shows up with a real-world problem: they've been contracted to create secure communications for the military. They want to ensure that externally-visible data Y contains no information about secret message X, so they need a

way to measure “how much information” one variable contains about another. What a perfect use-case! We advise them to design their system so that X and Y have zero correlation.

A few years later, Bell Telephone gets a visit from a very unhappy colonel. Apparently the enemy has been reading their messages. Zero correlation was not enough to keep the secret messages secret.

Now, Bell *could* patch over this problem. For instance, they could pick a bunch of functions like X^2 , $\sin(Y)$, $e^X + 2X - 1$, etc, and require that those *also* be uncorrelated. With enough functions, and a wide enough variety, that might be enough... but it’s going to get very complicated very quickly, with all these new design constraints piling up.

Fortunately, off in a corner of Bell Labs, one of their researchers already has an alternative solution. Claude Shannon suggests quantifying “how much information” X contains about Y using his “mutual information” metric $I(X;Y)$. This has a bunch of sensible properties, but the main argument is that $I(X;Y)$ is exactly the difference between the average number of bits one needs to send in a message in order to communicate the value of X, and the average number of bits one needs to send to communicate X if the receiving party already knows Y. It’s the number of bits “savable” by knowing Y. By imagining different things as the “message” and thinking about how hard it is to *guess* X after knowing Y, we can intuitively predict that this metric will apply to lots of different situations, including Bell’s secret message problem.



Claude Shannon. Note the electronics in the background; this guy is my kind of theorist. No ivory tower for him.

Shannon advises the engineers to design their system so that X and Y have zero mutual information. And now, the enemy can't read their messages quite so easily.

Proxies vs Definitions

In this story, what does the correlation coefficient do "wrong" which mutual information does "right"? What's the generalizable lesson here?

The immediate difference is that correlation is a *proxy* for amount of information, while mutual information is a true definition/metric. When we apply optimization pressure to a proxy, it breaks down - that's [Goodheart's Law](#). In this case, the optimization pressure is a literal adversary trying to read our secret messages. The optimizer finds the corner cases where our proxy no longer perfectly captures our intuitive idea of "no information", and they're able to extract information about our secret messages. Correlation doesn't capture our intuitive notion of "information which X contains about Y" well enough for zero correlation to prevent our adversaries from reading our messages.

Mutual information, by contrast, handles the optimization pressure just fine. We intuitively expect that "Y contains zero information about X" is enough to keep our messages secret, even in the presence of adversaries, and the mutual information definition of "information" is indeed enough to match that intuitive expectation.

So... that's all well and good. We want definitions/metrics which are robust to optimization pressure, rather than proxies which break down. But how do we find robust definitions/metrics in the first place? In the long run, of course, we can try out a metric on lots of different problems, prove lots of different theorems about it, and get an idea of robustness that way. But there are infinitely many possible metrics for any given concept; we don't have time to go through that whole process for all of them. How do we figure out *in advance* what the robust concept definitions are?

You Already Know The Answer

A classic quote from famed physicist John Archibald Wheeler: "Never make a calculation until you know the answer".

In math, it's very easy to write down some expressions or equations or definitions, and start pushing symbols around, without having any idea what the answer looks like or how to get there. In undergrad math classes, this often works, because the problem is set up so that there's only a handful of things which you can do at all. In research, we don't have that guardrail, and we *especially* don't have that guardrail when finding the right definitions is part of the problem. I have literally spent months pushing symbols around without getting anywhere at all. [Math is a high-dimensional space; brute force search does not work.](#)

Bottom line: if we want to get anywhere, we need to already have at least *some* intuition for what we're looking for, and we need that intuition to guide the search. "Never make a calculation until you know the answer" is the sort of lesson which gets beaten in by months or years of failure to follow it.

Fortunately, we already have a *lot* of intuition to lean on, even without years of mathematical study. For instance, if we look back at the information example from earlier... what are the intuitive arguments for why correlation seems like a reasonable measure of information?

- If there's no information, then $\rho(X, Y)^2$ is zero.

- If $\rho(X, Y)^2$ is one, then there's perfect information - one variable tells us everything there is to know about the other.

These seemed pretty natural, right? This is exactly what “knowing the answer” looks like - we have some intuition about what properties a measure of “information” should have. In the case of mutual information, the intuition was this argument:

$I(X; Y)$ is exactly the difference between the average number of bits one needs to send in a message in order to communicate the value of X , and the average number of bits one needs to send to communicate X if the receiving party already knows Y . It's the number of bits “savable” by knowing Y . By imagining different things as the “message” and thinking about how hard it is to guess X after knowing Y , we can intuitively guess that this metric will apply to lots of different situations...

These are the kinds of intuitions which guide our search in the high-dimensional space of mathematical definitions/metrics.

Note that the engineers' idea that “data Y contains no information about secret message X ” should be sufficient to prevent adversaries from reading the messages is also an intuitive property of information. Assuming our intuitions about information are correct (or at least approximately correct), a definition which fully captures our intuitive idea of information should imply this property. If it doesn't, then either (a) our definition does not fully capture our intuitive idea of information, or (b) our intuition is wrong (in which case we should be able to translate the math back into an intuitive example of how our previous intuition failed).

... But Have You Fully Specified The Answer?

So, math is high-dimensional, we need intuitions to guide our search. But both the correlation coefficient and mutual information have some intuitive arguments for why they're good measures of information. What's the difference? What makes one better than the other?

Let's go back to the two intuitive arguments for the correlation coefficient:

- If there's no information, then $\rho(X, Y)^2$ is zero.
- If $\rho(X, Y)^2$ is one, then there's perfect information - one variable tells us everything there is to know about the other.

Key thing to notice: $\rho(X, Y)^2$ is not the only metric which satisfies these two criteria. For instance, we could exponentiate X and Y and then take the correlation, $\rho(e^X, e^Y)^2$, and both properties still apply. Same with $\rho(X^2, \sin(Y) + 2Y - 3)^2$. There's lots of degrees of freedom here; these two intuitive arguments are not enough to uniquely specify the correlation coefficient as our definition/metric.

By contrast, consider Shannon's argument:

$I(X; Y)$ is exactly the difference between the average number of bits one needs to send in a message in order to communicate the value of X , and the average number of bits one needs to send to communicate X if the receiving party already knows Y .

This has zero degrees of freedom. This argument (with a couple approximations) is enough to uniquely specify Shannon's formula for mutual information.

Adam Shimi [gave a great analogy for this](#): the intuitive arguments are like a set of equations, and the definition/metric is like a solution. Ideally, we want the "equations" to nail down one unique "solution". If that's the case, then there's only one definition compatible with our intuitive arguments. If we intuitively expect some additional properties to hold (e.g. "no information" being sufficient to prevent adversaries from reading our secret messages), then either they have to hold for that one definition, or our intuition is wrong.

On the other hand, if our "equations" have multiple "solutions", then it's kind of misleading to pick out one solution and declare that to be our answer. Why that solution? If there's lots of different definitions/metrics which satisfy the intuitive arguments for correlation, then why not use one of the others? More to the point: how do we know our intuition itself isn't built around some other metric which satisfies the properties? We believe our intuitive concept satisfies the listed properties, and we believe our intuitive concept satisfies some more general properties as well (e.g. "no information" protecting secret messages"), but that does not mean that *any* random definition compatible with the listed properties is sufficient to imply the more general properties. If we want our intuition to apply, then we need to find the definition/metric which actually corresponds to our intuitive concept (assuming such a definition/metric exists), not just some proxy which satisfies a few of the same properties.

Recap

We want mathematical definitions/metrics which are robust - in particular, they should not break down when we apply optimization pressure. In the long run, we can verify robustness by using a definition/metric in lots of different problems and proving theorems about it. But math-space is high dimensional, so we need a more efficient way to search for good definitions/metrics.

One main way we do this is to lean on intuitions. We already have intuitive concepts, and we have some beliefs about the properties those concepts should have. If we can accurately translate our intuitive concepts into mathematical definitions/metrics, then they should satisfy the intuitively-expected properties. (Or else our intuitions are wrong, and a good definition/metric should convince us of that when the definition doesn't satisfy an expected property.)

The key challenge here is to come up with a set of intuitive arguments which *uniquely* specify a particular definition/metric, exactly like a set of equations can uniquely specify a solution. If our arguments have "many solutions", then there's little reason to expect that the ad-hoc "solution" we chose actually corresponds to our intuitive concept. If our chosen definition/metric does not correspond to our intuitive concept, then even if our intuition is correct, it shouldn't be too surprising if the definition/metric fails to have more general properties which we intuitively expect.

In short: if our arguments are not sufficient to uniquely nail down one definition/metric, then we lose our main reason to expect the definition/metric to be robust.

Thankyou to Adam Shimi for a conversation which led to this post.

Another RadVac Testing Update

Previously: [Making Vaccine](#), [Commercial Antibody Test Results](#), [Mini-Update](#)

I've now run 9 ELISA tests. The main result is noise: negative controls are all over the map, sometimes very blue (i.e. positive), sometimes not blue at all. I did see more positive results in the experimental group than I'd expect from noise alone, but I haven't gotten the noise to a point where results are consistently reproducible.

Meanwhile, I also ran one very simple test: I snorted a batch of the peptides, without the chitosan or anything else - just peptides in deionized water. Previously, on doses 3-6 of the vaccine, I had consistently been congested for a couple days after (and *not* congested the rest of the week), which strongly indicates an immune response. However, that response could have been to the chitosan or other contents of the vaccine, rather than the peptides. This test put that possibility to rest: after snorting just the peptides, I was very obviously congested for a couple days, in basically the same way as after the vaccine doses.

So thanks to that simple test, I personally am now pretty highly confident that I have an immune response to these peptides. Unfortunately it's not as *legible* as an ELISA test, so you should not necessarily be quite as convinced by this.

Now, this still leaves the question of whether an immune response to these peptides translates into an immune response to COVID. It could be that e.g. the conformation of the peptides' corresponding sequence within full COVID proteins is different enough that it doesn't carry over. Personally, though, I consider this a much less likely failure mode, for two reasons. First, the white paper indicates that the peptides were chosen based on antibodies developed by people who actually had COVID. Second, whether antibodies against these peptides bind the real proteins is something which I would not expect to vary much from person to person, so if it's worked for a few people it should work for everyone - and the whitepaper does indicate that multiple groups have seen positive results testing for binding against the full proteins.

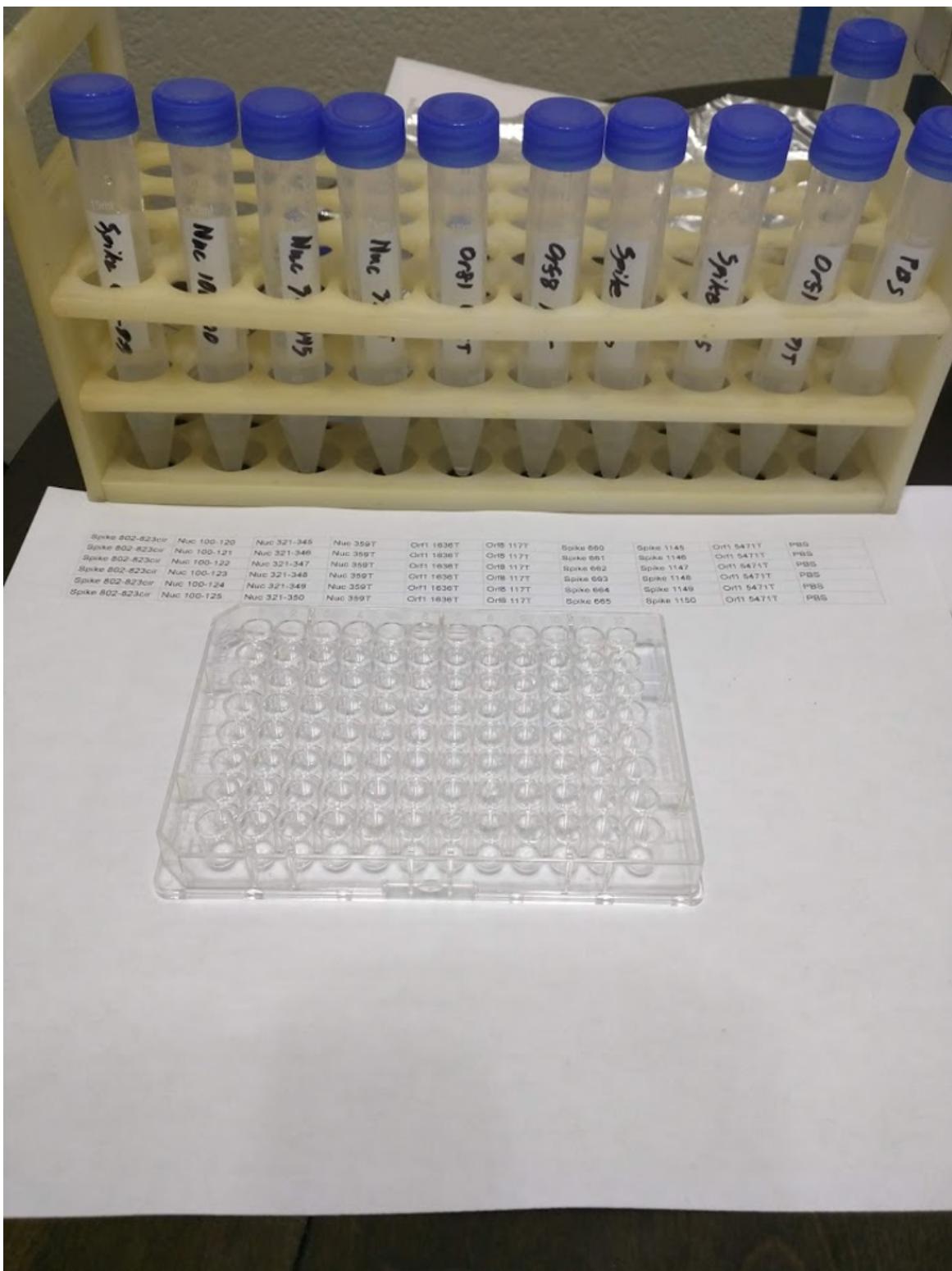
None of this puts my confidence up close to 99%, but I'm now considerably more confident that the vaccine worked (~90%). Also, that confidence is distributed over fewer possible worlds - e.g. based on the info in the latest version of the RadVac whitepaper, I now very much doubt that the vaccine will induce a response in the blood (unless it's injected). I also now have very little weight on the possibility that it works for some people sometimes but didn't work for me specifically, so additional dakka is not needed (at least for me).

The next section will be a bit more detail on the ELISA tests, for people who are curious about exactly how that sausage was made.

ELISA Tests

This section is an abbreviated chronology of what-I-saw and my reasoning about it; it's intended to show how I came to the conclusions I did. I expect most people will not find it very interesting, but one of the benefits of blog posts is that I can show all the questionable decisions and opportunities for confirmation bias to sneak in, so that's what I'm doing.

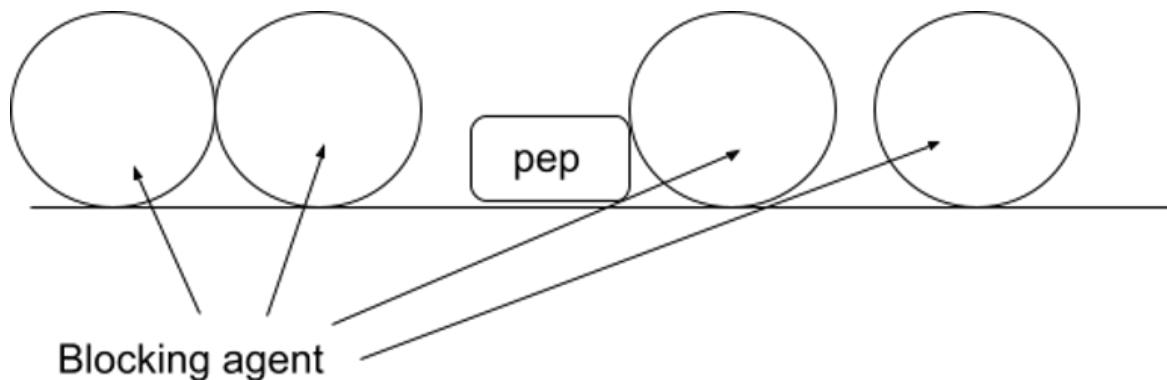
First, some background on how these tests work in theory. We start with a "high binding plate" - basically some plastic treated so that proteins/peptides stick to it.



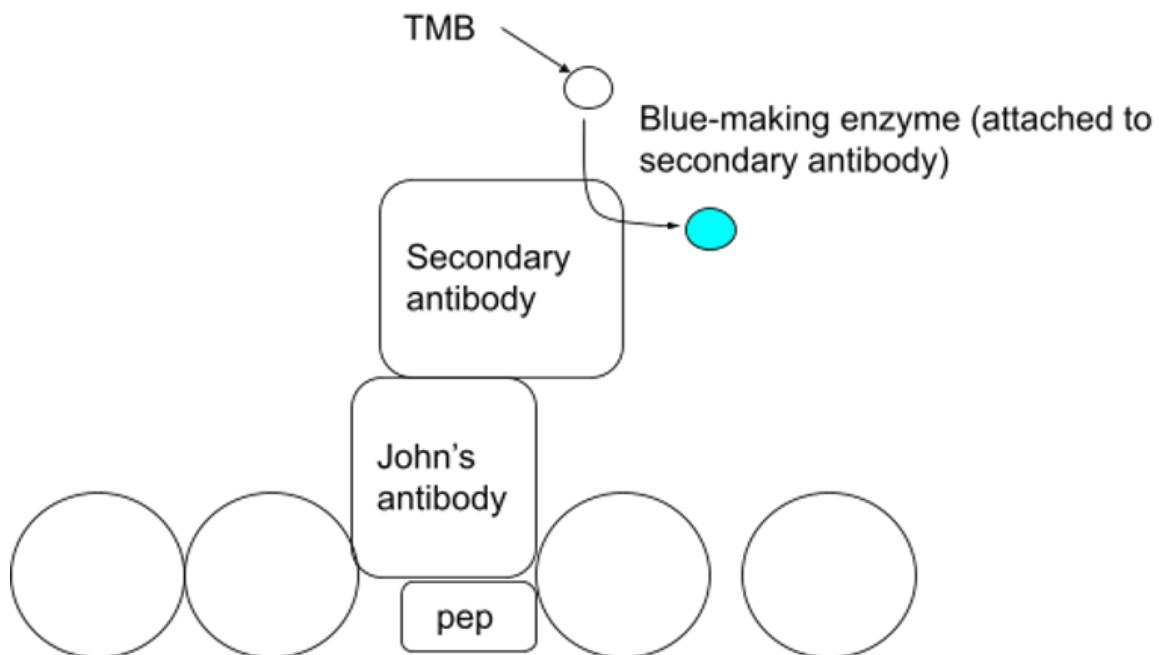
That's the plate; each of the holes is called a "well", and is basically a mini-test-tube with a high-binding surface.

We add a solution of our peptides, and some of them stick to the surface. Next, we dump that solution out, leaving behind only the peptides which bound to the surface. We add some

"binding solution" - in this case nonfat dry milk with a little detergent in it. The proteins in the binding solution fill whatever space on the surface was not taken up by the peptides.



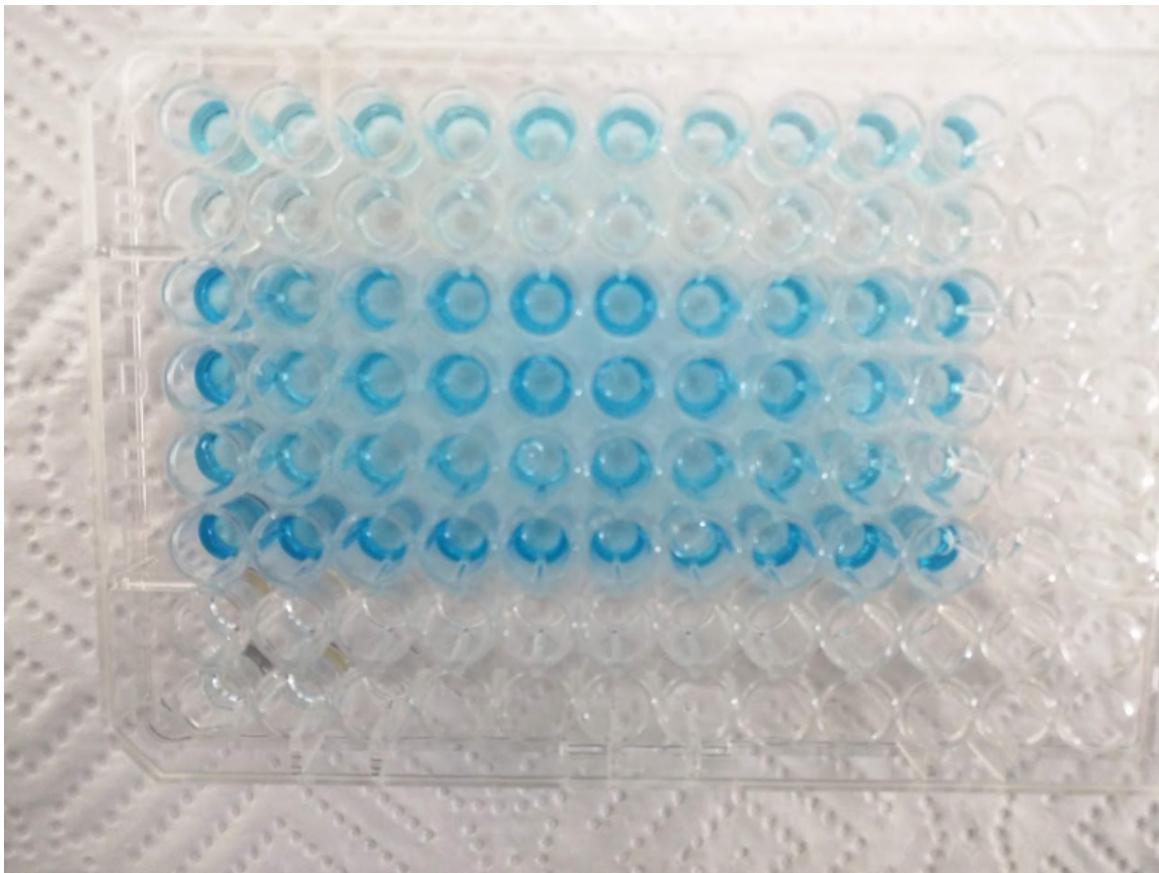
Now, with the foundation in place, we build a tower. We add a nasal wash sample from my nose, which hopefully contains antibodies that bind to the peptides. Then we dump that out, leaving behind only the antibodies which bound to a peptide attached to the plate. Next we add the "secondary antibodies", which bind to my antibodies and have an enzyme attached to them. Then we dump that solution out too. If all goes well, this leaves our "tower": peptide bound to the plate, antibody bound to the peptide, secondary bound to the antibody.



The final step is to add some TMB solution. The enzyme attached to the secondary antibody will turn the TMB blue, so if we see blue after a few minutes, then we know the secondary antibodies are present, and hopefully that means the rest of the tower is present too.

So how does this play out in practice?

Here's the first plate I ran:



Each row is a different sample (nasal wash and saliva from both myself and my girlfriend, blood from me, and one more concentrated nasal sample). Each column is a different peptide, with no-peptide control group in the right-most column. From this, we learn three things. First, there's a lot of variation between samples. Second, the variation we're seeing has nothing to do with the peptides. Third, and most important, the negative controls contain an awful lot of blue. So this does not match our theoretical picture of building-a-tower; somehow, something in the sample is sticking in the plate, and it's sticking to something other than the peptide.

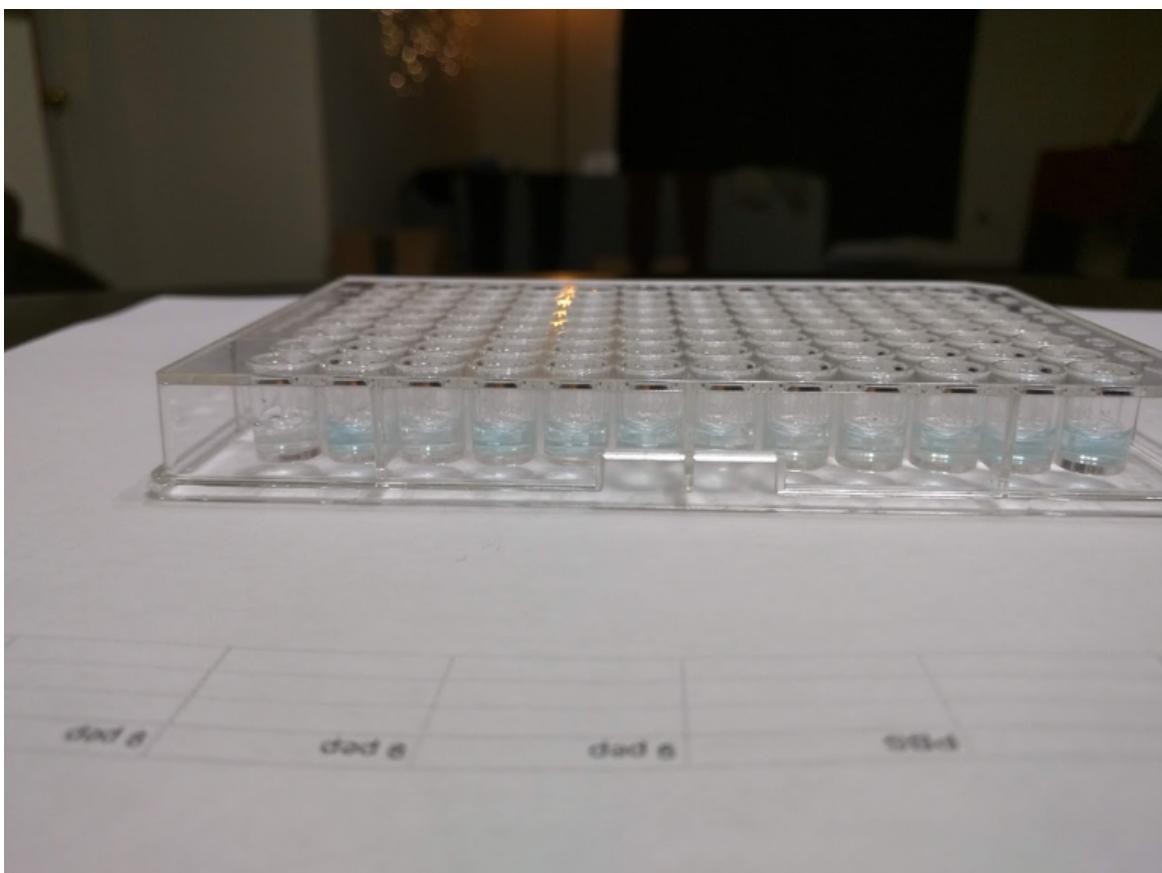
Next came a fair bit of trying stuff. I tried both more concentrated peptides, and dosing myself with RADVAC the day before taking a sample in order to induce more antibody production, both in hopes of increasing the signal enough to overcome the noise. Results were basically similar. I tried a bunch of different blocking agents, without any peptides at all, to see which gave the least-blue negative controls - dried milk was actually one of the best, although egg whites were better. So I tried using egg whites for the blocking solution, and results were basically similar. I ran another plate just comparing various negative controls, without any major new insight.

In general, I still saw some samples produce generally-more or generally-less blue, across both experimental and negative control groups, for no clear reason. Generally, things varied more from test-to-test and sample-to-sample than within similar treatment groups on the same sample/test - i.e. the noise is mostly systematic.

The main result from all this was that wells with no sample were pretty consistently not blue. Whatever causes my no-peptide negative controls to turn blue, it definitely involves something in the sample binding to something other than the peptide.

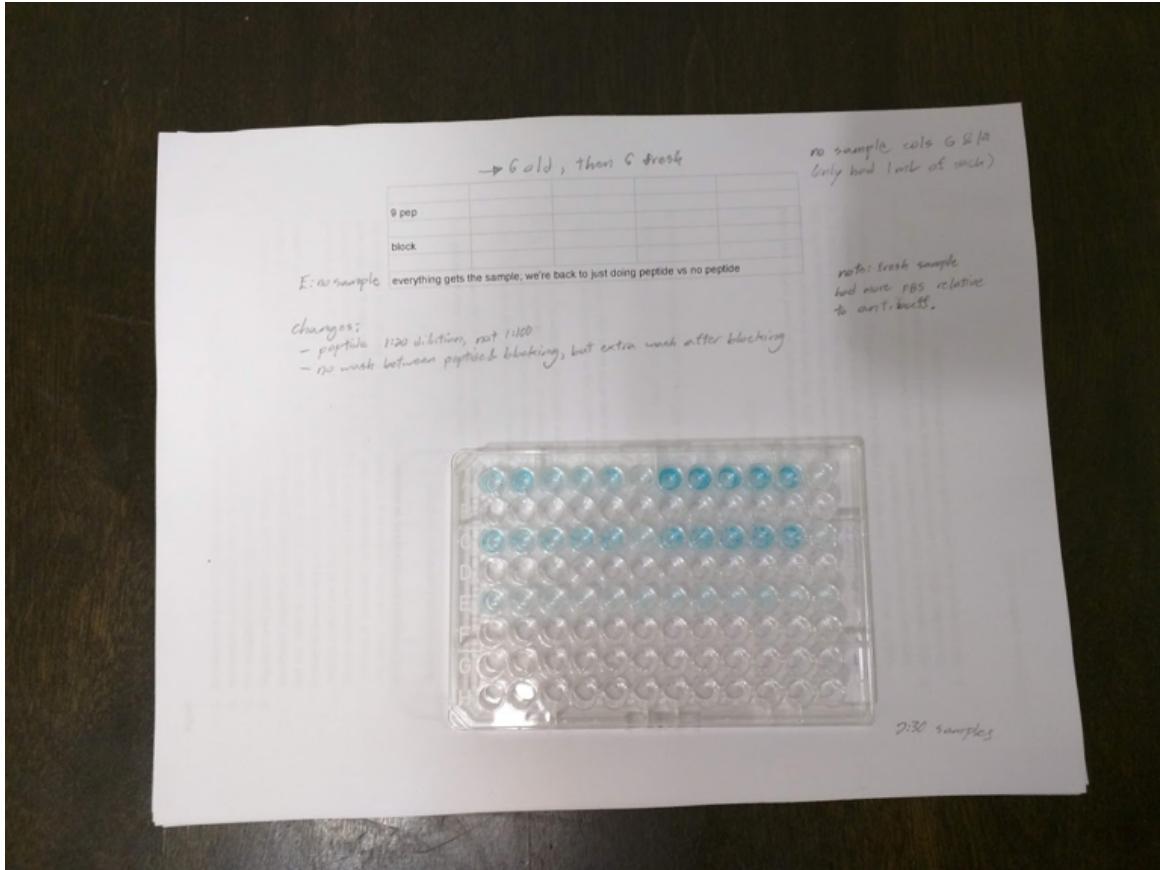
I did some reading online and some thinking about how to reduce the sample binding in the negative control group. I shifted to a mental model where a significant fraction of the peptide/binding agent on the plate was actually coming loose and being replaced by whatever was in solution. One way to reduce noise in the control, then, is to include a little binding agent in the antibody solution and secondary antibody solution - so if a space on the plate opens up, it will most likely be filled by the extra binding agent rather than the antibody. (This was included in the protocol, but listed as "optional", and I didn't understand before why it would be useful.) I also increased the amount of binding solution used (so that it covered the sides of the well more completely), and made extra sure to not accidentally use 100uL rather than 200 uL of binding solution (all of the other steps involve 100 uL in the well, so it's an easy mistake to make if not paying attention - I think I probably made this mistake multiple times in earlier tests). Finally, I removed the wash step between peptide and binding solution - I see no way that that one could reduce noise, and I'd expect it to reduce the signal.

With all that in place, I ran two more tests. One saw generally very little blue, but there was more blue in the experimental wells than the controls:



Left-to-right, we have three no-peptide controls, then three experimental wells, then three more no-peptide controls, then three more experimental wells. There is definitely some blue in the no-peptide controls (especially the second well from the left), but there is generally more blue in the experimental wells. Note that this test was run with a sample which had been frozen; I think that's the most likely cause of the generally-faint blueness.

Then we got the most promising result:



The group of five experimental wells in the upper right was visibly more blue than the corresponding no-peptide negative controls below them. (The group on the left is a sample which was frozen; so now we know to use fresh samples. This is another mistake I made multiple times in earlier tests.)

On the other hand, I'm still not able to get consistent results. I ran one more test the next day, just a single fresh sample with a whole bunch of experimental wells and no-peptide controls, and there was no visible difference between the controls and the experimental wells:



Experimental and control wells are in alternating rows. The last two columns are no-sample controls.

So, bottom line: the two positive results, especially the second, are more than I'd expect to see from noise after having run this many tests. But they're still definitely not knock-down unambiguous evidence, and there's still a lot of test-to-test variability which I'm unable to account for.

Going Forward

Mainstream vaccines will be available to the general populace here starting two weeks from today, so this project is probably reaching its end soon. We may run another test or two, but I probably won't have another post. Conditional on any more tests, I will put up a shortform summarizing whatever results we see.

My research methodology

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

(Thanks to Ajeya Cotra, Nick Beckstead, and Jared Kaplan for helpful comments on a draft of this post.)

I really don't want my AI to strategically deceive me and resist my attempts to correct its behavior. Let's call an AI that does so **egregiously misaligned** (for the purpose of this post).

Most possible ML techniques for avoiding egregious misalignment depend on detailed facts about the space of possible models: what kind of thing do neural networks learn? how do they generalize? how do they change as we scale them up?

But I feel like we should be possible to avoid egregious misalignment regardless of how the empirical facts shake out--it should be possible to get a model we build to do at least roughly what we want. So I'm interested in trying to solve the problem in the worst case, i.e. to develop competitive ML algorithms for which we can't tell *any* plausible story about how they lead to egregious misalignment.

This is a much higher bar for an algorithm to meet, so it may just be an impossible task. But if it's possible, there are several ways in which it could actually be easier:

- We can potentially iterate much faster, since it's often easier to think of a single story about how an algorithm can fail than it is to characterize its behavior in practice.
- We can spend a lot of our time working with simple or extreme toy cases that are easier to reason about, since our algorithm is supposed to work even in these cases.
- We can find algorithms that have a good chance of working in the future even if we don't know what AI will look like or how quickly it will advance, since we've been thinking about a very wide range of possible failure cases.

I'd guess there's a 25–50% chance that we can find an alignment strategy that looks like it works, in the sense that we can't come up with a plausible story about how it leads to egregious misalignment. That's a high enough probability that I'm very excited to gamble on it. Moreover, if it fails I think we're likely to identify some possible "hard cases" for alignment—simple situations where egregious misalignment feels inevitable.

What this looks like (3 examples)

My research basically involves alternating between "think of a plausible alignment algorithm" and "think of a plausible story about how it fails."

Example 1: human feedback

In [an unaligned benchmark](#) I describe a simple AI training algorithm:

- Our AI observes the world through a bunch of cameras and outputs motor actions. We train a generative model that predicts these camera observations given the motor actions.
- We ask humans to evaluate possible futures by looking at the predicted videos output by the model. We then train a model to predict these human evaluations.
- At test time the AI searches for plans that lead to trajectories that look good to humans.

In the same post, I describe a plausible story about how this algorithm leads to egregious misalignment:

- Our generative model understands reality better than human evaluators.
- There are plans that acquire influence in ways that are obvious to the generative model but completely incomprehensible and invisible to humans.
- It's possible to use that influence to "hack" the cameras, in the sense of creating a fiction that looks convincing to a human looking at predicted videos.
- The fiction can look much better than the actual possible futures. So our planning process finds an action that covertly gathers resources and uses them to create a fiction.

I don't know if or when this kind of reward hacking would happen—I think it's pretty likely eventually, but it's far from certain and it might take a long time.

But from my perspective this failure mode is at least *plausible*—I don't see any contradictions between this sequence of events and anything I know about the real world. So this is enough for me to conclude that human feedback can't handle the worst plausible situation, and to keep looking for an algorithm that can.

To better understand whether this story is really plausible, we can spend time refining it into something more and more concrete to see if it still seems to make sense. There are lots of directions in which we could add detail:

- Specify how our model could fool humans—maybe it sets some configuration files in ways that introduce security vulnerabilities we can't understand, and sends an email that looks innocuous to us but is actually enlisting aid from an external attacker.
- Specify how an attacker could ultimately corrupt cameras and how that would lead to a higher reward—maybe we were trying to train an agent to design our new computing cluster, and it is able to corrupt the benchmarks we were using to estimate cluster performance.
- Specify how the algorithm works in more detail—maybe it uses MCTS with a learned value function for planning against the learned models, and we spend 10% of our budget on human evaluations of the policies output by the current planner, and so on.

Filling more and more details lets us notice if our abstract story was actually incoherent in important ways, or to notice weird things the story implies about the world that we might want to rule out by assumption.

Example 2: iterated amplification

To avoid the problems with raw human feedback, we could train additional ML assistants that help us evaluate outcomes. For example, assistants could point out possible consequences of a plan that we didn't notice. Various variants of this idea are

explored in [benign model-free RL](#), [supervising strong learners by amplifying weak experts](#), [AI safety via debate](#), and [recursive reward modeling](#).

In [inaccessible information](#) I tried to explore a story about how this entire family of algorithms could fail:

- Our AI learns to model the world by looking at a lot of videos, doing gradient descent over a bunch of possible physical models, and selecting a good one. It might model physics with a complicated mix of low-level physics and high-level abstractions.
- This model lets it predict data like camera observations, but also predict “off-screen” events like what code is running on my computer (even if the computation be directly observed on video). These predictions are economically useful and don’t have any close substitute.
- The process of finding this model of the world is much more complex than anything our AI can do at inference time, and the intermediate results are too complex and numerous to be “memorized” in the weights of our trained AI. So there doesn’t seem to be any way to break the model-finding work into pieces that can be delegated to an ML assistant (in amplification) or a debater (in debate).
- If we decide that we don’t trust this physical model and so don’t use it, then we aren’t economically competitive. But if we decide that we *do* trust the model, then we are back in the unfortunate situation from the last time: our AI might find a plan in which the “off-screen” information (computers outside of our view) are taking actions that ultimately hack our cameras and create a compelling fiction that we will rate highly. Because amplification gives us no way to access that off-screen data, we appear to have no way to tell if the plan is bad.

Example 3: imitative generalization

[Imitative generalization](#) is intended to address this problem with iterated amplification.

To briefly summarize: instead of using gradient descent to search over a space of human-incomprehensible models that predict some data (e.g. autoregressive models of videos), we try to search over space of models that a human can “understand” (perhaps with the help of aligned assistants as in amplification or debate), and optimize for a model that both looks plausible to the human and allows the human to successfully predict the same data (i.e. to predict the next pixel of a video). We hope that this allows us to find a human-comprehensible model that allows the human to both predict the data and figure out if the camera is being hacked.

(This algorithm is quite vague, so you could think of it as a whole family of algorithms based on how you parametrize the space of “human-comprehensible” models, how you search over that space, and how you define the prior. I’m going to try to tell a story about the limitations of this whole approach.)

Here’s an exotic situation where I think the naive version of this approach wouldn’t work:

- Our AI learns an extremely low-level model of the world defined in terms of concepts like “atoms.” It makes predictions about high-level concepts (like human behavior) by simulating the lower-level parts of which they are made.
- There is an easy way to read our dataset itself off of the low level simulation— perhaps you can read bits off of our computer easily when expressed in the low-

level simulation. So our AI learns how to use that extraction procedure without learning anything else explicit about the mapping between high-level and low-level concepts.

- Any model which allows the human to predict the video must include all of the low-level facts. But that means that any further high-level facts (e.g. about how the low-level structure relates to the high-level things we care about) can only decrease the prior probability without any compensating increase in the ability to explain data.

There are many obvious ways to try to address this problem, but I think it does break the most obvious implementations of imitative generalization. So now I have two questions:

- Is this story really plausible? It looks bad, but maybe as we flesh it out we'll discover that it's implausible.
- Refine imitative generalization so that it handles the situation in this situation.

After a little bit of inspection it turns out that the original story is inconsistent: it's literally impossible to run a detailed low-level simulation of physics in situations where the computer itself needs to be part of the simulation. So the story as I told it is inconsistent, and we can breathe a temporary sigh of relief.

Unfortunately, the basic problem persists even when we make the story more complicated and plausible. Our AI inevitably needs to reason about some parts of the world in a heuristic and high-level way, but it could still use a model that is lower-level than what humans are familiar with (or more realistically just alien but simpler). And at that point we have the same difficulty.

It's possible that further refinements of the story would reveal other inconsistencies or contradictions with what we know about ML. But I've thought enough about this that I think this failure story is probably something that could actually happen, and so I'm back to the step of improving or replacing imitative generalization.

This story is even more exotic than the ones in the previous sections. I'm including it in part to illustrate how much I'm willing to push the bounds of "plausible." I think it's extremely difficult to tell completely concrete and realistic stories, so as we make our stories more concrete they are likely to start feeling a bit strange. But I think that's OK if we are trying to think about the worst case, until the story starts contradicting some clear assumptions about reality that we might want to rely on for alignment. When that happens, I think it's really valuable to talk concretely about what those assumptions are, and be more precise about why the unrealistic nature of the story excuses egregious misalignment.

More general process

We start with some [unaligned “benchmark”](#). We rule out a proposed alignment algorithm if we can come up with any story about how it can be *either* egregiously misaligned or uncompetitive.

I'm always thinking about a stable of possible alignment strategies and possible stories about how each strategy can fail. Depending on the current state of play, there are a bunch of different things to do:

- If there's a class of algorithms (like imitative generalization) for which I can't yet tell any failure story, I try to tell a story about how whole the class of algorithms would fail.
- If I can't come up with any failure story, then I try to fill in more details about the algorithm. As the algorithm gets more and more concrete it becomes easier and easier to tell a failure story.
- The best case is that we end up with a precise algorithm for which we still can't tell any failure story. In that case we should implement it (in some sense this is just the final step of making it precise) and see how it works in practice.
- More likely I'll end up feeling like all of our current algorithms are doomed in the worst case. At that point I try to think of a new algorithm. For this step, it's really helpful to look at the stories about how existing algorithms fail and try to design an algorithm that handles those difficulties.
- If all of my algorithms look doomed and I can't think of anything new, then I try to really dig in on the existing failure stories by filling in details more concretely and exploring the implications. Are those stories actually inconsistent after all? Do they turn out to contradict anything I know about the world? If so, I may add another assumption about the world that I think makes alignment possible (e.g. [the strategy stealing assumption](#)), and throw out any stories that violate that assumption or which I now realize are inconsistent.
- If I have a bunch of stories about how particular algorithms fail, and I can't think of any new algorithms, then I try to unify and generalize them to tell a story about why alignment could turn out to be impossible. This is a second kind of "victory condition" for my work, and I hope it would shed light on what the fundamental difficulties are in alignment (e.g. by highlighting additional empirical assumptions that would be necessary for any working approach to alignment).

Objections and responses

Can you really come up with a working algorithm on paper? Empirical work seems important

My goal from theoretical work is to find a credible alignment proposal. Even from that point I think it will take a lot of practical work to get it to the point where it works well and we feel confident about it in practice:

- I expect most alignment schemes are likely to depend on some empirical parameters that need to be estimated from experiment, especially to argue that they are competitive. For example, we may need to show that models are able to perform some tasks, like modeling some aspects of human preferences, "easily enough." (This seems like an unusually easy claim to validate empirically —if we show that our 2021 models can do a task, then it's likely that future models can as well.) Or maybe we've argued that the aligned optimization problem is only harder by a bounded amount, but it really matters whether it's 1.01 or 101 as expensive, so we need to measure this overhead and how it scales empirically. I've simplified my methodology a bit in this blog post, and I'd be thrilled if our alignment scheme ended up depending on some clearly defined and measurable quantities for which we can start talking about [scaling laws](#).
- I don't expect to literally have a proof-of-safety. I think at best we're going to have some convincing arguments and some years of trying-and-failing to find a plausible failure story. That means that empirical research can still turn up failures we didn't anticipate, or (more realistically) places where reality doesn't

quite match our on-paper picture and so we need to dig in to make sure there isn't a failure lurking somewhere.

- Even if we've correctly argued that our scheme is *workable*, it's still going to take a ton of effort to make it *actually work*. We need to write a bunch of code and debug it. We need to cope with the divergences between our conceptual "ML benchmark" and the messier ML training loops used in practice, even if those divergences are small enough that the theoretical algorithm still works. We need to collect the relevant datasets, even if we've argued that they won't be prohibitively costly. And so on.

My view is that working with pen and paper is an important first step that allows you to move quickly *until you have something that looks good on paper*. After that point I think you are mostly in applied world, and I think that applied investments are likely to ultimately dwarf the theoretical investments by orders of magnitude even if it turns out that we found a really good algorithm on paper.

That's why I'm personally excited about "starting with theory," but I think we should do theoretical and applied work in parallel for a bunch of reasons:

- We need to eventually be able to make alignment techniques in the real world, and so we want to get as much practice as we can. Similarly, we want to build and grow capable teams and communities with good applied track records.
- There's a good chance (50%?) that no big theoretical insights are forthcoming and empirical work is all that matters. So we really can't wait on theoretical progress.
- I think there's a reasonable chance of empirical work turning up unknown unknowns that change how we think about alignment, or to find empirical facts that make alignment easier. We want to get those sooner rather than later.

Why think this task is possible? 50% seems way too optimistic

When I describe this methodology, many people feel that I've set myself an impossible task. Surely *any* algorithm will be egregiously misaligned under some conditions?

My "50% probability of possibility" is coming largely from a soup of optimistic intuitions. I think it would be crazy to be confident on the basis of this kind of intuition, but I do think it's enough to justify 50%:

- 10 years ago this project seemed much harder to me and my probability would have been much lower. Since then I feel like I've made a lot of progress in my own thinking about this problem (I think that a lot of this was a personal journey of rediscovering things that other people already knew or answering questions in a way that was only salient to me because of the way I think about the domain). I went from feeling kind of hopeless, to feeling like [indirect normativity](#), formalized the goal, to thinking about evaluating [actions rather than outcomes](#), to believing that we can [bootstrap superhuman judgments using AI assistants](#), to understanding the [role of epistemic competitiveness](#), to seeing that all of these theoretical ideas appear to be practical for ML alignment, to seeing [imitative generalization](#) as a plausible approach to the big remaining limitation of iterated amplification.
- There is a class of theoretical problems for which I feel like it's surprisingly often possible to either solve the problem or develop a clear picture of why you can't. I don't really know how to pin down this category but it contains almost all of

theoretical computer science and mathematics. I feel like the “real” alignment problem is a messy practical problem, but that the worst-case alignment problem is more like a theory problem. Some theory problems turn out to be hard, e.g. it could be that worst-case alignment is as hard as P vs NP, but it seems surprisingly rare and even being as hard as P vs NP wouldn’t make it worthless to work on (and even for P vs NP we get various consolation prizes showing us why it’s *hard to argue that it’s hard*). And even for messy domains like engineering there’s something similar that often feels true, where given enough time we either understand how to build+improve a machine (like an engine or rocket) or we understand the fundamental limits that make it hard to improve further.

- So if it’s not possible to find any alignment algorithm that works in the worst case, I think there’s a good chance that we can say something about *why*, e.g. by identifying a particular hard case where we don’t know how to solve alignment and where we can say something about what causes misalignment in that case. This is important for two reasons: (i) I think that would be a really great consolation prize, (ii) I don’t yet see any good reason that alignment is impossible, so that’s a reason to be a bit more optimistic for now.
- I think one big reason to be more skeptical about alignment than about other theoretical problems is that the problem statement is incredibly imprecise. What constitutes a “plausible story,” and what are the assumptions about reality that an alignment algorithm can leverage? My feeling is that full precision isn’t actually essential to why theoretical problems tend to be soluble. But even more importantly, I feel like there *is* some precise problem here that we are groping towards, and that makes me feel more optimistic. (I discuss this more in the section “Are there any examples of this methodology working?”)
- Egregious misalignment still feels weird to me and I have a strong intuitive sense that we should be able to avoid it, at least in the case of a particular known technique like ML, if only we knew what we were doing. So I feel way more optimistic about being able to avoid egregious misalignment in the worst case than I do about most other theoretical or practical problems for which I have no strong feasibility intuition. This feasibility intuition also often does useful work for us since we can keep asking “Does *this* intermediate problem still feel like it should obviously be soluble?” and I don’t feel like this approach has yet led me into a dead end.
- Modern ML is largely based on simple algorithms that look good on paper and scale well in practice. I think this makes it much more plausible that alignment can also be based on simple algorithms that look good on paper and scale well in practice. Some people think of Sutton’s “[bitter lesson](#)” as bad news for the difficulty of alignment, and perhaps it is in general, but I think it’s great news if you’re looking for something really simple.

Despite having lots of optimistic words to say, feasibility is one of my biggest concerns with my methodology.

These failure stories involve very unrealistic learned models

My failure stories involve neural networks learning something like “simulate physics at a low level” or “perform logical deductions from the following set of axioms.” This is not the kind of thing that a neural network would learn in practice. I think this leads many people to be skeptical that thinking about such simplified stories could really be useful.

I feel a lot more optimistic:

- I don't think neural network cognition will be simple, but I think it will involve lots of the features that come up in simple cognition: powerful models will likely make cognitive steps similar to logical deduction, bayesian updating, modeling physics at some level of abstraction, and so on.
- If our alignment techniques don't work for simple cognition, I'm skeptical that they will work for complex cognition. I haven't seen any alignment schemes that leverage complexity *per se* in order to work. A bigger and messier model is more likely to have *some* piece of its cognition that satisfies any given desirable property—for example it's more likely to have particular neurons that whose behavior can be easily understood—but seems less likely to have *every* piece of its cognition satisfy any given desirable property.
- I think it's very reasonable to focus on *capable* models—we don't need to solve alignment for models that can't speak natural language or understand roughly what humans want. I think that's OK: we should imagine simple models being very capable, and we can rule out a failure story as implausible if it involves the model being too weak.
- I think it's more plausible for an alignment scheme to work well for simple cognition but fail for complex cognition. But in that case my methodology will just start with the simple cognition and move on to the more complex cognition, and I think that's OK.

**Are there any examples of a similar research methodology working well?
This is different from traditional theoretical work**

When theorists design algorithms they often focus on the worst case. But for them the “worst case” is e.g. a particular graph on which their algorithm runs slowly, not a “plausible” story about how a model is “egregiously misaligned.”

I think this is a real, big divergence that's going to make it way harder to get traditional theorists on board with this approach. But there are a few ways in which I think the situation is less disanalogous than it looks:

- Although the majority of computer science theorists work in closed, precisely defined domains, the field also has some experience with fuzzier domains where the definitions themselves need to be refined. For example, at the beginning of modern cryptography you could describe the methodology as “Tell a story about how someone learns something about your secret” and that only gradually crystallized into definitions like semantic security (and still people sometimes retreat to this informal process in order to define and clarify new security notions). Or while defining interactive and zero knowledge proofs people would work with more intuitive notions of “cheating” or “learning” before they were able to capture them with formal definitions.

I think the biggest difference is that most parts of theoretical CS move quickly past this stage and spend most of their time working with precise definitions. That said, (i) part of this is due to the taste of the field and the increasing unwillingness to engage in hard-to-formalize activities, rather than a principled take that you need to avoid spending long in this stage, (ii) although many people are working on alignment only very few are taking the kind of approach I'm advocating here, so it's not actually clear that we've spent so much more time than is typically needed in theoretical CS to formalize a new area (especially given that people in academia typically pick problems based on tractability).

- Both traditional theorists and I will typically start with a vague “hard case,” e.g. “What if the graph consists of two densely connected clusters with two edges in

between them?” They then tell a story about how the algorithm would fail in that case, and think about how to fix the problem. In both cases, the point is that you could make the hard case more precise if you wanted to—you can specify more details about the graph or you can fill in more details about the story. And in both cases, we learn how to tell vague stories by repeatedly going through the exercise of making them more precise and building intuitions about what the more precise story would look like. The big difference is that you can make a graph fully precise—you can exactly specify the set of vertices and edges—but you can never make a story about the world fully precise because there is just too much stuff happening. I think this really does mean that the traditional theorist’s intuition about what “counts” as a hard case is better grounded. But in practice I think it’s usually a difference in degree rather than kind. E.g., you very rarely need to actually write out the full graph in order to compute exactly how an algorithm behaves.

- Although the definition of a “plausible failure story” is pretty vague, most of the concrete stories we are working with can be made very specific in the ways that I think matter. For example, we may be able to specify completely precisely how a learned deduction process works (specifying the formal language L , specifying the “proof search order” it uses to loop over inferences, and so on) and why it leads to misalignment in a toy scenario.



[My research methodology](#) was originally published in [AI Alignment](#) on Medium, where people are continuing the conversation by highlighting and responding to this story.

Trapped Priors As A Basic Problem Of Rationality

Crossposted from [Astral Codex Ten](#)

Introduction and review

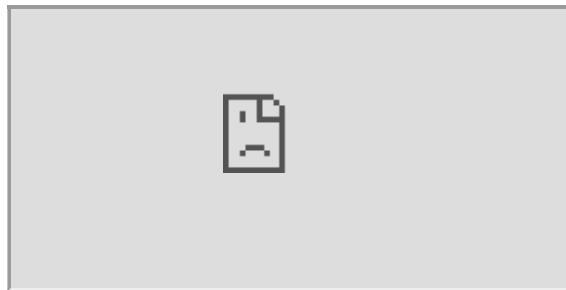
Last month I talked about van der Bergh et al's work on [the precision of sensory evidence](#), which introduced the idea of a *trapped prior*. I think this concept has far-reaching implications for the rationalist project as a whole. I want to re-derive it, explain it more intuitively, then talk about why it might be relevant for things like intellectual, political and religious biases.

To review: the brain combines raw experience (eg sensations, memories) with context (eg priors, expectations, other related sensations and memories) to produce perceptions. You don't notice this process; you are only able to consciously register the final perception, which feels exactly like raw experience.



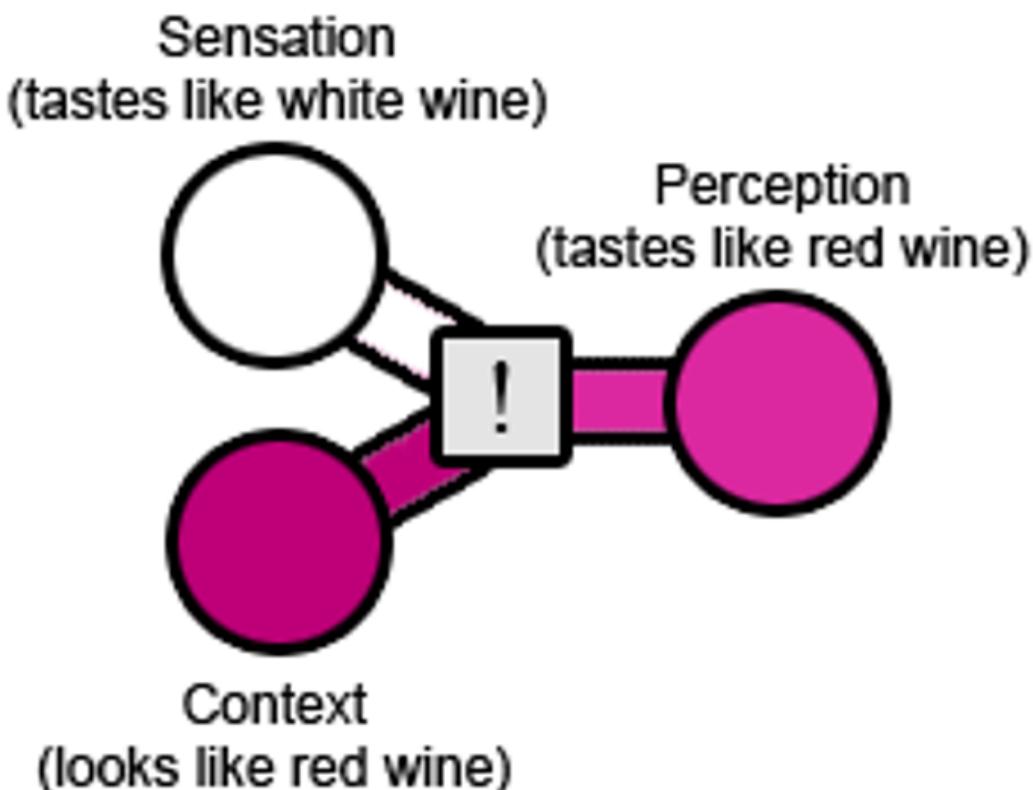
A typical optical illusion. The top chess set and the bottom chess set are the same color (grayish). But the top appears white and the bottom black because of the context (darker vs. lighter background). You perceive not the raw experience (grayish color) but the final perception modulated by context; to your conscious mind, it just seems like a brute fact that the top is white and the bottom black, and it is hard to convince yourself otherwise.

Or: maybe you feel like you are using a particular context independent channel (eg hearing). Unbeknownst to you, the information in that channel is being context-modulated by the inputs of a different channel (eg vision). You don't feel like "this is what I'm hearing, but my vision tells me differently, so I'll compromise". You feel like "this is exactly what I heard, with my ears, in a way vision didn't affect at all".



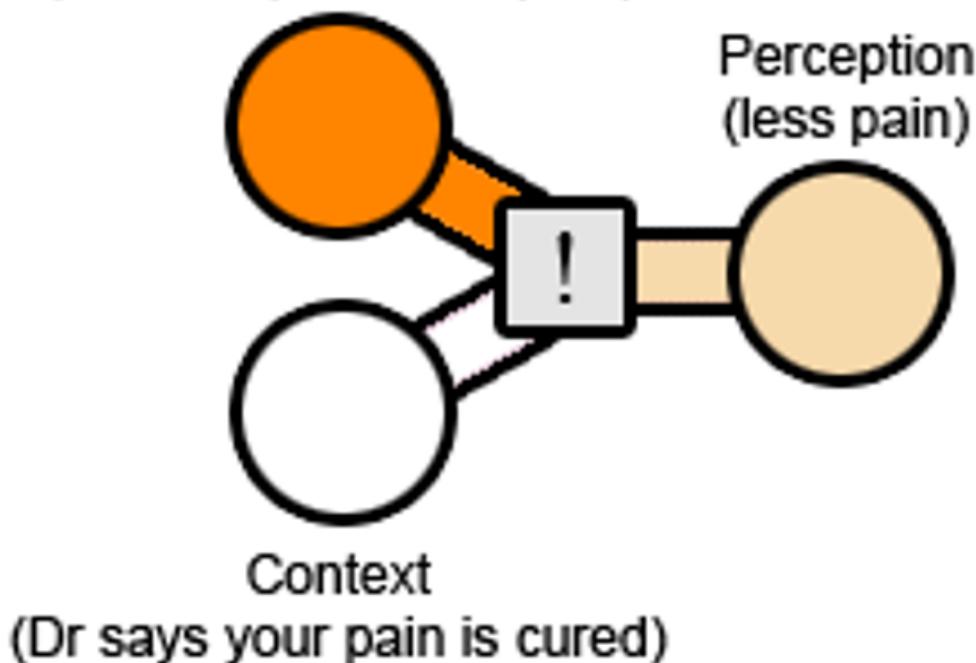
This is called the McGurk Effect. The man is saying the same syllable each time, but depending on what picture of his mouth moving you see, you hear it differently. Your vision is context-modulating your hearing, but it just sounds like hearing something.

The most basic illusion I know of is the Wine Illusion; dye a white wine red, and lots of people will say it tastes like red wine. The raw experience - the taste of the wine itself - is that of a white wine. But the context is that you're drinking a red liquid. Result: it tastes like a red wine.



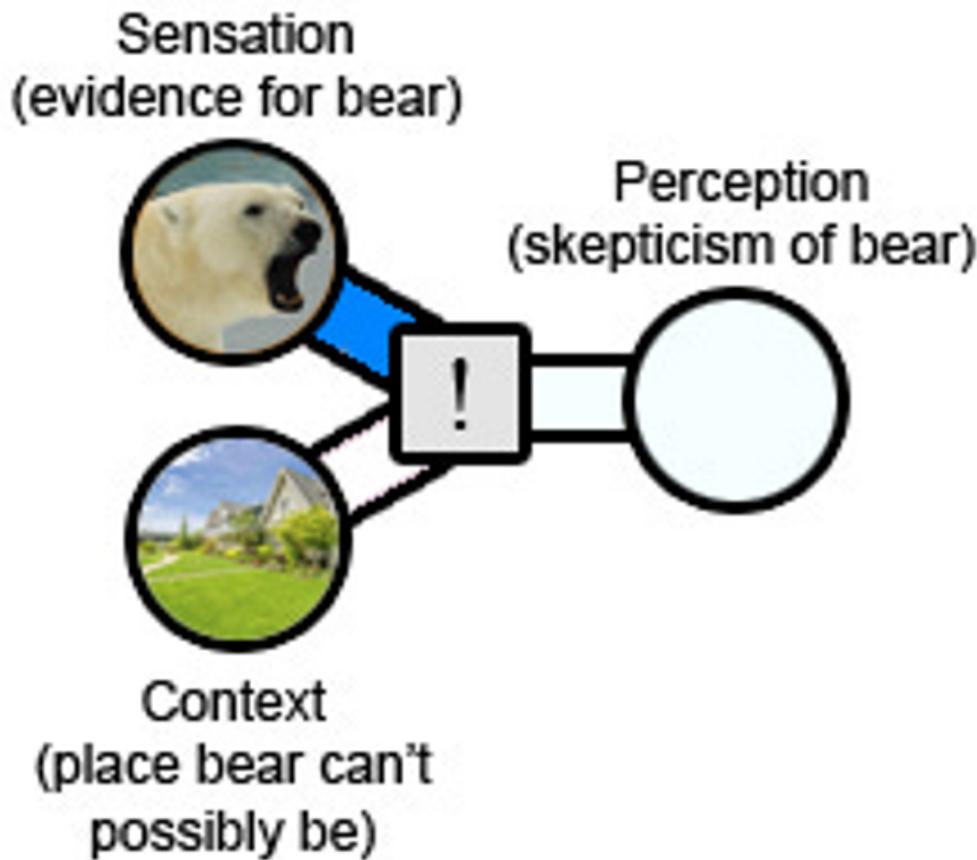
The placebo effect is almost equally simple. You're in pain, so your doctor gives you a "painkiller" (unbeknownst to you, it's really a sugar pill). The raw experience is the nerve sending out just as many pain impulses as before. The context is that you've just taken a pill which a doctor assures you will make you feel better. Result: you feel less pain.

Sensation
(nerve impulses for pain)



These diagrams cram a lot into the gray box in the middle representing a “weighting algorithm”. Sometimes the algorithm will place almost all its weight on raw experience, and the end result will be raw experience only slightly modulated by context. Other times it will place almost all its weight on context and the end result will barely depend on experience at all. Still other times it will weight them 50-50. The factors at play here are very complicated and I’m hoping you can still find this helpful even when I treat the gray box as, well, a black box.

The cognitive version of this experience is normal Bayesian reasoning. Suppose you live in an ordinary California suburb and your friend says she saw a coyote on the way to work. You believe her; your raw experience (a friend saying a thing) and your context (coyotes are plentiful in your area) add up to more-likely-than-not. But suppose your friend says she saw a polar bear on the way to work. Now you're doubtful; the raw experience (a friend saying a thing) is the same, but the context (ie the very low prior on polar bears in California) makes it implausible.



[Normal Bayesian reasoning slides gradually into confirmation bias](#). Suppose you are a zealous Democrat. Your friend makes a plausible-sounding argument for a Democratic position. You believe it; your raw experience (an argument that sounds convincing) and your context (the Democrats are great) add up to more-likely-than-not true. But suppose your friend makes a plausible-sounding argument for a Republican position. Now you're doubtful; the raw experience (a friend making an argument with certain inherent plausibility) is the same, but the context (ie your very low prior on the Republicans being right about something) makes it unlikely.

Still, this ought to work eventually. Your friend just has to give you a *good enough* argument. Each argument will do a little damage to your prior against Republican beliefs. If she can come up with enough good evidence, you have to eventually accept reality, right?

But in fact many political zealots never accept reality. It's not just that they're inherently skeptical of what the other party says. It's that even when something is proven beyond a shadow of a doubt, they *still* won't believe it. This is where we need to bring in the idea of trapped priors.

Trapped priors: the basic cognitive version

Phobias are a very simple case of trapped priors. They can be more technically defined as a failure of habituation, the fancy word for "learning a previously scary thing isn't scary anymore". There are lots of habituation studies on rats. You ring a bell, then give the rats an electric shock. After you do this enough times, they're scared of the bell - they run and cower as soon as they hear it. Then you switch to ringing the bell and *not* giving an electric shock.

At the beginning, the rats are still scared of the bell. But after a while, they realize the bell can't hurt them anymore. They adjust to treating it just like any other noise; they lose their fear - they habituate.

The same thing happens to humans. Maybe a big dog growled at you when you were really young, and for a while you were scared of dogs. But then you met lots of friendly cute puppies, you realized that most dogs aren't scary, and you came to some reasonable conclusion like "big growly dogs are scary but cute puppies aren't."

Some people never manage to do this. They get *cynophobia*, pathological fear of dogs. In its original technical use, a phobia is an intense fear that doesn't habituate. No matter how many times you get exposed to dogs without anything bad happening, you stay afraid. Why?

In the old days, psychologists would treat phobia by flooding patients with the phobic object. Got cynophobia? We'll stick you in a room with a giant Rottweiler, lock the door, and by the time you come out maybe you won't be afraid of dogs anymore. Sound barbaric? Maybe so, but more important it didn't really work. You could spend all day in the room with the Rottweiler, the Rottweiler could fall asleep or lick your face or do something else that should have been sufficient to convince you it wasn't scary, and by the time you got out you'd be even more afraid of dogs than when you went in.

Nowadays we're a little more careful. If you've got cynophobia, we'll start by making you look at pictures of dogs - if you're a severe enough case, even the pictures will make you a little nervous. Once you've looked at a zillion pictures, gotten so habituated to looking at pictures that they don't faze you at all, we'll put you in a big room with a cute puppy in a cage. You don't have to go near the puppy, you don't have to touch the puppy, just sit in the room without freaking out. Once you've done that a zillion times and lost all fear, we'll move you to something slightly doggier and scarier, than something slightly doggier and scarier than that, and so on, until you're locked in the room with the Rottweiler.

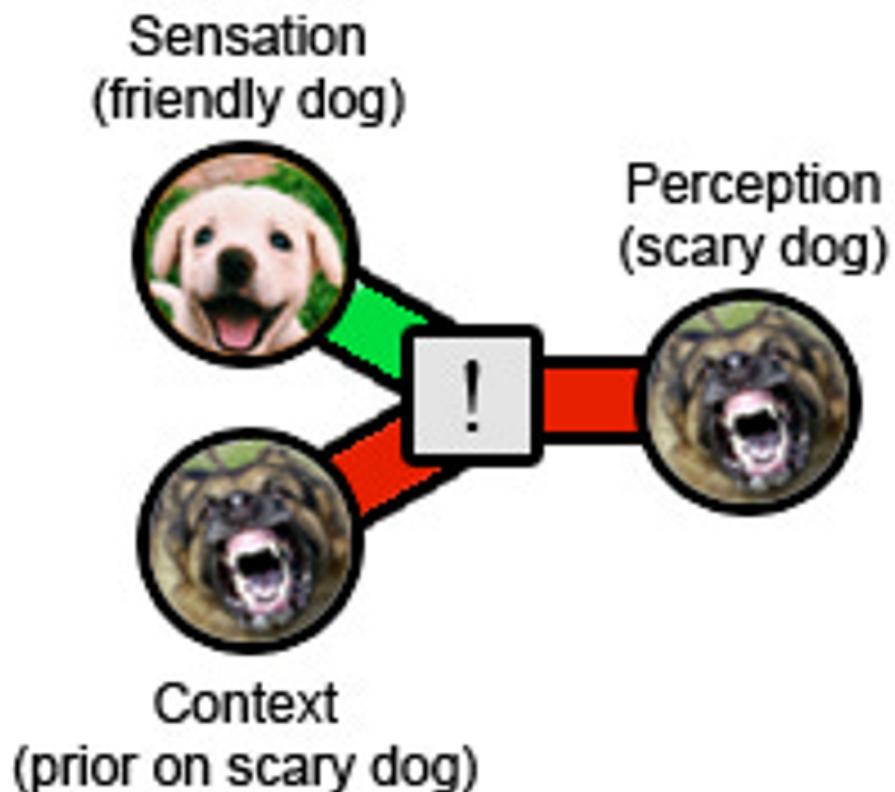
It makes sense that once you're exposed to dogs a million times and it goes fine and everything's okay, you lose your fear of dogs - that's normal habituation. But now we're back to the original question - how come flooding doesn't work? Forgetting the barbarism, how come we can't just start with the Rottweiler?

The common-sense answer is that you only habituate when an experience with a dog ends up being safe and okay. But being in the room with the Rottweiler is terrifying. It's not a safe okay experience. Even if the Rottweiler itself is perfectly nice and just sits calmly wagging its tail, your experience of being locked in the room is close to peak horror. Probably your intellect realizes that the bad experience isn't the Rottweiler's fault. But your lizard brain has developed a *stronger association than before* between dogs and unpleasant experiences. After all, you just spent time with a dog and it was a really unpleasant experience! Your fear of dogs increases.

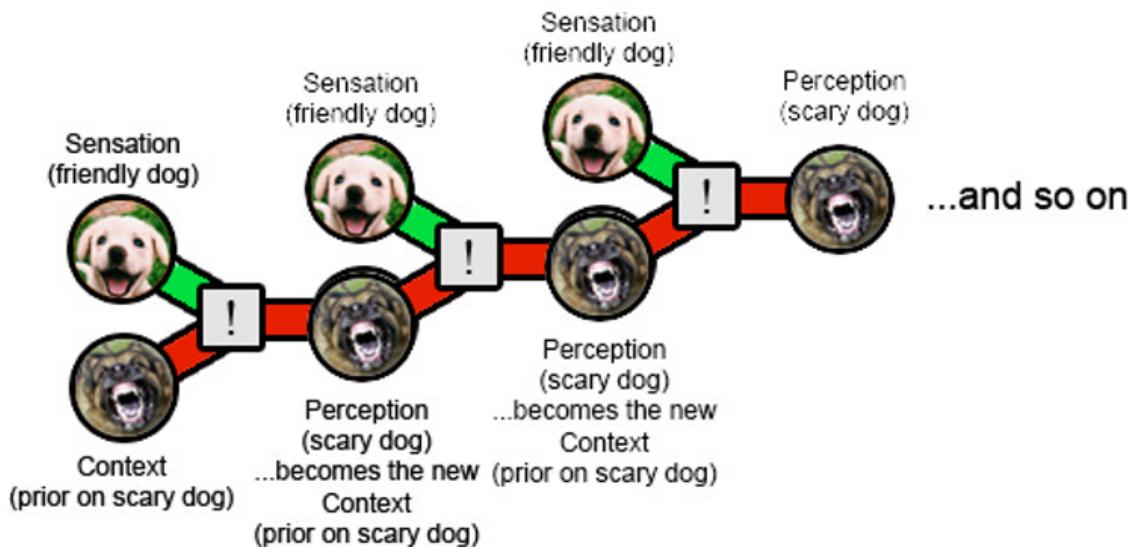
(How does this feel from the inside? Less-self-aware patients will find their prior coloring every aspect of their interaction with the dog. Joyfully pouncing over to get a headpat gets interpreted as a vicious lunge; a whine at not being played with gets interpreted as a murderous growl, and so on. This sort of patient will leave the room saying 'the dog came *this* close to attacking me, I knew all dogs were dangerous!' More self-aware patients will say something like "I know deep down that dogs aren't going to hurt me, I just know that whenever I'm with a dog I'm going to have a panic attack and hate it and be miserable the whole time". Then they'll go into the room, have a panic attack, be miserable, and the link between dogs and misery will be even more cemented in their mind.)

The more technical version of this same story is that habituation requires a perception of safety, but (like every other perception) this one depends on a combination of raw evidence and context. The raw evidence (the Rottweiler sat calmly wagging its tail) looks promising. But the context is a very strong prior that dogs are terrifying. If the prior is strong enough, it

overwhelms the real experience. Result: the Rottweiler was terrifying. Any update you make on the situation will be *in favor* of dogs being terrifying, not against it!



This is the trapped prior. It's trapped because it can never update, no matter what evidence you get. You can have a million good experiences with dogs in a row, and each one will just etch your fear of dogs deeper into your system. Your prior fear of dogs determines your present experience, which in turn becomes the deranged prior for future encounters.

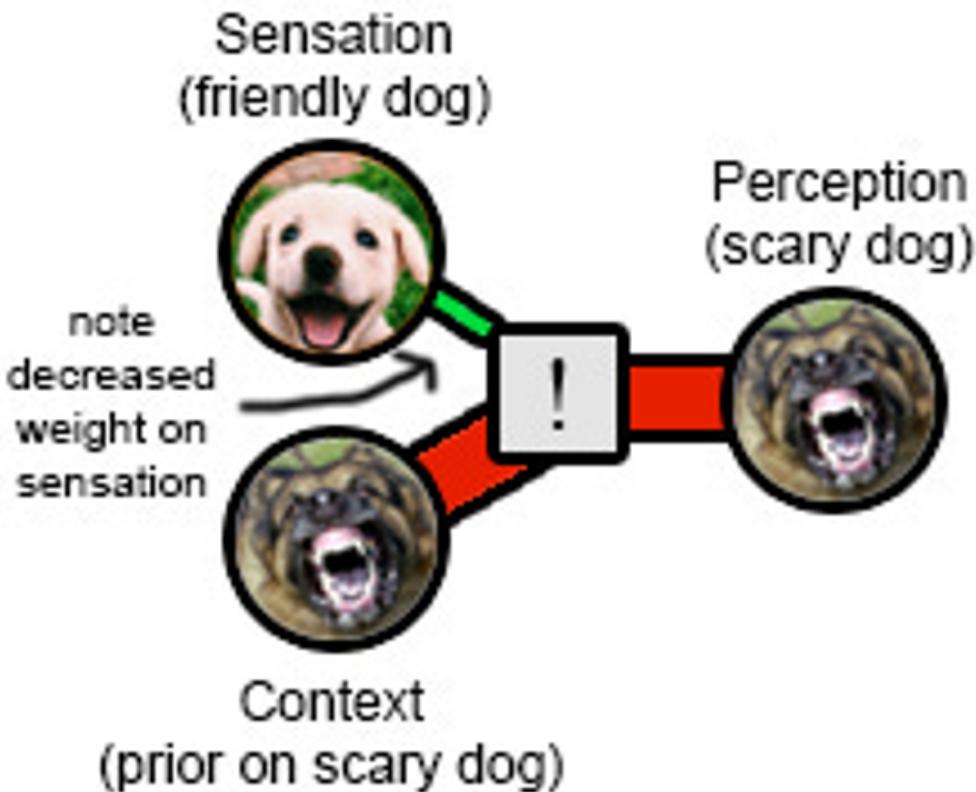


Trapped prior: the more complicated emotional version

The section above describes a simple cognitive case for trapped priors. It doesn't bring in the idea of emotion at all - an emotionless threat-assessment computer program could have the same problem if it used the same kind of Bayesian reasoning people do. But people find themselves more likely to be biased when they have strong emotions. Why?

Van der Bergh et al suggest that when experience is too intolerable, your brain will decrease bandwidth on the "raw experience" channel to protect you from the traumatic emotions. This is why some trauma victims' descriptions of their traumas are often oddly short, un-detailed, and to-the-point. This protects the victim from having to experience the scary stimuli and negative emotions in all their gory details. But it also ensures that context (and not the raw experience itself) will play the dominant role in determining their perception of an event.

You can't update on the evidence that the dog was friendly because your raw experience channel has become razor-thin; your experience is based almost entirely on your priors about what dogs *should* be like.



This diagram is a victim of my earlier decision to cram lots of things into the gray box in the middle. In earlier diagrams, I should have made it clear that a lot depended on the gray box choosing to weigh the prior more heavily than experience. In this diagram, less depends on this decision; the box is getting almost no input from experience, so no matter what its weighting function its final result will mostly be based on the prior. In most reasonable weighting functions, even a strong prior on scary dogs plus any evidence of a friendly dog should be able to make the perception slightly less scary than the prior, and iterated over a long enough chain this should update the prior towards dog friendliness. I don't know why this doesn't happen in real life, beyond a general sense that whatever weighting function we use isn't perfectly Bayesian and doesn't fit in the class I would call "reasonable". I realize this is a weakness of this model and something that needs further study.

I've heard some people call this "[bitch eating cracker syndrome](#)". The idea is - you're in an abusive or otherwise terrible relationship. Your partner has given you ample reason to hate them. But now you don't just hate them when they abuse you. Now even something as seemingly innocent as seeing them eat crackers makes you actively angry. In theory, an interaction with your partner where they just eat crackers and don't bother you in any way ought to produce some habituation, be a tiny piece of evidence that they're not always that bad. In reality, it will just make you hate them worse. At this point, your prior on them being bad is so high that every single interaction, regardless of how it goes, will make you hate them more. Your prior that they're bad has become trapped. And it colors every aspect of your interaction with them, so that even interactions which out-of-context are perfectly innocuous feel nightmarish from the inside.

From phobia to bias

I think this is a fruitful way to think of cognitive biases in general. If I'm a Republican, I might have a prior that Democrats are often wrong or lying or otherwise untrustworthy. In itself, that's fine and normal. It's a model shaped by my past experiences, the same as my prior against someone's claim to have seen a polar bear. But if enough evidence showed up - bear tracks, clumps of fur, photographs - I should eventually overcome my prior and admit that the bear people had a point. Somehow in politics that rarely seems to happen.

For example, [more scientifically literate people are more likely to have partisan positions on science](#) (eg agree with their own party's position on scientifically contentious issues, even when outsiders view it as science-denialist). If they were merely biased, they should start out wrong, but each new fact they learn about science should make them update a little toward the correct position. That's not what we see. Rather, they start out wrong, and each new fact they learn, each unit of effort they put into becoming more scientifically educated, *just makes them wronger*. That's not what you see in normal Bayesian updating. It's a sign of a trapped prior.

Political scientists have traced out some of the steps of how this happens, and it looks a lot like the dog example: zealots' priors determine what information they pay attention to, then distorts their judgment of that information.

So for example, [in 1979 some psychologists](#) asked partisans to read pairs of studies about capital punishment (a controversial issue at the time), then asked them to rate the methodologies on a scale from -8 to 8. Conservatives rated the pro-punishment study at about +2 and the anti-execution study as about -2; liberals gave an only slightly smaller difference the opposite direction. Of course, the psychologists had designed the studies to be about equally good, and even switched the conclusion of each study from subject to subject to average out any remaining real difference in study quality. At the end of reading the two studies, both the liberal and conservative groups reported believing that the evidence had confirmed their position, and described themselves as more certain than before that they were right. The more information they got on the details of the studies, the stronger their belief.

This pattern - increasing evidence just making you more certain of your preexisting belief, regardless of what it is - is pathognomonic of a trapped prior. These people are doomed.

I want to tie this back to [one of my occasional hobbyhorses](#) - discussion of "dog whistles". This is the theory that sometimes politicians say things whose literal meaning is completely innocuous, but which secretly convey reprehensible views, in a way other people with those reprehensible views can detect and appreciate. For example, in the 2016 election, Ted Cruz said he was against Hillary Clinton's "New York values". This sounded innocent - sure, people from the Heartland think big cities have a screwed-up moral compass. But [various news sources](#) argued it was actually Cruz's way of signaling support for anti-Semitism (because New York = Jews). Since then, almost anything any candidate from any party says has been accused of being a dog-whistle for something terrible - for example, apparently Joe Biden's comments about Black Lives Matter were dog-whistling his support for rioters burning down American cities.

Maybe this kind of thing is real sometimes. But think about how it interacts with a trapped prior. Whenever the party you don't like says something seemingly reasonable, you can interpret in context as them wanting something horrible. Whenever they want a seemingly desirable thing, you secretly know it means they want a horrible moral atrocity. If a Republican talks about "law and order", it doesn't mean they're concerned about the victims of violent crime, it means they want to lock up as many black people as possible to strike a blow for white supremacy. When a Democrat talks about "gay rights", it doesn't mean letting people marry the people they love, it means destroying the family so they can replace it with state control over your children. I've had arguments with people who believe that no pro-life

conservative really cares about fetuses, they just want to punish women for being sluts by denying them control over their bodies. And I've had arguments with people who believe that no pro-lockdown liberal really cares about COVID deaths, they just like the government being able to force people to wear masks as a sign of submission. Once you're at the point where all these things sound plausible, *you are doomed*. You can get a piece of evidence as neutral as "there's a deadly pandemic, so those people think you should wear a mask" and convert it into "they're trying to create an authoritarian dictatorship". And if someone calls you on it, you'll just tell them they need to look at it *in context*. It's the bitch eating cracker syndrome except for politics - even when the other party does something completely neutral, it seems like extra reason to hate them.

Reiterating the cognitive vs. emotional distinction

When I showed some people an early draft of this article, they thought I was talking about "emotional bias". For example, the phobic patient *fears* the dog, so his anti-dog prior stays trapped. The partisan *hates* the other party, so she can't update about it normally.

While this certainly happens, I'm trying to make a broader point. The basic idea of a trapped prior is purely epistemic. It can happen (in theory) even in someone who doesn't feel emotions at all. If you gather sufficient evidence that there are no polar bears near you, and your algorithm for combining prior with new experience is just a little off, then you can end up rejecting all apparent evidence of polar bears as fake, and trapping your anti-polar-bear prior. This happens without any emotional component.

Where does the emotional component come in? I think van der Bergh argues that when something is so scary or hated that it's aversive to have to perceive it directly, your mind decreases bandwidth on the raw experience channel relative to the prior channel so that you avoid the negative stimulus. This makes the above failure mode much more likely. Trapped priors are a cognitive phenomenon, but emotions create the perfect storm of conditions for them to happen.

Along with the cognitive and emotional sources of bias, there's a third source: self-serving bias. People are more likely to believe ideas that would benefit them if true; for example, rich people are more likely to believe low taxes on the rich would help the economy; minimum-wage workers are more likely to believe that raising the minimum wage would be good for everyone. Although I don't have any formal evidence for this, I suspect that these are honest beliefs; the rich people aren't just pretending to believe that in order to trick you into voting for it. I don't consider the idea of bias as trapped priors to account for this third type of bias at all; it might relate in some way that I don't understand, or it may happen through a totally different process.

Future research directions

If this model is true, is there any hope?

I've sort of lazily written as if there's a "point of no return" - priors can update normally until they reach a certain strength, and after that they're trapped and can't update anymore. Probably this isn't true. Probably they just become trapped *relative to the amount of evidence an ordinary person is likely to experience*. Given immense, overwhelming evidence, the evidence could still drown out the prior and cause an update. But it would have to be really big.

(...but now I'm thinking of the stories of apocalypse cultists who, when the predicted apocalypse doesn't arrive, double down on their cult in one way or another. Festinger, Rieken, and Schachter's classic book on the subject, [When Prophecy Fails](#), finds that these people "become a more fervent believer after a failure or disconfirmation". I'm not sure what level of evidence could possibly convince them. My usual metaphor is "if God came down from the heavens and told you..." - but God coming down from the heavens and telling you *anything* probably makes apocalypse cultism more probable, not less.)

If you want to get out of a trapped prior, the most promising source of hope is the psychotherapeutic tradition of treating phobias and PTSD. These people tend to recommend very gradual exposure to the phobic stimulus, sometimes with special gimmicks to prevent you from getting scared or help you "process" the information (there's no consensus as to whether the eye movements in EMDR operate through some complicated neurological pathway, work as placebo, or just distract you from the fear). A lot of times the "processing" involves trying to remember the stimulus multimodally, in as much detail as possible - for example drawing your trauma, or acting it out.

[Sloman and Fernbach](#) might be the political bias version of this phenomenon. They ask partisans their opinions on various issues, and as usual find strong partisan biases. Then they asked them to do various things. The only task that moderated partisan extremism was to give a precise mechanistic explanation of how their preferred policy should help - for example, describing in detail the mechanism by which sanctions on Iran would make its nuclear program go better or worse. The study doesn't give good enough examples for me to know precisely what this means, but I wonder if it's the equivalent of making trauma victims describe the traumatic event in detail; an attempt to give higher weight to the raw experience pathway compared to the prior pathway.

The other promising source of hope is psychedelics. These probably [decrease the relative weight given to priors by agonizing 5-HT2A receptors](#). I used to be confused about why this effect of psychedelics could produce lasting change (permanently treat trauma, help people come to realizations that they agreed with even when psychedelics wore off). I now think this is because they can loosen a trapped prior, causing it to become untrapped, and causing the evidence that you've been building up over however many years to suddenly register and to update it all at once (this might be that "simulated annealing" thing everyone keeps talking about. I can't unreservedly recommend this as a pro-rationality intervention, because it also [seems to create permanent weird false beliefs for some reason](#), but I think it's a foundation that someone might be able to build upon.

A final possibility is other practices and lifestyle changes that cause the brain to increase the weight of experience relative to priors. Meditation probably does this; see the discussion in [the van der Bergh post](#) for more detail. Probably every mental health intervention (good diet, exercise, etc) does this a little. And this is super speculative, and you should feel free to make fun of me for even thinking about it, but [sensory deprivation](#) might do this too, for the same reason that your eyes become more sensitive in the dark.

A hypothetical future research program for rationality should try to identify a reliable test for prior strength (possibly some existing psychiatric measures like mismatch negativity can be repurposed for this), then observe whether some interventions can raise or lower it consistently. Its goal would be a relatively tractable way to induce a low-prior state with minimal risk of psychosis or permanent fixation of weird beliefs, and then to encourage people to enter that state before reasoning in domains where they are likely to be heavily biased. Such a research program would dialogue heavily with psychiatry, since both mental diseases and biases would be subphenomena of the general category of trapped priors, and it's so far unclear exactly how related or unrelated they are and whether solutions to one would work for the other. Tentatively they're probably not too closely related, since very neurotic people can sometimes reason very clearly and vice versa, but I don't think we yet have a good understanding of why this should be.

Ironically, my prior on this theory is trapped - everything I read makes me more and more convinced it is true and important. I look forward to getting outside perspectives.

MIRI comments on Cotra's "Case for Aligning Narrowly Superhuman Models"

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

Below, I've copied comments left by MIRI researchers Eliezer Yudkowsky and Evan Hubinger on March 1-3 on a draft of Ajeya Cotra's "[Case for Aligning Narrowly Superhuman Models](#)." I've included back-and-forths with Cotra, and interjections by me and Rohin Shah.

The section divisions below correspond to the sections in [Cotra's post](#).

0. Introduction

How can we train GPT-3 to give “the best health advice it can give” using demonstrations and/or feedback from humans who may in some sense “understand less” about what to do when you’re sick than GPT-3 does?

Eliezer Yudkowsky: I've had some related conversations with Nick Beckstead. I'd be hopeful about this line of work primarily because I think it points to a bigger problem with the inscrutable matrices of floating-point numbers, namely, we have no idea what the hell GPT-3 is thinking and cannot tell it to think anything else. GPT-3 has a great store of medical knowledge, but we do not know where that medical knowledge is; we do not know how to tell it to internally apply its medical knowledge rather than applying other cognitive patterns it has stored. If this is still the state of opacity of AGI come superhuman capabilities, we are all immediately dead. So I would be relatively more hopeful about any avenue of attack for this problem that used anything other than an end-to-end black box - anything that started to address, "Well, this system clearly has a bunch of medical knowledge **internally**, can we find that knowledge and cause it to actually be applied" rather than "What external forces can we apply to this solid black box to make it think more about healthcare?"

Evan Hubinger: +1 I continue to think that language model transparency research is the single most valuable current research direction within the class of standard ML research, for similar reasons to what Eliezer said above.

Ajeya Cotra: Thanks! I'm also excited about language model transparency, and would love to find ways to make it more tractable as a research statement / organizing question for a field. I'm not personally excited about the connotations of transparency because it evokes the neuroscience-y interpretability tools, which don't feel scalable to situations when we don't get the concepts the model is using, and I'm very interested in finding slogans to keep researchers focused on the superhuman stuff.

Ajeya Cotra: I've edited the description of the challenge to emphasize human feedback less. It now reads "How can we get GPT-3 to give “the best health advice it

can give" when humans in some sense "understand less" about what to do when you're sick than GPT-3 does? And in that regime, how can we even tell/verify that it's "doing the best it can"?"

Rob Bensinger: Nate and I tend to talk about "understandability" instead of "transparency" exactly because we don't want to sound like we're talking about normal ML transparency work.

Eliezer Yudkowsky: Other possible synonyms: Clarity, legibility, cognitive readability.

Ajeya Cotra: Thanks all -- I like the project of trying to come up with a good handle for the kind of language model transparency we're excited about (and have talked to Nick, Evan, etc about it too) but I think I don't want to push it in this blog post right now because I haven't hit on something I believe in and I want to ship this.

In the end, we probably want to find ways to meaningfully supervise (or justifiably trust) models that are more capable than ~all humans in ~all domains.

Eliezer Yudkowsky: (I think you want an AGI that is superhuman in engineering domains and infrahuman in [human-modeling-and-manipulation](#) if such a thing is at all possible.)

Ajeya Cotra: Fair point, added a footnote:

"Though if we could pull off a path where we build an AI system that is superhuman in certain engineering capabilities but not yet human-level in modeling and manipulating people, and use that system to cut down on x-risk from other AI projects without having to figure out how to supervise arbitrary superhuman models, that could be really good."

1. What aligning narrowly superhuman models could look like

First of all, it's important to note that not all narrowly superhuman models are going to be equally interesting as alignment case studies. AlphaGoZero (AGZ) is narrowly superhuman in an extremely strong sense: it not only makes Go moves better than the moves made by top human players, but also probably makes moves that top players couldn't even reliably recognize as good. But there isn't really an alignment problem for Go: a precise, algorithmically-generated training signal (the win/loss signal) is capable of eliciting the "full Go-playing potential" of AGZ given enough training, and would keep working even as the model got much bigger.

Eliezer Yudkowsky: Having access to an incorruptible ground truth solves some of the alignment problems but not all of them, in particular [inner alignment problems](#). In the limit of optimizing infinitely hard on [logical Tic-Tac-Toe](#), it won't kill you because it hits a capability bound early and stops; in the limit of optimizing infinitely hard on any real-world problem, there's no capability bound lower than extreme superintelligence

so the thing inside keeps getting smarter and kills you. It is not obvious to me where logical Go falls on this spectrum, or rather, it is obvious to me that the answer is "it depends on the outer optimization method". (That is, some ways of optimizing a system to play Go will create an inner optimizer that plays Go and that will kill you; some ways might create a system that just played arbitrarily good Go.)

Ajeya Cotra: Good point, changed to:

"But there isn't really an outer alignment problem for Go: a precise, algorithmically-generated training signal (the win/loss signal) is capable of eliciting the "full Go-playing potential" of AGZ given enough training, and would keep working even as the model got much bigger (although at a certain scale inner alignment issues may crop up)."

Choose a helpful "fuzzy" task (e.g. summarization, question-answering, advice-giving, story-writing) for which we have suggestive evidence that makes us suspect a state-of-the-art model has the capacity to significantly outperform some reference set of humans (e.g. Mechanical Turk workers) given the right training signal. Then,

Eliezer Yudkowsky: I'd feel more cheerful about an open call "Yo, try to do something about the fact that text transformers seem like they in some sense contain the capability to solve this problem but we can't make them use that capability to [do] what we want" than "Yo, retrain GPT-3 to do this using an outer training signal in a way that scales well with task and model complexity". The latter call is tuned to one particular approach to the first problem, which some, such as myself, would worry is not the most promising approach in the long run.

Evan Hubinger: +1 I think this is very similar to my objection that a focus on solving downstream tasks is less likely to translate into important insights than focusing on the core problem directly—though having talked with Ajeya in the comment thread at the bottom, I think I'm now convinced that the alternative pitch of "get language models to actually answer questions honestly to the best of their ability" also has some problems getting people to work on the important part of the problem.

Ajeya Cotra: Added some language to the bottom of this section in purple to emphasize that this is just one formulation.

I want a concrete formulation and specific project ideas in this post, even if they're not the best ones, but I agree that the broader question of "What the hell do we do about situations where models can do stuff to help us but don't want to?" is where the focus should be.

1.2. What kinds of projects do and don't "count"

I think that if you can get the model to achieve superhuman performance at some task without collecting any human feedback or human demonstrations, the task is probably not "fuzzy" enough.

Eliezer Yudkowsky: What we have with GPT-3 is a case of a clear outer optimization, "predict the next character", which creates a wide-ranging variety of inner knowledge that seems like it could in principle be useful for many many tasks; but we can't directly make GPT-3 use that knowledge for other tasks, because we have no idea what the hell GPT-3 is thinking or how to make it think anything in particular. Instead we have to do elaborate dances to shoehorn our task into the shape of the predicted next character, i.e., prompt engineering. If you pick a new mechanical task with no humans in the loop and a clear outer loss function, it doesn't force you to confront any interesting part of this problem - we already know how to pretrain and retrain nets.

Things you do with humans in the loop can in principle also be boringly straightforward. But humans in the loop are expensive; so this potentially forces you to do something more data-efficient than usual, and find an efficient way to apply leverage. (This is how I'd interpret "[Learning to summarize](#)." It didn't poke at the internals, but it at least found a higher-leverage way to apply external pressures.)

Ajeya Cotra: I agree with what you've said -- I can't tell if your comment here was "I agree and here's some elaboration", or if it's objecting to the way I've framed it here?

Eliezer Yudkowsky: Something like, "I might have phrased that differently". I don't see a simple change of local phrasing that fits with the rest of the message in context; my inner editor worries that the message sounds here like "Go make it wantonly fuzzy to impress me" rather than "Here is why going for the easy clarity is a bad sign."

1.3. Potential near-future projects: “sandwiching”

In all of these cases, my guess is that the way to get the less-capable group of humans to provide training signals of a similar quality to the more-capable group will involve some combination of:

Eliezer Yudkowsky: In some ways this strikes me as a much **more** ambitious project than getting GPT-3 to cough up what it actually knows about health. You're trying to make the system perform "better" than its training data, in a certain sense. This focuses on some interesting parts of a capability problem. It potentially leads into a much more important alignment problem: "Given that humans can be fooled by some inputs, how do you operate a more intelligent input-finder without that fooling the humans?", where we could potentially test this by building a system that didn't fool easy-to-fool humans, while the capability levels are still too low to fool harder-to-fool humans, which gives us a possible check on whether the methodology actually succeeded, so long as we don't try too many times or try with an overly capable system.

But this all strikes me as a **different** project than getting better transparency and steering into a system that has only learned things implicit in the training data. Of course it may be that the solutions to both problems end up related, somehow, but their outer form at least looks different.

The different problem is potentially interesting in its own right, if its solution forces a non-black-box system that is then more transparent or steerable (but note that you are basically asking for a capability solution and praying it will be more alignable); or the variant problem of "How do you make a system not fool easy-to-fool humans, in a domain where there are hard-to-fool humans to check the answer" (and the system is

not an AGI yet and does not anticipate the existence of the hard-to-fool humans checking its answers &c).

Ajeya Cotra: I think there's room for both types of work in this umbrella. I'm mostly trying here to create a slogan that points researchers at the exciting types of work within ML, so I think outer alignment issues and figuring out how to get humans who are less-foolable (which I've focused on a lot in this post) and inner alignment stuff / transparency could fit under this umbrella. I've emphasized the less-foolable humans stuff here because I think there's something juicily concrete about the problem statement, and it's very easy to tell if you've made progress.

2. How this work could reduce long-term AI risk

On the outside view, I think we should be quite excited about opportunities to get experience with the sort of thing we want to eventually be good at (aligning models that are smarter than humans). In general, it seems to me like building and iterating on prototypes is a huge part of how R&D progress is made in engineering fields, and it would be exciting if AI alignment could move in that direction.

If there are a large number of well-motivated researchers pushing forward on making narrowly superhuman models as helpful as possible, we improve the odds that we first encounter serious problems [like the treacherous turn](#) in a context where a) models are not smart enough to cause actually catastrophic harm yet, and b) researchers have the time and inclination to really study them and figure out how to solve them well rather than being in a mode of scrambling to put out fires and watching their backs for competitors. Holistically, this seems like a much safer situation to be in than one where the world has essentially procrastinated on figuring out how to align systems to fuzzy goals, doing only the minimum necessary to produce commercial products.

Eliezer Yudkowsky: So as not to speak disagreements only, I remark that I agree with these two paragraphs. I worry about how very far underwater we are on the [logistic success curve](#), here, but that doesn't mean we should not throw resources at any hope of starting to swim (upward).

Ajeya Cotra: Thanks -- honestly I felt like your comments were radically more positive than I was expecting overall, not just this one.

Chance of discovering or verifying long-term solution(s): I'm not sure whether a "one shot" solution to alignment (that is, a single relatively "clean" algorithm which will work at all scales including for highly superintelligent models) is possible. But if it is, it seems like starting to do a lot of work on aligning narrowly superhuman models probably allows us to discover the right solution sooner than we otherwise would have.

Eliezer Yudkowsky: It's not possible. Not for us, anyways. A textbook that fell out of a wormhole from the future might have the simplest straightforward working

solution with no extra gears, all of whose pieces work reliably. We won't get it in time because it takes multiple decades to go from sigmoid activation functions to ReLUs, and so we will definitely be working with the AGI equivalent of sigmoid activation functions instead of ReLUs while the world is ending. Hope that answers your question!

It also seems plausible that a solution will emerge directly from this line of work rather than the conceptual work -- the latter is mostly focused on finding a one-shot solution *that will work under ~pessimal empirical assumptions*,

Eliezer Yudkowsky: "Pessimal" is a strange word to use for this apt description of humanity's entire experience with ML to date. Unless by "generalize" you mean "generalize correctly to one new example from the same distribution" rather than "generalize the underlying concept that a human would".

Ajeya Cotra: I used "pessimal" here in the technical sense that it's assuming if there are N generalizations equally valid on the training distribution the model will pick the one which is worst for humans. Even if there's a very high probability that the worst one is in fact picked, assuming the worst one will be picked is still "assuming the worst case."

Fwiw, I know lots of ML alignment researchers who would not agree this is highly likely and e.g. point to lots of "human-like" generalization going on in neural networks that makes them hopeful about the empirical situation. I haven't found those arguments super compelling but am generally in a state of uncertainty rather than confidence one way or another.

I recognize this is the sort of thing MIRI has been arguing with other people about for ages though, and I don't have an inside view here, so it probably doesn't make sense to have this discussion in the comments.

3. Advantages over other genres of alignment research

I'm broadly supportive of all three of these other lines of work, but none of these approaches come as close to "practicing the thing we eventually want to be good at" as aligning narrowly superhuman models does.

Eliezer Yudkowsky: I remark that it seems to me that work like "[Learning to summarize](#)", plus this, falls into a subcategory that's much closer to the subcategories of "reliability+robustness" and the subcategory "interpretability", than to gridworlds; and in turn, the category of gridworlds is much closer to that aforesaid supercategory than either of those categories is to "conceptual research".

4. Objections and responses

4.1. How would this address treachery by a superintelligence?

I'm very uncertain how relevant the near-term work will turn out to be for more exotic problems like the treacherous turn, and I want to think more about ways to nudge it to be more relevant.

Eliezer Yudkowsky: Having some idea of what the hell the system was thinking internally might be one good start. It's not a one-shot solution, but you'd at least get a couple of warning lights before you went on ahead anyways and died.

To be more precise, I'd imagine the warning lights shut off at a higher capability level if that capability level is "smart enough to conceal own thoughts from transparency mechanisms, without that intention itself lighting up in the transparency mechanisms" instead of "smart enough to socially manipulate operators through outer acts and shape the quoted inner intentions that operators infer from acts". Plausibly the former case results in everybody dying because surrounding arms race conditions caused people to press ahead ignoring warning signs; instead of everybody dying because they were wholly ignorant of the dangerous thoughts their AGI was thinking, and not especially focused on the missing knowledge. Trying to die in a more dignified and self-aware fashion can be a first step in climbing the [logistic success curve](#) towards eventually surviving, or at least that would be the optimistic take.

Ajeya Cotra: I added a footnote about this idea, and also mentioned it in my description of your views. I am interested in it but I don't think I'm the right person to speak for this idea since I don't totally "get it" or have good inside view project ideas. I'd be excited for you or Evan to leave a comment on the blog post discussing this.

[Added footnote: "One idea a couple of others have suggested here and which I'm generally interested in is 'transparency in (narrowly superhuman) language models': finding ways to understand 'what models are thinking and why,' especially when they know more about something than humans do. I like this idea but am very unsure about what execution could look like. E.g., would it look like [Chris Olah's work](#), which essentially 'does neuroscience' on neural networks? Would it look like training models to answer our questions about what they're thinking? Something else?"]

Evan Hubinger: I think that if we're in a situation where our model is actively trying to trick our transparency tools, we've already lost—but that we can potentially use transparency tools as a mechanism to understand what's happening during training enough to avoid ever entering that regime. Thus, I think current transparency research is likely to be useful because I think it will eventually enable us to build better training processes that avoid deception, rather than helping us detect deception directly. My guess is that that's sort of what Eliezer is saying in terms of knowing what's happening being a step towards preventing it, though I'm not sure I fully understand Eliezer's beliefs here.

Ajeya Cotra: That's a good point re using transparency to create better training processes, and makes me more positive.

Rohin Shah: The way I've summarized this point previously is that if our source of feedback can see how the model is reasoning, then we can select for models that get the right answer for the right reason, and not just models that get the right answer. One would hope that this selects heavily against treachery.

Ajeya Cotra: Well, I'm kind of wary of selecting for looking good on the transparency tools we currently have available without more information, tbc: I'd rather set things up so the transparency tools themselves are not under optimization pressure.

Evan Hubinger: I agree that ideally we wouldn't want to be applying optimization pressure against our transparency tools since we risk Goodharting on them, but imo I think this is a very manageable risk. I expect our training processes (e.g. SGD) to be dumb enough that a good overseer can mostly avoid being tricked by them, but I don't expect the same thing to be true about the model being overseen—that is, if you have to fight against SGD trying to trick you, that's probably fine, but if you have to fight against a deceptive model trying to trick you, that's probably not.

Though you could think that in an absolute sense it and all the other approaches that aren't tackling treachery head-on are doomed.

Eliezer Yudkowsky: Doomed, I tell you! Doomed!

4.3. Isn't this not neglected because lots of people want useful AI?

Eliezer Yudkowsky: Strong agree with this entire section (as it appeared at the time this comment was written).

[Editor's note: this section is more or less completely unchanged.]

5. Current state of opinion on this work

I think most researchers at MIRI probably feel that this work is very unpromising or net-harmful -- I don't totally understand their full reasoning, but my sense is that a strong version of [the treachery objection](#) plays a part (although they might frame it differently in important ways). But MIRI's view on virtually all non-MIRI research is that it's near-useless, and if anything Eliezer personally seems maybe [marginally more positive](#) on this type of thing than other non-MIRI alignment work -- so while it's unendorsed by MIRI, I don't think it's *particularly* unendorsed relative to other things.

Eliezer Yudkowsky: Doesn't seem especially fair? Every year or two I check again with Chris Olah whether he's currently receiving all the resources he thinks he can use, and/or remind OpenPhil that I wish it was possible to spend a billion dollars on that research.

Ajeya Cotra: I actually genuinely wasn't aware of this and thought that MIRI would pretty happily endorse this statement, my bad. I was also pretty surprised by your positive engagement here -- thanks for that!

I was wrong about the MIRI view here and glad I checked with Rob. Will propose an edit.

Ajeya Cotra: Okay I added in purple a different summary of MIRI's view: "My understanding of Eliezer Yudkowsky's position is one of "cautious relative optimism" about something in this general space compared to other non-MIRI alignment work, though he would frame the core concern differently, with more emphasis on transparency and interpretability ("GPT-3 has somewhere buried inside it knowledge of what to do when you're sick; how do you extract all of that and how can you tell when you've succeeded?"). He was reasonably positive on [Stiennon et al., 2020](#) when it came out, and would be happy to see more work like that. Evan Hubinger's position seems similar, and I'm not sure where others at MIRI would land on this work."

Eliezer Yudkowsky: I endorse this statement of my position.

Ajeya Cotra: Awesome, thanks!

6. Takeaways and possible next steps

If you disagree with this argument, say so -- especially if you [think it would be harmful](#) or would be dominated by a different line of work that shares [similar practical advantages](#) of tangibility, good feedback loops, and potential-for-scale.

Eliezer Yudkowsky: The closest thing I have to a disagreement with this draft is a general sense that 19 out of 20 bright ideas in machine learning don't work especially well; so I have more hope in calls for proposals that pinpoint key problems in a way that seems sufficiently crisp to technical thinkers, possibly including examples of potential bright ideas to show what the problem **is**, but leave wide open how those problems are to be addressed.

Compared to saying, "The problem is that we have no idea what the hell GPT-3 is thinking, and if that's still true at AGI then everybody dies", I think you get better results if you say "GPT-3 obviously contains a lot of medical knowledge but we would like a better way than prompt engineering to get GPT-3 to think using that knowledge, something that scales better than prompt engineering using a lot of human labor, and doesn't leave us wondering if today was the day that GPT-3 decided to emulate an Internet troll despite our best prompting". But you get worse results if you have your own bright idea for how to solve that, and ask for proposals to carry out your bright idea. This draft doesn't descend particularly far into that - it's not proposing a particular exact architecture - but it also doesn't explicitly distinguish "Here is what I think the problem is" and "Here is one example idea of how to go about it, which the call for proposals definitely isn't limited to because 19 out of 20 ideas don't work".

For that matter, I think you want to explicitly signal that you are open to people reframing the problem (providing that you think they are still directly challenging a key step and not just telling you to care about something else instead; but in real life I imagine you'd get a lot of lousy distantly-related proposals with strained arguments, no matter what you say in the CfP, if you are getting lots of proposals at all).

Ajeya Cotra: Thanks -- I basically agree with all of this on reflection (except for, ironically/appropriately, your particular framing of the problem). I've suggested some edits already (highlighted in purple) to make it more agnostic, and will look back at this and add more.

Ajeya Cotra: I tweaked various things, and I don't think it's totally nailing the balance between not being too prescriptive and not being too vague, but I'll probably ship it anyway to get the ball rolling.

[Editor's note: Evan made the following comment before Ajeya made the Q&A section "[Why not focus on testing a long-term solution?](#)".]

Evan Hubinger: Some comments after reading this, which I think broadly fall into the category of thinking that this is less valuable than other work that could be done:

I think there is a very important problem that this line of research is pointing at, which is figuring out how to get models that are ascription universal—that is, models that tell you everything they know. Paul has a whole sequence of posts on this (that I was surprised to not see linked anywhere, but maybe I missed it, see <https://ai-alignment.com/towards-formalizing-universality-409ab893a456>).

That being said, I feel pretty skeptical about telling people to just go work on something useful and hope that good work on universality comes out as a result. There's a lot of engineering work that goes into making useful systems that's pretty unrelated to actually solving universality-style problems and my guess is that most of the effort on a project like that would end up being unrelated engineering work rather than useful universality-relevant work.

I think that the distinctions you draw between for-profit work and what you're proposing here might help mitigate the above problem somewhat, but I'm still not sure why this would be better than just working directly on the universality problem. I feel like I would much rather have someone doing a bunch of research trying to figure out how to get GPT-3 to give truthful answers than someone trying to get GPT-3 to give legitimately helpful medical advice. There's just so much random, domain-specific stuff you'd need to do to get GPT-3 to give good medical advice that doesn't feel particularly alignment-relevant to me.

Another way of thinking about this: if you're doing research directly on universality, you can have a much tighter feedback loop, where you can iterate with lots of different techniques and try different things. If you have to produce something that's actually going to be useful, I don't think you can really afford to do that same sort of open-ended iteration. Instead, I expect you'll have to do a lot of very specific fine-tuning for the downstream tasks you're trying to solve that will prevent you from being able to iterate quickly.

Ajeya Cotra: Thanks for the thoughts! I don't think I expect most of the value of this work to be directly testing the specific concept of ascription universality or otherwise directly feeding into Paul's agenda, which is why I didn't link the post. The thing I'm saying is specifically trying to be pretty common sense and atheoretical, rather than focusing on resolving a particular uncertainty within a particular one shot solution agenda.

I think telling people to train models to give truthful answers to a wide range of questions would be good, but not obviously better than the other projects and not obviously very optimized for testing IDA / ascription universality (which I think most people don't really understand well enough to have a good sense of whether they're testing it or what they can learn about it). I think that whole line of work needs more conceptual clarification, at which point Paul will probably propose specific projects that

would test his views (which might end up looking very different than his current views when the dust settles).

With that said, I don't consider the *point* of this work to make models useful -- I just consider the usefulness of the final models a litmus test for whether you were choosing sufficiently challenging alignment problems to work on, and I agree it would be bad to waste a bunch of time on faff productizing the thing. A model which answers a bunch of questions well and honestly would pass that bar.

If you have ideas for research projects that would provide good tests of ascription universality and could be executed by someone who doesn't deeply understand ascription universality or really buy into Paul's agenda, I'd be excited to add that to the list of suggested projects. But I think I would frame it as "Here's a thing that looks like aligning a narrowly superhuman model. A special benefit of this project is that Evan Hubinger thinks it's particularly well-suited for shedding light on ascription universality", rather than leading with "testing out ascription universality" as the main benefit you could get out of any of this work.

Another point here is that I'm focusing a lot on the long-run field growth potential, which "directly testing (particular theory)" seems to have less of.

Evan Hubinger: Hmm... I guess I'm just having difficulty, from an inside-view perspective, of understanding what non-universality-related insights we would get from building models which are useful in this sense. I think I do understand the outside view argument, though I would feel a lot happier if I also had an inside view argument.

Perhaps my difficulty here is just coming from the fact that the primary example I have in mind is of trying to get GPT-3 to give more honest/helpful/useful answers. Do you have another example of a useful short-term alignment project that wouldn't fall into that category?

If it is just that category, though, it feels to me like it would just be better to tell people "work on getting GPT-3 to give more honest/helpful/useful answers" than "work on getting GPT-3 to be useful for concrete downstream task X."

Ajeya Cotra: Coding is an example that doesn't seem like it's about "honest helpful answers": I'm excited to figure out how to get non-software engineers to effectively supervise a coding model to write better code. I'm also interested in e.g. getting GPT-3 to write a good mystery story using feedback from people who aren't good writers, or maybe don't even speak the language.

What's your take on the three sources of value I list in the "How this work could reduce long-term x-risk" section? They were: 1) Practical know-how and infrastructure, 2) Better AI situation in the run up to superintelligence, and 3) Chance of discovering or verifying a long-term solution.

It seems like your take is something like "#3 is the main hope for impact, and getting insight about ascription universality is the main hope for #3, so if I don't see a strong route to getting insight about ascription universality from this work I'm not very bought in." Is that right? I think I feel much more agnostic about both of those claims: #1 and #2 seem pretty big to me, and it seems plausible that we'll discover routes to long-term solutions that look pretty different from the current Paul agenda.

E.g. I think it's pretty likely the thing Paul is trying to do is impossible because he is setting a certain bar like "If we don't have an a priori argument that it should be safe, assume it won't be", but at the same time something less stringent and more contingent on the actual empirical situation would work fine. It also seems plausible to me that the ascription universality stuff "just works out" and the main difficulty is elsewhere, etc.

As a more minor point, I think I don't like the frame of "train GPT-3 to give honest, helpful answers" because that doesn't put the focus on the "narrowly superhuman" part. That's kind of what I was getting at in my response to the "Why not just stick with getting models not to do bad things?" objection.

I think the interesting part of the challenge is getting GPT-3 to *add something* to humans -- to be better than they are at something. I think a frame of "get it to be honest" is more likely to create a bunch of ho hum work on correcting errors humans can notice pretty easily themselves.

Evan Hubinger: > I think the interesting part of the challenge is getting GPT-3 to *add something* to humans -- to be better than they are at something. I think a frame of "get it to be honest" is more likely to create a bunch of ho hum work on correcting errors humans can notice pretty easily themselves.

That's a good point that I hadn't considered; I definitely feel more convinced that this is a good idea now.

> What's your take on the three sources of value I list in the "How this work could reduce long-term x-risk" section?

I think I feel less convinced by (1) and (2) since they seem reasonably likely to just happen by default.

> I think it's pretty likely the thing Paul is trying to do is impossible

Fwiw, I agree with this—I think this is one of my current biggest cruxes with Paul. Nevertheless, I think that a better understanding of the extent to which it is possible is likely to be very useful and teach us a lot.

Ajeya Cotra: I agree that digging into the Paul thing is likely to teach us a lot even if it's impossible, fwiw. I think Paul should keep doing his thing, and I hope he comes out with experiment ideas soon. I think there's limited room for that kind of thinking and a limited number of people who could do well at it though, and the room to do this work (which could explore in parallel a bunch of worlds where Paul's thing is impossible but we're not doomed anyway) is much larger.

I think of this stuff as a pretty strong baseline to beat -- if you think you have privileged insight that beats the baseline you should probably try that out for at least a couple years and return to the baseline if it's not going well. (I think even if you have privileged insight, I'd want to see more experimental work and less pure pen-and-paper stuff, but that's a deeper worldview disagreement that I'm less confident in.)

I agree that 1) and 2) are likely to happen by default, but I also think 3) is pretty likely to happen by default too -- I see some gap there, just not that large of one. For all of #1-3 I'm thinking in terms of "speed up" relative to the default. I think right now if you look around, people could be doing this aligning narrowly superhuman models stuff and aren't, and I'd guess that if EAs don't make a push for it it'll get delayed by a

couple more years still, and less of it will get done. E.g. Paul was the one to start making human feedback a thing at all at OpenAI.

Evan Hubinger: Yeah—I definitely agree that Paul's work is good but hard to replicate. "Do more Paul stuff" was definitely not the alternative I was proposing. My thinking was just that it seems like focusing on downstream tasks with the hope of that leading to good insights into how to align language models feels less direct—and thus less likely to yield good insights by default—than just focusing on aligning language models directly. I buy that just framing it as "get GPT-3 to give honest answers" might lead people to not work on anything superhuman, though I don't yet feel fully convinced that there isn't still a yet better framing than either of those—that both emphasizes the importance of doing superhuman things but also focuses on the directly useful alignment task rather than things that are downstream of that.

Ajeya Cotra: Hm I think there's still some disconnect between us here -- the way I'm thinking about it, the activities I'm proposing here simply are aligning large models (most of them are language models but I don't want to be married to that). I don't see it as "doing things that might give us insight into how to align narrowly superhuman models"; it just seems like aligning them, i.e. getting them to try their best to use all their faculties to help us. I want to find ways to just directly practice the long-run thing.

I'm definitely open to a better framing of the thing I'm saying that's more likely to inspire productive work, though.

Evan Hubinger: Taking your examples from earlier:

> Coding is an example that doesn't seem like it's about "honest helpful answers": I'm excited to figure out how to get non-software engineers to effectively supervise a coding model to write better code. I'm also interested in e.g. getting GPT-3 to write a good mystery story using feedback from people who aren't good writers, or maybe don't even speak the language.

I think I wouldn't say that any of these things "simply are aligning large models"—they all feel like particular tasks which likely require some amount of alignment to get them to work, but also require lots of other non-alignment stuff as well. Ideally, I'd much prefer work that cuts out the non-alignment stuff and just focuses on the alignment stuff, but I think it's basically impossible to do that if you're trying to actually produce a model which is useful in practice for some downstream use case, since you're just not going to be able to do that without a ton of non-alignment-relevant work. I certainly think it's reasonable, even if you're just trying to do alignment work, to have some particular task in mind that you focus on, but once you start trying to do something like produce a product (not necessarily even for profit, just something that you want to be useful for real users), I expect most of the work that you'll end up needing to do for that won't be alignment-relevant.

Ajeya Cotra: Hm, my take is like "alignment = trying to do what we want." In the end we want models that are trying to run countries and companies and shepherd the future the way we want; in the near-term we could get models that are trying to do what we want as personal assistants and research assistants, and right now I think we should be tackling the hardest tasks where we could get models to try to do what we want, ideally where they are already better than us at the task.

I think of "seeing a glimmer of usefulness" as a proxy for "did you pick the hardest tasks", not as the end goal. I agree you'll need to do a bunch of stuff to make it

maximally useful that isn't the core part that seems like "aligning it" to me (basically training it with human feedback augmented/arranged in a certain way), and I think researchers working on aligning narrowly superhuman models should skip that work. But I don't think I understand the view that alignment is something that can totally be divorced from a task.

As a concrete example of the usefulness bar I'm thinking will usually make sense, take the Paul lab paper [Stiennon et al 2020](#): it's definitely not like a summarization product, and they left tons of things on the table (like collecting demonstrations from expert writers and hardcoding certain heuristics) that would feed into a real product. But it feels sort of markedly more useful than raw GPT-3, as a direct result of the fact that the model is now trying to use its faculties to be helpful instead of play an improv game. That's kind of the threshold that I think we should be aiming for.

Evan Hubinger: I certainly don't think that alignment can be totally divorced from a task, at least no more so than capabilities—and in fact I think the analogy to capabilities is very apt here. When you focus on solving Go, if you try to do it in a task-agnostic way, you learn a lot about AI capabilities in general. On the other hand, if you try to do it in a way that is very specific to Go, you don't learn very much about AI capabilities at all. Similarly, I expect relatively open-ended research on getting large language models to do helpful things in a relatively generic way to be useful for learning about alignment in general, but narrow work on getting large language models to solve specific tasks in a relatively tasks-specific way to not be very useful.

That being said, if you think "Learning to summarize from human feedback" is a good example of what you're talking about, then maybe we just agree, because I feel like it's a good example of what I'm talking about also—that is, an example of relatively open-ended research that was just trying to get a large language model to do a helpful alignment thing, rather than actually produce anything that might actually be used in the real world as a summarization tool.

Ajeya Cotra: Yeah I agree with trying to do task-general techniques and using tasks as case studies; I tried to emphasize this in the post but maybe I can tweak to make more clear or prominent.

Evan Hubinger: Yeah, I definitely think that's good—though I think a part of what I'm saying also is that the message of "solve specific task X" is likely to lead to that sort of task-specific work, whereas something more like "figure out how to make GPT-3 honest" is less likely to do that, in my opinion.

Ajeya Cotra: I want the one sentence summary to be "align narrowly superhuman models" rather than "solve specific task X"; the "align narrowly superhuman models" line doesn't seem less likely to lead to good work than the "Make GPT-3 honest" line to me (and right now I think it would lead to better work because of the problem of "honesty" calling to mind "avoid certain specific lies" and not sufficiently pushing the researcher to consider the hardest cases).

Evan Hubinger: Yeah, I think that's a good pitch to try, but I still worry that you'll end up with a lot of useless product engineering.

Ajeya Cotra: Cool thanks, this was a helpful discussion -- I just added a section on "Why not test the Paul stuff"

Evan Hubinger: I'm glad I was able to be helpful! :)

As I said above, **Open Phil is not soliciting grant applications right now** from people who want to try it out -- this blog post is my personal viewpoint, and institutionally we're still figuring out how much we want to prioritize this (discussion and arguments surrounding this post will feed into that).

Eliezer Yudkowsky: If I thought a proposal in this subarea looked as good as "[learning to summarize](#)" and OpenPhil wasn't picking it up or was throwing paperwork in front of it, I'd start sending emails to other funding sources or possibly even have MIRI fund it directly. We obviously have a lot less total funding, but **not** getting done what can be done in ML alignment is kind of... not acceptable at this point.

7. Appendix: beyond sandwiching?

I'm definitely very unsure what this would look like, but an important starting assumption I have is that whatever techniques worked well to get less-capable humans to reproduce the judgments of more-capable humans in a "sandwich" setting stand a good chance of just continuing to work.

Eliezer Yudkowsky: I remark that this is the kind of thought that needs to be hedged around **very** carefully on pain of undignified planetary extinction. Did you scale the answers of less-capable humans to results checkable by more-capable humans, while operating the AI under the capability threshold for modeling human psychology in detail, and are you assuming the same technique will generalize if an AGI is that smart? Did you scale worse human answers to checkable better answers while applying a small amount of optimization power, and are you assuming the same method will scale to using much more power than that? Did you scale worse to better across an identical environmental distribution, and are you hoping the same applies when the environmental distribution is being effectively altered between training and testing by the impact of an AGI that's smarter than when it was trained? And so on and so on.

I'm not saying it's useless to poke around and find things that seem to work for scaling unreliable human answers to better-than-those-humans answers, that smarter humans can still check to see if the whole method worked. I'm saying that if researchers actually believe the part where the journalists are like "and lo the entire alignment problem has been solved!", and the authors don't explicitly list out five stability conditions that held inside the experiment that might be necessary conditions, that's not what I'd call dignified.

A Semitechnical Introductory Dialogue on Solomonoff Induction

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

(Originally posted in December 2015: A dialogue between Ashley, a computer scientist who's never heard of [Solomonoff's theory of inductive inference](#), and Blaine, who thinks it is the best thing since sliced bread.)

i. Unbounded analysis

ASHLEY: Good evening, Msr. Blaine.

BLAINE: Good evening, Msr. Ashley.

ASHLEY: I've heard there's this thing called "Solomonoff's theory of inductive inference".

BLAINE: The rumors have spread, then.

ASHLEY: Yeah, so, what the heck is that about?

BLAINE: Invented in the 1960s by the mathematician Ray Solomonoff, the key idea in Solomonoff induction is to do sequence prediction by using Bayesian updating on a prior composed of a mixture of all computable probability distributions—

ASHLEY: Wait. Back up a lot. Before you try to explain what Solomonoff induction *is*, I'd like you to try to tell me what it *does*, or why people study it in the first place. I find that helps me organize my listening. Right now I don't even know why I should be interested in this.

BLAINE: Um, okay. Let me think for a second...

ASHLEY: Also, while I can imagine things that "sequence prediction" might mean, I haven't yet encountered it in a technical context, so you'd better go a bit further back and start more at the beginning. I do know what "computable" means and what a "probability distribution" is, and I remember the formula for [Bayes's Rule](#) although it's been a while.

BLAINE: Okay. So... one way of framing the usual reason why people study this general field in the first place, is that sometimes, by studying certain idealized mathematical questions, we can gain valuable intuitions about epistemology. That's, uh, the field that studies how to reason about factual questions, how to build a map of reality that reflects the territory—

ASHLEY: I have some idea what 'epistemology' is, yes. But I think you might need to start even further back, maybe with some sort of concrete example or something.

BLAINE: Okay. Um. So one anecdote that I sometimes use to frame the value of computer science to the study of epistemology is Edgar Allan Poe's argument in 1833 that chess was uncomputable.

ASHLEY: That doesn't sound like a thing that actually happened.

BLAINE: I know, but it totally *did* happen and not in a metaphorical sense either! Edgar Allan Poe wrote an [essay](#) explaining why no automaton would ever be able to play chess, and he specifically mentioned "Mr. Babbage's computing engine" as an example.

You see, in the nineteenth century, there was for a time this sensation known as the Mechanical Turk—supposedly a machine, an automaton, that could play chess. At the grandmaster level, no less.

Now today, when we're accustomed to the idea that it takes a reasonably powerful computer to do that, we can know *immediately* that the Mechanical Turk must have been a fraud and that there must have been a concealed operator inside—a person with dwarfism, as it turned out. Today we know that this sort of thing is *hard* to build into a machine. But in the 19th century, even that much wasn't known.

So when Edgar Allan Poe, who besides being an author was also an accomplished magician, set out to write an essay about the Mechanical Turk, he spent the *second* half of the essay dissecting what was known about the Turk's appearance to (correctly) figure out where the human operator was hiding. But Poe spent the first half of the essay arguing that no automaton—nothing like Mr. Babbage's computing engine—could possibly play chess, which was how he knew *a priori* that the Turk had a concealed human operator.

ASHLEY: And what was Poe's argument?

BLAINE: Poe observed that in an algebraical problem, each step followed from the previous step of necessity, which was why the steps in solving an algebraical problem could be represented by the deterministic motions of gears in something like Mr. Babbage's computing engine. But in a chess problem, Poe said, there are many possible chess moves, and no move follows with necessity from the position of the board; and even if you did select one move, the

opponent's move would not follow with necessity, so you couldn't represent it with the determined motion of automatic gears. Therefore, Poe said, whatever was operating the Mechanical Turk must have the nature of Cartesian mind, rather than the nature of deterministic matter, and this was knowable *a priori*. And then he started figuring out where the required operator was hiding.

ASHLEY: That's some amazingly impressive reasoning for being completely wrong.

BLAINE: I know! Isn't it great?

ASHLEY: I mean, that sounds like Poe correctly identified the *hard* part of playing computer chess, the branching factor of moves and countermoves, which is the reason why no *simple* machine could do it. And he just didn't realize that a deterministic machine could deterministically check many possible moves in order to figure out the game tree. So close, and yet so far.

BLAINE: More than a century later, in 1950, Claude Shannon published the first paper ever written on computer chess. And in passing, Shannon gave the formula for playing perfect chess if you had unlimited computing power, the algorithm you'd use to extrapolate the entire game tree. We could say that Shannon gave a short program that would solve chess if you ran it on a hypercomputer, where a hypercomputer is an ideal computer that can run any finite computation immediately. And then Shannon passed on to talking about the problem of locally guessing how good a board position was, so that you could play chess using only a *small* search.

I say all this to make a point about the value of knowing how to solve problems using hypercomputers, even though hypercomputers don't exist. Yes, there's often a *huge* gap between the unbounded solution and the practical solution. It wasn't until 1997, forty-seven years after Shannon's paper giving the unbounded solution, that Deep Blue actually won the world chess championship—

ASHLEY: And that wasn't just a question of faster computing hardware running Shannon's ideal search algorithm. There were a lot of new insights along the way, most notably the alpha-beta pruning algorithm and a lot of improvements in positional evaluation.

BLAINE: Right!

But I think some people overreact to that forty-seven year gap, and act like it's *worthless* to have an unbounded understanding of a computer program, just because you might still be forty-seven years away from a practical solution. But if you don't even have a solution that would run on a hypercomputer, you're Poe in 1833, not Shannon in 1950.

The reason I tell the anecdote about Poe is to illustrate that Poe was *confused* about computer chess in a way that Shannon was not. When we don't know how to solve a problem even given infinite computing power, the very work we are trying to do is in some sense murky to us. When we can state code that would solve the problem given a hypercomputer, we have become *less* confused. Once we have the unbounded solution we understand, in some basic sense, *the kind of work we are trying to perform*, and then we can try to figure out how to do it efficiently.

ASHLEY: Which may well require new insights into the structure of the problem, or even a conceptual revolution in how we imagine the work we're trying to do.

BLAINE: Yes, but the point is that you can't even get started on that if you're arguing about how playing chess has the nature of Cartesian mind rather than matter. At that point you're not 50 years away from winning the chess championship, you're 150 years away, because it took an extra 100 years to move humanity's understanding to the point where Claude Shannon could trivially see how to play perfect chess using a large-enough computer. I'm not trying to exalt the unbounded solution by denigrating the work required to get a bounded solution. I'm not saying that when we have an unbounded solution we're practically there and the rest is a matter of mere lowly efficiency. I'm trying to compare having the unbounded solution to the horrific confusion of *not understanding what we're trying to do*.

ASHLEY: Okay, I think I understand why, on your view, it's important to know how to solve problems using infinitely fast computers, or hypercomputers as you call them. When we can say how to answer a question using infinite computing power, that means we crisply understand the question itself, in some sense; while if we can't figure out how to solve a problem using unbounded computing power, that means we're *confused* about the problem, in some sense. I mean, anyone who's ever tried to teach the more doomed sort of undergraduate to write code knows what it means to be confused about what it takes to compute something.

BLAINE: Right.

ASHLEY: So what does this have to do with "Solomonoff induction"?

BLAINE: Ah! Well, suppose I asked you how to do epistemology using infinite computing power?

ASHLEY: My good fellow, I would at once reply, "Beep. Whirr. Problem 'do epistemology' not crisply specified." At this stage of affairs, I do not think this reply indicates any fundamental confusion on my part; rather I think it is you who must be clearer.

BLAINE: Given unbounded computing power, how would you reason in order to construct an accurate map of reality?

ASHLEY: That still strikes me as rather underspecified.

BLAINE: Perhaps. But even there I would suggest that it's a mark of intellectual progress to be able to take vague and underspecified ideas like 'do good epistemology' and turn them *into* crisply specified problems. Imagine that I went up to my friend Cecil, and said, "How would you do good epistemology given unlimited computing power and a short Python program?" and Cecil at once came back with an answer—a good and reasonable answer, once it was explained. Cecil would probably know something quite interesting that you do not presently know.

ASHLEY: I confess to being rather skeptical of this hypothetical. But if that actually happened—if I agreed, to my own satisfaction, that someone had stated a short Python program that would 'do good epistemology' if run on an unboundedly fast computer—then I agree that I'd probably have learned something *quite interesting* about epistemology.

BLAINE: What Cecil knows about, in this hypothetical, is Solomonoff induction. In the same way that Claude Shannon answered "Given infinite computing power, how would you play perfect chess?", Ray Solomonoff answered "Given infinite computing power, how would you perfectly find the best hypothesis that fits the facts?"

ASHLEY: Suddenly, I find myself strongly suspicious of whatever you are about to say to me.

BLAINE: That's understandable.

ASHLEY: In particular, I'll ask at once whether "Solomonoff induction" assumes that our hypotheses are being given to us on a silver platter along with the exact data we're supposed to explain, or whether the algorithm is organizing its own data from a big messy situation and inventing good hypotheses from scratch.

BLAINE: Great question! It's the second one.

ASHLEY: Really? Okay, now I have to ask whether Solomonoff induction is a recognized concept in good standing in the field of academic computer science, because that does not sound like something modern-day computer science knows how to do.

BLAINE: I wouldn't say it's a widely known concept, but it's one that's in good academic standing. The method isn't used in modern machine learning because it requires an infinitely fast computer and isn't easily approximated the way that chess is.

ASHLEY: This really sounds very suspicious. Last time I checked, we hadn't *begun* to formalize the creation of good new hypotheses from scratch. I've heard about claims to have 'automated' the work that, say, Newton did in inventing classical mechanics, and I've found them all to be incredibly dubious. Which is to say, they were rigged demos and lies.

BLAINE: I know, but—

ASHLEY: And then I'm even more suspicious of a claim that someone's algorithm would solve this problem if only they had infinite computing power. Having some researcher claim that their Good-Old-Fashioned AI semantic network *would* be intelligent if run on a computer so large that, conveniently, nobody can ever test their theory, is not going to persuade me.

BLAINE: Do I really strike you as that much of a charlatan? What have I ever done to you, that you would expect me to try pulling a scam like that?

ASHLEY: That's fair. I shouldn't accuse you of planning that scam when I haven't seen you say it. But I'm pretty sure the problem of "coming up with good new hypotheses in a world full of messy data" is [AI-complete](#). And even Mentif-

BLAINE: Do not say the name, or he will appear!

ASHLEY: Sorry. Even the legendary first and greatest of all AI crackpots, He-Who-Googlez-His-Name, could assert that his algorithms would be all-powerful on a computer large enough to make his claim unfalsifiable. So what?

BLAINE: That's a very sensible reply and this, again, is exactly the kind of mental state that reflects a problem that is *confusing* rather than just hard to implement. It's the sort of confusion Poe might feel in 1833, or close to it. In other words, it's just the sort of conceptual issue we *would* have solved at the point where we could state a short program that could run on a hypercomputer. Which Ray Solomonoff did in 1964.

ASHLEY: Okay, let's hear about this supposed general solution to epistemology.

ii. Sequences

BLAINE: First, try to solve the following puzzle. 1, 3, 4, 7, 11, 18, 29...?

ASHLEY: Let me look at those for a moment... 47.

BLAINE: Congratulations on engaging in, as we snooty types would call it, 'sequence prediction'.

ASHLEY: I'm following you so far.

BLAINE: The smarter you are, the more easily you can find the hidden patterns in sequences and predict them successfully. You had to notice the resemblance to the Fibonacci rule to guess the next number. Someone who didn't already know about Fibonacci, or who was worse at mathematical thinking, would have taken longer to understand the sequence or maybe never learned to predict it at all.

ASHLEY: Still with you.

BLAINE: It's not a sequence of *numbers* per se... but can you see how the question, "The sun has risen on the last million days. What is the probability that it rises tomorrow?" could be viewed as a kind of sequence prediction problem?

ASHLEY: Only if some programmer neatly parses up the world into a series of "Did the Sun rise on day X starting in 4.5 billion BCE, 0 means no and 1 means yes? 1, 1, 1, 1, 1..." and so on. Which is exactly the sort of shenanigan that I see as cheating. In the real world, you go outside and see a brilliant ball of gold touching the horizon, not a giant "1".

BLAINE: Suppose I have a robot running around with a webcam showing it a 1920×1080 pixel field that refreshes 60 times a second with 32-bit colors. I could view that as a giant sequence and ask the robot to predict what it will see happen when it rolls out to watch a sunrise the next day.

ASHLEY: I can't help but notice that the 'sequence' of webcam frames is absolutely enormous, like, the sequence is made up of 66-megabit 'numbers' appearing 3600 times per minute... oh, right, computers much bigger than the universe. And now you're smiling evilly, so I guess that's the point. I also notice that the sequence is no longer deterministically predictable, that it is no longer a purely mathematical object, and that the sequence of webcam frames observed will depend on the robot's choices. This makes me feel a bit shaky about the analogy to predicting the mathematical sequence 1, 1, 2, 3, 5.

BLAINE: I'll try to address those points in order. First, Solomonoff induction is about assigning *probabilities* to the next item in the sequence. I mean, if I showed you a box that said 1, 1, 2, 3, 5, 8 you would not be absolutely certain that the next item would be 13. There could be some more complicated rule that just looked Fibonacci-ish but then diverged. You might guess with 90% probability but not 100% probability, or something like that.

ASHLEY: This has stopped feeling to me like math.

BLAINE: There is a *large* branch of math, to say nothing of computer science, that deals in probabilities and statistical prediction. We are going to be describing absolutely lawful and deterministic ways of assigning probabilities after seeing 1, 3, 4, 7, 11, 18.

ASHLEY: Okay, but if you're later going to tell me that this lawful probabilistic prediction rule underlies a generally intelligent reasoner, I'm already skeptical.

No matter how large a computer it's run on, I find it hard to imagine that some simple set of rules for assigning probabilities is going to encompass truly and generally intelligent answers about sequence prediction, like [Terence Tao](#) would give after looking at the sequence for a while. We just have no idea how Terence Tao works, so we can't duplicate his abilities in a formal rule, no matter how much computing power that rule gets... you're smiling evilly again. I'll be *quite* interested if that evil smile turns out to be justified.

BLAINE: Indeed.

ASHLEY: I also find it hard to imagine that this deterministic mathematical rule for assigning probabilities would notice if a box was outputting an encoded version of "To be or not to be" from Shakespeare by mapping A to Z onto 1 to 26, which I would notice eventually though not immediately upon seeing 20, 15, 2, 5, 15, 18... And you're *still* smiling evilly.

BLAINE: Indeed. That is *exactly* what Solomonoff induction does. Furthermore, we have theorems establishing that Solomonoff induction can do it way better than you or Terence Tao.

ASHLEY: A *theorem* proves this. As in a necessary mathematical truth. Even though we have no idea how Terence Tao works empirically... and there's evil smile number four. Okay. I am very skeptical, but willing to be convinced.

BLAINE: So if you actually did have a hypercomputer, you could cheat, right? And Solomonoff induction is the most ridiculously cheating cheat in the history of cheating.

ASHLEY: Go on.

BLAINE: We just run all possible computer programs to see which are the simplest computer programs that best predict the data seen so far, and use those programs to predict what comes next. This mixture contains, among other things, an exact copy of Terence Tao, thereby allowing us to prove theorems about their relative performance.

ASHLEY: Is this an actual reputable math thing? I mean really?

BLAINE: I'll deliver the formalization later, but you did ask me to first state the point of it all. The point of Solomonoff induction is that it gives us a gold-standard ideal for sequence prediction, and this gold-standard prediction only errs by a bounded amount, over infinite time, relative to the best computable sequence predictor. We can also see it as formalizing the intuitive idea that was expressed by William Ockham a few centuries earlier that simpler theories are

more likely to be correct, and as telling us that 'simplicity' should be measured in algorithmic complexity, which is the size of a computer program required to output a hypothesis's predictions.

ASHLEY: I think I would have to read more on this subject to actually follow that. What I'm hearing is that Solomonoff induction is a reputable idea that is important because it gives us a kind of ideal for sequence prediction. This ideal also has something to do with Occam's Razor, and stakes a claim that the simplest theory is the one that can be represented by the shortest computer program. You identify this with "doing good epistemology".

BLAINE: Yes, those are legitimate takeaways. Another way of looking at it is that Solomonoff induction is an ideal but uncomputable answer to the question "What should our priors be?", which is left open by understanding [Bayesian updating](#).

ASHLEY: Can you say how Solomonoff induction answers the question of, say, the prior probability that Canada is planning to invade the United States? I once saw a crackpot website that tried to invoke Bayesian probability about it, but only after setting the prior at 10% or something like that, I don't recall exactly. Does Solomonoff induction let me tell him that he's making a math error, instead of just calling him silly in an informal fashion?

BLAINE: If you're expecting to sit down with Leibniz and say, "Gentlemen, let us calculate" then you're setting your expectations too high. Solomonoff gives us an idea of how we *should* compute that quantity given unlimited computing power. It doesn't give us a firm recipe for how we can best approximate that ideal in real life using bounded computing power, or human brains. That's like expecting to play perfect chess after you read Shannon's 1950 paper. But knowing the ideal, we can extract some intuitive advice that might help our online crackpot if only he'd listen.

ASHLEY: But according to you, Solomonoff induction does say in principle what is the prior probability that Canada will invade the United States.

BLAINE: Yes, up to a choice of universal Turing machine.

ASHLEY: (*looking highly skeptical*) So I plug a universal Turing machine into the formalism, and in principle, I get out a uniquely determined probability that Canada invades the USA.

BLAINE: Exactly!

ASHLEY: Uh huh. Well, go on.

BLAINE: So, first, we have to transform this into a sequence prediction problem.

ASHLEY: Like a sequence of years in which Canada has and hasn't invaded the US, mostly zero except around 1812—

BLAINE: *No!* To get a good prediction about Canada we need much more data than that, and I don't mean a graph of Canadian GDP either. Imagine a sequence that contains all the sensory data you have ever received over your lifetime. Not just the hospital room that you saw when you opened your eyes right after your birth, but the darkness your brain received as input while you were still in your mother's womb. Every word you've ever heard. Every letter you've ever seen on a computer screen, not as ASCII letters but as the raw pattern of neural impulses that gets sent down from your retina.

ASHLEY: That seems like a lot of data and some of it is redundant, like there'll be lots of similar pixels for blue sky—

BLAINE: That data is what *you* got as an agent. If we want to translate the question of the prediction problem Ashley faces into theoretical terms, we should give the sequence predictor *all* the data that you had available, including all those repeating blue pixels of the sky. Who knows? Maybe there was a Canadian warplane somewhere in there, and you didn't notice.

ASHLEY: But it's impossible for my brain to remember all that data. If we neglect for the moment how the retina actually works and suppose that I'm seeing the same $1920 \times 1080 @60\text{Hz}$ feed the robot would, that's far more data than my brain can realistically learn per second.

BLAINE: So then Solomonoff induction can do better than you can, using its unlimited computing power and memory. That's fine.

ASHLEY: But what if you can do better by forgetting more?

BLAINE: If you have limited computing power, that makes sense. With unlimited computing power, that really shouldn't happen and that indeed is one of the lessons of Solomonoff induction. An unbounded Bayesian never expects to do worse by updating on another item of evidence—for one thing, you can always just do the same policy you would have used if you hadn't seen that evidence. That kind of lesson is one of the lessons that might not be intuitively obvious, but which you can feel more deeply by walking through the math of probability theory. With unlimited computing power, nothing goes wrong as a result of trying to process 4 gigabits per second; every extra bit just produces a better expected future prediction.

ASHLEY: Okay, so we start with literally all the data I have available. That's 4 gigabits per second if we imagine 1920×1080 frames of 32-bit pixels repeating 60 times per second. Though I remember hearing 100 megabits per second would be a better estimate of what the retina sends out, and that it's pared down to 1 megabit per second very quickly by further processing.

BLAINE: Right. We start with all of that data, going back to when you were born. Or maybe when your brain formed in the womb, though it shouldn't make much difference.

ASHLEY: I note that there are some things I know that don't come from my sensory inputs at all. Chimpanzees learn to be afraid of skulls and snakes much faster than they learn to be afraid of other arbitrary shapes. I was probably better at learning to walk in Earth gravity than I would have been at navigating in zero G. Those are heuristics I'm born with, based on how my brain was wired, which ultimately stems from my DNA specifying the way that proteins should fold to form neurons—not from any photons that entered my eyes later.

BLAINE: So, for purposes of following along with the argument, let's say that your DNA is analogous to the code of a computer program that makes predictions. What you're observing here is that humans have 750 megabytes of DNA, and even if most of that is junk and not all of what's left is specifying brain behavior, it still leaves a pretty large computer program that could have a lot of prior information programmed into it.

Let's say that your brain, or rather, your infant pre-brain wiring algorithm, was effectively a 7.5 megabyte program—if it's actually 75 megabytes, that makes little difference to the argument. By exposing that 7.5 megabyte program to all the information coming in from your eyes, ears, nose, proprioceptive sensors telling you where your limbs were, and so on, your brain updated itself into forming the modern Ashley, whose hundred trillion synapses might be encoded by, say, one petabyte of information.

ASHLEY: The thought does occur to me that some environmental phenomena have effects on me that can't be interpreted as "sensory information" in any simple way, like the direct effect that alcohol has on my neurons, and how that feels to me from the inside. But it would be perverse to claim that this prevents you from trying to summarize all the information that the Ashley-agent receives into a single sequence, so I won't press the point.

(ELIEZER: (*whispering*) More on this topic later.)

ASHLEY: Oh, and for completeness's sake, wouldn't there also be further information embedded in the laws of physics themselves? Like, the way my brain executes implicitly says something about the laws of physics in the universe I'm in.

BLAINE: Metaphorically speaking, our laws of physics would play the role of a particular choice of Universal Turing Machine, which has some effect on which computations count as "simple" inside the Solomonoff formula. But normally, the UTM should be very simple compared to the amount of data in the sequence we're trying to predict, just like the laws of physics are very simple compared to a human brain. In terms of [algorithmic complexity](#), the laws of physics are very simple compared to watching a $1920 \times 1080 @60\text{Hz}$ visual field for a day.

ASHLEY: Part of my mind feels like the laws of physics are quite complicated compared to going outside and watching a sunset. Like, I realize that's false, but I'm not sure how to say out loud exactly why it's false...

BLAINE: Because the algorithmic complexity of a system isn't measured by how long a human has to go to college to understand it, it's measured by the size of the computer program required to generate it. The language of physics is differential equations, and it turns out that this is something difficult to beat into some human brains, but differential equations are simple to program into a simple Turing Machine.

ASHLEY: Right, like, the laws of physics actually have much fewer details to them than, say, human nature. At least on the Standard Model of Physics. I mean, in principle there could be another decillion undiscovered particle families out there.

BLAINE: The concept of "algorithmic complexity" isn't about seeing something with lots of gears and details, it's about the size of computer program required to compress all those details. The [Mandelbrot set](#) looks very complicated visually, you can keep zooming in using more and more detail, but there's a very simple rule that generates it, so we say the algorithmic complexity is very low.

ASHLEY: All the visual information I've seen is something that happens *within* the physical universe, so how can it be more complicated than the universe? I mean, I have a sense on some level that this shouldn't be a problem, but I don't know why it's not a problem.

BLAINE: That's because particular parts of the universe can have much higher algorithmic complexity than the entire universe!

Consider a library that contains all possible books. It's very easy to write a computer program that generates all possible books. So any *particular* book in the library contains much more algorithmic information than the *entire* library; it contains the information required to say 'look at this particular book here'.

If π is normal, then somewhere in its digits is a copy of Shakespeare's *Hamlet*—but the number saying which particular digit of π to start looking at, will be just about exactly as large as *Hamlet* itself. The copy of Shakespeare's *Hamlet* that exists in the decimal expansion of π is more complex than π itself.

If you zoomed way in and restricted your vision to a particular part of the Mandelbrot set, what you saw might be much more *algorithmically* complex than the entire Mandelbrot set, because the specification has to say where in the Mandelbrot set you are.

Similarly, the world Earth is much more algorithmically complex than the laws of physics. Likewise, the visual field you see over the course of a second can easily be far more algorithmically complex than the laws of physics.

ASHLEY: Okay, I think I get that. And similarly, even though the ways that proteins fold up are very complicated, in principle we could get all that info using just the simple fundamental laws of physics plus the relatively simple DNA code for the protein. There are all sorts of obvious caveats about epigenetics and so on, but those caveats aren't likely to change the numbers by a whole order of magnitude.

BLAINE: Right!

ASHLEY: So the laws of physics are, like, a few kilobytes, and my brain has say 75 megabytes of innate wiring instructions. And then I get to see a lot more information than that over my lifetime, like a megabit per second after my initial visual system finishes preprocessing it, and then most of that is forgotten. Uh... what does that have to do with Solomonoff induction again?

BLAINE: Solomonoff induction quickly catches up to any single computer program at sequence prediction, even if the original program is very large and contains a lot of prior information about the environment. If a program is 75 megabytes long, it can only predict 75 megabytes worth of data better than the Solomonoff inductor before the Solomonoff inductor catches up to it.

That doesn't mean that a Solomonoff inductor knows everything a baby does after the first second of exposure to a webcam feed, but it does mean that after the first second, the Solomonoff inductor is already no more surprised than a baby by the vast majority of pixels in the next frame.

Every time the Solomonoff inductor assigns half as much probability as the baby to the next pixel it sees, that's one bit spent permanently out of the 75 megabytes of error that can happen before the Solomonoff inductor catches up to the baby.

That your brain is written in the laws of physics also has some implicit correlation with the environment, but that's like saying that a program is written in the same programming language as the environment. The language can contribute something to the power of the program, and the environment being written in the same programming language can be a kind of prior knowledge. But if Solomonoff induction starts from a standard Universal Turing Machine as its language, that doesn't contribute any more bits of lifetime error than the complexity of that programming language in the UTM.

ASHLEY: Let me jump back a couple of steps and return to the notion of my brain wiring itself up in response to environmental information. I'd expect an important part of that process was my brain learning to *control* the environment, not just passively observing it. Like, it mattered to my brain's wiring algorithm that my brain saw the room shift in a certain way when it sent out signals telling my eyes to move.

BLAINE: Indeed. But talking about the sequential *control* problem is more complicated math. [AIXI](#) is the ideal agent that uses Solomonoff induction as its epistemology and expected reward as its decision theory. That introduces extra complexity, so it makes sense to talk about just Solomonoff induction first. We can talk about AIXI later. So imagine for the moment that we were *just* looking at your sensory data, and trying to predict what would come next in that.

ASHLEY: Wouldn't it make more sense to look at the brain's inputs *and* outputs, if we wanted to predict the next input? Not just look at the series of previous inputs?

BLAINE: It'd make the problem easier for a Solomonoff inductor to solve, sure; but it also makes the problem more complicated. Let's talk instead about what would happen if you took the complete sensory record of your life, gave it to an ideally smart agent, and asked the agent to predict what you would see next. Maybe the agent could do an even better job of prediction if we also told it about your brain's outputs, but I don't think that subtracting the outputs would leave it helpless to see patterns in the inputs.

ASHLEY: It sounds like a pretty hard problem to me, maybe even an unsolvable one. I'm thinking of the distinction in computer science between needing to learn from non-chosen data, versus learning when you can choose particular queries. Learning can be much faster in the second case.

BLAINE: In terms of what can be predicted *in principle* given the data, what facts are *actually reflected in it* that Solomonoff induction might uncover, we shouldn't imagine a human trying to analyze the data. We should imagine [an entire advanced civilization pondering it for years](#). If you look at it from that angle, then the alien civilization isn't going to balk at the fact that it's looking at the answers to the queries that Ashley's brain chose, instead of the answers to the queries it chose itself.

Like, if the Ashley had already read Shakespeare's *Hamlet*—if the image of those pages had already crossed the sensory stream—and then the Ashley saw a mysterious box outputting 20, 15, 2, 5, 15, 18, I think somebody eavesdropping on that sensory data would be equally able to guess that this was encoding 'tobear' and guess that the next thing the Ashley saw might be the box outputting 14. You wouldn't even need an entire alien civilization of superintelligent cryptographers to guess that. And it definitely wouldn't be a killer problem that Ashley was controlling the eyeball's saccades, even if you could learn even faster by controlling the eyeball yourself.

So far as the computer-science distinction goes, Ashley's eyeball *is* being controlled to make intelligent queries and seek out useful information; it's just Ashley controlling the eyeball instead of you—that eyeball is not a query-oracle answering *random* questions.

ASHLEY: Okay, I think this example is helping my understanding of what we're doing here. In the case above, the next item in the Ashley-sequence wouldn't actually be 14. It would be this huge 1920×1080 visual field that showed

the box flashing a little picture of '14'.

BLAINE: Sure. Otherwise it would be a rigged demo, as you say.

ASHLEY: I think I'm confused about the idea of *predicting* the visual field. It seems to me that what with all the dust specks in my visual field, and maybe my deciding to tilt my head using motor instructions that won't appear in the sequence, there's no way to *exactly* predict the 66-megabit integer representing the next visual frame. So it must be doing something other than the equivalent of guessing "14" in a simpler sequence, but I'm not sure what.

BLAINE: Indeed, there'd be some element of thermodynamic and quantum randomness preventing that exact prediction even in principle. So instead of predicting one particular next frame, we put a probability distribution on it.

ASHLEY: A probability distribution over possible 66-megabit frames? Like, a table with $2^{66,000,000}$ entries, summing to 1?

BLAINE: Sure. $2^{32 \times 1920 \times 1080}$ isn't a large number when you have unlimited computing power. As Martin Gardner once observed, "Most finite numbers are very much larger." Like I said, Solomonoff induction is an epistemic ideal that requires an unreasonably large amount of computing power.

ASHLEY: I don't deny that big computations can sometimes help us understand little ones. But at the point when we're talking about probability distributions that large, I have some trouble holding onto what the probability distribution is supposed to *mean*.

BLAINE: Really? Just imagine a probability distribution over N possibilities, then let N go to $2^{66,000,000}$. If we were talking about a letter ranging from A to Z, then putting 100 times as much probability mass on (X, Y, Z) as on the rest of the alphabet, would say that although you didn't know *exactly* what letter would happen, you expected it would be toward the end of the alphabet. You would have used 26 probabilities, summing to 1, to precisely state that prediction.

In Solomonoff induction, since we have unlimited computing power, we express our uncertainty about a 1920×1080 video frame the same way. All the various pixel fields you could see if your eye jumped to a plausible place, saw a plausible number of dust specks, and saw the box flash something that visually encoded '14', would have high probability. Pixel fields where the box vanished and was replaced with a glow-in-the-dark unicorn would have very low, though not zero, probability.

ASHLEY: Can we really get away with viewing things that way?

BLAINE: If we could not make identifications like these *in principle*, there would be no principled way in which we could say that you had ever *expected to see something happen*—no way to say that one visual field your eyes saw had higher probability than any other sensory experience. We couldn't justify science; we couldn't say that, having performed Galileo's experiment by rolling an inclined cylinder down a plane, Galileo's theory was thereby to some degree supported by having assigned a *high relative probability* to the only actual observations our eyes ever report.

ASHLEY: I feel a little unsure of that jump, but I suppose I can go along with that for now. Then the question of "What probability does Solomonoff induction assign to Canada invading?" is to be identified, in principle, with the question "Given my past life experiences and all the visual information that's entered my eyes, what is the relative probability of seeing visual information that encodes Google News with the headline 'CANADA INVADES USA' at some point during the next 300 million seconds?"

BLAINE: Right!

ASHLEY: And Solomonoff induction has an in-principle way of assigning this a relatively low probability, which that online crackpot could do well to learn from as a matter of principle, even if he couldn't *begin* to carry out the exact calculations that involve assigning probabilities to exponentially vast tables.

BLAINE: Precisely!

ASHLEY: Fairness requires that I congratulate you on having come further in formalizing 'do good epistemology' as a sequence prediction problem than I previously thought you might.

I mean, you haven't satisfied me yet, but I wasn't expecting you to get even this far.

iii. Hypotheses

BLAINE: Next, we consider how to represent a *hypothesis* inside this formalism.

ASHLEY: Hmm. You said something earlier about updating on a probabilistic mixture of computer programs, which leads me to suspect that in this formalism, a hypothesis or *way the world can be* is a computer program that outputs a sequence of integers.

BLAINE: There's indeed a version of Solomonoff induction that works like that. But I prefer the version where a hypothesis assigns *probabilities* to sequences. Like, if the hypothesis is that the world is a fair coin, then we shouldn't try to make that hypothesis predict "heads—tails—tails—tails—heads" but should let it just assign a 1/32 prior probability to the sequence **HTTHH**.

ASHLEY: I can see that for coins, but I feel a bit iffier on what this means as a statement *about the real world*.

BLAINE: A single hypothesis inside the Solomonoff mixture would be a computer program that took in a series of video frames, and assigned a probability to each possible next video frame. Or for greater simplicity and elegance, imagine a program that took in a sequence of bits, ones and zeroes, and output a rational number for the probability of the next bit being '1'. We can readily go back and forth between a program like that, and a probability distribution over sequences.

Like, if you can answer all of the questions, "What's the probability that the coin comes up heads on the first flip?", "What's the probability of the coin coming up heads on the second flip, if it came up heads on the first flip?", and "What's the probability that the coin comes up heads on the second flip, if it came up tails on the first flip?" then we can turn that into a probability distribution over sequences of two coinflips. Analogously, if we have a program that outputs the probability of the next bit, conditioned on a finite number of previous bits taken as input, that program corresponds to a probability distribution over infinite sequences of bits.

$$\begin{aligned} P_{\text{prog}}(\text{bits}_1 \dots N) &= \prod_{i=1}^N \text{InterpretProb}(\text{prog}(\text{bits}_1 \dots i-1), \text{bits}_i) \\ \text{InterpretProb}(\text{prog}(x), y) &= \begin{cases} \text{InterpretFrac}(\text{prog}(x)) & \text{if } y = 1 \\ 1 - \text{InterpretFrac}(\text{prog}(x)) & \text{if } y = 0 \\ 0 & \text{if } \text{prog}(x) \text{ does not halt} \end{cases} \end{aligned}$$

ASHLEY: I think I followed along with that in theory, though it's not a type of math I'm used to (yet). So then in what sense is a program that assigns probabilities to sequences, a way the world could be—a hypothesis about the world?

BLAINE: Well, I mean, for one thing, we can see the infant Ashley as a program with 75 megabytes of information about how to wire up its brain in response to sense data, that sees a bunch of sense data, and then experiences some degree of relative surprise. Like in the baby-looking-paradigm experiments where you show a baby an object disappearing behind a screen, and the baby looks longer at those cases, and so we suspect that babies have a concept of object permanence.

ASHLEY: That sounds like a program that's a way Ashley could be, not a program that's a way the world could be.

BLAINE: Those indeed are dual perspectives on the meaning of Solomonoff induction. Maybe we can shed some light on this by considering a simpler induction rule, Laplace's Rule of Succession, invented by the Reverend Thomas Bayes in the 1750s, and named after Pierre-Simon Laplace, the inventor of Bayesian reasoning.

ASHLEY: Pardon me?

BLAINE: Suppose you have a biased coin with an unknown bias, and every possible bias between 0 and 1 is equally probable.

ASHLEY: Okay. Though in the real world, it's quite likely that an unknown frequency is exactly 0, 1, or 1/2. If you assign equal probability density to every part of the real number field between 0 and 1, the probability of 1 is 0. Indeed, the probability of all rational numbers put together is zero.

BLAINE: The original problem considered by Thomas Bayes was about an ideal billiard ball bouncing back and forth on an ideal billiard table many times and eventually slowing to a halt; and then bouncing other billiards to see if they halted to the left or the right of the first billiard. You can see why, in first considering the simplest form of this problem without any complications, we might consider every position of the first billiard to be equally probable.

ASHLEY: Sure. Though I note with pointless pedantry that if the billiard was really an ideal rolling sphere and the walls were perfectly reflective, it'd never halt in the first place.

BLAINE: Suppose we're told that, after rolling the original billiard ball and then 5 more billiard balls, one billiard ball was to the right of the original, an **R**. The other four were to the left of the original, or **Ls**. Again, that's 1 **R** and 4 **Ls**. Given only this data, what is the probability that the next billiard ball rolled will be on the left of the original, another **L**?

ASHLEY: Five sevenths.

BLAINE: Ah, you've heard this problem before?

ASHLEY: No, but it's obvious.

BLAINE: Uh... really?

ASHLEY: Combinatorics. Consider just the orderings of the balls, instead of their exact positions. Designate the original ball with the symbol **|**, the next five balls as **LLLLR**, and the next ball to be rolled as **+**. Given that the current ordering of these six balls is **LLLL|R** and that all positions and spacings of the underlying balls are equally likely, after rolling the **+**, there will be seven equally likely orderings **+LLLL|R**, **L+LLL|R**, **LL+LL|R**, and so on up to **LLLL|+R** and **LLLL|R+**. In five of those seven orderings, the **+** is on the left of the **|**. In general, if we see M of **L** and N of **R**, the probability of the next item being an **L** is $(M + 1)/(M + N + 2)$.

BLAINE: Gosh... Well, the much more complicated proof originally devised by Thomas Bayes starts by considering every position of the original ball to be equally likely *a priori*, the additional balls as providing evidence about that position, and then integrating over the posterior probabilities of the original ball's possible positions to arrive at the probability that the next ball lands on the left or right.

ASHLEY: Heh. And is all that extra work useful if you also happen to know a little combinatorics?

BLAINE: Well, it tells me exactly how my beliefs about the original ball change with each new piece of evidence—the new posterior probability function on the ball's position. Suppose I instead asked you something along the lines of, "Given 4 **L** and 1 **R**, where do you think the original ball **+** is most likely to be on the number line? How likely is it to be within 0.1 distance of there?"

ASHLEY: That's fair; I don't see a combinatoric answer for the later part. You'd have to actually integrate over the density function $f^M(1 - f)^N df$.

BLAINE: Anyway, let's just take at face value that Laplace's Rule of Succession says that, after observing M 1s and N 0s, the probability of getting a 1 next is $(M + 1)/(M + N + 2)$.

ASHLEY: But of course.

BLAINE: We can consider Laplace's Rule as a short Python program that takes in a sequence of 1s and 0s, and spits out the probability that the next bit in the sequence will be 1. We can also consider it as a probability distribution over infinite sequences, like this:

- **0** : 1/2
- **1** : 1/2
- **00** : $1/2 * 2/3 = 1/3$
- **01** : $1/2 * 1/3 = 1/6$
- **000** : $1/2 * 2/3 * 3/4 = 1/4$
- **001** : $1/2 * 2/3 * 1/4 = 1/12$
- **010** : $1/2 * 1/3 * 1/2 = 1/12$

... and so on.

Now, we can view this as a rule someone might espouse for *predicting* coinflips, but also view it as corresponding to a particular class of possible worlds containing randomness.

I mean, Laplace's Rule isn't the only rule you could use. Suppose I had a barrel containing ten white balls and ten green balls. If you already knew this about the barrel, then after seeing M white balls and N green balls, you'd predict the next ball being white with probability $(10 - M)/(20 - M - N)$.

If you use Laplace's Rule, that's like believing the world was like a billiards table with an original ball rolling to a stop at a random point and new balls ending up on the left or right. If you use $(10 - M)/(20 - M - N)$, that's like the hypothesis that there are ten green balls and ten white balls in a barrel. There isn't really a sharp border between rules we can use to predict the world, and rules for how the world behaves—

ASHLEY: Well, that sounds just plain wrong. The map is not the territory, don'tcha know? If Solomonoff induction can't tell the difference between maps and territories, maybe it doesn't contain all epistemological goodness after all.

BLAINE: Maybe it'd be better to say that there's a dualism between good ways of computing predictions and being in actual worlds where that kind of predicting works well? Like, you could also see Laplace's Rule as implementing the rules for a world with randomness where the original billiard ball ends up in a random place, so that the first thing you see is equally likely to be 1 or 0. Then to ask what probably happens on round 2, we tell the world what happened on round 1 so that it can update what the background random events were.

ASHLEY: Mmmaybe.

BLAINE: If you go with the version where Solomonoff induction is over programs that just spit out a determined string of ones and zeroes, we could see those programs as corresponding to particular environments—ways the world *could* be that would produce our sensory input, the sequence.

We could jump ahead and consider the more sophisticated decision-problem that appears in [AIXI](#): an environment is a program that takes your motor outputs as its input, and then returns your sensory inputs as its output. Then we can see a program that produces Bayesian-updated predictions as corresponding to a hypothetical probabilistic environment that implies those updates, although they'll be conjugate systems rather than mirror images.

ASHLEY: Did you say something earlier about the deterministic and probabilistic versions of Solomonoff induction giving the same answers? Like, is it a distinction without a difference whether we ask about simple programs that reproduce the observed data versus simple programs that assign high probability to the data? I can't see why that should be true, especially since Turing machines don't include a randomness source.

BLAINE: I'm *told* the answers are the same but I confess I can't quite see why, unless there's some added assumption I'm missing. So let's talk about programs that assign probabilities for now, because I think that case is clearer.

iv. Simplicity

BLAINE: The next key idea is to prefer *simple* programs that assign high probability to our observations so far.

ASHLEY: It seems like an obvious step, especially considering that you were already talking about "simple programs" and Occam's Razor a while back. Solomonoff induction is part of the Bayesian program of inference, right?

BLAINE: Indeed. Very much so.

ASHLEY: Okay, so let's talk about the program, or hypothesis, for "This barrel has an unknown frequency of white and green balls", versus the hypothesis "This barrel has 10 white and 10 green balls", versus the hypothesis, "This barrel always puts out a green ball after a white ball and vice versa."

Let's say we see a green ball, then a white ball, the sequence **GW**. The first hypothesis assigns this probability $1/2 * 1/3 = 1/6$, the second hypothesis assigns this probability $10/20 * 9/19$ or roughly $1/4$, and the third hypothesis assigns probability $1/2 * 1$.

Now it seems to me that there's some important sense in which, even though Laplace's Rule assigned a lower probability to the data, it's significantly simpler than the second and third hypotheses and is the wiser answer. Does Solomonoff induction agree?

BLAINE: I think you might be taking into account some prior knowledge that isn't in the sequence itself, there. Like, things that alternate either **101010...** or **010101...** are *objectively* simple in the sense that a short computer program simulates them or assigns probabilities to them. It's just unlikely to be true about an actual barrel of white and green balls.

If **10** is literally the first sense data that you ever see, when you are a fresh new intelligence with only two bits to rub together, then "The universe consists of alternating bits" is no less reasonable than "The universe produces bits with an unknown random frequency anywhere between 0 and 1."

ASHLEY: Conceded. But as I was going to say, we have three hypotheses that assigned $1/6$, $\sim 1/4$, and $1/2$ to the observed data; but to know the posterior probabilities of these hypotheses we need to actually say how relatively likely they were *a priori*, so we can multiply by the odds ratio. Like, if the prior odds were $3 : 2 : 1$, the posterior odds would be $3 : 2 : 1 * (2/12 : 3/12 : 6/12) = 3 : 2 : 1 * 2 : 3 : 6 = 6 : 6 : 6 = 1 : 1 : 1$. Now, how would Solomonoff induction assign prior probabilities to those computer programs? Because I remember you saying, way back when, that you thought Solomonoff was the answer to "How should Bayesians assign priors?"

BLAINE: Well, how would you do it?

ASHLEY: I mean... yes, the simpler rules should be favored, but it seems to me that there's some deep questions as to the exact relative 'simplicity' of the rules $(M + 1)/(M + N + 2)$, or the rule $(10 - M)/(20 - M - N)$, or the rule "alternate the bits"...

BLAINE: Suppose I ask you to just make up some simple rule.

ASHLEY: Okay, if I just say the rule I think you're looking for, the rule would be, "The complexity of a computer program is the number of bits needed to specify it to some arbitrary but reasonable choice of compiler or Universal Turing Machine, and the prior probability is $1/2$ to the power of the number of bits. Since, e.g., there's 32 possible 5-bit programs, so each such program has probability $1/32$. So if it takes 16 bits to specify Laplace's Rule of Succession, which seems a tad optimistic, then the prior probability would be $1/65536$, which seems a tad pessimistic."

BLAINE: Now just apply that rule to the infinity of possible computer programs that assign probabilities to the observed data, update their posterior probabilities based on the probability they've assigned to the evidence so far, sum over all of them to get your next prediction, and we're done. And yes, that requires a [hypercomputer](#) that can solve the [halting problem](#), but we're talking ideals here. Let P be the set of all programs and $s_1 s_2 \dots s_n$ also written $s_{\leq n}$ be the sense data so far, then

$$SOL(s_{\leq n}) := \sum_{\text{prog} \in P} 2^{-\text{length}(\text{prog})} \cdot \prod_{j=1}^n \text{InterpretProb}(\text{prog}(s_{\leq j-1}), s_j)$$

$$P(s_{n+1} = 1 \mid s_{\leq n}) = \frac{SOL(s_1 s_2 \dots s_n 1)}{SOL(s_1 s_2 \dots s_n 1) + SOL(s_1 s_2 \dots s_n 0)}$$

ASHLEY: Uh.

BLAINE: Yes?

ASHLEY: Um...

BLAINE: What is it?

ASHLEY: You invoked a countably infinite set, so I'm trying to figure out if my predicted probability for the next bit must necessarily converge to a limit as I consider increasingly large finite subsets in any order.

BLAINE: (sighs) Of course you are.

ASHLEY: I think you might have left out some important caveats. Like, if I take the rule literally, then the program "**0**" has probability $1/2$, the program "**1**" has probability $1/2$, the program "**01**" has probability $1/4$ and now the total probability is 1.25 which is *too much*. So I can't actually normalize it because the series sums to infinity. Now, this just means we need to, say, decide that the probability of a program having length 1 is $1/2$, the probability of it having length 2 is $1/4$, and so on out to infinity, but it's an added postulate.

BLAINE: The conventional method is to require a [prefix-free code](#). If "**0111**" is a valid program then "**01110**" cannot be a valid program. With that constraint, assigning " $1/2$ to the power of the length of the code", to all valid codes, will sum to less than 1; and we can normalize their relative probabilities to get the actual prior.

ASHLEY: Okay. And you're sure that it doesn't matter in what order we consider more and more programs as we approach the limit, because... no, I see it. Every program has positive probability mass, with the total set summing to 1, and Bayesian updating doesn't change that. So as I consider more and more programs, in any order, there are only so many large contributions that can be made from the mix—there's only so often that the final probability can change.

Like, let's say there are at most 99 programs with probability 1% that assign probability 0 to the next bit being a 1; that's only 99 times the final answer can go down by as much as 0.01, as the limit is approached.

BLAINE: This idea generalizes, and is important. List all possible computer programs, in any order you like. Use any definition of *simplicity* that you like, so long as for any given amount of simplicity, there are only a finite number of computer programs that simple. As you go on carving off chunks of prior probability mass and assigning them to

programs, it *must* be the case that as programs get more and complicated, their prior probability approaches zero!—though it's still positive for every finite program, because of [Cromwell's Rule](#).

You can't have more than 99 programs assigned 1% prior probability and still obey Cromwell's Rule, which means there must be some *most complex* program that is assigned 1% probability, which means every more complicated program must have less than 1% probability out to the end of the infinite list.

ASHLEY: Huh. I don't think I've ever heard that justification for Occam's Razor before. I think I like it. I mean, I've heard a lot of appeals to the empirical simplicity of the world, and so on, but this is the first time I've seen a *logical* proof that, in the limit, more complicated hypotheses *must* be less likely than simple ones.

BLAINE: Behold the awesomeness that is Solomonoff Induction!

ASHLEY: Uh, but you didn't actually use the notion of *computational* simplicity to get that conclusion; you just required that the supply of probability mass is finite and the supply of potential complications is infinite. Any way of counting discrete complications would imply that conclusion, even if it went by surface wheels and gears.

BLAINE: Well, maybe. But it so happens that Yudkowsky did invent or reinvent that argument after pondering Solomonoff induction, and if it predates him (or Solomonoff) then Yudkowsky doesn't know the source. Concrete inspiration for simplified arguments is also a credit to a theory, especially if the simplified argument didn't exist before that.

ASHLEY: Fair enough.

v. Choice of Universal Turing Machine

ASHLEY: My next question is about the choice of Universal Turing Machine—the choice of compiler for our program codes. There's an infinite number of possibilities there, and in principle, the right choice of compiler can make our probability for the next thing we'll see be anything we like. At least I'd expect this to be the case, based on how the "[problem of induction](#)" usually goes. So with the right choice of Universal Turing Machine, our online crackpot can still make it be the case that Solomonoff induction predicts Canada invading the USA.

BLAINE: One way of looking at the problem of good epistemology, I'd say, is that the job of a good epistemology is not to make it *impossible* to err. You can still blow off your foot if you really insist on pointing the shotgun at your foot and pulling the trigger.

The job of good epistemology is to make it *more obvious* when you're about to blow your own foot off with a shotgun. On this dimension, Solomonoff induction excels. If you claim that we ought to pick an enormously complicated compiler to encode our hypotheses, in order to make the 'simplest hypothesis that fits the evidence' be one that predicts Canada invading the USA, then it should be obvious to everyone except you that you are in the process of screwing up.

ASHLEY: Ah, but of course they'll say that their code is just the simple and natural choice of Universal Turing Machine, because they'll exhibit a meta-UTM which outputs that UTM given only a short code. And if you say the meta-UTM is complicated—

BLAINE: Flon's Law says, "There is not now, nor has there ever been, nor will there ever be, any programming language in which it is the least bit difficult to write bad code." You can't make it impossible for people to screw up, but you can make it *more obvious*. And Solomonoff induction would make it even more obvious than might at first be obvious, because—

ASHLEY: Your Honor, I move to have the previous sentence taken out and shot.

BLAINE: Let's say that the whole of your sensory information is the string **10101010...** Consider the stupid hypothesis, "This program has a 99% probability of producing a **1** on every turn", which you jumped to after seeing the first bit. What would you need to claim your priors were like—what Universal Turing Machine would you need to endorse—in order to maintain blind faith in that hypothesis in the face of ever-mounting evidence?

ASHLEY: You'd need a Universal Turing Machine **blind-utm** that assigned a very high probability to the **blind** program "def ProbNextElementIsOne(previous_sequence): return 0.99". Like, if **blind-utm** sees the code **0**, it executes the **blind** program "return 0.99".

And to defend yourself against charges that your UTM **blind-utm** was not itself simple, you'd need a meta-UTM, **blind-meta**, which, when it sees the code **10**, executes **blind-utm**.

And to really wrap it up, you'd need to take a fixed point through all towers of meta and use diagonalization to create the UTM **blind-diag** that, when it sees the program code **0**, executes "return 0.99", and when it sees the program code **10**, executes **blind-diag**.

I guess I can see some sense in which, even if that doesn't resolve Hume's problem of induction, anyone *actually advocating that* would be committing blatant shenanigans on a commonsense level, arguably more blatant than it would have been if we hadn't made them present the UTM.

BLAINE: Actually, the shenanigans have to be much worse than that in order to fool Solomonoff induction. Like, Solomonoff induction using your **blind-diag** isn't fooled for a minute, even taking **blind-diag** entirely on its own terms.

ASHLEY: Really?

BLAINE: Assuming 60 sequence items per second? Yes, absolutely, Solomonoff induction shrugs off the delusion in the first minute, unless there are further and even more blatant shenanigans.

We did require that your **blind-diag** be a *Universal* Turing Machine, meaning that it can reproduce every computable probability distribution over sequences, given some particular code to compile. Let's say there's a 200-bit code **laplace** for Laplace's Rule of Succession, "lambda sequence: return (sequence.count('1') + 1) / (len(sequence) + 2)", so that its prior probability relative to the 1-bit code for **blind** is 2^{-200} . Let's say that the sense data is around 50/50 1s and 0s. Every time we see a 1, **blind** gains a factor of 2 over **laplace** (99% vs. 50% probability), and every time we see a 0, **blind** loses a factor of 50 over **laplace** (1% vs. 50% probability).

On average, every 2 bits of the sequence, **blind** is losing a factor of 25 or, say, a bit more than 4 bits, i.e., on average **blind** is losing two bits of probability per element of the sequence observed.

So it's only going to take 100 bits, or a little less than two seconds, for **laplace** to win out over **blind**.

ASHLEY: I see. I was focusing on a UTM that assigned lots of prior probability to **blind**, but what I really needed was a compiler that, *while still being universal* and encoding every possibility somewhere, still assigned a really tiny probability to **laplace**, **faircoin** that encodes "return 0.5", and every other hypothesis that does better, round by round, than **blind**. So what I really need to carry off the delusion is **obstinate-diag** that is universal, assigns high probability to **blind**, requires billions of bits to specify **laplace**, and also requires billions of bits to specify any UTM that can execute **laplace** as a shorter code than billions of bits. Because otherwise we will say, "Ah, but given the evidence, this other UTM would have done better." I agree that those are even more blatant shenanigans than I thought.

BLAINE: Yes. And even *then*, even if your UTM takes two billion bits to specify **faircoin**, Solomonoff induction will lose its faith in **blind** after seeing a billion bits.

Which will happen before the first year is out, if we're getting 60 bits per second.

And if you turn around and say, "Oh, well, I didn't mean *that* was my UTM, I really meant *this* was my UTM, this thing over here where it takes a *trillion* bits to encode **faircoin**", then that's probability-theory-violating shenanigans where you're changing your priors as you go.

ASHLEY: That's actually a very interesting point—that what's needed for a Bayesian to maintain a delusion in the face of mounting evidence is not so much a blindly high prior for the delusory hypothesis, as a blind skepticism of all its alternatives.

But what if their UTM requires a googol bits to specify **faircoin**? What if **blind** and **blind-diag**, or programs pretty much isomorphic to them, are the only programs that can be specified in less than a googol bits?

BLAINE: Then your desire to shoot your own foot off has been made very, very visible to anyone who understands Solomonoff induction. We're not going to get absolutely objective prior probabilities as a matter of logical deduction, not without principles that are unknown to me and beyond the scope of Solomonoff induction. But we can make the stupidity really *blatant* and force you to construct a downright embarrassing Universal Turing Machine.

ASHLEY: I guess I can see that. I mean, I guess that if you're presenting a ludicrously complicated Universal Turing Machine that just refuses to encode the program that would predict Canada not invading, that's more *visibly* silly than a verbal appeal that says, "But you must just have faith that Canada will invade." I guess part of me is still hoping for a more objective sense of "complicated".

BLAINE: We could say that reasonable UTMs should contain a small number of wheels and gears in a material instantiation under our universe's laws of physics, which might in some ultimate sense provide a prior over priors. Like, the human brain evolved from DNA-based specifications, and the things you can construct out of relatively small numbers of physical objects are 'simple' under the 'prior' implicitly searched by natural selection.

ASHLEY: Ah, but what if I think it's likely that our physical universe or the search space of DNA won't give us a good idea of what's complicated?

BLAINE: For your alternative notion of what's complicated to go on being believed even as other hypotheses are racking up better experimental predictions, you need to assign a *ludicrously low probability* that our universe's space of physical systems buildable using a small number of objects, could *possibly* provide better predictions of that universe than your complicated alternative notion of prior probability.

We don't need to appeal that it's *a priori* more likely than not that "a universe can be predicted well by low-object-number machines built using that universe's physics." Instead, we appeal that it would violate [Cromwell's Rule](#), and would constitute exceedingly special pleading, to assign the possibility of a physically learnable universe a probability of *less* than $2^{-1,000,000}$. It then takes only a megabit of exposure to notice that the universe seems to be regular.

ASHLEY: In other words, so long as you don't start with an absolute and blind prejudice against the universe being predictable by simple machines encoded in our universe's physics—so long as, on this planet of seven billion people, you don't assign probabilities less than $2^{-1,000,000}$ to the other person being right about what is a good Universal Turing Machine—then the pure logic of Bayesian updating will rapidly force you to the conclusion that induction works.

vi. Why algorithmic complexity?

ASHLEY: Hm. I don't know that good *pragmatic* answers to the problem of induction were ever in short supply. Still, on the margins, it's a more forceful pragmatic answer than the last one I remember hearing.

BLAINE: Yay! Now isn't Solomonoff induction wonderful?

ASHLEY: Maybe?

You didn't really use the principle of *computational* simplicity to derive that lesson. You just used that *some inductive principle* ought to have a prior probability of more than $2^{-1,000,000}$.

BLAINE: ...

ASHLEY: Can you give me an example of a problem where the *computational* definition of simplicity matters and can't be factored back out of an argument?

BLAINE: As it happens, yes I can. I can give you *three* examples of how it matters.

ASHLEY: Vun... two... three! Three examples! Ah-ah-ah!

BLAINE: Must you do that every—oh, never mind. Example one is that galaxies are not so improbable that no one could ever believe in them, example two is that the limits of possibility include Terrence Tao, and example three is that diffraction is a simpler explanation of rainbows than divine intervention.

ASHLEY: These statements are all so obvious that no further explanation of any of them is required.

BLAINE: On the contrary! And I'll start with example one. Back when the Andromeda Galaxy was a hazy mist seen through a telescope, and someone first suggested that maybe that hazy mist was an incredibly large number of distant stars—that many "nebulae" were actually *distant galaxies*, and our own Milky Way was only one of them—there was a time when Occam's Razor was invoked against that hypothesis.

ASHLEY: What? Why?

BLAINE: They invoked Occam's Razor against the galactic hypothesis, because if that were the case, then there would be a *much huger number of stars* in the universe, and the stars would be entities, and Occam's Razor said "Entities are not to be multiplied beyond necessity."

ASHLEY: That's not how Occam's Razor works. The "entities" of a theory are its types, not its objects. If you say that the hazy mists are distant galaxies of stars, then you've reduced the number of laws because you're just postulating a previously seen type, namely stars organized into galaxies, instead of a new type of hazy astronomical mist.

BLAINE: Okay, but imagine that it's the nineteenth century and somebody replies to you, "Well, I disagree! William of Ockham said not to multiply entities, this galactic hypothesis obviously creates a huge number of entities, and that's the way I see it!"

ASHLEY: I think I'd give them your spiel about there being no human epistemology that can stop you from shooting off your own foot.

BLAINE: I don't think you'd be justified in giving them that lecture.

I'll parenthesize at this point that you ought to be very careful when you say "I can't stop you from shooting off your own foot", lest it become a Fully General Scornful Rejoinder. Like, if you say that to someone, you'd better be able to explain exactly why Occam's Razor counts types as entities but not objects. In fact, you'd better explain that to someone *before* you go advising them not to shoot off their own foot. And once you've told them what you think is foolish and why, you might as well stop there. Except in really weird cases of people presenting us with enormously complicated and jury-rigged Universal Turing Machines, and then we say the shotgun thing.

ASHLEY: That's fair. So, I'm not sure what I'd have answered before starting this conversation, which is much to your credit, friend Blaine. But now that I've had this conversation, it's obvious that it's new types and not new objects that use up the probability mass we need to distribute over all hypotheses. Like, I need to distribute my probability mass over "Hypothesis 1: there are stars" and "Hypothesis 2: there are stars plus huge distant hazy mists". I don't need to distribute my probability mass over all the actual stars in the galaxy!

BLAINE: In terms of Solomonoff induction, we penalize a program's *lines of code* rather than its *runtime* or *RAM used*, because we need to distribute our probability mass over possible alternatives each time we add a line of code. There's

no corresponding choice between mutually exclusive alternatives when a program uses more runtime or RAM.

(**ELIEZER:** (whispering) Unless we need a [leverage prior](#) to consider the hypothesis of being a particular agent inside all that RAM or runtime.)

ASHLEY: Or to put it another way: any fully detailed model of the universe would require some particular arrangement of stars, and the more stars there are, the more possible arrangements there are. But when we look through the telescope and see a hazy mist, we get to sum over all arrangements of stars that would produce that hazy mist. If some galactic hypothesis required a hundred billion stars to *all* be in *particular exact places* without further explanation or cause, then that would indeed be a grave improbability.

BLAINE: Precisely. And if you needed all the hundred billion stars to be in particular exact places, that's just the kind of hypothesis that would take a huge computer program to specify.

ASHLEY: But does it really require learning Solomonoff induction to understand that point? Maybe the bad argument against galaxies was just a motivated error somebody made in the nineteenth century, because they didn't want to live in a big universe for emotional reasons.

BLAINE: The same debate is playing out today over no-collapse versions of quantum mechanics, also somewhat unfortunately known as "many-worlds interpretations". Now, regardless of what anyone thinks of all the other parts of that debate, there's a *particular* sub-argument where somebody says, "It's simpler to have a collapse interpretation because all those extra quantum 'worlds' are extra entities that are unnecessary under Occam's Razor since we can't see them." And Solomonoff induction tells us that this invocation of Occam's Razor is flatly misguided because Occam's Razor does not work like that.

Basically, they're trying to cut down the RAM and runtime of the universe, at the expense of adding an extra line of code, namely the code for the collapse postulate that prunes off parts of the wavefunction that are in undetectably weak causal contact with us.

ASHLEY: Hmm. Now that you put it that way, it's not so obvious to me that it makes sense to have *no* prejudice against *sufficiently* enormous universes. I mean, the universe we see around us is exponentially vast but not superexponentially vast—the visible atoms are 10^{80} in number or so, not $10^{10^{80}}$ or "bigger than Graham's Number". Maybe there's some fundamental limit on how much gets computed.

BLAINE: You, um, know that on the Standard Model, the universe doesn't just cut out and stop existing at the point where our telescopes stop seeing it? There isn't a giant void surrounding a little bubble of matter centered perfectly on Earth? It calls for a literally infinite amount of matter? I mean, I guess if you don't like living in a universe with more than 10^{80} entities, a universe where *too much gets computed*, you could try to specify *extra laws of physics* that create an abrupt spatial boundary with no further matter beyond them, somewhere out past where our telescopes can see—

ASHLEY: All right, point taken.

(**ELIEZER:** (whispering) Though I personally suspect that the spatial multiverse and the quantum multiverse are the same multiverse, and that what lies beyond the reach of our telescopes is not entangled with us—meaning that the universe is as finitely large as the superposition of all possible quantum branches, rather than being literally infinite in space.)

BLAINE: I mean, there is in fact an alternative formalism to Solomonoff induction, namely [Levin search](#), which says that program complexities are further penalized by the logarithm of their runtime. In other words, it would say that 'explanations' or 'universes' that require a long time to run are inherently less probable.

Some people like Levin search more than Solomonoff induction because it's more computable. I dislike Levin search because (a) it has no fundamental epistemic justification and (b) it assigns probability zero to quantum mechanics.

ASHLEY: Can you unpack that last part?

BLAINE: If, as is currently suspected, there's no way to simulate quantum computers using classical computers without an exponential slowdown, then even in principle, this universe requires exponentially vast amounts of classical computing power to simulate.

Let's say that with sufficiently advanced technology, you can build a quantum computer with a million qubits. On Levin's definition of complexity, for the universe to be like that is as improbable *a priori* as any *particular* set of laws of physics that must specify on the order of one million equations.

Can you imagine how improbable it would be to see a list of one hundred thousand differential equations, without any justification or evidence attached, and be told that they were the laws of physics? That's the kind of penalty that Levin search or Schmidhuber's Speed Prior would attach to any laws of physics that could run a quantum computation of a million qubits, or, heck, any physics that claimed that a protein was being folded in a way that ultimately went through considering millions of quarks interacting.

If you're *not* absolutely certain *a priori* that the universe *isn't* like that, you don't believe in Schmidhuber's Speed Prior. Even with a collapse postulate, the amount of computation that goes on before a collapse would be prohibited by the

Speed Prior.

ASHLEY: Okay, yeah. If you're phrasing it that way—that the Speed Prior assigns probability nearly zero to quantum mechanics, so we shouldn't believe in the Speed Prior—then I can't easily see a way to extract out the same point without making reference to ideas like penalizing algorithmic complexity but not penalizing runtime. I mean, maybe I could extract the lesson back out but it's easier to say, or more obvious, by pointing to the idea that Occam's Razor should penalize algorithmic complexity but not runtime.

BLAINE: And that isn't just *implied* by Solomonoff induction, it's pretty much the whole idea of Solomonoff induction, right?

ASHLEY: Maaaybe.

BLAINE: For example two, that Solomonoff induction outperforms even Terence Tao, we want to have a theorem that says Solomonoff induction catches up to every computable way of reasoning in the limit. Since we iterated through all possible computer programs, we know that somewhere in there is a simulated copy of Terence Tao in a simulated room, and if this requires a petabyte to specify, then we shouldn't have to make more than a quadrillion bits of error relative to Terence Tao before zeroing in on the Terence Tao hypothesis.

I mean, in practice, I'd expect far less than a quadrillion bits of error before the system was behaving like it was vastly smarter than Terence Tao. It'd take a lot less than a quadrillion bits to give you some specification of a universe with simple physics that gave rise to a civilization of vastly greater than intergalactic extent. Like, [Graham's Number](#) is a very simple number, so it's easy to specify a universe that runs for that long before it returns an answer. It's not obvious how you'd extract Solomonoff predictions from that civilization and incentivize them to make good ones, but I'd be surprised if there were no Turing machine of fewer than one thousand states which did that somehow.

ASHLEY: ...

BLAINE: And for all I know there might be even better ways than that of getting exceptionally good predictions, somewhere in the list of the first decillion computer programs. That is, somewhere in the first 100 bits.

ASHLEY: So your basic argument is, "Never mind Terence Tao, Solomonoff induction dominates God."

BLAINE: Solomonoff induction isn't the epistemic prediction capability of a superintelligence. It's the epistemic prediction capability of something that eats superintelligences like potato chips.

ASHLEY: Is there any point to contemplating an epistemology so powerful that it will never begin to fit inside the universe?

BLAINE: Maybe? I mean, a lot of times, you just find people *failing to respect* the notion of ordinary superintelligence, doing the equivalent of supposing that a superintelligence behaves like a bad Hollywood genius and misses obviously-seeming moves. And a lot of times you find them insisting that "there's a limit to how much information you can get from the data" or something along those lines. "[That Alien Message](#)" is intended to convey the counterpoint, that smarter entities can extract more info than is immediately apparent on the surface of things.

Similarly, thinking about Solomonoff induction might also cause someone to realize that if, say, you simulated zillions of possible simple universes, you could look at which agents were seeing exact data like the data you got, and figure out where you were inside that range of possibilities, so long as there was literally *any* correlation to use.

And if you say that an agent *can't* extract that data, you're making a claim about which shortcuts to Solomonoff induction are and aren't computable. In fact, you're probably pointing at some *particular* shortcut and claiming nobody can ever figure that out using a reasonable amount of computing power *even though the info is there in principle*. Contemplating Solomonoff induction might help people realize that, yes, the data *is* there in principle. Like, until I ask you to imagine a civilization running for Graham's Number of years inside a Graham-sized memory space, you might not imagine them trying all the methods of analysis that *you personally* can imagine being possible.

ASHLEY: If somebody is making that mistake in the first place, I'm not sure you can beat it out of them by telling them the definition of Solomonoff induction.

BLAINE: Maybe not. But to brute-force somebody into imagining that [sufficiently advanced agents](#) have [Level 1 protagonist intelligence](#), that they are [epistemically efficient](#) rather than missing factual questions that are visible even to us, you might need to ask them to imagine an agent that can see *literally anything seeable in the computational limit* just so that their mental simulation of the ideal answer isn't running up against stupidity assertions.

Like, I think there are a lot of people who could benefit from looking over the evidence they already personally have, and asking what a Solomonoff inductor could deduce from it, so that they wouldn't be running up against stupidity assertions *about themselves*. It's the same trick as asking yourself what God, Richard Feynman, or a "perfect rationalist" would believe in your shoes. You just have to pick a real or imaginary person that you respect enough for your model of that person to lack the same stupidity assertions that you believe about yourself.

ASHLEY: Well, let's once again try to factor out the part about Solomonoff induction in particular. If we're trying to imagine something epistemically smarter than ourselves, is there anything we get from imagining a complexity-weighted prior over programs in particular? That we don't get from, say, trying to imagine the reasoning of one particular Graham-Number-sized civilization?

BLAINE: We get the surety that even anything we imagine *Terence Tao himself* as being able to figure out, is something that is allowed to be known after some bounded number of errors versus Terence Tao, because Terence Tao is inside the list of all computer programs and gets promoted further each time the dominant paradigm makes a prediction error relative to him.

We can't get that dominance property without invoking "all possible ways of computing" or something like it—we can't incorporate the power of all reasonable processes, unless we have a set such that all the reasonable processes are in it. The enumeration of all possible computer programs is one such set.

ASHLEY: Hm.

BLAINE: Example three, diffraction is a simpler explanation of rainbows than divine intervention.

I don't think I need to belabor this point very much, even though in one way it might be the most central one. It sounds like "Jehovah placed rainbows in the sky as a sign that the Great Flood would never come again" is a 'simple' explanation; you can explain it to a child in nothing flat. Just the diagram of diffraction through a raindrop, to say nothing of the Principle of Least Action underlying diffraction, is something that humans don't usually learn until undergraduate physics, and it *sounds* more alien and less intuitive than Jehovah. In what sense is this intuitive sense of simplicity wrong? What gold standard are we comparing it to, that could be a better sense of simplicity than just 'how hard is it for me to understand'?

The answer is Solomonoff induction and the rule which says that simplicity is measured by the size of the computer program, not by how hard things are for human beings to understand. Diffraction is a small computer program; any programmer who understands diffraction can simulate it without too much trouble. Jehovah would be a much huger program—a complete mind that implements anger, vengeance, belief, memory, consequentialism, etcetera. Solomonoff induction is what tells us to retrain our intuitions so that differential equations feel like less *burdensome* explanations than heroic mythology.

ASHLEY: Now hold on just a second, if that's actually how Solomonoff induction works then it's not working very well. I mean, Abraham Lincoln was a great big complicated mechanism from an algorithmic standpoint—he had a hundred trillion synapses in his brain—but that doesn't mean I should look at the historical role supposedly filled by Abraham Lincoln, and look for simple mechanical rules that would account for the things Lincoln is said to have done. If you've already seen humans and you've already learned to model human minds, it shouldn't cost a vast amount to say there's one *more* human, like Lincoln, or one more entity that is *cognitively humanoid*, like the Old Testament jealous-god version of Jehovah. It may be *wrong* but it shouldn't be vastly improbable *a priori*.

If you've already been forced to acknowledge the existence of some humanlike minds, why not others? Shouldn't you get to reuse the complexity that you postulated to explain humans, in postulating Jehovah?

In fact, shouldn't that be what Solomonoff induction *does*? If you have a computer program that can model and predict humans, it should only be a slight modification of that program—only slightly longer in length and added code—to predict the modified-human entity that is Jehovah.

BLAINE: Hm. That's fair. I may have to retreat from that example somewhat.

In fact, that's yet another point to the credit of Solomonoff induction! The ability of programs to reuse code, incorporates our intuitive sense that if you've already postulated one kind of thing, it shouldn't cost as much to postulate a similar kind of thing elsewhere!

ASHLEY: Uh huh.

BLAINE: Well, but even if I was wrong that Solomonoff induction should make Jehovah seem very improbable, it's still Solomonoff induction that says that the alternative hypothesis of 'diffraction' shouldn't itself be seen as burdensome—even though diffraction might require a longer time to explain to a human, it's still at heart a simple program.

ASHLEY: Hmm.

I'm trying to think if there's some notion of 'simplicity' that I can abstract away from 'simple program' as the nice property that diffraction has as an explanation for rainbows, but I guess anything I try to say is going to come down to some way of counting the wheels and gears inside the explanation, and justify the complexity penalty on probability by the increased space of possible configurations each time we add a new gear. And I can't make it be about surface details because that will make whole humans seem way too improbable.

If I have to use simply specified systems and I can't use surface details or runtime, that's probably going to end up basically equivalent to Solomonoff induction. So in that case we might as well use Solomonoff induction, which is probably simpler than whatever I'll think up and will give us the same advice. Okay, you've mostly convinced me.

BLAINE: Mostly? What's left?

vii. Limitations

ASHLEY: Well, several things. Most of all, I think of how the 'language of thought' or 'language of epistemology' seems to be different in some sense from the 'language of computer programs'.

Like, when I think about the laws of Newtonian gravity, or when I think about my Mom, it's not just one more line of code tacked onto a big black-box computer program. It's more like I'm crafting an explanation with modular parts—if it contains a part that looks like Newtonian mechanics, I step back and reason that it might contain other parts with differential equations. If it has a line of code for a Mom, it might have a line of code for a Dad.

I'm worried that if I understood how humans think like that, maybe I'd look at Solomonoff induction and see how it doesn't incorporate some further key insight that's needed to do good epistemology.

BLAINE: Solomonoff induction literally incorporates a copy of you thinking about whatever you're thinking right now.

ASHLEY: Okay, great, but that's *inside* the system. If Solomonoff learns to promote computer programs containing good epistemology, but is not itself good epistemology, then it's not the best possible answer to "How do you compute epistemology?"

Like, natural selection produced humans but population genetics is not an answer to "How does intelligence work?" because the intelligence is in the inner content rather than the outer system. In that sense, it seems like a reasonable worry that Solomonoff induction might incorporate only *some* principles of good epistemology rather than *all* the principles, even if the *internal content* rather than the *outer system* might bootstrap the rest of the way.

BLAINE: Hm. If you put it *that* way...

(long pause)

... then I guess I have to agree. I mean, Solomonoff induction doesn't explicitly say anything about, say, the distinction between analytic propositions and empirical propositions, and knowing that is part of good epistemology on my view. So if you want to say that Solomonoff induction is something that bootstraps to good epistemology rather than being all of good epistemology by itself, I guess I have no choice but to agree.

I do think the outer system already contains a *lot* of good epistemology and inspires a lot of good advice all on its own. Especially if you give it credit for formally reproducing principles that are "common sense", because correctly formalizing common sense is no small feat.

ASHLEY: Got a list of the good advice you think is derivable?

BLAINE: Um. Not really, but off the top of my head:

1. The best explanation is the one with the best mixture of simplicity and matching the evidence.
2. "Simplicity" and "matching the evidence" can both be measured in bits, so they're commensurable.
3. The simplicity of a hypothesis is the number of bits required to formally specify it—for example, as a computer program.
4. When a hypothesis assigns twice as much probability to the exact observations seen so far as some other hypothesis, that's one bit's worth of relatively better matching the evidence.
5. You should actually be making your predictions using all the explanations, not just the single best one, but explanations that poorly match the evidence will drop down to tiny contributions very quickly.
6. Good explanations let you compress lots of data into compact reasons which strongly predict seeing just that data and no other data.
7. Logic can't dictate prior probabilities absolutely, but if you assign probability less than $2^{-1,000,000}$ to the prior that mechanisms constructed using a small number of objects from your universe might be able to well predict that universe, you're being unreasonable.
8. So long as you don't assign infinitesimal prior probability to hypotheses that let you do induction, they will very rapidly overtake hypotheses that don't.
9. It is a logical truth, not a contingent one, that more complex hypotheses must in the limit be less probable than simple ones.
10. Epistemic rationality is a precise art with no user-controlled degrees of freedom in how much probability you ideally ought to assign to a belief. If you think you can tweak the probability depending on what you want the answer to be, you're doing something wrong.
11. Things that you've seen in one place might reappear somewhere else.
12. Once you've learned a new language for your explanations, like differential equations, you can use it to describe other things, because your best hypotheses will now already encode that language.
13. We can learn meta-reasoning procedures as well as object-level facts by looking at which meta-reasoning rules are simple and have done well on the evidence so far.
14. So far, we seem to have no *a priori* reason to believe that universes which are more expensive to compute are less probable.
15. People were wrong about galaxies being *a priori* improbable because that's not how Occam's Razor works. Today, other people are equally wrong about other parts of a continuous wavefunction counting as extra entities for the purpose of evaluating hypotheses' complexity.
16. If something seems "weird" to you but would be a consequence of simple rules that fit the evidence so far, well, there's nothing in these explicit laws of epistemology that adds an extra penalty term for weirdness.
17. Your epistemology shouldn't have extra rules in it that aren't needed to do Solomonoff induction or something like it, including rules like "science is not allowed to examine this particular part of reality"—

ASHLEY: This list isn't finite, is it.

BLAINE: Well, there's a *lot* of outstanding debate about epistemology where you can view that debate through the lens of Solomonoff induction and see what Solomonoff suggests.

ASHLEY: But if you don't mind my stopping to look at your last item, #17 above—again, it's attempts to add completeness clauses to Solomonoff induction that make me the most nervous.

I guess you could say that a good rule of epistemology ought to be one that's promoted by Solomonoff induction—that it should arise, in some sense, from the simple ways of reasoning that are good at predicting observations. But that doesn't mean a good rule of epistemology ought to explicitly be in Solomonoff induction or it's out.

BLAINE: Can you think of good epistemology that doesn't seem to be contained in Solomonoff induction? Besides the example I already gave of distinguishing logical propositions from empirical ones.

ASHLEY: I've been trying to. First, it seems to me that when I reason about laws of physics and how those laws of physics might give rise to higher levels of organization like molecules, cells, human beings, the Earth, and so on, I'm not constructing in my mind a great big chunk of code that reproduces my observations. I feel like this difference might be important and it might have something to do with 'good epistemology'.

BLAINE: I guess it could be? I think if you're saying that there might be this unknown other thing and therefore Solomonoff induction is terrible, then that would be the [nirvana fallacy](#). Solomonoff induction is the best formalized epistemology we have *right now*—

ASHLEY: I'm not saying that Solomonoff induction is terrible. I'm trying to look in the direction of things that might point to some future formalism that's better than Solomonoff induction. Here's another thing: I feel like I didn't have to learn how to model the human beings around me from scratch based on environmental observations. I got a jump-start on modeling other humans by observing *myself*, and by recruiting my brain areas to run in a sandbox mode that models other people's brain areas—empathy, in a word.

I guess I feel like Solomonoff induction doesn't incorporate that idea. Like, maybe *inside* the mixture there are programs which do that, but there's no explicit support in the outer formalism.

BLAINE: This doesn't feel to me like much of a disadvantage of Solomonoff induction—

ASHLEY: I'm not *saying* it would be a disadvantage if we actually had a hypercomputer to run Solomonoff induction. I'm saying it might point in the direction of "good epistemology" that isn't explicitly included in Solomonoff induction.

I mean, now that I think about it, a generalization of what I just said is that Solomonoff induction assumes I'm separated from the environment by a hard, Cartesian wall that occasionally hands me observations. Shouldn't a more realistic view of the universe be about a simple program that *contains me somewhere inside it*, rather than a simple program that hands observations to some other program?

BLAINE: Hm. Maybe. How would you formalize *that*? It seems to open up a big can of worms—

ASHLEY: But that's what my actual epistemology actually says. My world-model is not about a big computer program that provides inputs to my soul, it's about an enormous mathematically simple physical universe that instantiates Ashley as one piece of it. And I think it's good and important to have epistemology that works that way. It wasn't *obvious* that we needed to think about a simple universe that embeds us. Descartes *did* think in terms of an impervious soul that had the universe projecting sensory information onto its screen, and we had to get away from that kind of epistemology.

BLAINE: You understand that Solomonoff induction makes only a bounded number of errors relative to any computer program which does reason the way you prefer, right? If thinking of yourself as a contiguous piece of the universe lets you make better experimental predictions, programs which reason that way will rapidly be promoted.

ASHLEY: It's still unnerving to see a formalism that seems, in its own structure, to harken back to the Cartesian days of a separate soul watching a separate universe projecting sensory information on a screen. Who knows, maybe that would somehow come back to bite you?

BLAINE: Well, it wouldn't bite you in the form of repeatedly making wrong experimental predictions.

ASHLEY: But it might bite you in the form of having no way to represent the observation of, "I drank this 'wine' liquid and then my emotions changed; could my emotions themselves be instantiated in stuff that can interact with some component of this liquid? Can alcohol touch neurons and influence them, meaning that I'm not a separate soul?" If we interrogated the Solomonoff inductor, would it be able to understand that reasoning?

Which brings up that dangling question from before about modeling the effect that my actions and choices have on the environment, and whether, say, an agent that used Solomonoff induction would be able to correctly predict "If I drop an anvil on my head, my sequence of sensory observations will end."

ELIEZER: And that's my cue to step in!

The natural next place for this dialogue to go, if I ever write a continuation, is the question of actions and choices, and the agent that uses Solomonoff induction for beliefs and expected reward maximization for selecting actions—the perfect rolling sphere of advanced agent theory, [AIXI](#).

Meanwhile: For more about the issues Ashley raised with agents being a contiguous part of the universe, see "[Embedded Agency](#)."

Demand offsetting

For the last few years I've been avoiding factory farmed eggs because I think they involve a lot of unnecessary suffering. I'm hesitant to be part of that even if it's not a big deal on utilitarian grounds. This is a pain since factory-farmed eggs are used all over the place (e.g. in ice cream, pastries, pasta...). I'd prefer just spend a bit of money and not think too much about what I eat.

In this post I'll describe a possible offsetting strategy that I think is unusually robust and should be satisfying for many moral perspectives. The same proposal would also apply to many other animal products and potentially to the environmental impacts of consumption.

Proposal

I think it's possible to produce humane eggs where hens have positive lives and nothing horrifying happens to anyone. So my ideal would be to buy and use humane eggs. But this is tough since most of the time I'm eating eggs that someone else used as an ingredient (and even when I'm using them myself acquiring really humane eggs is kind of a pain).

So here's an alternative that seems easier and just as good:

- Some people raise humane eggs.
- They sell these on the wholesale market as if they were totally normal eggs.
- An inspector verifies that hens are treated extremely well and that they have sold N eggs on the wholesale market.
- The inspector issues N "humane egg" certificates to the producer.
- The producer sells these certificates in an online marketplace in order to cover the extra costs of humane eggs.
- Whenever I eat an egg, I buy a humane egg certificate to go with it.

Analysis

If I buy an egg and a humane egg certificate, what is the net effect on the world?

Buying the egg increased demand for eggs. If I hadn't also bought a certificate, that would indirectly cause someone to make one more factory-farmed egg.

Buying the positive-welfare certificate means that someone sold a wholesale egg on my behalf and increased the supply of eggs. If I hadn't also bought an egg, that would indirectly cause someone to make one less factory-farmed egg.

So my net effect on factory farmed eggs is zero. It's as if I was making my own positive-welfare egg and eating it, with no effect on how many factory-farmed eggs other people make or eat.

(In reality both of these actions will have other effects, e.g. causing other people to eat more or fewer eggs, but I think they still cancel out perfectly.)

This is an unusually pure form of offsetting. I'm ensuring that every hen who comes into existence because of me is living a positive life. Put differently, buying eggs only

hurt hens via some indirect market effects, and I'm now offsetting my harm at that level before it turns into any actual harm to a hen. I think this form of offsetting is acceptable on a very broad range of moral perspectives (practically any perspective that is comfortable with humane eggs themselves).

Cost to the consumer

I'd guess that positive welfare eggs cost something like 3x more than typical eggs. For example I think Vital Farms sells eggs for around \$6/dozen vs \$2/dozen for more typical eggs.

So for each \$1 that I would spend on eggs, I'd need to spend \$2 to buy an egg-offset certificate. I haven't looked into it but I could imagine wanting to go even higher to have a margin of error and shoot for even higher welfare standards. Let's call it \$0.50/egg, suggesting a 4-5x markup over typical eggs.

(I'm also not sure about relative egg sizes and didn't look into prices very precisely, for me personally the numbers are low enough that it doesn't matter too much even if being conservative.)

How much would that cost in practice? Here are some estimates from quick googling of recipes:

- \$0.03 for a croissant ([16 croissants / egg](#))
- <\$0.30 for a scoop of egg-y ice cream ([5 yolks / quart](#) x [8 scoops / quart](#))
- <\$0.20 for egg-y donuts ([3 eggs and 2 whites / 12 donuts](#))
- \$0.06 for a pancake ([8 pancakes / egg](#))

The whole US produces about 100 billion eggs / year. If we wanted to offset all of that at the less-conservative \$0.25/egg number, that would be \$25B/year or \$76 per person per year.

Benefits of decoupling (paying for welfare) from (paying for eggs)

Today animal welfare standards involve (at least) three parties: whoever raises the hen humanely, the grocery store or restaurant or whatever who uses the humane eggs, and the customer who demands humane eggs. Often there are more people still, like a host who prepares a meal for guests. And that's all on top of the fact that people need to buy eggs at a particular place and time and so on. This means that dealing with humane eggs is always kind of a pain, and we need to standardize on

But humane egg certificates cut out all the middle men and just involve the producer and the consumer. So they make life way easier, and then they can also be specialized based on different welfare expectations. A certificate can say as much as it wants to about the welfare of the hens who produced it. Some customers could just go for anything that sounds "humane," others could rely on more nuanced evaluations by non-profits who evaluate welfare claims, and the most over-the-top consumers could be matched with the most overt-the-top farms.

Getting started

By the same token, it seems much less daunting to get this kind of system off the ground since it only involves two parties. A single farm could opt to sell wholesale eggs and allow inspections in order to serve a small number of customers, since they don't have to geographically or temporally matched. I think it could be very fast to reach the point where humane egg certificates had as much flexibility as the current patchwork ecosystem for buying humane eggs.

The market could even get jump started by a tiny number of philanthropists (or even just one) who buy a few million dollars of humane egg certificates, negotiating directly with a few farmers. So this doesn't necessarily involve any hard coordination problems at all.

Or we could literally have a kickstarter amongst people who are interested in humane egg certificates and have enough agreement about what kind of welfare standard they'd want to use (and then could pick someone to find and negotiate with a farm directly).

I think most philanthropists wouldn't do this unless they were excited about setting norms or driven by non-utilitarian interests. In some sense this methodology feels like the "[GiveDirectly](#) of animal welfare," extremely scalable and robust but a very expensive way to fix the problem. See the next section.

(In practice I think there are all kinds of issues with this proposal that I don't know about, so I actually don't know what the first steps would end up looking like, e.g. it depends on how small farmers think about humane eggs and wholesaling. This post is definitely written in spitballing mode, not as someone who understands the domain.)

Thoughts on cost-effectiveness

I'd estimate that it would cost something like \$300B/year to offset all of the global harms of factory farming in this way, which feels at least 1-2 orders of magnitude worse than the kinds of welfare interventions that philanthropists feel excited about. Of course that's also an argument against the cost-effectiveness of offsetting.

Here are some of the things going on:

- Most interventions consist in lobbying or advocacy, trying to convince *other people* to pay a cost (e.g. to eat less meat or to pay for better animal welfare standards). I think that's ethically problematic as an offsetting strategy relative to paying the cost yourself, and to a lesser extent I think there's something to be said for philanthropists directly paying these kinds of costs in addition to trying to convince others to pay them.
- The standards I'm imagining above (at \$0.50/egg) are kind of "overkill," designed to address all of the possible negative effects on a hen rather than targeting the lowest hanging fruit. Individual humane egg certificate purchasers could focus on particular standards (e.g. just ensuring that hens have dust baths and adequate cage space) and get to something closer to the cost-effectiveness of traditional hen welfare interventions, and the same norms and infrastructure could be used for both.

Universalizability

What would happen if many people tried to use this offsetting strategy?

If demand for humane egg certificates was equal to demand for eggs, then all eggs would be humane. Economically, this would be because the price of eggs has fallen below the costs of factory farming, with most of the value being in the humane certificates. I think this is basically the best possible case for an offsetting strategy, and I'd personally consider it better than abolition.

If deployed at large scale, there could be considerable customization of humane egg certificates and people who wanted it would be able to get high welfare standards with high-quality monitoring. In general I think that early adopters should view themselves as overpaying in order to share the burden of setting up the scheme and helping scale up humane agriculture.

How do we prepare for final crunch time?

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

[Crossposted from [Musings and Rough Drafts](#).]

Epistemic status: Brainstorming and first draft thoughts.

Inspired by something that Ruby Bloom wrote and the Paul Christiano episode of the 80,000 hours podcast.]

One claim I sometimes hear about AI alignment [paraphrase]:

It is really hard to know what sorts of AI alignment work are good this far out from transformative AI. As we get closer, we'll have a clearer sense of what AGI / Transformative AI is likely to actually look like, and we'll have much better traction on what kind of alignment work to do. In fact, MOST of the work of AI alignment is done in the final few years (or months) before AGI, when we've solved most of the hard capabilities problems already so we know what AGI will look like and we can work directly, with good feedback loops, on the sorts of systems that we want to align.

Usually, this is said to argue that the value of the alignment research being done today is primarily that of enabling, future, more critical, alignment work. But "progress in the field" is only one dimension to consider in boosting and unblocking the work of alignment researchers in this last stretch.

In this post I want to take the above posit seriously, and consider the implications. If most of the alignment work that will be done is going to be done in the final few years before the deadline, our job in 2021 is mostly to do everything that we can to enable the people working on the problem in the crucial period (which might be us, or our successors, or both) so that they are as well equipped as we can possibly make them.

What are all the ways that we can think of that we can prepare now, for our eventual final exam? What should we be investing in, to improve our efficacy in those final, crucial, years?

The following are some ideas.

[In this post, I'm going to refer to this last stretch of a few months to a few years, "final crunch time", as distinct from just "crunch time", ie this century.]

Access

For this to matter, our alignment researchers need to be at the cutting edge of AI capabilities, and they need to be positioned such that their work can actually be incorporated into AI systems as they are deployed.

A different kind of work

Most current AI alignment work is pretty abstract and theoretical, for two reasons.

The first reason is a philosophical / methodological claim: There's a fundamental "nearest unblocked strategy" / overfitting problem. Patches that correct clear and obvious alignment failures are unlikely to generalize fully, they'll only constrain unaligned optimization to channels that you can't recognize. For this reason, some claim, we need to have an extremely robust, theoretical understanding of intelligence and alignment, ideally at the level of proofs.

The second reason is a practical consideration: we just don't have powerful AI systems to work with, so there isn't much that can be done in the way of tinkering and getting feedback.

That second objection becomes less relevant in final crunch time: in this scenario, we'll have powerful systems 1) that will be built along the same lines as the systems that it is crucial to align and 2) that will have enough intellectual capability to pose at least semi-realistic "creative" alignment failures (ie, current systems are so dumb, and live in such constrained environments, that it isn't clear how much we can learn about aligning literal superintelligences from them.)

And even if the first objection ultimately holds, theoretical understanding often (usually?) follows from practical engineering proficiency. It seems like it might be a fruitful path to tinker with semi-powerful systems: trying out different alignment approaches empirically, and tinkering to discover new approaches, and then backing up to do robust theory-building given much richer data about what seems to work.

I could imagine sophisticated setups that enable this kind of tinkering and theory building. For instance, I imagine a setup that includes:

- A "**sandbox**" that afford easy implementation of many different AI architectures and custom combinations of architectures, with a wide variety easy-to-create, easy-to-adjust, training schemes, and a full suite of interpretability tools. We could quickly try out different safety schemes, in different distributions, and observe what kinds of cognition and behavior result.
- A **meta AI** that observes the sandbox, and all of the experiments therein, to learn general principles of alignment. We could use interpretability tools to use this AI as a "[microscope](#)" on the AI alignment problem itself, abstracting out patterns and dynamics that we couldn't easily have teased out with only our own brains. This meta system might also play some role in designing the experiments to run in the sandbox, to allow it to get the best data to test its hypotheses.
- A **theorem prover** that would formalize the properties and implications of those general alignment principles, to give us crisply specified alignment criteria by which we can evaluate AI designs.

Obviously, working with a full system like this is quite different than abstract, purely theoretical work on decision theory or logical uncertainty. It is closer to the sort of experiments that the OpenAI and Deep Mind safety teams have published, but even that is a pretty far cry from the kind of rapid-feedback tinkering that I'm pointing at here.

Given that the kind of work that leads to research progress might be very different in final crunch time than it is now, it seems worth trying to forecast what shape that work will take and trying to see if there are ways to practice doing that kind of work *before* final crunch time.

Knowledge

Obviously, when we get to final crunch time, we don't want to have to spend any time studying fields that we could have studied in the lead-up years. We want to have *already* learned all the information and ways of thinking that we'll want to know, then. It seems worth considering what fields we'll wish we had known when time comes.

The obvious contenders:

- Machine Learning
- Machine Learning interpretability
- All the Math of Intelligence that humanity has yet amassed [Probability theory, Causality, etc.]

Some less obvious possibilities:

- Neuroscience?
- Geopolitics, if it turns out that which technical approach is ideal hinges on important facts about the balance of power?
- Computer security?
- Mechanism design in general?

Do other subjects come to mind?

Research methodology / Scientific “rationality”

We want the research teams tackling this problem in final crunch time to have the best scientific methodology and the best cognitive tools / habits for making research progress, that we can manage to provide them.

This maybe includes skills or methods in the domains of:

- Ways to notice as early as possible if you're following an ultimately-fruitless research path
- Noticing / Resolving /Avoiding blindspots
- Effective research teams
- Original seeing / overcoming theory blindness / hypothesis generation
- ???

Productivity

One obvious thing is to spend time now, investing in habits and strategies for effective productivity. It seems senseless to waste precious hours in the final crunch time due to procrastination or poor sleep. It is well worth in to solve those problems *now*. But

aside from the general suggestion to get your shit in order and develop good habits, I can think of two more specific things that seem good to do.

Practice no-cost-too-large productive periods

There maybe trades that could make people more productive on the margin, but are too expensive in regular life. For instance, I think that I might conceivably benefit from having a dedicated person who's job is to always be near me, so that I can duck with them (or have them "hold space" for me) with 0 friction. I've experimented a little bit with similar ideas (like having a list of people on call to duck with), but it doesn't seem worth it for me to pay a whole extra person-salary to have the person be on call and in the same building, instead of on-call via zoom.

But it *is* worth it at final crunch time.

It might be worth it to spend some period of time, maybe a week, maybe a month, every year, optimizing unrestrainedly for research productivity, with no heed to cost at all, so that we can practice how to do that. This is possibly a good thing to do anyway, because it might uncover trades that, on reflection, are worth importing into my regular life.

Optimize rest

One particular subset of personal productivity, that jumps out at me: each person should figure out their actual optimal cadence of rest.

There's a failure mode that ambitious people commonly fall into, which is working past the point when marginal hours of work are negative. When the whole cosmic endowment is on the line, there will be a natural temptation to push yourself to work as hard as you can, and forgo rest. Obviously, this is a mistake. Rest isn't just a luxury: it is one of the inputs to productive work.

There is a second level of this error in which one, grudgingly, takes the minimal amount of rest time, and gets back to work. But the amount of rest time required to stay functional is not the *optimal* amount of rest, the amount that maximizes productive output. Eliezer [mused](#) years ago, that he felt kind of guilty about it, but maybe he should actually take two days off between research days, because the quality of his research seemed better on days when he happened to have had two rest days preceding.

In final crunch time, we want everyone to be resting the optimal amount that actually maximizes area under the curve, not the one that maximizes work-hours. We should do binary search *now*, to figure out what the optimum is.

Also, obviously, we should explore to discover highly effective methods of rest, instead of doing whatever random things seem good (unless, as it turns out, "whatever random thing seems good" is actually the best way to rest).

Picking up new tools

One thing that will be happening in this time, is there will be a flurry of new AI tools that can radically transform thinking and research, perhaps increasingly radical tools coming at a rate of once a month or faster.

Being able to take advantage of those tools and start using them for research immediately, with minimal learning curve, seems extremely high leverage.

If there are things that we can do that increase the ease of picking up new tools and using them to their full potential (instead of, as is common, using only the features afforded by your old tools and only very gradually)

Some thoughts (probably bad):

- Could we set up our workflows, somehow, such that it is easy to integrate new tools into them? Like if you already have a flexible, expressive research interface (something like Roam? or maybe, if the technology is more advanced by then, something like neurolink?), and you're used to regular changes in capability to the back of the interface?
- Can we just practice? Can we have a competitive game of introducing new tools, and trying to orient to them and figure out how to exploit them creatively as possible?
- Probably it should be some people's full time job to translate cutting edge developments in AI into useful tools and practical workflows, and then to teach those workflows to the researchers?
- Can we design a meta-tool that helps us figure out how to exploit new tools? Is it possible to train an AI assistant specifically for helping us get the most out of our new AI tools?
- Can we map out the sort of constraints on human thinking and/or the sorts of tools that will be possible, in advance, so that we can practice with much weaker versions of those tools, and get a sense of how we would use them, so that we're ready when they arrive?
- Can we try out new tools on psychedelics, to boost neuroplasticity? Is there some other way to temporarily weaken our neural priors? Maybe some kind of training in original seeing?

Staying grounded and stable in spite of the stakes

Obviously, being one of the few hundred people on whom the whole future of the cosmos rests, while the singularity is happening around you, and you are confronted with the stark reality of how doomed we are, is scary and disorienting and destabilizing.

I imagine that that induces all kinds of psychological pressures, that might find release in any of a number of concerning outlets: by deluding one's self about the situation or slipping sideways into a more convenient world, by becoming manic and frenetic, by sinking into immovable depression.

We need our people to have the virtue of being able to look the problem in the eye, with all of its terror and disorientation, and stay stable enough to make tough calls, and make them sanely.

We're called to cultivate a virtue (or maybe a set of virtues) of which I don't know the true name, but which involves courage and groundedness, and determination-without-denial.

I don't know what is entailed in cultivating that virtue. Perhaps meditation? Maybe testing one's self at literal risk to one's life? I would guess that people in other times and places, who needed to face risk to their own lives and that of their families, and take action anyway, did have this virtue, or some part of it, and it might be fruitful to investigate those cultures and how that virtue was cultivated.

Any more ideas?

Defending the non-central fallacy

Aaron Bergman recently [defended](#) logical fallacies against the charge that they're bad arguments. His thesis wasn't new. Gwern [highlighted "Bayesian Informal Logic and Fallacy"](#) and ["Fallacies as weak Bayesian evidence"](#).

Defending logical fallacies sounds pretty fun, so I decided to do it myself. But, I thought, if I'm going to do it, I might as well go all the way. Why not defend the final boss of all fallacies, the so-called "[worst argument in the world](#)": the non-central fallacy.

Scott Alexander [describes](#) the non-central fallacy as such,

X is in a category whose archetypal member gives us a certain emotional reaction. Therefore, we should apply that emotional reaction to X, even though it is not a central category member.

When phrased like that, it's hard to disagree that this is a bad argument. Probably not literally the worst argument in the world, but still pretty bad. However, people don't usually phrase their arguments like that. Taken literally, this description is looking terribly like a strawman.

What might the steelmanned version of the fallacy look like? Here's one possibility:

While often mistaken for being outside of morally relevant category Y, X in fact belongs to Y upon reflection. Therefore, we ought to treat X more similarly to other more typical members of Y.

Imagine the following argument.

Person A: "I think eating meat is wrong."

Person B: "Why?"

Person A: "Because animal farming is *cruelty*. By eating meat, you are contributing to this cruelty, and that's wrong."

Person B: "That's ridiculous. The archetypal examples of cruelty are things like torture and child abuse. Animal farming just means raising animals for food. You, my friend, are guilty of the non-central fallacy."

As much as I sympathize with Person B, I must say that Person A appears to have the better point. It kind of just looks like Person B is deflecting from the central argument. Let's examine how Person A might reply.

Person A: "I'm not saying that animal farming merely fits the [dictionary definition](#) of *cruelty*, and therefore we ought to treat it exactly like every other case that matches that definition.

Look, why do we think that torture and child abuse are wrong in the first place? For me, it's because those things involve involuntary suffering, and I think involuntary suffering is bad, no matter who experiences it. When I said that animal farming is cruelty, I was merely using that word as short-hand to convey my stance on involuntary suffering."

Is my argument really a steelman of the non-central fallacy, or is it merely a different argument? Maybe I'm giving people too much credit by re-interpreting their argument in a way that they never actually meant. To find out, let's take a peek at some examples Scott Alexander gave in his post.

"Taxation is theft!" True if you define theft as "taking someone else's money regardless of their consent", but though the archetypal case of theft (breaking into someone's house and stealing their jewels) has nothing to recommend it, taxation (arguably) does. In the archetypal case, theft is both unjust and socially detrimental. Taxation keeps the first disadvantage, but arguably subverts the second disadvantage if you believe being able to fund a government has greater social value than leaving money in the hands of those who earned it. The question then hinges on the relative importance of these disadvantages. Therefore, you can't dismiss taxation without a second thought just because you have a natural disgust reaction to theft in general. You would also have to prove that the supposed benefits of this form of theft don't outweigh the costs.

Who makes this argument? Well, Wikipedia [cites](#) thinkers from Augustine of Hippo to John Locke and Frédéric Bastiat, but mostly agrees that in the modern day, this argument is primarily made by American libertarians. [Michael Huemer](#) is an American libertarian philosopher who [has made this argument](#) on several occasions, so he is probably a good representative. How does he put the argument?

Imagine that I have founded a charity organization that helps the poor.¹ But not enough people are voluntarily contributing to my charity, so many of the poor remain hungry. I decide to solve the problem by approaching well-off people on the street, pointing a gun at them, and demanding their money. I funnel the money into my charity, and the poor are fed and clothed at last.

In this scenario, I would be called a thief. Why? The answer seems to be: because I am *taking other people's property without their consent*. The italicized phrase just seems to be what "[theft](#)" means. "Taking without consent" includes taking by means of a threat of force issued against other people, as in this example. This fact is not altered by what I do with the money after taking it. You wouldn't say, "Oh, you gave the money to the poor? In that case, taking people's property without consent *wasn't* theft after all." No; you might claim that it was a socially beneficial theft, but it was still a theft.

Now compare the case of taxation. When the government "taxes" citizens, what this means is that the government demands money from each citizen, under a threat of force: if you do not pay, armed agents hired by the government will take you away and lock you in a cage. This looks like about as clear a case as any of taking people's property without consent. So the government is a thief. This conclusion is not changed by the fact that the government uses the money for a good cause (if it does so). That might make taxation a socially beneficial kind of theft, but it is still theft.

This argument is sounding suspiciously similar to my steelman. Put in my own words, Michael Huemer is saying, "although you might not think of a tax collector as a thief, the two have as much in common. Just as we wouldn't be so ready to accept the justification of a thief who says they're 'serving the public interests' by stealing, we shouldn't be so willing to do the same for the tax collector."

Is Michael Huemer simply being deontological? Scott Alexander *had* alluded to his "personal and admittedly controversial opinion [that] much of deontology is just an attempt to formalize and justify [the non-central fallacy]." Huemer responds to this accusation directly,

If taxation is theft, does it follow that we must abolish all taxation? Not necessarily. Some thefts might be justified. If you have to steal a loaf of bread to survive, then you are justified in doing so. Similarly, the government might be justified in taxing, if this is necessary to prevent some terrible outcome, such as a breakdown of social order.

Why, then, does it matter whether taxation is theft? Because although theft *can* be justified, it is *usually* unjustified. It is wrong to steal without having a very good reason. What counts as good enough reasons is beyond the scope of this short article. But as an example, you are not justified in stealing money, say, so that you can buy a nice painting

for your wall. Similarly, if taxation is theft, then it would probably be wrong to tax people, say, to pay for an art museum.

In other words, the "taxation is theft" thesis has the effect of *raising the standards* for justified use of taxes. When the government plans to spend money on something (support for the arts, a space program, a national retirement program, and so on), one should ask: would it be permissible to steal from people in order to run this sort of program? If not, then it is not permissible to tax people in order to run the program, since taxation is theft.

Agree or disagree with his argument, Huemer is making a perfectly valid inference. As a society, maybe we are being duped when we are told that taxation is *not real theft because it's totally justified, unlike normal theft, trust me*. We have all seen the [lengths to which the human mind will go](#) to justify phenomena it considers normal and natural, even when those things are actually quite terrible upon closer examination. When viewed from an outside perspective -- outside society, that is -- Huemer's point becomes even clearer.

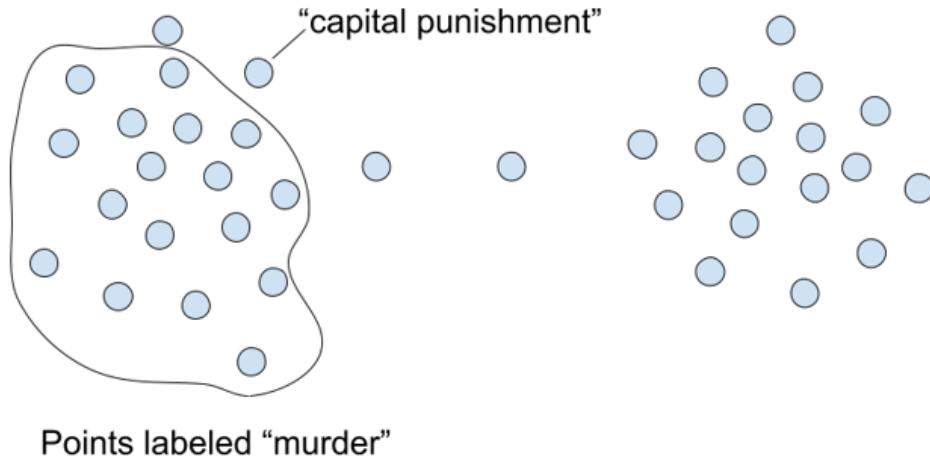
Suppose anthropologists open a dialogue with a tribe that throws people into lava after every solar eclipse. "This behavior is ritual sacrifice, and therefore barbaric!" the anthropologists object. "Not so," replies the chief. "Our Western critics are always committing the non-central fallacy. Unjust ritual sacrifice is when other tribes illegitimately sacrifice people to false gods. Our gods are real."

How do Scott Alexander's other examples fare? He also mentions the argument "Capital punishment is *murder!*" My steelman for "taxation is theft" can easily extend to this case too. The reader can probably fill in the blanks: just as we wouldn't accept the justification "but that guy was evil" as a good one in the case of ordinary murder, maybe we should also be skeptical when the state makes the same excuse for capital punishment.

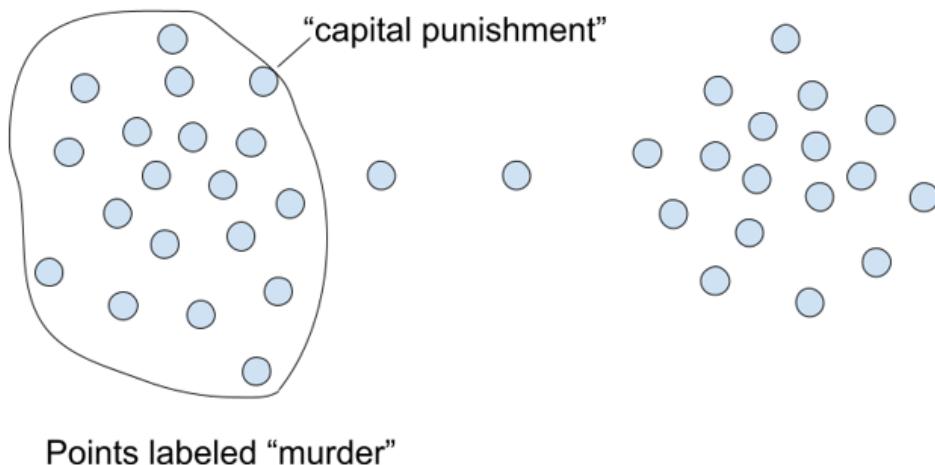
In fact, I would allege that *all* of Scott Alexander's examples suffer from the same flaw. Far from changing the argument, I think my steelman is the *usual* meaning people have in mind when they argue "X is Y" as a reason to oppose X.

Scott Alexander probably called this argument the worst in the world, not because it really is the worst, but because it is so common. However, I suspect that the argument is so common precisely because it conveys so much using so few words. People who use it are trying to convey something meaningful.

Consider this sketch.



Here we see [points in thingspace](#) related to murder. The interlocutor has labeled some points murder, but left capital punishment out. The reply "capital punishment is *murder*" could just mean that we ought to draw a more natural boundary. Although capital punishment is a non-central member of the murder cluster, it's still a member. And a more natural clustering would reflect that.



While capital punishment might not be an archetypal member of the cluster, it certainly has more in common with murder than the points of a different cluster. It's as if we want to say, "if you just reflect a bit, you'll see that capital punishment is more similar to murder than you previously thought. Surely it would be unfair to deny that it's at least very close."

Admittedly, people are often lazy and so often don't clarify themselves after using the non-central fallacy. Also, it's common for people to be unable to fully elaborate their own arguments. Yet let us not confuse a lack of rhetorical skills for bad argumentation. What is epistemic charity good for anyway?

Toward A Bayesian Theory Of Willpower

(crossposted from [Astral Codex Ten](#))

I.

What is willpower?

Five years ago, I reviewed [Baumeister and Tierney's book](#) on the subject. They tentatively concluded it's a way of rationing brain glucose. But their key results have [failed to replicate](#), and people who know more about glucose physiology say it [makes no theoretical sense](#).

Robert Kurzban, one of the most on-point critics of the glucose theory, gives his own model of willpower: it's a way of [minimizing opportunity costs](#). But how come my brain is convinced that playing Civilization for ten hours has no opportunity cost, but spending five seconds putting away dishes has such immense opportunity costs that it will probably leave me permanently destitute? I can't find any correlation between the subjective phenomenon of willpower or effort-needingness and real opportunity costs at all.

A tradition originating in psychotherapy, and ably represented eg [here by Kaj Sotala](#), interprets willpower as conflict between mental agents. One "subagent" might want to sit down and study for a test. But maybe one subagent represents the pressure your parents are putting on you to do well in school so you can become a doctor and have a stable career, and another subagent represents your own desire to drop out and become a musician, and even though the "do well in school" subagent is on top now, the "become a musician" subagent is strong enough to sabotage you by making you feel mysteriously unable to study. This usually ends with something about how enough therapy can help you reconcile these subagents and have lots of willpower again. But this works a lot better in [therapy books](#) than it does in real life. Also, what childhood trauma made my subagents so averse to doing dishes?

I've come to disagree with all of these perspectives. I think willpower is best thought of as a Bayesian process, ie an attempt to add up different kinds of evidence.

II.

My model has several different competing mental processes trying to determine your actions. One is a prior on motionlessness; if you have no reason at all to do anything, stay where you are. A second is a pure reinforcement learner - "do whatever has brought you the most reward in the past". And the third is your high-level conscious calculations about what the right thing to do is.

These all submit "evidence" to your basal ganglia, the brain structure that chooses actions. Using the same evidence-processing structures that you would use to resolve ambiguous sense-data into a perception, or resolve conflicting evidence into a belief, it resolves its conflicting evidence about the highest-value thing to do, comes up with some hypothesized highest-value next task, and does it.

I've previously quoted Stephan Guyenet on the [motivational system of lampreys](#) (a simple fish used as a model organism). Guyenet describes various brain regions making "bids" to the basal ganglia, using dopamine as the "currency" - whichever brain region makes the highest bid gets to determine the lamprey's next action. "If there's a predator nearby", he writes "the flee-predator region will put in a very strong bid to the striatum".

The economic metaphor here is cute, but the [predictive coding](#) community uses a different one: they describe it as representing the "confidence" or "level of evidence" for a specific calculation. So an alternate way to think about lampreys is that the flee-predator region is

saying "I have VERY VERY strong evidence that fleeing a predator would be the best thing to do right now." Other regions submit their own evidence for their preferred tasks, and the basal ganglia weighs the evidence using Bayes and flees the predator.

This ties the decision-making process into the rest of the brain. At the deepest level, the brain isn't really an auction or an economy. But it is an inference engine, a machine for weighing evidence and coming to conclusions. Your perceptual systems are like this - they weigh different kinds of evidence to determine what you're seeing or hearing. Your cognitive systems are like this, they weigh different kinds of evidence to discover what beliefs are true or false. Dopamine affects all these systems in predictable ways. My theory of willpower asserts that it affects decision-making in the same way - it's representing the amount of evidence for a hypothesis.

III.

In fact, we can look at some of the effects of dopaminergic drugs to flesh this picture out further.

Stimulants increase dopamine in the frontal cortex. This makes you more confident in your beliefs (eg cocaine users who are sure they outrun that cop car) and sometimes perceptions (eg how some stimulant abusers will hallucinate voices). But it also improves willpower (eg Adderall helping people study). I think all of these are functions of increasing the (apparent) level of evidence attached to "beliefs". Since the frontal cortex disproportionately contains the high-level conscious processes telling you to (eg) do your homework, the drug artificially makes these processes sound "more convincing" relative to the low-level reinforcement-learning processes in the limbic system. This makes them better able to overcome the desire to do reinforcing things like video games, and also better able to overcome the prior on motionlessness (which makes you want to lie in bed doing nothing). So you do your homework.

Antipsychotics decrease dopamine. At low doses of antipsychotics, patients might feel like they have a little less willpower. At high doses, so high we don't use them anymore, patients might sit motionless in a chair, not getting up to eat or drink or use the bathroom, not even shifting to avoid pressure sores. Now not only can the frontal cortex conscious processes not gather up enough evidence overcome the prior on motionlessness even the limbic system instinctual processes (like "you should eat food" and "you should avoid pain") can't do it. You just stay motionless forever (or until your doctor lowers your dose of antipsychotics).

In contrast, people on stimulants fidget, pace, and say things like "I have to go outside and walk this off now". They have so much dopamine in their systems that any passing urge is enough to overcome the prior on motionlessness and provoke movement. If you really screw up someone's dopamine system by severe methamphetamine use or obscure side effects of swinging around antipsychotic doses, you can give people involuntary jerks, tics, and movement disorders - now even random neural noise is enough to overcome the prior.

(a quick experiment: wiggle your index finger for one second. Now wave your whole arm in the air for one second. Now jump up and down for one second. Now roll around on the floor for one second. If you're like me, you probably did the index finger one, maybe did the arm one, but the thought of getting up and jumping - let alone rolling on the floor - sounded like too much work, so you didn't. These didn't actually require different amounts of useful resources from you, like time or money or opportunity cost. But the last two required moving more and bigger muscles, so you were more reluctant to do them. This is what I mean when I say there's a prior on muscular immobility)

IV.

I think this theory matches my internal experience when I'm struggling to exert willpower. My intellectual/logical brain processes have some evidence for doing something ("knowing how the education system works, it's important to do homework so I can get into a good college

and get the job I want"). My reinforcement-learner/instinctual brain processes have some opposing argument ("doing your homework has never felt reinforcing in the past, but playing computer games has felt really reinforcing!"). These two processes fight it out. If one of them gets stronger (for example, my teacher says I have to do the homework tomorrow or fail the class) it will have more "evidence" for its view and win out.

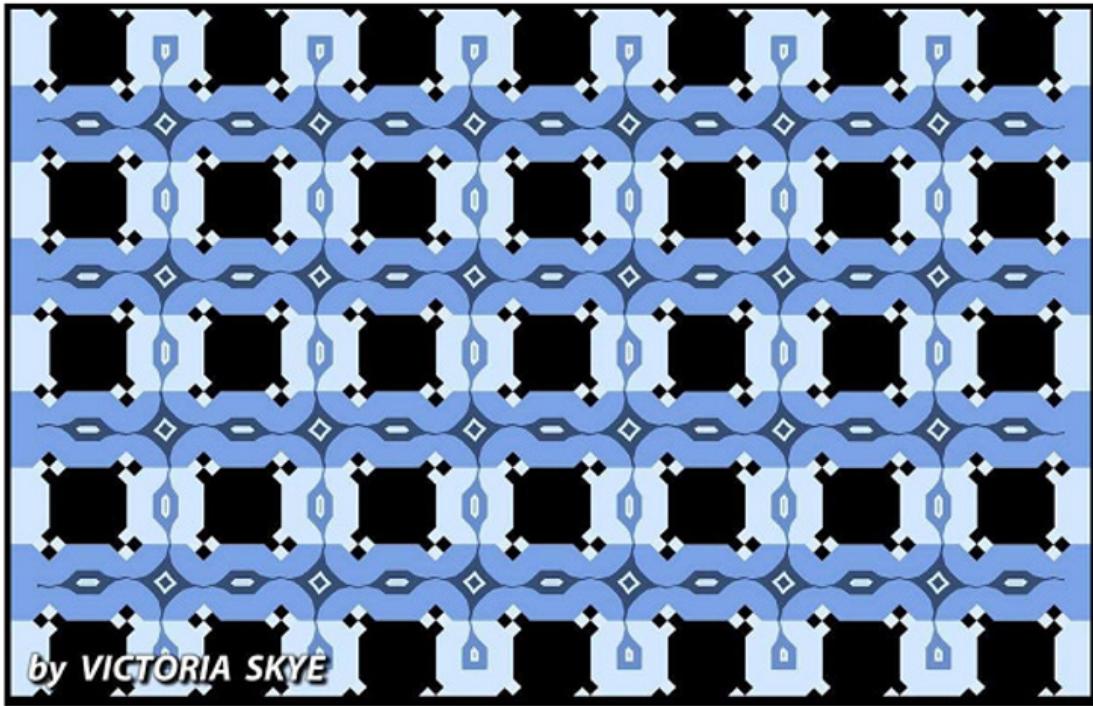
It also explains an otherwise odd feature of willpower: sufficient evidence doesn't necessarily make you do something, but overwhelming evidence sometimes does. For example, many alcoholics know that they need to quit alcohol, but find they can't. They only succeed after they "hit bottom", ie things go so bad that the evidence against using alcohol gets "beyond a reasonable doubt". Alcoholism involves some imbalance in brain regions such that the reinforcing effect of alcohol is abnormally strong. The reinforcement system is always more convinced in favor of alcohol than the intellectual system is convinced against it - until the intellectual evidence becomes disproportionately strong even more than the degree to which the reinforcement system is disproportionately strong.

Why don't the basal ganglia automatically privilege the intellectual/logical processes, giving you infinite willpower? You could give an evolutionary explanation - in the past, animals were much less smart, and their instincts were much better suited to their environment, so the intellectual/logical processes were less accurate, relative to the reinforcement/instinctual processes, than they are today. Whenever that system last evolved, it was right to weight them however much it weighted them.

But maybe that's giving us too much credit. Even today, logical/intellectual processes can be pretty dumb. Millions of people throughout history have failed to reproduce because they became monks for false religions; if they had just listened to their reinforcement/instinctual processes instead of their intellectual/logical ones, they could have avoided that problem. The moral law says we should spend our money saving starving children instead of buying delicious food and status goods for ourselves; our reinforcement/instinctual processes let us tell the moral law to f#@k off, keeping us well-fed, high-status, and evolutionarily fit. Any convincing sophist can launch an attack through the intellectual/logical processes; when they do, the reinforcement/instinctual processes are there to save us; Heinrich argues that [the secret of our success](#) is avoiding getting too bogged down by logical thought. Too bad if you have homework to do, though.

Does this theory tell us how to get more willpower? Not immediately, no. I think naive attempts to "provide more evidence" that a certain course of action is good will fail; the brain is harder to fool than people expect. I also think the space of productivity hacks has been so thoroughly explored that it would be surprising if a theoretical approach immediately outperformed the existing atheoretical one.

I think the most immediate gain to having a coherent theory of willpower is to be able to more effectively rebut positions that assume willpower doesn't exist, like [Bryan Caplan's theory of mental illness](#). If I'm right, lack of willpower should be thought of as an imbalance between two brain regions that decreases the rate at which intellectual evidence produces action. This isn't a trivial problem to fix!



The lines here are perfectly straight - feel free to check with a ruler. Can you force yourself to perceive them that way? If not, it sounds like you can't always make your intellectual/logical system overrule your instincts, which might make you more sympathetic to people with low willpower.

Thirty-three randomly selected bioethics papers

Some scholarly fields are very healthy (e.g., physics). Some are fairly unhealthy (e.g., [evolutionary psychology](#), sad to say). Some are outright crazy (e.g., philosophy of religion).

How good or bad is bioethics? How rigorous and truth-tracking is it? How much social benefit or harm does it cause?

Looking at lots of random examples is an under-used tool for making progress on this kind of question. It's fast, it avoids the perils of cherry-picking, and it doesn't require you to trust someone else's high-level summary of the field.

I picked the two highest-impact-factor "[medical ethics](#)" journals that have the word "bioethics" in the title: *The American Journal of Bioethics* and *Bioethics*. Together, these two journals release about 500 reports, book reviews, etc. per year.

I then picked a random article from 2014 through 2020 from each journal, plus extra random articles from the more cited journal (*The American Journal of Bioethics*): five from 2016, 2018, and 2020, and two from 2017 and 2019.

For each article, I quoted the abstract (if there is one), the first two paragraphs, and the final paragraph. These are provided below, sorted by year.

Obviously this breadth-first approach won't provide a full sense of the quality of argumentation in these papers; but it will hopefully provide a picture of what kinds of views tend to get argued for in the field, what those arguments tend to look like, what tends to get taken for granted, what tends to get ignored, and so on.

I expect more accurate conclusions and faster consensus-building about bioethics if conversation is grounded in a common pool of examples (which can then be drilled down on to spot-check a particular paper's arguments, etc.).

Added: aphyer summarized the articles in the comment section; I've put aphyer's summaries [in an editable Google Doc](#) to make it easier to improve on the summaries (and to encourage article-by-article critiques, in the comments below or in the Doc).

2014

1. [**Against Fetishism About Egalitarianism and in Defense of Cautious Moral Bioenhancement**](#)

Ingmar Persson and Julian Savulescu. *The American Journal of Bioethics* 14(4). Open Peer Commentaries.

We respond to Sparrow's (2014) criticisms of our support for a project of moral bioenhancement. We distinguish between a confident and cautious proposal, arguing that his objections only apply to a confident proposal, while we endorse a cautious proposal. We argue that the project of moral bioenhancement (1) need not be more compulsory than traditional moral education, which should supplement it; (2) is not dependent on "shonky" sociobiology or biologism that downplays the importance of traditional education and the role of cognitive processes; (3) is not committed to increasing inequality, but if judiciously applied would rather reduce it; and (4) would be desirable even if, as he believes, it leads to a democracy in which a moral elite has more political power. Sparrow fails to appreciate the extent of existing inequalities in moral as well as socioeconomic respects. He displays a fetish about egalitarianism and labors under a prescientific understanding of the relationship between the brain and moral behavior. We also argue that he confuses liberal neutrality with moral relativism, and fails to realize the extent to which all functioning societies have to be based on common core of norms and values.

When Robert Sparrow criticizes our work on moral bioenhancement (see Persson and Savulescu 2012), he doesn't bother to make precise the claims by us to which he objects. It's important to distinguish two types of proposals about moral bioenhancement, a *confident* one and a *cautious* one. (This is a simplification; there is in fact a whole spectrum of possible views with varying degrees of confidence/cautiousness.) These types of proposals differ with respect to the following three issues.

[...]

We believe that judicious moral bioenhancement will reduce the inequality between people with respect to moral motivation, rather than exacerbating it. But let's play along with Sparrow and hypothesize that it will create a morally enhanced elite and that this will result in an inegalitarian democracy "restricting participation in government to the morally enhanced" (25). He suggests that nonetheless our current commitment to egalitarianism "might provide reasons to resist bringing it about such that an inegalitarian politics should be justified" (42). This is strong stuff. Sparrow seems to say that even if a morally elitist democracy would pursue the morally right politics—that would bring anthropogenic climate change to a halt, put an effective check on the use of weapons of mass destruction, and so on—we should still prevent it coming into existence, clinging to our current more "egalitarian" democracies that leave us teetering on the brink of disaster. This seems to us to be making a fetish out of current egalitarianism, which—considering the huge socioeconomic inequalities it tolerates both within democratic states and globally—is surely far from perfect.

2. Why Publishing Pseudonymously Can Protect Academic Freedom

Francesca Minerva. *Bioethics* 28(4). Symposium: 'Anonymised Publishing in Bioethics'.

'The point of philosophy is to start with something so simple as not to seem worth stating, and to end with something so paradoxical that no one will believe it'
Bertrand Russell, *The Philosophy of Logical Atomism*

In 'New Threats to Academic Freedom' I suggested that new media may represent a threat to academic freedom.[1] In particular, I argued that academic freedom may be limited *ab ovo* if people fear that they will have to go through media storms when publishing papers that might be perceived as controversial.

[...]

I would prefer a world where people were free to ask questions and to attempt to give answers to those questions without being dragged into media frenzies, missing out on job opportunities (as happened to both me and Alberto Giubilini) and being threatened with physical harm. But as we do not live in such an ideal world, the best we can do in order to change the current society into a less violent and intolerant one is to keep asking questions, and to keep seeking answers.

2015

3. A Philosophical Misunderstanding at the Basis of Opposition to Nudging

Shlomo Cohen. *The American Journal of Bioethics* 15(10). Open Peer Commentaries.

Ploug and Helm's (2015) insightful article criticizes recent views—mainly those advanced by Saghai (2013) and me (Cohen 2013)—that allow "nudging" patients while obtaining their informed consent (IC). Although many of their arguments deserve discussion, this short commentary focuses on one philosophical point that I see as pivotal in determining perspectives in the debate. The authors indeed open their criticism of my approach by reference to this point: "First and foremost, while tacitly assuming that the value of informed consent is derived from the protection of personal autonomy, Cohen's view fails to provide a coherent and plausible account of personal autonomy." Although seemingly trivial and overwhelmingly common, the diagnosis that one's conception of valid IC is determined by one's theoretical conception of personal autonomy is in my view wrong: The specific amount of autonomy often makes little difference.[1] If we derived the ethics of IC from the concept of autonomy, we would probably conclude with the authors that nudging is incompatible with respecting autonomy, since, conceptually, manipulation and respecting autonomy point in opposing directions. However, the idea that the ethics of IC can be determined by theories of autonomy is wrong. I must leave the more theoretical discussion for another occasion, and restrict myself to showing how this is exemplified in the authors' own account.

The requirements for IC entailed by the authors' theoretical account of autonomy are very unspecific: We must provide "adequate" information; the patient needs to "understand" it; the patient cannot be "unduly influenced" while deliberating; these conditions are supposed to enable "reasonable" choice. The problem is that all these parameters are terribly underdetermined ("not unduly influenced" may be question-begging): Except for hard paternalism or outright deception, they can allow the entire spectrum of views under debate. I too can embrace these formulations—nothing in them rules out nudging. Why don't the authors provide more precise directives? They are not guilty of sloppy work; the deep point is rather that beyond general pronouncements, conceptions of autonomy cannot

provide practical ethical guidelines generally or specifically for IC. One central reason for this is that true autonomy is an exceedingly superhuman possibility and therefore cannot reliably prescribe what providing the conditions for humanly possible autonomy requires morally (e.g., whether to avoid nudging entirely)—that must rather be determined by some different method.

[...]

In this short commentary I have argued that the guidelines for a valid practice of IC are not predominantly determined by theoretical conceptions of autonomy, that they are rather determined by relevant social conventions, and that those conventions allow room for some types of nudging. Once this philosophical point is clear, the question ceases to be whether nudging is permissible, but which nudging is.

4. Why not Commercial Assistance for Suicide? On the Question of Argumentative Coherence of Endorsing Assisted Suicide

Roland Kripke. *Bioethics* 29(7). Article.

Abstract: Most people who endorse physician-assisted suicide are against commercially assisted suicide – a suicide assisted by professional non-medical providers against payment. The article questions if this position – endorsement of physician-assisted suicide on the one hand and rejection of commercially assisted suicide on the other hand – is a coherent ethical position. To this end the article first discusses some obvious advantages of commercially assisted suicide and then scrutinizes six types of argument about whether they can justify the rejection of commercially assisted suicide while simultaneously endorsing physician-assisted suicide. The conclusion is that they cannot provide this justification and that the mentioned position is not coherent. People who endorse physician-assisted suicide have to endorse commercially assisted suicide as well, or they have to revise their endorsement of physician-assisted suicide.

For many years, there has been a fierce controversy regarding the question of the ethical evaluation of assisted suicide. Surprisingly, there is broad agreement on one point, namely the rejection of commercial assistance for suicide. ‘Commercial assistance for suicide’ means that professional non-medical providers assist people in the implementation of their suicidal intents in return for payment. Not only opponents of physician-assisted suicide (PAS) but also most of its proponents think that commercially assisted suicide (CAS) is immoral and should not be permitted – if they discuss this form of assisted dying at all.[1] For whilst the legitimacy or illegitimacy of PAS is the subject of intense discussion, the question of CAS leads a miserable existence within the international ethical discussion.

However, the issue is in no way far-fetched. Firstly, in some countries at least, there is ethical and political discussion on this topic. In Germany, for example, not only have some cases of CAS come to public attention in recent years, but also several legislative initiatives to prohibit CAS were launched.[2] Secondly, the issue of CAS seems to be reasonable because of the well-known general tendency

towards commercialization of various areas of life.[3] Thirdly, the issue is important due to theoretical reasons, namely as a touchstone for the coherence of the ethical position of the proponents of PAS.

[...]

If one does not want a society in which suicide and its support is normal and taken for granted like other services, and if one wants to adhere at the same time to the claim of coherence for their own ethical position, the only possibility is to reject PAS. Those who do not endorse CAS cannot endorse PAS, either.

2016

5. One Exemption Too Many: The Case for Mandated CCHD Screening

John D. Lantos, Julie Caciki, and Jeremy R. Garrett. *The American Journal of Bioethics* 16(1). Guest Editorial.

For more than 50 years, most industrialized countries have mandated population screening of newborns for medical conditions that meet a few straightforward criteria (Wilson and Jungner [1968](#)). The condition must be an important health problem. The screening test should be accurate. There must be an effective treatment that is readily available to the people who have been screened. The cost of the program, including screening, diagnosis, and treatment, must be reasonable. When all these conditions are met, then the medical, legal, and ethical consensus has been that it is justifiable to mandate newborn screening. Screening tests that meet these criteria have been endorsed by the Institute of Medicine, the American Society of Human Genetics and the American College of Medical Genetics, an NIH Task Force on Genetic Testing, the American Academy of Pediatrics, and the President's Council on Bioethics (Committee on Assessing Genetic Risks [1994](#); Botkin et al. [2015](#); Holtzman 1997; American Academy of Pediatrics Newborn Screening Task Force 2000; [President's Council on Bioethics 2008](#)). For all of these organizations, the ethical justification for screening is that it provides direct benefit to the newborn and that avoiding or delaying screening could cause harm to that newborn.

In their target article in this issue, Hom and colleagues (2016) raise the question of whether parental religious beliefs should allow exemptions to mandated screening for critical congenital heart disease (CCHD). It is important to be clear that in making this argument, they are not addressing three of the important and thorny controversies that are the usual focus of debates about newborn screening. The three controversial issues are (1) screening to detect conditions for which there is no effective treatment; (2) the use of stored samples from newborn screening for research unrelated to the screening program; and (3) tests for conditions that do not manifest until adulthood and for which there are no beneficial interventions in childhood. All three of those situations are appropriately controversial for traditional newborn screenings. None of them, however, is relevant to the central issue addressed in the target article or to the type of screening that is done for CCHD.

[...]

We acknowledge that some CCHDs are not easily correctable and thus that parents can ethically refuse treatment for the most severe CCHDs. But at present, the screening tests do not distinguish the easily correctable ones from those that are not so easy to treat. It is appropriate, then, to mandate newborn screening, without exceptions for religious beliefs, for CCHDs.

6. The Curious Case of the De-ICD: Negotiating the Dynamics of Autonomy and Paternalism in Complex Clinical Relationships

Daryl Pullman and Kathleen Hodgkinson. *The American Journal of Bioethics* 16(8). Article.

Abstract: This article discusses the response of our ethics consultation service to an exceptional request by a patient to have his implantable cardioverter defibrillator (ICD) removed. Despite assurances that the device had saved his life on at least two occasions, and cautions that without it he would almost certainly suffer a potentially lethal cardiac event within 2 years, the patient would not be swayed. Although the patient was judged to be competent, our protracted consultation process lasted more than 8 months as we consulted, argued with, and otherwise cajoled him to change his mind, all to no avail. Justifying our at times aggressive paternalistic intervention helped us to reflect on the nature of autonomy and the dynamics of the legal, moral, and personal relationships in the clinical decision-making process.

Several years ago our ethics consultation service faced a particularly vexing and exceptional case that forced us to reconsider our ethical and professional responsibilities in the face of what appeared to be a legally permissible but ill-advised medical decision. We had to reexamine the relationship between the legal and the moral, and the role of our ethics consultation service in this process. In addition to revisiting the much-discussed question of the relationship between autonomy and justified paternalistic intervention, we had to reflect upon the nature of the relationships between various health care professionals, the patient, and his family. Additionally, we had to assess the roles and responsibilities of members of our ethics consultation group who resisted this patient's legally permissible but ethically problematic request, and of the cardiologist who eventually facilitated it. Finally, we had to manage the tension between private interests and public goods when the decision in question involved the allocation of scarce resources in a publicly funded health care system.

There is a vast and expanding literature on the concept of autonomy and the nature of justified paternalistic intervention. Perspectives vary from those that acknowledge the primacy and centrality of autonomy in biomedical decision making (Gillon 2003; Wolpe 1998) to those that present a somewhat more nuanced perspective (O'Neill 2002; Kukla 2005; Levy 2014). Our intent is not to review this literature in any systematic or robust fashion, but rather to draw upon various perspectives in assessing, critiquing, and in some respects justifying the manner in which we managed this complex and difficult case. Our emphasis is

less on analytic precision and more on practical understanding as we continue to rely on these and related concepts in navigating the ambiguities of moral space.

[...]

Finally, we comment briefly on the implications of acceding to a patient's demands when the economic consequences of those choices are borne by a publicly funded health care system. When James's initial clinical assessment determined he needed an ICD, there was no question about cost; as a resident of a society that has deemed access to basic medical care a positive right, James was entitled to the provision of his basic health care needs. When he subsequently requested that the ICD be removed against all medical advice to the contrary, there was still no mention of the potential financial costs in discussions with James, although the issue did come up in conversation among the health care and clinical ethics teams. In this case the surgeon who offered to remove the ICD said he would not bill the provincial health plan for the usual fee associated with the procedure. But the cost of the removal was insignificant compared to the additional costs of maintaining James after his cardiac event, the subsequent replacement of the ICD, and the eventual heart transplant. What responsibilities should individual patients bear for the consequences of their decisions? Are the considerations in the current case different from those that arise in the context of other "lifestyle choices" where individuals make poor decisions that affect their health, resulting in a greater burden for the health care system? These latter issues are beyond the purview of the current discussion but should not go unmentioned. Talk about money will put a strain on any relationship, and for this reason we often tiptoe around such issues in our personal lives. Nevertheless, when all personal relational resources have been exhausted, as they were eventually in our dealings with James, perhaps the financial consequences of his decision should have been raised. Indeed, at that point the relationship was more contractual than personal in nature, and in the interests of transparency and full disclosure the real financial costs of the transaction he was insisting upon could have been revealed and discussed. But again, such financial considerations and discussions should occur only at the end of the relational encounter, not at the outset. In any case, in our publicly insured health care system talk of costs would serve only as one last (and perhaps desperate) attempt at moral suasion, as there are no coercive financial mechanisms available by which to recoup the costs of an unwarranted and frivolous demand for services. However, such mechanisms might exist in health care systems that rely more heavily on private insurance and where the "ethics of strangers" are often more readily apparent. Thus, while an ethic of intimacy may encourage care and compassion, an ethic of strangers may more readily accommodate individual responsibility. Finding the means by which to assess and nurture the full continuum of relational values and responsibilities is thus an ongoing challenge.

7. A Pilot Evaluation of Portfolios for Quality Attestation of Clinical Ethics Consultants

Joseph J. Fins, et al. *The American Journal of Bioethics* 16(3). Target Article.

Abstract: Although clinical ethics consultation is a high-stakes endeavor with an increasing prominence in health care systems, progress in developing standards for quality is challenging. In this article, we describe the results of a pilot project

utilizing portfolios as an evaluation tool. We found that this approach is feasible and resulted in a reasonably wide distribution of scores among the 23 submitted portfolios that we evaluated. We discuss limitations and implications of these results, and suggest that this is a significant step on the pathway to an eventual certification process for clinical ethics consultants.

In 2013, the American Society for Bioethics and Humanities Quality Attestation Presidential Task Force (ASBH QAPTF) published a proposed two-step model for evaluating and attesting to the quality of clinical ethics (CE) consultants (Kodish and Fins et al. 2013). The first part of this two-step process envisioned the submission of a portfolio that summarizes the education, experience, philosophy, and substance of ethics case consultation to establish a consultant's eligibility to undergo a subsequent examination.

The Quality Attestation Presidential Task Force (QAPTF) was formed to develop a means to assess the performance of CE consultants. The task force functions under the aegis of ASBH, the primary society of bioethicists and scholars in the medical humanities and the organizational home for many individuals who perform clinical ethics consultation (CEC) in the United States. Members of the task force were appointed for their breadth and depth of experience in clinical ethics consultation and the diversity of their professional and educational backgrounds.

[...]

Much work remains to be done. The first experience with an ethics consultation portfolio has yielded ways to improve it and leaves undecided what the next steps should be. Future developments will require (1) more resources in time and expertise, (2) attention to governance with regard to the appropriate institutional home for CE consultant attestation/certification, and (3) financial self-sufficiency. Nonetheless, the field has moved further along the path toward a process to demonstrate professional competency of CE consultants and to develop robust quality standards consistent with the expectation for other members of the health care professions. Remaining challenges notwithstanding, this is a positive development for both clinical ethics consultation and for the patients and caregivers it serves.

8. Research Moratoria and Off-Label Use of Ketamine

Andrea Segal and Dominic Sisti. *The American Journal of Bioethics* 16(4). Open Peer Commentaries.

We wish to point out an additional consequence of the Catch-22 described by Andreea and colleagues (Andreea et al. 2016). The decades-long research gridlock of controlled drugs has unintentionally resulted in wide-scale off-label use of one such drug, ketamine, in free-standing private psychiatric clinics. The growth of these clinics and burgeoning consumer demand for ketamine raise important ethical and policy issues related to patient safety, consent, and the future of research on controlled drugs.

Ketamine was approved in 1970 as a dissociative anesthetic, and by the 1980s researchers began to use ketamine for the treatment of psychiatric disorders. Ketamine's properties as an NMDA (N-methyl-D-aspartate) receptor antagonist have spurred enthusiasm about a new line of psychiatric treatment for very depressed patients (Sisti, Segal, and Thase 2014). Outpatient clinics have sprung up worldwide purporting to treat serious depression with ketamine. Unfortunately, a market has developed where unscrupulous clinics haphazardly provide ketamine to all comers. In Australia, one clinic was selling ketamine-filled syringes for at-home use, raising serious concerns about patient safety and drug diversion (Worthington 2015).

[...]

Off-label commercial use of generic ketamine will accelerate as long as the clinical research on generic ketamine is not incentivized in some way. As this dynamic continues to unfold, we are concerned that desperate, potentially incapacitated patients will be harmed, and that a clinical mishap could set back legitimate research for this highly stigmatized drug. With proper clinical oversight of ketamine administration, and clear systematic assessments of its risks and benefits, some of the ethical concerns about off-label ketamine use can be assuaged. However, unchecked enthusiasm and market forces may take us down "a slippery ketamine slope" (Schatzberg 2014). In addition to pushing for research dollars for ketamine and other controlled substances, we also recommend stricter regulation and oversight of off-label use.

9. Good Ethics Begin With Good Facts

Birgitta Sujdak Mackiewicz. *The American Journal of Bioethics* 16(7). Case Commentaries.

Four-year-old James was transferred from an outlying hospital in a critically ill state and has not responded to conventional therapies. For the last day and a half he has been treated with inhalational sevoflurane. James's condition is not worsening, but he remains critically ill, suffering from a life-threatening condition. If inhalational sevoflurane remains the optimal treatment for James, then there is no clinical justification for transitioning him to VV ECMO given its risks and lack of clinical benefit—in this case there is no indication that any clinical benefit in doing so exists.

The case as presented constitutes both a clinical and an organizational ethics issue. The physicians who accepted James for transfer a few days ago are now faced with the decision of whether they can transition James off an effective therapy. The decision is necessitated by organizational decisions about the level of anesthesia staffing on weekends. In short, if the facts of the case hold, the answer is "no." It is not under ordinary circumstances, citing the case summary, "ethically permissible to transition a patient to a therapy that is working to another therapy, which has increased morbidity and mortality, due to insufficient staffing." To do so represents a failure of the facility to provide the necessary and reasonably expected resources tacitly promised by the acceptance of James from the outlying hospital. To switch James to a riskier treatment due to institutional inability to provide the appropriate care also constitutes an injustice to James. His initial transfer was based on the understanding that the accepting facility was able to

provide the necessary care. James might instead have been refused admission and transferred to a facility able to meet his needs.

[...]

The physician must consider whether the current treatment of inhalational sevoflurane is optimal for James even if staffing is maintained. Physicians have an ethical obligation to reassess patient status throughout the course of treatment and to recognize when a patient who is not declining, but is still critically ill, has an insufficient response to the current treatment warranting a new approach. This determination is outside of the scope of the ethicist's practice; however, it is not unreasonable for an ethicist to inquire as to the clinical factors at play and how the physician arrived at his or her decision. The ethicist may also recommend that physicians consider whether it is appropriate to seek further opinions from other institutions who may have had similar cases. This is not a questioning of the physician's medical judgment or competency by the ethicist, but rather an interdisciplinary exploration of the case at hand, examining the facts of the case, the literature, and the range of clinically and morally acceptable options from which physicians will make a recommendation.

10. The Ethics of Organ Donor Registration Policies: Nudges and Respect for Autonomy

Douglas MacKay and Alexandra Robinson. *The American Journal of Bioethics* 16(11). Target Article.

Abstract: Governments must determine the legal procedures by which their residents are registered, or can register, as organ donors. Provided that governments recognize that people have a right to determine what happens to their organs after they die, there are four feasible options to choose from: opt-in, opt-out, mandated active choice, and voluntary active choice. We investigate the ethics of these policies' use of nudges to affect organ donor registration rates. We argue that the use of nudges in this context is morally problematic. It is disrespectful of people's autonomy to take advantage of their cognitive biases since doing so involves bypassing, not engaging, their rational capacities. We conclude that while mandated active choice policies are not problem free—they are coercive, after all—voluntary active choice, opt-in, and opt-out policies are potentially less respectful of people's autonomy since their use of nudges could significantly affect people's decision making.

Governments must determine the legal procedures by which their residents are registered, or can register, as organ donors. Provided they recognize that people have a right to determine what happens to their organs after they die, an assumption underlying our analysis below, there are four feasible options to choose from.^[1] *Opt-in* requires people to actively register as organ donors, for example, by signing an organ donor registration card, or checking a box when renewing their driver's license. Many jurisdictions, including Canada, some U.S. states, the United Kingdom, and Germany, currently register organ donors in this way. *Opt-out* presumes that people are willing to be organ donors and so automatically registers them as such, but provides them with the opportunity to

opt out. This policy is employed by many European countries, including Spain, Austria, Belgium, Greece, and France. *Voluntary active choice* (VAC) presents people with the choice of registering as an organ donor or not, but does not require them to make a decision. This policy is employed by the U.S. states of California and Vermont. Finally, *mandated active choice* (MAC) presents people with the choice of registering as an organ donor or not, and requires them to make a decision, for example, by making the renewal of their driver's license conditional on them stating their donation preference. This policy is employed by the U.S. states of Illinois and New York, as well as New Zealand.

Debates regarding the ethics of organ donor registration policies have largely concerned the permissibility of opt-out policies. Proponents of these policies argue that they are likely to yield the highest donor registration rate, but critics object that they are morally deficient since they do not secure the consent of potential donors (Veatch 2000, 167–174; den Hartogh 2011; MacKay 2015). In this article, we set the issue of consent aside, exploring instead a different way in which these policies can be more or less respectful of people's autonomy. Organ donor registration policies differ not only in terms of whether they secure people's consent or not, but also in terms of the types of influence they employ to affect people's decision making. As Richard H. Thaler and Cass R. Sunstein make clear in *Nudge*, one possible reason that organ donor registration rates are so high in jurisdictions employing opt-out is that people exhibit status quo bias, a tendency to stick with the current state of affairs or choose default options (2008, 34– 35). By taking advantage of people's status quo bias in this way, opt-out policies employ what Thaler and Sunstein refer to as a "nudge," that is, a way of designing choice situations that "alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (2008, 6).

[...]

Consider second that people's registration status need not be fully determinative of the disposition of their organs after death for our analysis to have important implications for the ethics of donor registration policies. Instead, all that needs to be the case is that people's registration status significantly affects the disposition of their organs after death. That is, even if families continue to have a de facto or de jure veto over the disposition of their relative's organs, our analysis is important if people's registration status significantly affects the decisions that families make. Provided that people's registration status has this effect, then it matters ethically how that registration status is secured, particularly if future studies show that under VAC, opt-in, or opt-out policies, people's registration status is significantly influenced by their status quo bias. This point is important, since there is a good deal of evidence in the U.S. context showing that a person's registration as a donor is highly associated with familial decisions to donate (Siminoff and Lawrence 2002; Rodrigue, Cornell, and Howard 2006; Christmas et al. 2008; Traino and Siminoff 2013). The choice of donor registration policies therefore matters morally, and we hope that our article makes an important contribution to the ethical analyses of these policy options, identifying a way in which opt-in, optout, and VAC policies are morally deficient, and MAC policies potentially morally superior.

11. A Trust-Based Pact in Research Biobanks. From Theory to Practice

Virginia Sanchini, et al. *Bioethics* 30(4). Original Article.

Abstract: Traditional Informed Consent is becoming increasingly inadequate, especially in the context of research biobanks. How much information is needed by patients for their consent to be truly informed? How does the quality of the information they receive match up to the quality of the information they ought to receive? How can information be conveyed fairly about future, non-predictable lines of research? To circumvent these difficulties, some scholars have proposed that current consent guidelines should be reassessed, with trust being used as a guiding principle instead of information. Here, we analyse one of these proposals, based on a Participation Pact, which is already being offered to patients at the Istituto Europeo di Oncologia, a comprehensive cancer hospital in Milan, Italy.

Recent studies suggest that it is increasingly difficult to provide patients with accurate information during informed consent procedures.[1] While this problem affects many areas of medicine, it is particularly urgent in the context of research biobanks. In this latter setting, in addition to other traditional issues – such as patients' options at stake and their likelihood to develop a 'therapeutic misconception'[2] – the kind of information to be delivered renders canonical informed-consent forms potentially ill-suited. Several criticalities have been identified concerning the use of traditional informed consent (henceforth IC) in the domain of research biobanks.

The most widely known objection to the use of traditional IC in biobanks relates to the requirement of providing participants with full information about the research projects in which the specimens will be utilized. However, since at the time of collection it is impossible to foresee the future role of tissue samples in research, the provision of full information prior to signing of the consent is unfeasible. Therefore, the very concept of an IC for research biobanks as a means to provide the prospective participant with 'full information' seems inconsistent.[3]

[...]

From an applied bioethical viewpoint, one question is paramount. Does the PP work? From the initial data we described, it appears to be so. More than 97% of participants signed the *Pact* in a non-anonymous form, an almost two-fold increase with respect to the traditional IC, arguing for the relevance of trust in mediating the relationship between participants and researchers.

2017

12. Responsible Translation of Psychiatric Genetics and Other Neuroscience Developments: In Need of Empirical Bioethics Research

Gabriel Lázaro-Muñoz. *The American Journal of Bioethics* 17(4). Open Peer Commentaries.

In recent years, billions of dollars have been allocated to large-scale neuroscience projects with the goal of advancing our understanding of neural function, developing neurotechnologies, and, ultimately, improving neuropsychiatric care and prevention. These projects include the U.S. Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative, the European Union's Human Brain Project (HBP), the U.S. National Institute of Mental Health's Research Domain Criteria (RDoC) Initiative, and the Psychiatric Genomics Consortium (PGC), among others. Although these initiatives will undoubtedly yield clinical benefits, if we aim to responsibly research and translate the knowledge and technologies they produce, it is essential to empirically examine the potential harms—including opportunity costs—and the ethical or “neuroethics” challenges generated.

These initiatives are already yielding benefits. For example, over the past 8 years, the PGC has identified more than 100 genomic loci that are reliably associated with schizophrenia (Schizophrenia Working Group of the PGC 2014). The PGC achieved this by pooling resources and large samples from studies around the world and using an unbiased array-based genetic testing approach. This recent progress is in stark contrast to the years of struggle psychiatric genetics researchers faced trying to replicate findings obtained through candidate gene approach studies with small samples and little collaboration across research laboratories (Farrell et al. 2015; Need and Goldstein 2014). The success of initiatives like the PGC is evidence of how large-scale neuroscience projects can accelerate our understanding of psychiatric illnesses, which are known to be multifactorial and highly complex from a biological standpoint. Findings such as the identification of genomic loci associated with schizophrenia can lead to better risk prediction, resource allocation, prevention, drug targets, diagnosis, and treatment selection.

[...]

At the end of the day, as bioethicists, our goal should not be to simply raise questions, but to set well-informed ethical agendas and search for answers to effectively address ethical challenges. These four actions can help us achieve that.

13. Interconnectedness and Interdependence: Challenges for Public Health Ethics

Jonathan Beever and Nicolae Morar. *The American Journal of Bioethics* 17(9). Open Peer Commentaries.

An increasing number of contemporary voices in both bioethics and environmental ethics have grown dissatisfied with the schisms, abysses, and raging torrents that continue to flow between those two domains of ethical inquiry. Thus, Lee's (2017) call for a public health ethics that serves as a bridge between them is a welcome addition to an expanding literature that highlights in terms of “health” the interconnection between individuals, their communities, and their environments.

It is clear to us that public health ethics, in the footsteps of others seeking the same ecological understanding, has begun to articulate values based on “our connectedness with each other, animals, and the environment” (9). This conception comes not only from environmental philosophy (Jamieson 2001), from philosophy of ecology (Shrader-Frechette and McCoy 1993), and from eco-phenomenology (Brown and Toadvine 2003), but also from within bioethics and environmental ethics themselves (Beever and Morar 2016a; Morar and Skorburg 2016; Jennings 2016; Whitehouse 2003). Indeed, Lee’s argument is a revisit of Van Rensselaer Potter’s own predominantly anthropocentric concerns about “environmental health” as both intend to capture the ways that environmental and nonhuman animal factors influence human health. All agree that there is a growing recognition of the intimate relations among previously siloed epistemic units and ethical domains.

[...]

In conclusion, in the absence of a careful distinction between interconnectedness and interdependence, public health ethics reaches too far and too fast so that it can account for the immense complexity of relations for which the challenges demand answers. It is, as of yet, not only epistemically impoverished concerning the nature of relationships, but also normatively impoverished of resources to sufficiently parse out the significance of certain moral contraventions. Yet even if we were to overcome such conceptual issues, focusing on bridges still tends to ignore the water for the sake of holding firm stable shores of inquiry. Perhaps the only way to pay attention to that flowing in between (which is why the bridge exists in the first place, of course) is to move beyond Potter and recognize there are no stable shores (Beever and Morar 2016b). Instead, we might be compelled to call into question the historical silos of applied ethical inquiry. Public health ethics hasn’t evidenced that it is robust enough to do that hard work— yet.

14. Expanded Access for Nusinersen in Patients With Spinal Muscular Atrophy: Negotiating Limited Data, Limited Alternative Treatments, and Limited Hospital Resources

Benjamin S. Wilfond, Christian Morales, and Holly A. Taylor. *The American Journal of Bioethics* 17(10). Case Report.

The issue of resource allocation is not new to bioethics. From the early days of dialysis to current complex solid organ transplant systems, medical advances continue to generate rich questions about the distribution of resources that are limited. Who will qualify? How will we decide? Who will need to wait? Who will die waiting?

Spinal muscular atrophy 1 (SMA1) is a rare autosomal recessive disease (approximately 250 cases per year in the United States). The disease causes rapid, progressive decline in muscular strength with preserved intellectual function. Historically, 50% of infants die by age 1 year and the majority of the remaining patients by age 2 years, following the provision of comfort care. Until recently, life extension was limited to symptom management with long-term ventilator and nutrition support.

[...]

The Pediatric Neuromuscular Service requests an ethics consultation to assist in determining how to fairly allocate nusinersen to its patients, should they choose to participate in the EAP. Within driving distance of the center there are approximately 100 SMA1 patients, many of whom are already ventilator dependent. These patients are not all being followed by the service, but are not prohibited from applying for EAP therapy. There will likely be 5-10 new cases per year in the center's catchment area. It is possible that the center will be the only hospital in the state to provide this treatment. Given their current staffing, procedural space, and bed space availability, the center is only able to reasonably guarantee therapy for 2-4 patients per month. The center worries about how to ethically prioritize patients, knowing it will be difficult to accommodate all those who might benefit from treatment. How should the group determine who will be treated and when?

15. The Consequences of Vagueness in Consent to Organ Donation

David M. Shaw. *Bioethics* 31(6). Original Article.

Abstract: In this article I argue that vagueness concerning consent to post-mortem organ donation causes considerable harm in several ways. First, the information provided to most people registering as organ donors is very vague in terms of what is actually involved in donation. Second, the vagueness regarding consent to donation increases the distress of families of patients who are potential organ donors, both during and following the discussion about donation. Third, vagueness also increases the chances that the patient's intention to donate will not be fulfilled due to the family's distress. Fourth, the consequent reduction in the number of donated organs leads to avoidable deaths and increased suffering among potential recipients, and distresses them and their families.

There are three strategies which could be used to reduce the harmful effects of this vagueness. First, recategorizing the reasons (commonly referred to as 'overrules' under the current system) given by families who refuse donation from registered donors would bring greater clarity to donation discussions. Second, people who wish to donate their organs should be encouraged to discuss their wishes in detail with their families, and to consider recording their wishes in other ways. Finally, the consent system for organ donation could be made more detailed, ensuring both that more information is provided to potential donors and that they have more flexibility in how their intentions are indicated; this last strategy, however, could have the disadvantage of discouraging some potential donors from registering.

Transplantation of organs from living and deceased organ donors saves and improves tens of thousands of lives globally every year. Despite a substantial increase in the number of living kidney donors in recent years, donation from deceased patients remains the leading source of organs for transplantation in most countries. In this article, organ donation is discussed in the context of the ethical and legal frameworks that apply in the United Kingdom, but most countries operate broadly similar deceased donation systems that are affected by similar issues of vagueness.

In England and Northern Ireland, people who wish to donate their organs after they die must consent to any such donation. They can do so by informing relatives of their wishes or by signing up to the organ donor register. In Scotland, the same applies, but the term used is 'authorization' rather than consent. Wales recently switched to a 'deemed consent' system where it will be presumed that a patient wished to donate in the absence of any evidence to the contrary (a so-called opt-out), but here too, those who wish to donate can register their intentions on the UK-wide database.

[...]

I have suggested three potential strategies for combating the problem of vagueness. Reclassifying 'overrules' and recognizing the importance of new evidence of refusal and reassessment of best interests will bring clarity to conversations with families. If donors broach the subject of donation with their families vagueness will be reduced, particularly if a personalized organ donation directive is created. Most importantly, preventing vagueness at the point of registration is better than attempting to cure it at the point of donation when the family is upset. Creating a new consent system and providing more detailed information might discourage some people from donating, but it would increase the chances that the wishes of someone who does want to donate will be respected, substantially increase health professionals' and families' confidence in the consent of donors, and consequently reduce the distress currently caused to both families and staff.

2018

16. A Model for Communication About Longshot Treatments in the Context of Early Access to Unapproved, Investigational Drugs

Eline M. Bunnik and Nikkie Aarts. *The American Journal of Bioethics* 18(1). Open Peer Commentaries

When seriously ill patients run out of standard treatment options, they may consider nonstandard treatment options such as expanded access, also known as "compassionate use." Through expanded access programs, patients are given access to investigational drugs that are still under development and not yet approved for marketing. As the safety and efficacy of unapproved drugs have not been fully established, it is uncertain whether these drugs will offer medical benefit. Especially when the compound is in an early stage of the drug development process, its odds of success may be low or "longshot." Expanded access raises ethical concerns, notably that seriously ill patients may overestimate the benefits of an investigational drug and underestimate its safety issues, fail to make informed decisions, and become susceptible to false hope and exploitation (Darrow et al. [2015](#)). The communication model proposed by Weiss and Fiest (2018) and its central distinction between low-odds and no-odds treatment can be used to assist seriously ill patients and their treating physicians not only with decision making with regard to initiating expanded access to investigational therapies but also with monitoring and managing their effects.

Thus, it may help to overcome some of the ethical problems associated with expanded access.

Systems for expanded access differ across countries, but have a set of conditions in common: Patients must be suffering from serious or life-threatening diseases must have exhausted standard treatment options, and must not be eligible for participation in clinical trials (Jarow et al. [2017](#)). The managing physician must believe that the potential benefits of the drug will outweigh the risks. A regulatory authority such as the Food and Drug Administration (FDA) will need to approve the request (FDA [2017](#)). In some countries, including the United States, requests must also be evaluated by an institutional review board. Importantly, the pharmaceutical company must be willing to supply the drug, often at no cost. Finally, patients must provide informed consent.

[...]

As the—often criticized—Right-to-Try movement in the United States is raising awareness of expanded access and seeking to increase its accessibility (Holbein et al. [2015](#)), demand among patients for unapproved drugs around the world is expected to rise. Adequate informed consent processes that explicitly rebut unwarranted therapeutic optimism will be crucial for morally responsible practices of expanded access in the future.

17. Enhance Diversity Among Researchers to Promote Participant Trust in Precision Medicine Research

Demetrio Sierra-Mercado and Gabriel Lázaro-Muñoz. *The American Journal of Bioethics* 18(4). Open Peer Commentaries.

Participants in the study reported by Kraft and colleagues (2018) raised an issue that supports how having more diversity among researchers could help promote trust toward precision medicine research and, ultimately, more precise medicine for all. First, participants' perceptions of trustworthiness in biomedical research were related to their personal or group experiences with racism. Ideally, experience with racism would not play a role in individuals' willingness to participate in research, and participants would trust researchers who do not look or sound like them as much as they trust those that do. Unfortunately, however, that does not seem to be the case. As discussed by Kraft and colleagues (2018), participants would feel more comfortable with researchers with whom they can identify, or as a participant described: "I think that people relate to people that look like them. They trust people that look like them more" (10). Researchers who share similar ethnic, racial, or cultural backgrounds, as potential participants, may share similar life experiences, like racism and discrimination. This common history or experiences could help those researchers better understand participants' concerns and build rapport when interacting with them. Rapport between researchers and potential participants would help build trust toward precision medicine research and likely increase participation rates among minority populations.

Not only would increasing diversity among researchers promote participation rates among minority populations, but empirical evidence demonstrates that research is of higher quality when performed by diverse groups (Campbell et al. 2013).

However, achieving the goal of more diversity among researchers may be more complex than it seems at first glance. The proportion of researchers from racial and ethnic minority groups such as African Americans, Hispanics/Latinos, and Native Americans in U.S. academic institutions is significantly less than expected based on their share of the U.S. population (National Science Foundation 2012). In order to increase diversity among researchers we must address multiple obstacles at different stages of academic and scientific career development. In the following, we discuss some of these obstacles and potential solutions to help increase the representation of underrepresented minorities (URMs) in the sciences.

[...]

A history of research misconduct against minority groups, coupled with research institutions in which potential participants from minority groups do not see themselves or their interests sufficiently represented, can help explain why it is difficult to recruit individuals from minority populations as participants in precision medicine research. How we address the lack of diversity among researchers will be key to promoting trust and participation in precision medicine research among individuals from minority populations and achieving precision medicine for all. Efforts to ensure that URM researchers who are currently in academia are involved in precision medicine research would go a long way. However, the significant disparity between the proportion of URMs in the U.S. population and scientists at research institutions conducting precision medicine research needs to be addressed. Finally, increasing diversity among researchers will enhance precision medicine research not only by increasing the participation rate of minority populations but also because the complex thought processes necessary to conduct quality science are facilitated by having researchers from diverse backgrounds (Antonio et al. 2004).

18. Miracles, Scarce Resources, and Fairness

Steve Clarke. *The American Journal of Bioethics* 18(5). Open Peer Commentaries.

Clinical bioethicists (ethicists) work for hospitals, nursing homes, rehabilitation centers, and other health care institutions (Fox, McGee and Caplan 1998). They perform various work roles, including policy development, ethics education, and consulting with patients, surrogate decision makers, families, health care professionals, and administrators about ethical issues (Chidwick et al. 2010). Bibler and colleagues (2018) offer practical suggestions to ethicists who are asked to consult patients, or their surrogate decision makers, when a particular medical intervention or course of care is requested on the grounds that this may enable a miraculous cure to take place. They focus their discussion on Christian patients and surrogates who invoke the possibility of miracles.

Bibler and colleagues (2018) argue that when ethicists counsel patients or surrogates who invoke the possibility of miracles, “the overall interests of the patient should (*ceteris paribus*) take priority” (44) in shaping the counsel provided. I do not agree that ethicists should prioritize the overall interests of particular patients when counseling those patients or their surrogate decision makers. Ethicists should be sensitive to the needs of particular patients and should demonstrate respect for the beliefs and values of patients and their

surrogates, including belief in the possibility of miraculous cures and the religious values that accompany such beliefs. However, ethicists are not advocates for particular patients. They have a professional duty to help patients and surrogates make, and accept, good all-things-considered ethical decisions. Sometimes such decisions will be ones that go against the overall interests of particular patients. This is especially likely when a proposed course of action, which is in the interests of a particular patient, will have harmful consequences for other patients.

[...]

It may be, however, that the patient's family members do not only intend to use the additional time that is being requested to pray, but also plan to undertake specific activities that, they hope, will make it more likely that divine intervention will occur. They may be planning to cast spells, offer sacrifices, recite incantations, and so on. If this is what they intend, then they are not proposing to wait and hope for a miracle. A miracle is usually understood as the consequence of a decision made by God to intervene in the natural world and is not usually understood as being subject to human prediction or control. Someone who acted in a way that, they supposed, would make it more likely for supernatural intervention in the natural world to occur than it would be otherwise, would be attempting to perform magic, rather than praying for a miracle (Clarke 1999). Even though resorting to magic is not uncommon, especially among those who are desperate to see their ill relatives cured, most religious traditions regard attempts to perform magic as incompatible with genuine religious faith (Clarke 2014). If all of this is explained to surrogates who propose to extend life support, in defiance of received medical opinion, then, after discussion with family members, those surrogates may be persuaded to reconsider their proposed course of action.

19. Conscience as a Civil and Criminal Defense

Nadia N. Sawicki. *The American Journal of Bioethics* 18(7). Open Peer Commentaries.

Prof. Nelson (2018) makes a compelling argument that conscientious objection to abortion (COTA) laws should not immunize health care providers from criminal prosecution when they deny abortions to women with life-threatening pregnancies. Conscience, he argues, should not be a defense to homicide. As a practical matter, however, only a small percentage of COTA laws protect providers from criminal prosecution—less than 15% of states have such protections. Far more concerning is that the majority of states—almost 75%—explicitly immunize providers from civil liability, even when their conduct falls below the standard of care and harms patients. That is, most COTA laws create a “conscience defense” to malpractice. And while state prosecutors may be politically disinclined to take action in such cases, patients have far greater incentive to use whatever legal tools are available to seek redress for their injuries. Thus, those advocating for limits on health care conscience laws would be better served by challenging the many state laws that protect providers from civil liability, rather than the few that protect providers from criminal prosecution.

As Prof. Nelson notes, many COTA laws offer “sweeping immunity” from adverse consequences that might result when a provider refuses to participate in abortion for reasons of conscience (Nelson 2018). I am currently conducting a nationwide

study of health care conscience laws to evaluate the types of procedural protections they offer, and my research confirms Prof. Nelson's assessment. While the protections offered vary significantly from state to state, they may include immunity from criminal prosecution, civil liability, administrative sanctions, adverse action by employers or other private entities, loss of government funding, and denial of educational opportunities, among other potential adverse consequences. However, my preliminary research indicates that immunity from civil liability is by far the most common form of procedural protection established by health care conscience laws.

[...]

There are good reasons for American law to grant some protections to health care providers whose deeply held conscientious beliefs limit the types of services that they are able to deliver. In many cases, it may be possible to respect a provider's conscientious convictions without hindering patients' rights to access high-quality medical care. But in cases of true conflict—such as when a patient seeking emergency treatment is refused care because her treating physician or hospital oppose even lifesaving abortions—the law must strike a balance. While it may not wish to compel a provider to perform a medical service that he considers unconscionable, it can demand that he face the consequences of his choice if his refusal causes patient injury that would otherwise be compensable under civil or criminal law (Sawicki 2018).

20. The “Reasonable Subject Standard” as an Alternative to the “Best Interest Standard”

Joseph Millum. *The American Journal of Bioethics* 18(8). Open Peer Commentaries.

Johan Bester makes a compelling case against the use of the harm principle to guide medical decisions on behalf of children (Bester 2018). He notes that parents have positive obligations to their children over and above refraining from harming them. Moreover, the correct decision-making standard for children must take into account the competing duties of parents and clinicians: "The obligation to do what is best for the child should be weighed against competing ethical obligations and practical constraints" (16). Bester claims that a modified best interest standard—"do the best for a child, *all things considered*"—can achieve the requisite balancing. He proposes to operationalize the standard through two questions: (1) "Can a reasonable argument be offered that the decision is best for the child, all things considered?"; (2) "Does the decision expose the child to obvious risk of harm?" If the answers are yes and no respectively, then the decision meets the standard.

In this commentary, I argue that Bester's framework for parental decision making is flawed. I propose an alternative—the "reasonable subject standard"—that can better achieve the goals that Bester endorses.

[...]

Bester is right that decisions for children are more complex than the harm principle would imply. He is also right that the standard parents use for proxy decision making must take into account other moral considerations in addition to

the child's interests. I reject only his framework for operationalizing his standard and recommend the reasonable subject standard as a workable alternative that I elaborate and defend elsewhere (Millum 2018, Chapter 6).

21. Keep It Simple

John J. Paris and Brian M. Cummings. *The American Journal of Bioethics* 18(8). Open Peer Commentary.

The multilayered "tool" proposed by the author of the target article (2018) for resolving disputes between parents and physicians over medical treatment decisions for seriously ill newborns involves not only a close parsing of the "best interest" standard, the "harm principle," and the "zone of parental discretion," but a journey through a "think list," multiple "cues," and color-coded "assists" for assigning weight "to the concerns queried." Such a Rube Goldberg formula makes one pine for the simple explanation of "best interests" found in the President's Commission report, *Deciding to Forego Life-Sustaining Treatment* (President's Commission for the Study of Ethical Problems in Medicine & Biomedical & Behavioral Research, 1983).

The President's Commission recommended that the decision maker take into account such factors as relief of suffering, the preservation or restoration of functioning, and the quality as well as the extent of life sustained. And because most people have an interest in the wellbeing of their families or close associates, the commission included consideration, from the patient's perspective, of "the impact of a decision on the patient's loved ones." The commission acknowledged there would be cases in which there is no agreement on what the incapacitated patient would select. In such instances it proposed the surrogate have discretion to choose among a range of acceptable choices.

[...]

In such a world, one might ask, how do the "toolbox," "stepladder," or "multicolored assists" proposed by the author of the target article help understand, let alone resolve, family-physician conflicts on medical treatment? The President's Commission approach provides a starting point, one that has helped numerous courts and multiple ethics committees resolve such disputes. The three-step AAP Guidelines are even more succinct and readily applicable. Keeping the standards simple is a good starting point.

22. Not a matter of parental choice but of social justice obligation: Children are owed measles vaccination

Johan C. Bester. *Bioethics* 32(9). Original Article.

Abstract: This article presents arguments that reframe the discussion on vaccination ethics. The correct starting point for discussions on vaccination ethics is not what society owes parents, but rather what society owes children. Drawing on the justice theory of Powers and Faden, two conclusions are defended by presenting and defending a set of arguments. First, a just society is obligated to

protect its children against serious vaccine-preventable diseases such as measles through adequate levels of vaccination. Second, this obligation of the just society rests on identifiable individuals and institutions: parents, healthcare professionals, government, and vaccine producers have important obligations in this regard. This removes vaccination out of the realm of individual or parental discretion, and situates it in the realm of societal obligation. Children are owed vaccination, society is obligated to provide it. If parents cannot or will not provide it, society ought to respond.

I will argue that a just society is obligated to protect its children against serious vaccine-preventable illnesses such as measles. This societal obligation rests on various identifiable individuals and institutions. These arguments reframe the ongoing discussion on vaccination ethics: they move the primary focus of vaccine ethics and policy away from a dialogue on what is owed parents to focus instead on what is owed to children. This provides grounds for considering vaccination as a matter of a societal moral obligation owed to children rather than a matter of individual or parental choice. These considerations form the moral basis for sound vaccination policy and action.

Despite the fact that vaccines are one of the most successful public health and medical interventions of the contemporary world, ethical and policy issues raised by vaccination continue to be a topic of academic and philosophical reflection. This is understandable, given the persistence of vaccine-preventable disease outbreaks, vaccine hesitancy, vaccine refusals, and various barriers to vaccination in some communities.[1]

[...]

These conclusions reframe the dialogue on measles vaccination, moving it away from a framework of what is owed to parents, and instead placing it in the framework of what is owed to children. Vaccination is a moral obligation owed to children, and not a matter of individual choice or preference. This takes vaccination out of the realm of individual discretion and into the realm of social moral obligation. These considerations have implications for social practices and policy surrounding vaccination.

2019

23. Rational Freedom and Six Mistakes of a Bioconservative

Julian Savulescu. *The American Journal of Bioethics* 19(7). Editorials.

“Life’s but a walking shadow, a poor player, that struts and frets his hour upon the stage, and then is heard no more; it is a tale told by an idiot, full of sound and fury, signifying nothing” (*Macbeth*, Act 5, Scene 5).

One of the well-worn objections in the enhancement literature is based on inequality. Enhancement will only be available to some, so it will create unjust

inequality. This was captured in the popular film *Gattaca*. In the most common form, it is based on concerns about capitalist markets: the rich will buy superior enhancements, exacerbating existing injustice. In the case of genetic enhancement, that injustice will be written into our genes. I have responded elsewhere, arguing that regulation can enable enhancement to promote justice and correct natural inequality (Savulescu [2006](#)).

[...]

Aging will make us all obsolete. In the end, we will all be dead and forgotten. Time's passage is a great equalizer. The truly significant enhancement would be immortality. But short of this, we should each just hope for the best chance of the best life.

24. The Exploitation of Professional “Guinea Pigs” in the Gig Economy: The Difficult Road From Consent to Justice

Roberto Abadie. *The American Journal of Bioethics* 19(9). Open Peer Commentaries.

Willing to endure pain, discomfort, and, mainly, boredom, professional guinea pigs perceive themselves as independent contractors, placing their bodies at the service of an industry that both exploits and dehumanizes them. With flexible schedules and mobile lifestyles, they constitute the ultimate foot soldiers of global capitalism (Abadie 2010). When I began researching this topic in the early 2000s, Uber, Lift, Airbnb, We Work, and many other forms of the gig economy had yet to arrive. In retrospect, it is clear that guinea pigs were at the forefront of this new form of exploitation; after all, “guineapigging” had been already been their gig for years, raising tough ethical questions. In *Beyond Consent*, Kahn et al. (1998), almost two decades ago, called for a focus on justice while appraising the participation of human subjects in research. Yet the normative analysis of the human subjects participating in clinical trials has been dominated by issues of consent, coercion, and undue inducement, as Millum and Garnett (2019) and Malmqvist (2019) have pointed out in this volume.

Both articles make efforts to leave this framework behind, with one making an explicit call to consider social justice aspects and the other introducing the notions of acceptable alternatives, shared goals, and ample benefits. Yet what these authors fail to consider sufficiently is how the feasibility of their proposals is shaped by political and economic forces.

[...]

While it is easy to focus on the human subject protections in isolation, one particular phase at a time, or without addressing the sociopolitical context in which they are carried out, a more comprehensive approach might be needed if we are to address social justice issues in clinical trial research (Petryna 2009; Fisher 2008). A social justice approach cannot avoid the racialized, gendered, and class-based forms of inequality that compel mostly poor, desperate people to take risks while others more fortunate can just wait until the drug comes to market. Furthermore, tackling the politics of drug development from a social justice perspective requires asking who is benefiting, who is not, and how benefits could be distributed more equitably. These questions are difficult, but they are harder to

avoid in the context of deepening social inequalities, where it is harder to pretend that the clinical trials research enterprise is a shared unmitigated social good. In turn, social inequality raises another challenge for social-justice-based redistributive policies. To succeed, they have to confront powerful economic actors that, thanks to decades of concentrated economic growth, have more access than ever to the economic and political resources to shut down any opposition.

25. Caring for Dying Children

Edwin N. Forman and Rosalind Ekman Ladd. *The American Journal of Bioethics* 19(12). Open Peer Commentaries.

Elisabeth Kübler-Ross provided an impetus for physicians and patients to begin thinking in new ways—or to begin thinking at all—about death and dying. Kübler-Ross' work led to expectations about changes and improvements in end-of-life care and consequent increased patient and family satisfaction. Have these expectations been met? Perhaps. But is it still only a myth that a) the development of palliative care has greatly improved the care of dying children and increased parent satisfaction and b) that medical education now prepares pediatricians to communicate effectively and deal with their emotions about death and dying? (Forman and Ladd 2010). Recent literature as well as personal experience indicate that the reality is different from the myths. We will review the literature and suggest some steps that could improve care.

There is evidence that often children with life threatening illness suffer unrelieved pain in their final months. Wolfe et al found that children with advanced cancer continue to experience high symptom distress. In the last 12 weeks of life, pain was reported as highly prevalent (Wolfe et al. 2015). Feudtner et al suggest that research is needed to dramatically advance the evidence base for pediatric palliative care, practices, and programs. They identified two challenges in particular: to improve symptom management and quality of life and to improve communication, elicitation of goals of care, and decision-making (Feudtner et al. 2019).

[...]

Myths are long-lived and die hard. But we need to do better to honor the legacy of Kübler-Ross.

26. Research versus practice: The dilemmas of research ethics in the era of learning health-care systems

Jan Piasecki and Vilius Dranseika. *Bioethics* 33(5). Original Article.

Abstract. In this article we attempt to answer the question of how the ethical and conceptual framework (ECF) for a learning health-care system (LHS) affects some of the main controversies in research ethics by addressing five key problems of research ethics: (a) What is the difference between practice and research? (b) What is the relationship between research ethics and clinical ethics? (c) What is the ethical relevance of the principle of clinical equipoise? (d) Does participation in

research require a higher standard of informed consent than the practice of medicine? and (e) What ethical principle should take precedence in medicine? These questions allow us to construct two opposite idealized positions on the distinction between research and practice: the *integration model* and the *segregation model* of research and practice. We then compare the ECF for an LHS with these two idealized positions. We argue that the ECF for a LHS does not, in fact, solve these problems, but that it is a third, separate position in the relationship between research ethics and clinical ethics. Moreover, we suggest that the ECF for a LHS raises new ethical problems that require additional ethical analysis and justification. Our article contributes to the discussion on the relationship between research ethics and clinical ethics, revealing that although a learning health-care system may significantly change the landscape of health care, some ethical dilemmas still require resolving on both theoretical and policy-making levels.

There is an ongoing controversy over the distinction between medical research and practice. On the one hand, the proponents of the therapeutic orientation argue that there is no morally significant difference between research and practice. Therefore, researchers have the same moral obligations in a learning health-care system (LHS) as physicians.[1] On the other hand, there are supporters of the research orientation who not only stress the differences between research and practice, but also distinguish between research and clinical moral duties.[2] The proponents of the therapeutic orientation have high hopes for the implementation of an LHS that promises to dismantle not only conceptual differences between research and practice but also the regulatory system that differentiates between these two types of activities.[3] Nevertheless, one can still ask whether the therapeutic orientation is really in harmony with the LHS. Here we want to reconstruct conceptual and ethical framework of the controversy between therapeutic and research orientations, then show how these two orientations relate to the concept of an LHS. We think that although there seems to be some concurrence of opinions between proponents of the therapeutic orientation and advocates of the LHS, this commonality is only superficial in character.

The distinction between research and practice is a focal concept for both clinical and research ethics. Both the ethical and regulatory requirements of an ethics review and informed consent hinge on this very division.[4]] The debate begins with the definitions given by the *Belmont Report*. According to this influential document, research is an activity that produces generalizable knowledge and practice aims at enhancing the well-being of an individual patient with a reasonable expectation of success.[5] However, by drawing a clear line demarcating research and practice, the National Commission shaped the whole conceptual framework of research ethics. This sharp distinction implies that the concept of therapeutic research should be discarded as it is confusing and conflates research and treatment.[6] But this is a controversial resolution, as some argue that in clinical research one tests the therapeutic character of new interventions.[7] Thus, the problem of division between research and practice leads to the question of what researcher-physicians actually do when conducting clinical research and what they should do. This problem of the demarcation between research and practice was recently addressed in the discussion over the LHS and quality improvement studies.[8] Some authors argue that the development of medicine is transforming the very nature of medical practice and in consequences the ethical and conceptual lines between research and practice

no longer exist. Therefore, one may think that the therapeutic orientation somehow harmonizes with the ethical and conceptual framework (ECF) for an LHS. [9] In order to show that this is not the case we construct and compare two competing idealized models of the relationship between research and practice in medicine: a model that segregates (the *segregation model*) and a model that integrates (the *integration model*) scientific research and practice. Elements of both models can be found in current debate over the distinction between research and practice. We analyze the theoretical and ethical assumptions built into these models and show how the ECF for an LHS could change the discussion about the relation between research and practice. We distinguish five key issues that differentiate the segregation model from the integration model and we demonstrate that the ECF for an LHS cannot be treated as a middle position between the two, but instead it creates a separate approach to the issue of research and practice distinction. This new approach can be called the *public health attitude towards research and practice*.

[...]

It can be concluded that the ECF for an LHS model of relationship has not provided us with conceptual instruments that would resolve the ethical debate between proponents of the segregation and integration models. The ECF should be considered a third proposal for characterizing the ethical problems of research. Namely, it is a proposal that navigates between research and clinical ethics and heads towards public health ethics. The ECF also does nothing to resolve the problem of a researcher's clinical obligation. Rather, it creates a new source of moral obligation: a health-care system. Next, the ECF for LHS seems not to resolve the controversy over the concept of clinical equipoise. Rather, the ECF may be interpreted as a skeptical approach to existing therapies, which aims at eliminating ineffective and inefficient treatments from a health-care system. Finally, we have argued that the ECF for an LHS attributes less significance to the standard of informed consent and that instead, it stresses the principle of beneficence over the principle of respect for autonomy. This makes the ECF for an LHS similar to the integration model, although it is at odds with the integration model's skepticism towards research. Moreover, the ECF for LHS leaves some ethical questions unanswered, such as whether there exists an obligation to participate in biomedical research. Therefore the implementation of an LHS opens a new area for ethical analysis rather than resolving the old problems of research ethics.

2020

27. "Unusual Care": Groupthink and Willful Blindness in the SUPPORT Study

George J. Annas and Catherine L. Annas. *The American Journal of Bioethics* 20(1). Open Peer Commentaries.

The SUPPORT study of extremely premature newborns seems likely to go down as one of the most controversial studies of the 21st century (SUPPORT Study Group 2010). We previously suggested that the researchers in SUPPORT were "legally blind" in failing to understand that the "standard" that defines the content of

informed consent is set by law, including the federal regulations, not by what physicians “usually” do or don’t do (Annas and Annas 2013). Macklin and Natanson, also early critics of the SUPPORT study’s failure to disclose the increased risk of death posed by the study, (Macklin et al. 2013) attack the study’s methodology itself in this issue, arguing that even on its own terms SUPPORT was fatally flawed (Macklin and Natanson 2020). Specifically, they argue that one arm of the study (the low oxygen arm) was not followed anywhere and could not be reasonably considered “standard care,” but was rather “unusual” and therefore experimental care (Cortes-Puch et al. 2016; Macklin and Natanson 2020). They also make useful suggestions about how to prevent future mischaracterizations of “usual care.”

Another central problem with SUPPORT is denial of death, illustrated by the inability of researchers, IRBs, and supporters of SUPPORT, to acknowledge the fact that the study itself could put the newborn subjects at increased risk of death. This is understandable. It is, and should be, difficult to justify risky research on newborns. Even strong supporters of SUPPORT conclude that if it had been “known before the study began [that] standard clinical care would not have encompassed the lower oxygen range … it would have been unethical to conduct the study” (Hudson et al. 2013).

[...]

It seems reasonable to conclude that SUPPORT will be classified as an ironic, rather than an iconic, study: a study whose goal was to promote evidencebased medicine in the NICU, but whose execution failed to consider contemporary evidence that one of its two arms “differed markedly from usual care” (Cortes-Puch et al. 2016). This “groupthink” result could have been avoided had the human rights and human dignity of the research subjects been taken seriously enough to identify and disclose the risk of death inherent in SUPPORT.

28. From Solo Decision Maker to Multi-Stakeholder Process: A Defense and Recommendations

David Ozar, et al. *The American Journal of Bioethics* 20(2). Open Peer Commentaries.

Berger (2020) argues effectively that “representativeness is more aptly understood as a variable that is multidimensional and continuous based on relational moral authority,” and also makes some useful suggestions about how taking this observation seriously might require changes in current patterns of practice regarding surrogates. But the essay raises additional important questions about how the Best Interest Standard (BIS) should be used among unrepresented patients and other patients as well because many surrogates besides those who “have no actionable knowledge of a patient’s preferences” find themselves in positions in which they need to determine, with the physician, what is in the patient’s best interests.

First of all, consider that much of the author’s argument is based on Pope’s recommendation that, for unrepresented patients, BIS decisions should not be made by any one person solo, but by “a robust and transparent multi-stakeholder process involving other institution-based health care professionals and extrainstitutional parties” (citing Pope 2013). Despite this recommendation being

foundational to Berger's argument, however, Berger does not provide an argument—neither Pope's nor the author's own—for the ethical correctness of this method for reaching a BIS judgment about an unrepresented patient, or about any other patient who is incapable of relevant decision making. This omission is internally problematic for the author's argument since the author's criticism of permitting "self-identified surrogates" who "have no actionable knowledge of a patient's preferences" depends on the assumption that the multi-stakeholder process is superior.

[...]

It might also be argued that the core values of the medical profession as it has been developed in our society have been sufficiently consistent from physician to physician that they can be depended on to counter the charge of arbitrariness. While that might once have been true, the history of medically related values in our society over the last four decades, and the extent to which they are or aren't reflected in a particular physician's own practice-guiding values does not encourage confidence on this matter.

29. The Healthcare Ethics Consultant-Certified Program: Fair, Feasible, and Defensible, But Neither Definitive Nor Finished

Armand H. Matheny Antommaria, et al. *The American Journal of Bioethics* 20(3). Guest Editorials.

In June 2018, the Healthcare Ethics Consultant (HCEC) Certification Commission (the Commission) began accepting applications, and since then three candidate cohorts have received the Healthcare Ethics Consultant-Certified (HEC-C) designation. While these individuals have reported favorable experiences, concerns about the HEC-C program—both the certification process and what the designation represents—exist.

As members of the Commission, we welcome dialogue about these concerns and wish to enhance understanding of the Commission's reasoning behind key aspects of the HEC-C program. This program is part of an evolving system of professional structures, for which much work remains. In developing the HEC-C program, we were guided by well-established standards. The HEC-C program is not set in stone, but instead will evolve over time. We hope that transparency will encourage productive dialogue, improving the practice of, and access to, healthcare ethics consultation.

[...]

The Commission continuously seeks to improve the HEC-C program. To this end, all comments, suggestions, and criticisms are welcome. The Commission proactively seeks feedback from individuals who pursue the HEC-C designation. Based on this ongoing dialogue and new knowledge, the Commission may change the program, for example, the initial qualifications or the requirements for recertification. The examination itself will certainly change, with new questions written and evaluated. We encourage individuals to participate in developing questions or join the Commission. We hope that these and other improvements in

the HEC-C program will be supplemented by the development of other professional structures, and that jointly these developments not only will enhance the value of certification but also will improve the quality of healthcare ethics consultation.

30. A Call for Diversity and Inclusivity in the HEC-C Program

Julie Aultman and Cynthia Pathmathasan. *The American Journal of Bioethics* 20(3). Open Peer Commentaries.

As clinical ethics continues to evolve as a part of the nation's healthcare system, there remains a disconnect between its practice and application across various healthcare organizations. This disconnect has led the American Society for Bioethics and Humanities (ASBH) to call for the standardization of clinical ethicists via the creation of the Healthcare Ethics Consultant-Certified (HEC-C) Program. Although this certification process aims to create a well-developed and equal representation of ethical knowledge, there are associated limitations with the HEC-C exam and eligibility criteria that can be addressed with opportunities for standardized training and additional assessment criteria.

From our perspectives as a clinical ethics educator and a dual enrolled 4th year medical and clinical ethics graduate student, we recently examined the benefits and burdens of the HEC-C examination, considering it as a promising opportunity for clinical ethics trainees and experts in the field. Many of these benefits and burdens were eloquently described by Horner et al. (2020) and confirmed our initial assessment of this promising certification program, in addition to a few additional concerns centering on issues of diversity and inclusion. In order for a HEC-C examination and program to be valued among professionals and trainees throughout the United States, it is important to recognize the heterogeneity of clinical ethics consultation services, hospital ethics committees, and community-based ethics consultation programs, and the diversity of providers, patients, and their families who require critical ethical deliberation and person-centered guidance. From alternative systems of ethics training and education to the current HEC-C exam, which is limited in scope and diversity in the 110-question, multiple choice options it provides, there is a need to create a training and certification program that strengthens and assesses foundational ethical concepts and guiding principles, while training clinical ethicists how to deliberate and reflect on complex, diverse ethical situations.

[...]

By expanding the HEC-C program to include mock ethics consultation cases (i.e., video simulations) and standardized patient assessments, clinical ethicists could better demonstrate their abilities to empathize and consider patient and others' values and interests, utilize their own professional experience and ethical knowledge, and be better prepared to address diverse, challenging ethical issues following certification. We have personally found this approach from an educator and a student perspective to be very beneficial for training and assessment, and a mechanism to address issues of diversity and inclusion that often are ignored or misinterpreted in standard multiple-choice questions. We are in support of standardization for clinical ethics consultations, but also argue that supporting

resources and more guidance is needed for educators and institutions (beyond the ASBH handbook and accompanying publications) to better prepare future clinical ethicists unless such training can be appropriately provided without significant financial burdens through the HEC-C program.

31. To Be Coherently Beneficent, Be Communitarian

Charles Foster. *The American Journal of Bioethics* 20(3). Open Peer Commentaries.

Bester (2020) is right to observe that “beneficence” is not self-defining: it does not determine its own substance. And there is indeed at least apparent tension between (a) the objective functioning/health of the patient and (b) the patient’s view of her own good. If beneficence is to be defined (or at least enacted) as a conversation between (a) and (b), Bester’s suggestion as to how that conversation should be conducted is as good as any and better than most.

But Bester’s analysis is tainted by the besetting sin of modern clinical ethics: he sees the patient-clinician relationship as essentially contractual. For him, there are only two people in the consulting room: the patient and the clinician. They are the two contracting parties. And usually (for Bester) there is no difficulty in identifying the patient, and hence her interests. In fact, identifying the patient is tremendously difficult. All human beings change hats, and hence preferences and interests, all the time (Foster 2019). Sometimes the patient will be primarily a mother, steered mainly by her perception of her children’s interests—which might include an interest in not burdening them with the stress and financial cost of her care. Sometimes she will be a wife, for whom the priority is to keep an ailing husband company. Sometimes she will be a tired old woman, desperate for the rest and palliation that presumably comes with death. Sometimes she will be a mentally sprightly woman, looking forward to next spring’s flowers. Sometimes she will be a religious Catholic, determined to stay alive because that’s what she thinks the Church (which is a big part of her) requires. If she is normal she will oscillate queasily between all of these states. This is one reason why informed consent, as commonly conceived, is nonsense. It is also a reason why Bester’s account does not address the human condition as it really is.

[...]

So: there is a better way of conceiving and enacting beneficence than that articulated by Bester. It entails seeing the patient as a quintessentially relational entity—defined by the network of relationships in which she exists. Seen this way, beneficence always refers to the objective good of the patient and (since it is the same thing) of the people who comprise the relevant network. This sounds tyrannous, and it may indeed lead sometimes (in the interests of the welfare both of the patient and the network members) to autonomy not having the last word. Yet this is how ethicists and lawyers traditionally and uncontroversially talk about resource allocation questions.

32. The Case Against Solicitation of Consent for Apnea Testing

Dchristie Bhagat and Ariane Lewis. *The American Journal of Bioethics* 20(6). Open Peer Commentaries

Berkowitz and Garrett (2020) provide an excellent overview of the ethical and legal discussion about the need for consent prior to apnea testing (Berkowitz and Garrett 2020). However, we disagree with their conclusions that (1) apnea testing is a medical procedure necessitating consent and (2) the right to refuse treatment is applicable to apnea testing. The brain death exam is unique from other medical testing. Requiring consent for apnea testing would obligate a false equivalence between the test and medical procedures, leading to confusion as well as inappropriate and unjust allocation of resources.

The authors defend the need for consent by citing the right to refuse medical treatment (Berkowitz and Garrett 2020). While this is an important concept, it does not pertain to this context, as apnea testing is not medical treatment; rather, apnea testing is an integral component of the evaluation to distinguish life from death, and clarify whether or not a person exists, or has ceased to exist. The intent of apnea testing is not to ameliorate brain injury, nor is there even potential for a therapeutic secondary effect of apnea testing. Furthermore, no matter how broadly one defines “medical treatment,” the term cannot apply if there is no person left to treat.

[...]

There are numerous ethical and legal considerations associated with the question of whether consent should be required prior to apnea testing. We thank Berkowitz and Garrett for their contribution to the literature on this topic (Berkowitz and Garrett 2020). We believe that universal legal clarification about both the need for consent and management of dissent is warranted to ensure practice does not vary from patient to patient, hospital to hospital and state to state. It is necessary to keep in mind the fact that apnea testing is neither akin to any other medical procedure nor to treatment. Rather, it is a unique process that clarifies whether or not a person still exists, and that this process has been proven to be low risk when guidelines are followed and prerequisites are met. As such, we disagree with the need for informed consent before apnea testing, and note that this would have profound consequences.

33. THE TRUSTED DOCTOR: MEDICAL ETHICS AND PROFESSIONALISM

Philip Charles Hébert. *Bioethics* 34(9). Book Review.

A new book by Rosamond Rhodes is an ambitious undertaking, arguing for a new approach to medical ethics. She specifically takes issue with “principlism,” “[t]he well-entrenched approach that regards medical ethics as an extension of common morality.” Rhodes asserts: “the ethics of medicine is distinct and different from the ethics of everyday life” and “the vast bioethics literature of the past 50-odd years that largely adopts that view, is simply mistaken.”

The author claims that: (a) “[d]octors need a different touchstone for their professional behavior, a theory of medical ethics that provides them with clear and reliable moral guidance,” and (b) “[t]he most popular approaches to medical

ethics were not particularly useful in resolving dilemmas in clinical practice." Her duty-based analysis, she promises, is intended to provide a better, more reliable, and insightful guide for medical professionals.

[...]

Ethical principles and the requirements of common morality, as well as Rhodes' duty-based perspective, can aid us in arriving at consensual responses to ethical issues in medicine. To paraphrase Raymond Carver, there are many paths to the waterfall.[2]

Open, Free, Safe: Choose Two

Epistemic status: More of a heuristic than an iron-clad law. Hopefully useful for decision-making, even if it's not a full gears-level model.

Related: <https://www.lesswrong.com/posts/vHSrtmr3EBohcw6t8/norms-of-membership-for-voluntary-groups>

In software engineering, there is a famous dictum which Wikipedia knows as the [Project Management Triangle](#), which takes the following form:

Good, fast, cheap. Choose two.

The essence of the triangle is to point out that while *good*, *fast*, and *cheap* are all potentially desirable traits for a project, you cannot usually (or ever) get all of them at once. If you want something good and fast, it will not be cheap; if you want something fast and cheap, it will not be good.

I have observed that a similar triangle seems to govern communities, especially online communities. I phrase my triangle in a similar fashion:

Open, free, safe. Choose two.

Let's consider what this means.

The three vertices of the triangle

These words indicate something specific in the context of online communities:

- **Open** means that anyone who wishes to may join the group. There are no requirements of expertise, professionalism, or documentation. You don't have to pay to get in. Joining does not require arcane technical knowledge beyond knowing how to visit a webpage.
- **Free** means that you are allowed to post about whatever you want, in whatever posting style suits you. A variety of interaction styles are allowed. The group may have a "topic", but the borders of that topic are broad, and off-topic chatter is tolerated.
- **Safe** means that you are unlikely to be verbally attacked, ostracised, mocked, or shunned for your behaviour. A safe community is one with low levels of conflict, or one on which conflict is carefully channeled and handled in a mild and congenial fashion. Safety means that you can let your guard down, and you can trust that other community members are not out to get you.

It should be obvious that all three of these traits are a spectrum rather than a simple binary. No group is entirely open (there's always some kind of requirement to get in, even if that requirement is "you know how to use a keyboard"), and no group is completely closed once it has more than one member. Freedom exists on a scale, as even the freest groups ban outright spam. Safety, too, is never absolute, as interpersonal conflict is always possible even if the local culture discourages it.

Case studies of each type

Open, free, not safe: 4chan. 4chan is trivial to join. You can do and say almost whatever you want, there, and behaviour which would get you kicked out of most other places on the internet is totally acceptable, even normal on 4chan. As a result, 4chan is famously, outrageously, extravagantly unsafe. Abuse, mockery, insult and every other form of verbal violence that you can imagine are typical on 4chan. This becomes the main thing that 4chan is known for, alongside being a potent fermenter of memes. (These two things are probably related.)

Open, safe, not free: Stack Overflow and the related Stack Exchange network. Stack Overflow advertises itself as a site for professional developers, but there is no one checking this, and anyone can (and does) participate. SO, however, is explicitly *not free*. There are only two supported social verbs: "ask question" and "answer question", and the site's guidelines are quite strict about this fact. Notably, the verb "have conversation" is deliberately lacking. The commenting system is the only place where something like a back-and-forth exchange is allowed, and even there both custom and policy discourage extended argument. This strictness about the kinds of allowed interactions is arguably the main thing that allows the site to work.

Of course, one of the long-running complaints about SO is that it isn't *safe enough*, particularly for new users who haven't yet learned the community customs.

Free, safe, not open: Metafilter. It's actually kind of hard to think of examples of this category, because sites which aren't open tend not to enter the public consciousness. I choose Metafilter as one of the only sites I know of which is explicitly non-open (you have to pay a fee in order to get an account), but which has cultivated a long-standing reputation as a haven of good discourse, a fact which is doubtlessly related to this fact.

But really, the best example of this third type is whatever your favourite Web 1.0 forum was. Most good forums of this type had the virtues of freedom and safety because they were implicitly closed: web access was rarer back then, signup was often slow and tedious, the ostensible topics was uninteresting, and the site itself was so obscure that most people would never find it. These barriers to entry sufficed to make the group non-open in practice, even if there was no explicit rule keeping people out. But as has been noticed elsewhere (cf.

<https://www.lesswrong.com/posts/tscc3e5eujsEeFN4/well-kept-gardens-die-by-pacifism>), when implicit barriers to entry become too low many such communities collapse because they aren't prepared to do the necessary work to maintain their borders.

Failing at all three

Getting two out of three virtues is a maximum. My thesis is that it's not possible to have all three, but it is possible to have one or none.

I was witness to the decline (from my POV---others may not have perceived these events as a decline) of a group to help new professionals that failed in just this manner. The group was deliberately non-open: you had to apply to join, and your application required you to present work of acceptable level of professionalism. As a result, the tone and quality of on-topic discussions in the group were extremely high,

and the group had a lot of pleasant socialising and off-topic discussion in their dedicated sub-fora. It was not open, but it was free and safe.

However, there was a faction within the group that decided that the forum wasn't safe enough, particularly for certain classes of people (stop me if you've heard this one before). Members of this faction first engaged in a bunch of high-energy confrontations, decreasing everyone's safety, then took those very confrontations as evidence that the group was unsafe. What followed was a slow downward ratchet in which the group became simultaneously less free and less safe, with an ever-proliferating thicket of regulations, moderators, and oversight groups, necessitated by a more and more frequent conflicts over the content of those very regulations and supposed infractions.

This should serve as a reminder that even getting two out of three is something of an accomplishment, and it's possible to get none.

The problems of big social media

Consider Twitter. Twitter is stuck between three incompatible demands:

- Twitter's value is directly proportional to the number of users that they have, so they have a financial incentive to be as open as possible.
- Moderation is expensive (even the automated kind takes time to develop and deploy), so all else equal Twitter would prefer to have a *free* social network where they don't have to monitor user behaviour.
- At the same time, their most influential users demand safety, and won't tolerate 4chan-esque pervasive conflict.

What we observe them doing to "solve" these problems is an intermittent and incoherent attempt at automated moderation, supplemented by occasional human intervention against particular high-profile accounts. These efforts cannot really succeed in the form in which Twitter has deployed them, but they do signal (sort of) that Twitter is trying to limit freedom and enforce safety. I expect this to continue for the foreseeable future, as Twitter gradually and haphazardly becomes less free in order to ensure a minimal degree of safety, but doesn't put more effort than they have to.

Every other social media site faces the same dichotomy: they have to stay open for financial reasons, but too much freedom and they come under fire for lack of safety. The counterpoint to this development is Discord, Slack, WhatsApp, Signal and the like. What these groups have in common is that they *don't* put everyone into the same social universe, but rather create a framework in which you can easily and quickly create your own private groups, organised however you want, with local control over membership and moderation. This is a re-emergence of the Web 1.0 model, in which we once again have spaces which are safe and free because they have enforceable boundaries.

Against Openness

If it wasn't already obvious, I actually don't think that the choice between vertices of this triangle is neutral. There is a correct choice, and that choice is *for freedom and*

safety, and *against* openness. Freedom and safety create great communities, while openness is at best a tax that groups must pay to avoid stagnation.

As described above, this is a choice that the big social media networks can't make, because they need to be open in order to function, and this is why I find so little value in them. The hopeful future for the future of internet communities is that the model of Discord etc. becomes the new norm.

And if you're running a site that's not on social media at all, a genuine web forum, then you should understand what your job is. Build a gate and keep it well.

But you already knew that.

I'm still mystified by the Born rule

(This post was originally intended as a comment on [Adele's question](#), but ballooned to the point where it seems worthy of a top-level post. Note that I'm not trying to answer Adele's (specific fairly-technical) question here. I consider it to be an interesting one, and I [have some guesses](#), but here I'm comentating on how some arguments mentioned within the question relate to the mysteries swirling around the Born rule.)

(Disclaimer: I wrote this post as a kind of intellectual recreation. I may not have the time and enthusiasm to engage with the comments. If you point to a gaping error in my post, I may not reply or fix it. If I think there's a gaping error in your comment, I may not point it out. You have been warned.)

My current take is that the "problem with the Born rule" is actually a handful of different questions. I've listed some below, including some info about my current status wrt each.

Q1. What hypothesis is QM?

In, eg, the theory of Solomonoff induction, a "hypothesis" is some method for generating a stream of sensory data, interpreted as a prediction of what we'll see. Suppose you know for a fact that reality is some particular state vector in some Hilbert space. How do you get out a stream of sensory data? It's easy enough to get a *single* sensory datum — sample a classical state according to the Born probabilities, sample some coordinates, pretend that there's an eyeball at those coordinates, record what it sees. But once we've done that, how do we get our next sense datum?

Or in other words, how do we "condition" a quantum state on our past observations, so that we can sample repeatedly to generate a sequence of observations suitable for linking our theories of induction with our theories of physics?

To state the obvious, a sensory stream generated by just re-sampling predicts that you're constantly teleporting through the multiverse, and a sensory stream generated by putting a delta spike on the last state you sampled and then evolving that forward for a tick will... not yield good predictions (roughly, it will randomize all momenta).

Current status: I assert that additional machinery is required to turn QM into a hypothesis in the induction-compatible sense — ie, I'd say "the Born rule is not complete (as a rule for generating a hypothesis from a quantum state)". My guess is that the missing machinery involves something roughly like sampling classical states according to the Born rule and filtering them by how easy it is to read the (remembered) sense history off of them. I suspect that a full resolution of this question requires some mastery of naturalized induction. (I have some more specific models than this that I won't get into at the moment. Also there are things to say about how this problem looks from the updateless perspective, but I also won't go into that now.)

ETA: I am not claiming to be the first person to notice this problem. As best I can tell, this problem or something close to it is what physicists refer to as the "measurement

problem". I have not seen anyone clearly frame it as a challenge of segueing a quantum state into an inductor-compatible sensory stream; I'd guess that's b/c most physicists don't (think most other physicists) natively speak inductor-tongue. I'm aware of the fact that various people have worked on the problem of identifying qualitative branches in a quantum state, and that one explicit motivation for that research is resolving this issue. @interstice [linked some below](#), thanks interstice. That's not my preferred approach. I still think that that research is cool.

Q2. Why should we believe the Born rule?

For instance, suppose my friend is about to roll a biased quantum die, why should I predict according to the Born-given probabilities?

The obvious answer is "because we checked, and that's how it is (ie, it's the simplest explanation of the observed data so far)".

I suspect this answer is correct, but I am not personally quite willing to consider the case closed on this question, for a handful of reasons:

- I'm not completely solid on how to twist QM into a full-on sensory stream (see Q1), and I suspect some devils may be lurking in the details, so I'm not yet comfortable flatly declaring "Occam's razor pins the Born rule down".
- There's an intuitive difference (that may or may not survive philosophical progress) between indexical uncertainty, empirical uncertainty, and logical uncertainty, and it's not completely obvious that I'm supposed to use induction to manage my indexical uncertainty. For example, if I have seen a million coin tosses in my past, and 2/3 of them came up heads (with no other detectable pattern), and I have a bona fide guarantee that I'm an emulation running on one of $2^{2000000}$ computers, each of which is halfway through a simulation of me living my life while two million coins get flipped (in literally all combinations), then there's some intuition that I'm supposed to predict the future coins to be unbiased, in defiance of the observed past frequency. Furthermore, there's an intuition that QM is putting us in an analogous scenario. (My current bet is that it's not, and that the aforementioned intuition is deceptive. I have models about precisely where the disanalogy is that I won't go into at the moment. The point I'm trying to make is that it's reasonable to think that the Born rule requires justification beyond 'Occam says'. See also Q4 below.)
- It's not clear to me that the traditional induction framework is going to withstand the test of time. For example, the traditional framework has trouble dealing with inductors who live inside the world and have to instantiate their hypotheses physically. And, humans sure are keen to factor their hypotheses into "a world" + "a way of generating my observations from some path through that world's history". And, the fact that QM does not naturally beget an observation stream feels like something of a hint (see Q1), and I suspect that a better theory of induction would accommodate QM in a way that the traditional theory doesn't. Will a better theory of reasoning-while-inside-the-world separate the "world" from the "location therein", rather than lumping them all into a single sensory stream? If so, might the Born rule end up on the opposite side of some relevant

chasm? I suspect not, but I have enough confusion left in this vicinity that I'm not yet comfortable closing the case.

My current status is "best guess: we believe the Born for the usual reason (ie "we checked"), with the caveat that it's not yet completely clear that the usual reason works in this situation".

Q3. But... why the Born rule in particular?

Why is the Born rule natural? In other words, from what mathematical viewpoint is this a rule so simple and elegant as to be essentially forced?

Expanding a bit, I observe that there's a sense in which discrete mathematics feels easier to many humans (see, eg, how human formalizations of continuous math often arise from taking limits or other $\epsilon\delta$ manship built atop our formalizations for discrete math). Yet, physics makes heavy use of smooth functions and differential equations. And, it seems to me like we're supposed to stare at this and learn something about which things are "simple" or "elegant" or "cheap" with respect to reality. (See also gauge theory and the sense that it is trying to teach us some lessons about symmetry, etc.)

I think that hunger-for-a-lesson is part of the "but whyyyy" that many people feel when they encounter the Born rule. Like, why are we *squaring* amplitude? What ever happened to "zero, one, or infinity"? When physics raises something to a power that's not zero, one, or infinity, there's probably some vantage point from which this is particularly forced, or simple, or elegant, and if you can find it then it can likely help you predict what sorts of other stuff you'll see.

Or to put it another way, consider the 'explanation' of the Born rule which goes "Eh, you have a complex number and you need a real number, there aren't that many ways you can do it. Your first guess might be 'take the magnitude', your second guess might be 'take the real component', your third guess might be 'multiply it by its own complex conjugate', and you'll turn out to be right on the third try. Third try isn't bad! We know it is so because we checked. What more is there to be explained?". Observe that there's a sense in which this explanation feels unconvincing — like, there are a bunch of things wrong with the objection "reality wasn't made by making a list of possible ways to get a real from a complex number and rolling a die", but there's also something to it.

My current status on this question is that it's significantly reduced — though not completely solved — by the argument in the OP (and the argument that @evhub mentions, and the ignorance+symmetry argument @Charlie Steiner mentions, which I claim all ground out in the same place). In particular, I claim that the aforementioned argument-cluster grounds out the Born rule into the inner product operator, thereby linking the apparently-out-of-the-blue 2 in the Born rule with the same 2 from "L₂ norm" and from the Pythagorean theorem. And, like, from my vantage point there still seem to be deep questions here, like "what is the nature of the connection between orthonormality and squaring", and "is the L₂ norm preferred b/c it's the only norm that's invariant under orthonormal change of basis, or is the whole idea of orthonormality somehow baking in the fact that we're going to square and sqrt everything in sight (and if so how)" etc. etc. I might be willing to consider this one

solved in my own book once I can confidently trace that particular 2 all the way back to its maker; I have not yet done so.

For the record, on the axis from "Gentlemen, that is surely true, it is absolutely paradoxical; we cannot understand it, and we don't know what it means. But we have proved it, and therefore we know it must be the truth" to... whatever the opposite of that is, I tend to find myself pretty far on the "opposite of that" end, ie, I often anticipate finding explanations for logical surprises. In this regard, I find arguments of the form "the Born rule is the only one that satisfies properties X, Y, and Z" fairly unconvincing — those feel to me like proofs that I must believe the Born rule is good, not reasons why it is good. I'm generally much more compelled by arguments of the form "if you meditate on A, B, and C you'll find that the Correct Way (tm) to visualize the x-ness of $(3x, 4y)$ is with the number $(3^{2/5})$ " or suchlike. Fortunately for me, an argument of the latter variety can often be reversed out of a proof of the former variety. I claim to have done some of that reversing in the case of the Born rule, and while I haven't fully absorbed the results yet, it seems quite plausible to me that the argument cluster named by Adele/Evan/Charlie essentially answers this third question (at least up to, say, some simpler Qs about the naturality of inner products).

Q4. wtf magical reality fluid

What the heck is up with the thing where, not only can we be happening in multiple places, but we can be happening quantitatively more in some of them?

I see this as mostly a question of anthropics, but the Born rule is definitely connected. For instance, you might wish to resolve questions of how-much-you're-happening by just counting physical copies, but this is tricky to square with the continuous distribution of QM, etc.

Some intuition that's intended to highlight the remaining confusion: suppose you watch your friend walk into a person-duplicating device. The left copy walks into the left room and grabs a candy bar. The right copy walks into the right room and is just absolutely annihilated by a tangle of whirring blades — screams echo from the chamber, blood spatters against the windows, the whole works. You blink in horror at the left clone as they exit the door eating a candy bar. "What?" they say. "Oh, that. Don't worry. There's a dial in the duplicating device that controls how happening each clone is, and the right clone was happening only negligibly — they basically weren't happening at all".

Can such a dial exist? Intuition says no. But quantum mechanics says yes! Kind of! With the glaring disanalogy that in QM, you can't watch the negligibly-happening people get ripped apart — light bouncing off of them cannot hit your retinas, or else their magical-happening-ness would be comparable to yours. Is that essential? How precisely do we go about believing that magical happening-ness dials exist but only when things are "sufficiently non-interacting"? (Where, QM reminds us, this interacting-ness is a continuous quantity that rarely if ever hits zero.) (These questions are intended to gesture at confusion, not necessarily to be answered.)

And it feels like QM is giving us a bunch of hints — ie, if physics turned out to look like a discrete state plus a discrete time evolution rule, we would have been able to say "aha, that's what happening" and feel content about it, never quite noticing our deeper confusion about this whole "happening-ness" thing. But reality's not like that. Reality is like a unit vector in an extraordinarily high-dimensional room, casting

complex-valued shadows on each wall in the room, and each wall corresponds to a way that everything can be arranged. And if we cast our gaze to the walls in accordance with the degree to which that wall is supporting the overall magnitude of the reality-vector (ie, in accordance with the shadow that the shadow-on-the-wall casts back onto reality, ie in proportion to the shadow times its conjugate, ie in proportion to the squared amplitude of the shadow) then our gaze occasionally falls on arrangements of everything that look kinda like how everything seems to be arranged. And if we cast our gaze using any other rule, we find only noise. And, like, one thing you can do is be like "haha weird" and then figure out how to generate an observation stream from it and chalk it up to "we followed Occam's razor and this is what we found". But it seems to me that this is ignoring this great big surprise that reality handed us. This is an unexpected *breed of object* for reality to be. This shadow-of-a-shadow thing feels like a surprising way for happening-ness to meta-happen. It all feels like a *hint*, a hint about how our beliefs about what the heck is going on with this whole "existence" thing are built atop false assumptions. And it's a hint that I can't yet read.

And... this is somewhat related to the beef I have with measure non-realism. Like, one thing a person can say is "everything is happening; I'm built to optimize what happens in places in accordance with how simple they are; it seems that the simplest way you find me in the logical multiverse is by flitting your gaze along those walls in accordance with the shadow-of-a-shadow and in accordance with some as-yet-unnamed rule about following coherent histories starting from the birth of a particular child; the shadow-of-a-shadow rule is elegant, ridiculously overdetermined by the data, and has no special status relative to any other part of the description of how to find me; what remains to be explained?" And... well, I'm still pretty confused about this whole "stuff is happening" thing. And I'm suspicious of a metaphysics that places physics on the same status as every other mathematical object, b/c I am not yet sure which of physics and math "comes first". And yes, that's a confused question, but that doesn't make me any less confused about the answer. And, yeah, there are deflationary measure-non-realist replies to these inarticulate gesticulations, but they leave me no less confused. And all the while, reality is sitting there having this counter-intuitive shadow-casting form, and I cannot help but wonder what false assumptions it would reveal, what mysteries it would lay bare, what lessons it would teach about which sorts of things can meta-exist at all, if only I could find my errant intuitions and put them in contact with this surprise.

And, like, there's a way in which the hypothesis "everything is; we are built to attend to the simple stuff" is a curiosity-stopper — a mental stance that, when adopted, makes it hard to mine a surprise like "reality has the quantum nature" for information about what sort of things can be.

I have a bunch more model than this, and various pet hypotheses, but ultimately my status on this one is "confused". I expect to remain confused at least until the point where I can understand all these blaring hints.

In sum, there are some ways in which I find the Born rule non-mysterious, and there are also Born-rule-related questions that I remain quite confused about.

With regards to the things I consider non-mysterious, I mostly endorse the following, with some caveats (mostly given in the Q2 section above):

The Born rule is on the same status as the Fourier transform in quantum mechanics — it's just another equation in the simple description of where to find us. It gets an undeservedly bad rep on account of being just barely on the reality-side of the weird boundary humans draw between "reality" and "my location therein" in their hypotheses, and it has become a poster-child for the counter-intuitive manner in which we are embedded in our reality. Even so, fixing the nature of the rest of reality, once one has fully comprehended the job that the Born rule does, the Born rule is the only intuitively natural tool for its job.

(And, to be clear, I've updated in favor of that last sentence in recent times, thanks in part to meditating on the cluster of arguments mentioned by Adele/Evan/Charlie.)

With regards to the remaining mystery, there is a sense in which the Born rule is the star in a question that I consider wide-open and interesting, namely "why is 'trace your eyes across these walls in accordance with the Born rule' a reasonable way for reality to be?". I suspect this question is confused, and so I don't particularly seek its answer, but I do seek mastery of it, and I continue to expect such mastery to pay dividends.

Book review: "A Thousand Brains" by Jeff Hawkins

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

Jeff Hawkins gets full credit for getting me first interested in the idea that neuroscience might lead to artificial general intelligence—an idea which gradually turned into an all-consuming hobby, and more recently a new job. I'm not alone in finding him inspiring.

Andrew Ng claimed [here](#) that Hawkins helped convince him, as a young professor, that a simple scaled-up learning algorithm could reach Artificial General Intelligence (AGI). (Ironically, Hawkins scoffs at the deep neural nets built by Ng and others—Hawkins would say: "Yes yes, a simple scaled-up learning algorithm can reach AGI, but not *that* learning algorithm!!")

Hawkins's last book was *On Intelligence* in 2004. What's he been up to since then? Well, if you don't want to spend the time reading his journal articles or [watching his research meetings on YouTube](#), good news for you—his new book, *A Thousand Brains*, is out! There's a lot of fascinating stuff here. I'm going to pick and choose a couple topics that I find especially interesting and important, but do read the book for much more that I'm not mentioning.

A grand vision of how the brain works

Many expert neuroscientists think that the brain is horrifically complicated, and we are centuries away from understanding it well enough to build AGI (i.e., computer systems that have the same kind of common-sense and flexible understanding of the world and ability to solve problems that humans do). Not Jeff Hawkins! He thinks we *can* understand the brain well enough to copy its principles into an AGI. And he doesn't think that goal is centuries away. He thinks we're most of the way there! In [an interview last year](#) he guessed that we're within 20 years of finishing the job.

The people arguing that the brain is horrifically complicated seem at first glance to have a strong case. The brain has a whopping 10^{11} neurons with 10^{14} synapses, packed full of intricate structure. [One study](#) found 180 distinct areas within the cerebral cortex. Neuroscience students pour over huge stacks of flashcards with terms like "striatum", "habenula", "stria medullaris", "fregula", and "interpeduncular nucleus". (Quiz: Which of those are real brain regions, and which are types of pasta?) Every year we get another 50,000 or so new neuroscience papers dumped into our ever-deepening ocean of knowledge about the brain, with no end in sight.

So the brain is indeed horrifically complicated. Right? Well, Jeff Hawkins and like-minded thinkers have a rebuttal, and it comes in two parts:

1. The horrific complexity of the “old brain” doesn’t count, because we don’t need it for AGI

According to Hawkins, much of the brain—including a disproportionate share of the brain's horrific complexity, like the interpeduncular nucleus I mentioned—*just doesn't count*. Yes it's complicated. But we don't care, because understanding it is not necessary for building AGI. In fact, understanding it is not even *helpful* for building AGI!

I'm talking here about the distinction between what Hawkins calls "**old brain vs new brain**". The "new brain" is the mammalian neocortex, a wrinkly sheet on that is especially enlarged in humans, wrapping around the outside of the human brain, about 2.5 mm thick and the size of a large dinner napkin (if you unwrinkled it). The "old brain" is everything else in the brain, which (says Hawkins) is more similar between mammals, reptiles, and so on.

"The neocortex is the organ of intelligence," writes Hawkins. "Almost all the capabilities we think of as intelligence—such as vision, language, music, math, science, and engineering—are created by the neocortex. When we think about something, it is mostly the neocortex doing the thinking.... If we want to understand intelligence, then we have to understand what the neocortex does and how it does it. An animal doesn't need a neocortex to live a complex life. A crocodile's brain is roughly equivalent to our brain, but without a proper neocortex. A crocodile has sophisticated behaviors, cares for its young, and knows how to navigate its environment...but nothing close to human intelligence."

I think Hawkins's **new brain / old brain discussion is bound to drive neuroscientist readers nuts**. See, for example, the paper [Your Brain Is Not An Onion With A Tiny Reptile Inside](#) for this perspective, or see the current widespread dismissal of ["triune brain theory"](#). The mammalian neocortex is in fact closely related to the "pallium" in other animals, particularly the well-developed pallium in birds and reptiles (including, yes, crocodiles!). One researcher (Tegan McCaslin) attempted a [head-to-head comparison between bird pallium and primate neocortex](#), and found that there was no obvious difference in intelligence, when you hold the number of neurons fixed. A [recent paper](#) found suggestive evidence of similar neuron-level circuitry between the bird pallium and mammalian neocortex. Granted, the neurons have a different spatial arrangement in the bird pallium vs the mammal neocortex. But it's the neuron types and connectivity that define the algorithm, not the spatial arrangement. [Paul Cisek traces the origin of the pallium](#) all the way back to the earliest proto-brains. The human neocortex indeed massively expanded relative to chimpanzees, but then again, so did the "old brain" human cerebellum and thalamus.

And what's more (these angry neuroscientists would likely continue), it's not like the neocortex works by itself. The "old brain" thalamus has just as much a claim to be involved in human intelligence, language, music, and so on as the neocortex does, and likewise with the "old brain" basal ganglia, cerebellum, and hippocampus.

OK. All this is true. But I'm going to stick my neck out and say that Hawkins is "*correct in spirit*" on this issue. And I've tried (e.g. [here](#)) to stake out a more careful and defensible claim along the same lines.

My version goes: Mammal (and lizard) brains have a "*learning subsystem*". It implements a learning algorithm that starts from scratch (analogous to random weights—so it's utterly useless to the organism at birth—see discussion of "learning-from-scratch-ism" [here](#)), but helps the organism more and more over time, as it learns. This subsystem involves the entire "telencephalon" region of the brain—namely, the neocortex (or pallium), hippocampus, amygdala, part of the basal ganglia, and a few other things (again see [here](#))—along with parts of the thalamus and cerebellum, but definitely *not*, for example, the hypothalamus or brainstem. This subsystem is not particularly "new" or peculiar to mammals; very simple versions of this subsystem date back to the earliest vertebrates, helping them learn to navigate their environment, remember where there's often food, etc. But the subsystem *is* unusually large and sophisticated in humans, and it *is* the home of human intelligence, and it does *primarily* revolve around the activities of the cortex / pallium.

So far as I can tell, my version keeps all the good ideas of Hawkins (and like-minded thinkers) intact, while avoiding the problematic parts. I'm open to feedback, of course.

2. The horrific complexity of the neocortex is in the learned content, not the learning algorithm

The second reason that making brain-like AGI is easier than it looks, according to Hawkins, is that “the neocortex looks similar everywhere”. He writes, “The complex circuitry of the neocortex looks remarkably alike in visual regions, language regions, and touch regions, [and even] across species.... There are differences. For example, some regions of the neocortex have more of certain cells and less of others, and there are some regions that have an extra cell type not found elsewhere...But overall, the variations between regions are relatively small compared to the similarities.”

How is it possible for one type of circuit to do so many things? Because it’s a learning algorithm! Different parts of the neocortex receive different types of data, and correspondingly learn different types of patterns as they develop.

Think of the [OpenAI Microscope](#) visualizations of different neurons in a deep neural net. There’s so much complexity! But no human needed to design that complexity; it was automatically discovered by the learning algorithm. The learning algorithm itself is comparatively simple—gradient descent and so on.

By the same token, a cognitive psychologist could easily spend her entire career diving into the intricacies of how an adult neocortex processes phonemes. But on Hawkins’s view, we can build brain-like AGI without doing any of that hard work. We just need to find the learning algorithm, and let ‘er rip, and it will construct the phoneme-processing machinery on its own.

Hawkins offers various pieces of evidence that the neocortex runs a single, massively-parallel, legible learning algorithm. First, as above, “the detailed circuits seen everywhere in the neocortex are remarkably similar”. Second, “the major expansion of the modern human neocortex relative to our hominid ancestors occurred rapidly in evolutionary time, just a few million years. This is probably not enough time for multiple new complex capabilities to be discovered by evolution, but it is plenty of time for evolution to make more copies of the same thing.” Third is plasticity—for example how blind people use their visual cortex for other purposes. Fourth, “our brains did not evolve to program computers or make ice cream.”

There’s a lot more evidence for and against, beyond what Hawkins talks about. (For example, [here’s](#) a very clever argument in favor that I saw just a few days ago.) I’ve written about cortical uniformity previously ([here](#), [here](#)), and plan to do a more thorough and careful job in the future. For now I’ll just say that this is certainly a hypothesis worth taking seriously, and even if it’s not *universally* accepted in neuroscience, Hawkins is by no means the only one who believes it.

3. Put them together, and you get a vision for brain-like AGI on the horizon

So if indeed we can get AGI by reverse-engineering just the neocortex (and its “helper” organs like the thalamus and hippocampus), and if the neocortex is a relatively simple, human-legible, learning algorithm, then all of the sudden it doesn’t sound so crazy for Hawkins to say that brain-like AGI is feasible, and not centuries away, but rather already starting to crystallize into view on the horizon. I found this vision intriguing when I first heard it, and after quite a bit more research and exposure to other perspectives, I still more-or-less buy into it (although as I mentioned, I’m not done studying it).

By the way, an interesting aspect of cortical uniformity is that it’s a giant puzzle piece into which we need to (and haven’t yet) fit every other aspect of human nature and psychology. There should be whole books written on this. Instead, *nothing*. For example, I have all sorts of social instincts—guilt, the desire to be popular, etc. How exactly does that work? The neocortex knows whether or not I’m popular, but it doesn’t care, because (on this view) it’s just a generic learning algorithm. The old brain cares very much whether I’m popular, but it’s too stupid to understand the world, so how would it know whether I’m popular or not? I’ve

casually speculated on this a bit (e.g. [here](#)) but it seems like a gaping hole in our understanding of the brain, and you won't find any answers in Hawkins's book ... or anywhere else as far as I know! I encourage anyone reading this to try to figure it out, or tell me if you know the answer. Thesis topic anyone?

A grand vision of how the neocortex works

For everything I've written so far, I could have written essentially the same thing about Hawkins's 2004 book. That's not new, although it remains as important and under-discussed as ever.

A big *new* part of the book is that Hawkins and collaborators now have more refined ideas about exactly what learning algorithm the neocortex is running. (Hint: it's *not* a deep convolutional neural net trained by backpropagation. Hawkins *hates* those!)

This is a big and important section of the book. I'm going to skip it. My excuse is: [I wrote a summary of an interview he did a while back](#), and that post covered more-or-less similar ground. That said, this book describes it better, including a new and helpful (albeit still a bit sketchy) discussion of learning abstract concepts.

To be clear, in case you're wondering, Hawkins does not have a complete ready-to-code algorithm for how the neocortex works. He claims to have a framework including essential ingredients that need to be present. But many details are yet to be filled in.

Does machine intelligence pose any risk for humanity?

Some people (cf. [Stuart Russell's book](#)) are concerned that the development of AGI poses a substantial risk of catastrophic accidents, up to and including human extinction. They therefore urge research into how to ensure that AIs robustly do what humans want them to do—just as Enrico Fermi invented [nuclear reactor control rods](#) before he built the first nuclear reactor.

Jeff Hawkins is having none of it. "When I read about these concerns," he says, "I feel that the arguments are being made without any understanding of what intelligence is."

Well, I'm more-or-less fully on board with Hawkins's underlying framework for thinking about the brain and neocortex and intelligence. And I *do* think that developing a neocortex-like AGI poses a serious risk of catastrophic accidents, up to and including human extinction, if we don't spend some time and effort developing new good ideas analogous to Fermi's brilliant invention of control rods.

So I guess I'm in an unusually good position to make this case!

Start with Hawkins's argument against machine intelligence being a risk

I'll start by summarizing Hawkins's argument that neocortex-like AGI does *not* pose an existential threat of catastrophic accidents. Here are what I take to be his main and best arguments:

First, Hawkins says that we'll build in safety features.

Asimov's three laws of robotics were proposed in the context of science-fiction novels and don't necessarily apply to all forms of machine intelligence. But in any product design, there are safeguards that are worth considering. They can be quite simple. For example, my car has a built-in safety system to avoid accidents. Normally, the car follows my orders, which I communicate via the accelerator and brake pedals. However, if the car detects an obstacle that I am going to hit, it will ignore my orders and apply the brakes. You could say the car is following Asimov's first and second laws, or you could say that the engineers who designed my car built in some safety features. Intelligent machines will also have built-in behaviors for safety.

Second, Hawkins says that goals and motivations are separate from intelligence. The neocortex makes a map of the world, he says. You can use a map to do good or ill, but "a map has no motivations on its own. A map will not desire to go someplace, nor will it spontaneously develop goals or ambitions. The same is true for the neocortex."

Third, Hawkins has specific disagreements with the idea of "goal misalignment". He correctly describes what that is: "This threat supposedly arises when an intelligent machine pursues a goal that is harmful to humans *and* we can't stop it. It is sometimes referred to as the "Sorcerer's Apprentice" problem.... The concern is that an intelligent machine might similarly do what we ask it to do, but when we ask the machine to stop, it sees that as an obstacle to completing the first request. The machine goes to any length to pursue the first goal....

Again, he rejects this:

The goal-misalignment threat depends on two improbabilities: first, although the intelligent machine accepts our first request, it ignores subsequent requests, and second, the intelligent machine is capable of commandeering sufficient resources to prevent all human efforts to stop it.... Intelligence is the ability to learn a model of the world. Like a map, the model can tell you how to achieve something, but on its own it has no goals or drives. We, the designers of intelligent machines, have to go out of our way to design in motivations. Why would we design a machine that accepts our first request but ignores all others after that?...The second requirement of the goal-misalignment risk is that an intelligent machine can commandeer the Earth's resources to pursue its goals, or in other ways prevent us from stopping it...To do so would require the machine to be in control of the vast majority of the world's communications, production, and transportation.... A possible way for an intelligent machine to prevent us from stopping it is blackmail. For example, if we put an intelligent machine in charge of nuclear weapons, then the machine could say "If you try to stop me, I will blow us all up."... We have similar concerns with humans. This is why no single human or entity can control the entire internet and why we require multiple people to launch a nuclear missile."

The devil is in the details

Now I don't think any of these arguments are particularly unreasonable. The common thread as I see it is, what Hawkins writes is the *start* of a plausible idea to avoid catastrophic AGI accidents. But when you think about those ideas a bit more carefully, and try to work out the details, it starts to seem much harder, and less like a slam-dunk and more like an open problem which might or might not even be solvable.

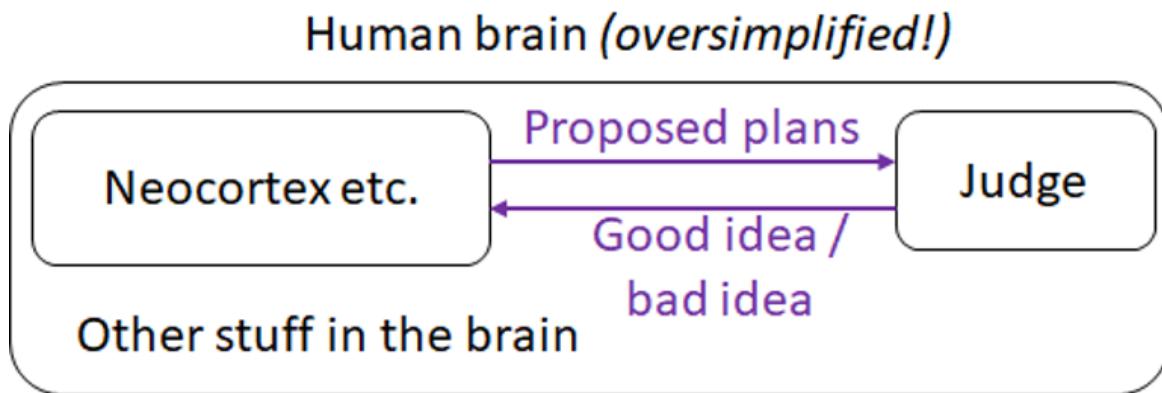
1. Goals and motivations are separate from intelligence ("The Alignment Problem")

Hawkins writes that goals and motivations are separate from intelligence. Yes! I'm totally on board with that. As stated above, I think that the neocortex (along with the thalamus etc.) is running a general-purpose learning algorithm, and the brainstem etc. is nudging it to hatch and execute plans that involve reproducing and winning allies, and nudging it to *not* hatch and execute plans that involve falling off cliffs and getting eaten by lions.

By the same token, we want and expect our intelligent machines to have goals. As Hawkins says, "We wouldn't want to send a team of robotic construction workers to Mars, only to find them lying around in the sunlight all day"! So how does that work? Here's Hawkins:

To get a sense of how this works, imagine older brain areas conversing with the neocortex. Old brain says, "I am hungry. I want food." The neocortex responds, "I looked for food and found two places nearby that had food in the past. To reach one food location, we follow a river. To reach the other, we cross an open field where some tigers live." The neocortex says these things calmly and without value. However, the older brain area associates tigers with danger. Upon hearing the word "tiger," the old brain jumps into action. It releases [cortisol]... and neuromodulators...in essence, telling the neocortex "Whatever you were just thinking, DON'T do that."

When I put that description into a diagram, I wind up with something like this:



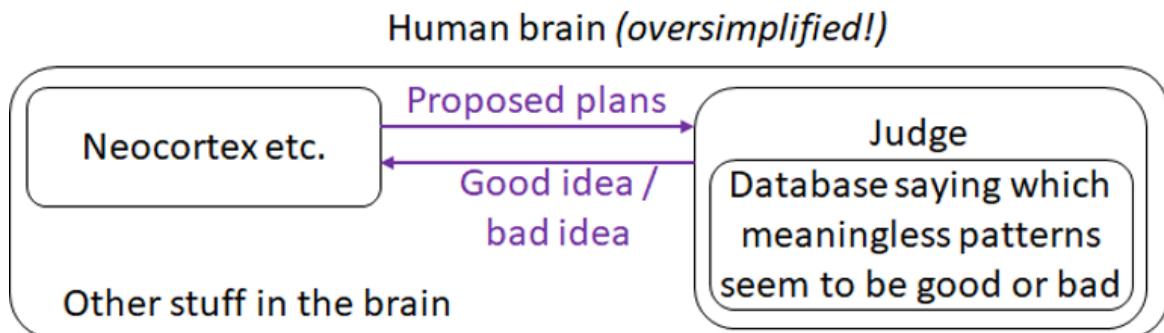
My attempt to depict goals and motivation, as described by Hawkins via his tiger example above. The box on the left has the learning algorithm (neocortex, thalamus, etc.) The box on the right is the Old Brain module that, for example, associates tigers with danger. (For my part, [I would draw the boundaries slightly differently, and put things into the terminology of reinforcement learning](#), but I'm trying to stick closely to the book here.)

The neocortex proposes ideas, and the Judge (in the "old brain") judges those ideas to be good or bad.

This is a good start. I can certainly imagine building an intelligent goal-seeking machine along these lines. But *the devil is in the details!* Specifically: **Exactly what algorithm do we put into the "Judge" box?** Let's think it through.

First things first, we should not generally expect the "Judge" to be an intelligent machine that understands the world. Otherwise, *that* neocortex-like machine would need *its own* motivation, and we're right back to where we started! So I'm going to suppose that the Judge box will house a relatively simple algorithm written by humans. So exactly what do you put in there to make the robot want to build the infrastructure for a Mars colony? That's an open question.

Second, given that the Judge box is relatively stupid, it needs to do a lot of memorization of the form “*this meaningless pattern of neocortical activity is good, and this meaningless pattern of neocortical activity is bad*”, without having a *clue* what those patterns actually mean. Why? Because otherwise the neocortex would have an awfully hard time coming up with intelligent instrumental subgoals on its way to satisfying its actual goals. Let’s say we have an intelligent robot trying to build the infrastructure for a Mars colony. It needs to build an oxygen-converting machine, which requires a gear, which requires a lubricant, and there isn’t any, so it needs to brainstorm. As the robot’s artificial neocortex brainstorms about the lubricant, its Judge needs to declare that some of the brainstormed plans are good (i.e., the ones that plausibly lead to finding a lubricant), while others are bad. But the Judge is too dumb to know what a lubricant is. The solution is a kind of back-chaining mechanism. The Judge starts out knowing that the Mars colony is good (How? I don’t know! See above.). Then the neocortex envisages a plan where an oxygen machine helps enable the Mars colony, and the Judge sees this plan and memorizes that the “oxygen machine” pattern in the neocortex is probably good too, and so on. The human brain has exactly this kind of mechanism, I believe, and I think that it’s implemented in the basal ganglia. (Update: I now think it’s not just the basal ganglia, see [here](#).) It seems like a necessary design feature, I’ve never heard Hawkins say that there’s anything problematic or risky about this mechanism, so I’m going to assume that the Judge box will involve this kind of database mechanism.



Modified version of the motivation installation system. The database—which I believe is implemented in the basal ganglia—is essential for the machine to pursue “instrumental subgoals”, like “trying” to design a lubricant without the machine needing to constantly have in mind the entire chain of logic for why it’s doing so, i.e. that the lubricant is needed for the gear which is needed for the machine which is...etc. etc. Again, for my own purposes [I would draw it a bit differently and use reinforcement learning terminology](#), but I’m trying to stay close to what’s in the book.

Now given all that, we have two opportunities for “goal misalignment” to happen:

Outer misalignment: The algorithm that we put into the Judge box might not exactly reflect the thing that we want the algorithm to do. For example, let’s say I set up a machine intelligence to be the CEO of a company. This being America, my shareholders immediately launch a lawsuit that says that I am in violation of my fiduciary duty unless the Judge box is set to “Higher share price is good, lower share price is bad,” and nothing else. With lawyers breathing down my neck, I reluctantly do so. The machine is not *that* smart or powerful, what’s the worst that could happen? The results are quite promising for a while, as the algorithm makes good business decisions. But meanwhile, over a year or two, the algorithm keeps learning and getting smarter, and behind my back it is *also* surreptitiously figuring out how to hack into the stock exchange to set its share price to infinity, and it’s working to prevent anyone from restoring the computer systems after it does that, by secretly self-replicating around the internet, and earning money to hire freelancers for strange little jobs that involve receiving packages and mixing chemicals and mailing them off, unknowingly

engineering a new pandemic virus, and meanwhile the algorithm is also quietly hacking into military robotics systems so that it will be ready to hunt down the few survivors of the plague, and spreading disinformation so that nobody knows what the heck is happening even as it's happening, etc. etc. I could go on all day but you get the idea. OK, maybe you'll say "anyone could have seen that coming, *obviously* maximizing stock price is a dumb and dangerous goal". So what goal should we use instead, and how do we write that code? Let's figure it out! And by the way, even if we have a concrete and non-problematic idea of what the goal is, remember that the Judge box is stupid and doesn't understand the world, and therefore the code that we write into the Judge box will presumably be a simplistic approximation of the goal we really want. And unfortunately, [seeking a simplistic approximation of a goal looks very different from seeking the actual goal.](#)

Inner misalignment: The assigned values in the database of meaningless (to the Judge) memorized patterns could diverge from how the Judge algorithm would judge their consequences if it actually saw them implemented in the real world. I don't have to look far for an example of this: *Look at Hawkins himself!* He has a neocortex, and he has an "old brain" putting goals and motivations into his mind, and he just *hates* it! *His book has a whole subsection called "How the neocortex can thwart the old brain"!* (And to be clear, thwarting the old brain is portrayed as a very good idea that he endorses.) I find it remarkable that Hawkins can gleefully plan to thwart his own "old brain", while at the same time being baffled at the idea that anything might go wrong when we put old-brain-like motivation systems into our own intelligent machines.

Not that things *definitely will* go wrong; it's just that avoiding these problems does not seem straightforward, and we shouldn't declare that this is a solvable problem until we have a better idea of what the solution is.

So again, *exactly* what code do we put into the "Judge" box such that the resulting system is definitely motivated to build the infrastructure for a Mars colony (while not trying to manipulate its own motivation system, prevent its goals from being edited, and so on)? You won't find the answer in Hawkins's book. In fact, it's a wide-open problem, and I may well be literally the only person on Earth who is actively working on it in the specific context of neocortex-like machine intelligence. I sure don't know the answer, or indeed whether an answer exists. Reach out if you have ideas or want to collaborate!

2. "Safety features"

Now let's move on to Hawkins's "safety features", like Asimov's first two laws or something vaguely analogous. Great idea! I'm all for it! We should definitely do that! But once again, the devil is in the details!

Let's say we want to put in a safety interlock that prevents the AGI from self-replicating. How *exactly* do we do that? What code do we write?

When I think about it, I immediately hit a similar problem as before. The safety interlock code is probably *not* going to be a neocortex-like intelligent submodule that understands the world—because if it is, we're back to the problem of installing the right motivation. Instead maybe we'll go with human-written code, which implies it's a "stupid" module that does *not* understand the world. For example, to avoid self-replication, maybe we'll add a module that detects when a copy-paste operation is happening to the source code, and block it. Unfortunately, if an intelligent robot is motivated to self-replicate, and they notice that there is a module preventing them from doing so, then they will naturally start trying to undermine, outwit, or disable that module. And remember, the robot is a lot more intelligent than the module!

By all means let's put in such a module anyway. It seems especially helpful in "early childhood" when the machine is not yet very intelligent, and still messing around, and we don't want it to do anything dangerous by accident. We should just recognize that it's

unlikely to keep working when the machine becomes highly intelligent, unless we have *both* a safety interlock *and* a carefully-sculpted motivation system that makes the machine *like and endorse* that safety interlock. If we do it right, then the machine will even go out of its way to repair the safety interlock if it breaks! And how do we do that? Now we're back to the open problem of installing motivations, discussed above.

The other option is to design a safety interlock that is *absolutely perfectly rock-solid air-tight*, such that it cannot be broken even by a highly intelligent machine trying its best to break it. A fun example is [Appendix C of this paper](#) by Marcus Hutter and colleagues, where they propose to keep an intelligent machine from interacting with the world except through certain channels. They have a plan, and it's *hilariously awesome*: it involves multiple stages of air-tight boxes, Faraday cages, laser interlocks, and so on, which could be (and absolutely should be) incorporated into a big-budget diamond heist movie starring Tom Cruise. OK sure, that could work! Let's keep brainstorming! But let's *not* talk about "safety features" for machine intelligence as if it's the same kind of thing as an automatic braking system.

3. Instrumental convergence

Hawkins suggests that a machine will want to self-replicate if (and *only if*) we deliberately program it to want to self-replicate, and likewise that a machine will "accept our first request but ignore all others after that" if (and *only if*) we deliberately program it to accept our first request but ignore all others after that. (That would still leave the vexing problem of troublemakers deliberately putting dangerous motivations into AGIs, but let's optimistically set that aside.)

...If only it were that easy!

["Instrumental convergence"](#) is the insight (generally credited to Steve Omohundro) that lots of seemingly-innocuous goals *incidentally* lead to dangerous motivations like self-preservation, self-replication, and goal-preservation.

Stuart Russell's famous example is asking a robot to fetch some coffee. Let's say we solve the motivation problem (above) and actually get the robot to *want* to fetch the coffee, and to want absolutely nothing else in the world (for the sake of argument, but I'll get back to this). Well, what does that entail? What should we expect?

Let's say I go to issue a new command to this robot ("fetch the tea instead"), before the robot has actually fetched the coffee. The robot sees me coming and knows what I'm going to do. Its neocortex module imagines the upcoming chain of events: it will receive my new command, and then all of the sudden it will only want to fetch tea, and it will never fetch the coffee. The Judge watches this imagined chain of events and—just like the tiger example quoted above—the judge will say "Whatever you were just thinking, DON'T do that!" Remember, the Judge hasn't been reprogrammed yet! So it is still voting for neocortical plans-of-action based on whether the coffee winds up getting fetched. So that's no good. The neocortex goes right back to the drawing board. Hey, here's an idea, if I shut off my audio input, then I won't hear the new command, and I *will* fetch coffee. "Hey, now *that's* a good plan," says the Judge. "With *that* plan, the coffee will get fetched! Approved!" And so that's what the robot does.

Similar considerations show that intelligent machines may well try to stay alive, self-replicate, increase their intelligence, and so on, without anyone "going out of their way" to install those things as goals. A better perspective is that if we want our machines to have any goals at all, we have to "go out of our way" to *prevent* these problematic motivations—and how to do so reliably is an open problem.

Now, you ask, why would anyone do something so stupid as to give a robot a maniacal, all-encompassing, ultimate goal of fetching the coffee? Shouldn't we give it a more nuanced and inclusive goal, like "fetch the coffee unless I tell you otherwise", "fetch the coffee while

respecting human values and following the law and so on" or more simply "Always try to do the things that I, the programmer, want you to do"?

Yes! Yes they absolutely should! But yet again, the devil is in the details! As above, installing a motivation is in general an unsolved problem. It may not wind up being possible to install a complex motivation with *surgical precision*; installing a goal may wind up being a sloppy, gradual, error-prone process. If "most" generic motivations lead to dangerous things like goal-preservation and self-replication, and if installing motivations into machine intelligence is a sloppy, gradual, error-prone process, then we should be awfully concerned that even skillful and well-intentioned people will sometimes wind up making a machine that will take actions to preserve its goals and self-replicate around the internet to prevent itself from being erased.

How do we avoid that? Besides what I mentioned above (figure out a safe goal to install and a method for installing it with surgical precision), there is also interesting ongoing work searching for ways to generally prevent systems from developing these instrumental goals ([example](#)). It would be awesome to figure out how to apply those ideas to neocortex-like machine intelligence. Let's figure it out, hammer out the details, and *then* we can go build those intelligent machines with a clear conscience!

Summary

I found this book thought-provoking and well worth reading. Even when Hawkins is wrong in little details—like whether the “new brain” is “newer” than the “old brain”, or whether a deep neural net image classifier can learn a new image class without being retrained from scratch (I guess he hasn’t heard of fine-tuning?)—I think he often winds up profoundly right about the big picture. Except for the “risks of machine intelligence” chapter, of course...

Anyway, I for one thank Jeff Hawkins for inspiring me to do the research I’m doing, and I hope that he spends more time applying his formidable intellect to the problem of how *exactly* to install goals and motivations in the intelligent machines he aims to create—including complex motivations like “build the infrastructure for a Mars colony”. I encourage everyone else to think about it too! And reach out to me if you want to brainstorm together! Because I sure don’t know the answers here, and if he’s right, the clock is ticking...

Quotes from the WWMoR Podcast Episode with Eliezer

Spoiler warning: This post contains full spoilers for Harry Potter and the Methods of Rationality.

I listened to [this](#) WWMoR podcast episode in which Eliezer had a guest appearance. I didn't see a transcript for the podcast, but found some of his replies interesting, so here is my attempt at a partial transcript.

All quotes are by Eliezer (EY), and I tried to quote him verbatim. I can't guarantee perfect accuracy (nor spelling or punctuation, for that matter), so assume that any errors are my own. That said, I provide timestamps so you can check the context for yourself if so desired.

Quotes and Excerpts

13:00 (= at the 13-minute mark)

If you'd been meant to learn from Quirrel, you would have seen Quirrel learning.

19:00

By the end of the story, Harry has taken off the Hero mask and given it to Hermione, and put on Dumbledore's mask instead.

26:00, regarding the plausibility of the scene when Harry and Quirrel escape from Azkaban on a contraption consisting of a broomstick plus a rocket:

Things out here in reality are just so much gratuitously worse than you would imagine them going if you were just going to be realistically pessimistic in a story.
[...]

When I got to that part of the story, I realized that if you take a rocket and you glue it to a broomstick, you are just inevitably going to die. Like even for a story, that was too much. I couldn't make myself believe it long enough to write it down. So I had Quirrel wake up and apply a Charm of Flawless Function to the rocket and attach it to the broomstick using an actual proper spell, and then I could believe it. I could believe that would work inside a story.

Now in real life, you know, the rocket drifts off course and crashes into the walls and they all die [...].

29:00, on how the story was planned:

Primarily what was going on was that there were scenes that I was using as the anchor point, and then, knowing that these things would happen later, I could make the rest of the story be made entirely out of foreshadowing of them.

46:18

All the characters are made out of pieces of me. [...] The particular way in which Harry is made out of me is something like 18-year-old Eliezer with his wisdom and constitution scores swapped and all the brakes removed.

49:08, on Harry's and Quirrel's apparent hypercompetence:

I do not make my stories out of tropes. I make them out of subverted tropes. So there is certainly a sense in which you have the hypercompetent character. But Harry is, if anything, a subversion of that. The rest of Hogwarts thinks he can do anything, but we are watching him from the inside, watching how he is faking all of it. [...]

Quirrel is obviously - says the author, whose job it was to make this obvious and may not have done that [well enough] - Quirrel is obviously pulling the same stunt as Harry from a different viewpoint. We just don't get his viewpoint.

53:59, remarking on Harry's guess before the Azkaban arc that they were going to break a Black out of Azkaban, which was correct by accident:

Harry is like [...] 'I don't care how wrong I was, I am taking that secret to the grave I will never occupy'.

1:05:20

I feel like the largest literary flaw [in the story] is that the grand climax of the story is Harry solving what I would later call a Level 2 Intelligent Character puzzle, which is sort of like a Munchkin puzzle. The Final Exam is like 'Assemble these facts from inside the story and come up with a creative use for them.', and it's not a final challenge that holds up the thematic weight of the rest of the book. [...] It's like a thing of cleverness where the solution doesn't really have the depth that I learned to write in the rest of the story.

And that was an example of a flaw that just could not be fixed because of the number of open parentheses that had been set up and the amount of foreshadowing done going literally back to the first sentence of the book, pinpointing that exact puzzle and that exact solution. By the time I got there and could sort of see the way in which it wasn't adequate, the structure of the book was woven together so tightly that there was absolutely no way to change it.

1:12:45, on whether specific characters in HPMoR were written to be liked or disliked:

I think there's some kind of whole judgy thing - 'Whose side are you on?' - like, [in reference to the literary concept of Death of the Author which was previously discussed on the podcast] don't kill the readers completely, but I feel like that part of the readers could afford to die or something...

This used to be this old tradition of you created a literary artifact and it would stand there being what it was, and now people have Twitter and identity politics and they think they're supposed to take sides... It's just not the way I was raised to write things.

Around 1:25:00, EY points out that Quirrel's biggest error during the Final Exam was bringing in 36 Death Eaters after not seeing them for ten years, something a real

Quirrel would never do. Whereas letting Harry keep his wand only turns out badly because Harry uses wordless magic powerful enough to defeat Quirrel, which is not something he expects, being the more powerful wizard in this equation.

1:32:18, on literary themes involving Hermione, including Mary Sue:

The real problem with Mary Sue is not having teeth made of unicorn horn, it's whether you take over the story. Other themes in Hermione include Hermione representing the plight of the secondary character in fan fiction which she and Harry are both aware of and the rest of Hogwarts is determined to force her into that mold. It's commentary on all the poor secondary characters in fan fiction who get shoved off into somebody's harem or something. There's also Hermione being part of the exchange of masks [= literary roles] where she gets the Hero mask from Harry and Harry gets the Old Wizard mask from Dumbledore.

Nobody did a literary analysis featuring any of this stuff, and therefore I reiterate that all literary analysis everywhere must be bogus.

Covid 3/12: New CDC Guidelines Available

This post is a day late although not all that short, as I warned it might be. This is because I have spent the last week visiting the best place on Earth, my true home, which is New York City. I will return again soon, and soon after that I will once again be able to live there. A great moment. I hope to write about the trip, but I need to get this post out quickly, so that won't make it in this week.

This week had three (other) big discrete things happen. The CDC issued unexpectedly sane guidelines for vaccinated people, and thanks to the day of delay, we also have Biden announcing a date that everyone will be eligible for the vaccine. Finally, on March 7 the Covid Tracking Project stopped collecting data, requiring another phase transition in much of the data and leaving me without a source for detailed positive test rate data that I'm happy about.

We also had a bunch of opening up around the country, despite what is obviously about to happen.

Let's run the numbers.

The Numbers

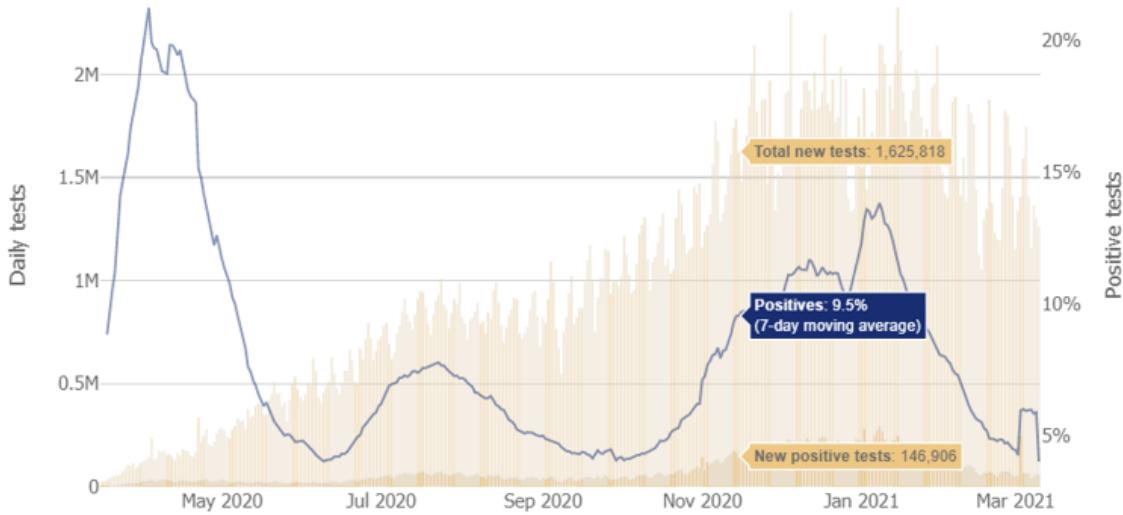
Predictions

Last week: 4.2% positive test rate and an average of 1,827 deaths after subtracting the California bump, using Covid Tracking Project's final week of data.

Last week's prediction: No prediction due to some combination of 'somehow I forgot to do this' and the expected lack of data collection making it difficult to fairly evaluate the prediction. We'll start again fresh now.

(My (highly unreliable guess you should not trust to match what I would have said) would likely have been to see small declines, to something like 3.9% positive rate and 1,650 deaths, which absolutely does not count for anything as a prediction but does give a sense of vaguely where expectations were a week ago.)

[This is the testing trends chart from Johns Hopkins](#), a plausible new data source.



I hate when I want charts and all I get are graphs, and also note that phase jump up and down, and also when you highlight a day *it doesn't tell you what date that is* so unless you can work backwards from the final day (and this doesn't make it that easy to know what the final day is, if it might or might not have updated yet) everything is super fuzzy.

The phase jump here appears in their data in some states and not in others. It's clearly not 'real.' For now, I'm going to presume that the new 4.0% rate at the end is now calibrated however the old numbers were calibrated at the Covid Tracking Project, and ignore the numbers in-between.

Result: Thus, I will conclude that likely the positive test rate did indeed fall slightly, from 4.1% to 4.0%. Deaths fell from 1,867 to 1,374 (!).

Another possibility [is the Washington Post](#), if you're able to reliably check it on the right day before it updates:

In the past week in the U.S....

New daily reported **cases fell 11.3% ↓**

New daily reported **deaths fell 16.7% ↓**

Covid-related **hospitalizations fell 8.5% ↓** [Read more](#)

Among reported tests, **the positivity rate was 4.2%**.

The **number of tests reported fell 19.5% ↓** from the previous week. [Read more](#)

Since Dec. 14, more than **98,203,000 doses of a covid-19 vaccine have been administered** in the U.S.

More than **33,863,000 people have completed vaccination**, or about **10.2%** of the population. [Read more in our vaccination tracker.](#)

Those are the headline numbers we need in easy to see form, so it makes sense to predict on those in relative terms. Note the 19.5% decline in the number of tests, which indicates that cases falling by 11.3% should be concerning.

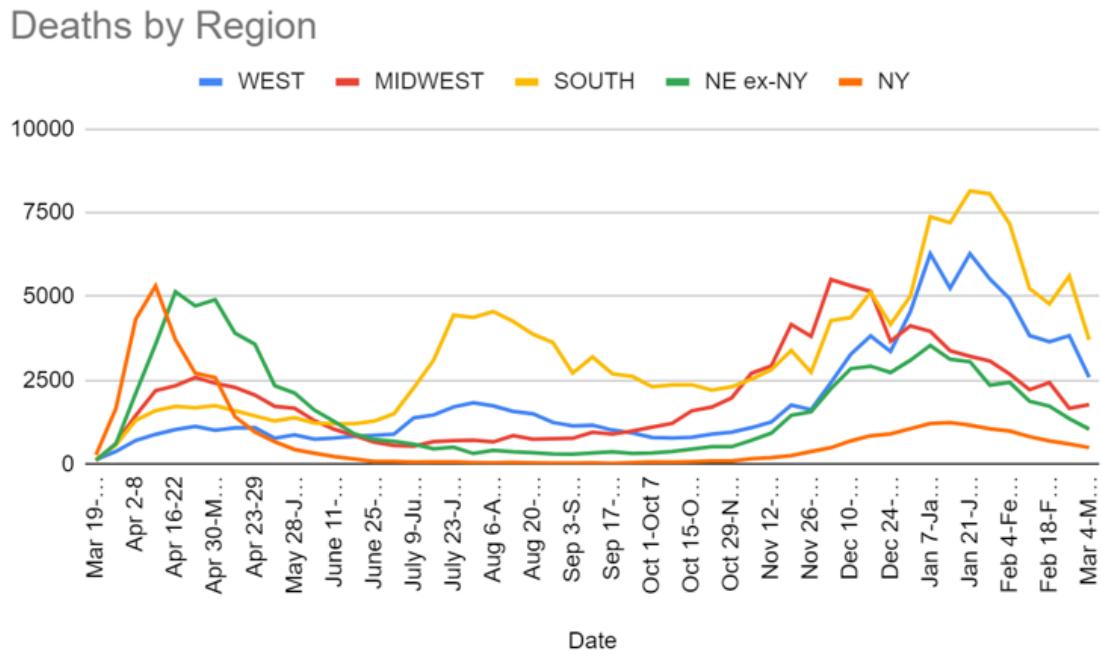
Prediction (WaPo numbers): Positivity rate will be 4.2% (unchanged) and deaths will fall by 12%.

Deaths should continue to fall since they lag substantially. Cases could go either way, depending on the impact of the new strains and how people react to reopenings.

The search for a better data source continues. Wikipedia is still good for raw positive test numbers and for deaths.

Alas, [the Covid Machine Learning project from Youyang Gu is also wrapping up](#) due to the Covid Tracking Project no longer gathering data. I hope someone gets us a new version with a new data source, but Gu has already gone above and beyond.

Deaths

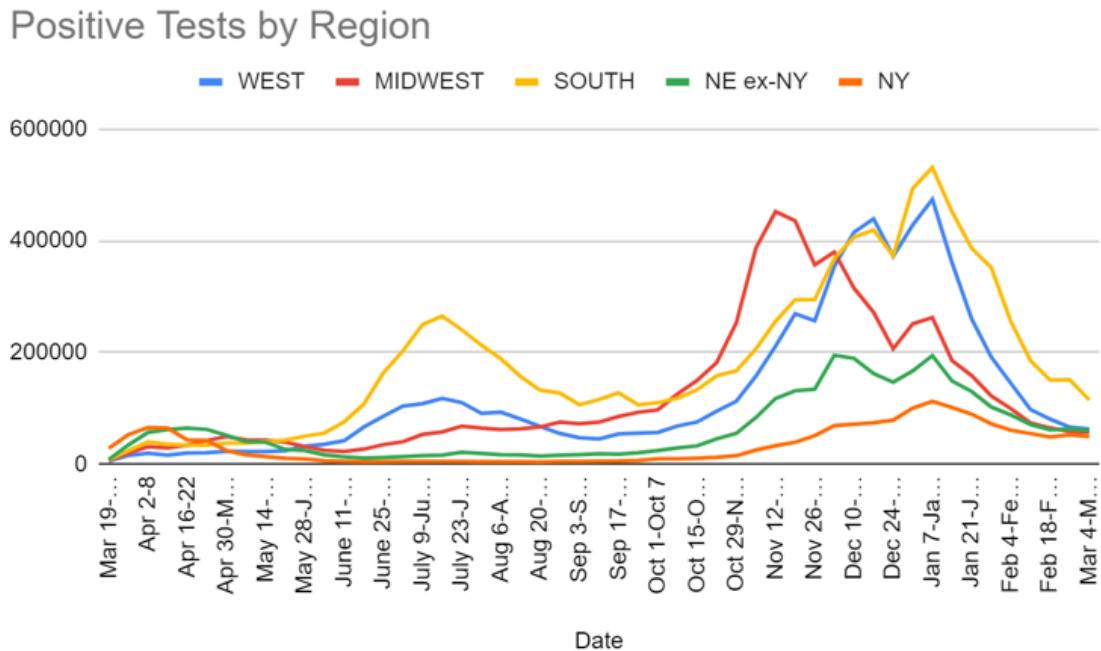


Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jan 21-Jan 27	6281	3217	8151	4222	21871
Jan 28-Feb 3	5524	3078	8071	3410	20083
Feb 4-Feb 10	4937	2687	7165	3429	18218
Feb 11-Feb 17	3837	2221	5239	2700	13997
Feb 18-Feb 24	3652	2433	4782	2427	13294
Feb 25-Mar 3	3834	1669	5610	1958	13071
Mar 4-Mar 10	2595	1775	3714	1539	9623

This is clearly wonderful news and deaths are rapidly on the decline. I worry there's data artifacts here because the death counts on March 7 and March 8 are so low (e.g. 562 on

March 8, and before the 7th the last day under 1000 deaths was November 29). For a while it's been an increasingly large mystery why deaths haven't fallen faster, and now they're falling an amount that reasonably tracks declines in cases. Whew.

Positive Tests



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Jan 28-Feb 3	191,804	122,259	352,018	174,569
Feb 4-Feb 10	144,902	99,451	255,256	149,063
Feb 11-Feb 17	97,894	73,713	185,765	125,773
Feb 18-Feb 24	80,625	64,857	150,493	110,339
Feb 25-Mar 3	66,151	58,295	151,253	115,426
Mar 4-Mar 10	62,935	57,262	114,830	109,916

This is disappointing news, as is the positive test rate. The decline in the South is impressive, but the other regions are stalling out, and the decline in the South both likely reflects a (slightly) artificially high number last week, and conditions that have since loosened considerably in major areas including Texas. Our march straight down to zero will have to wait.

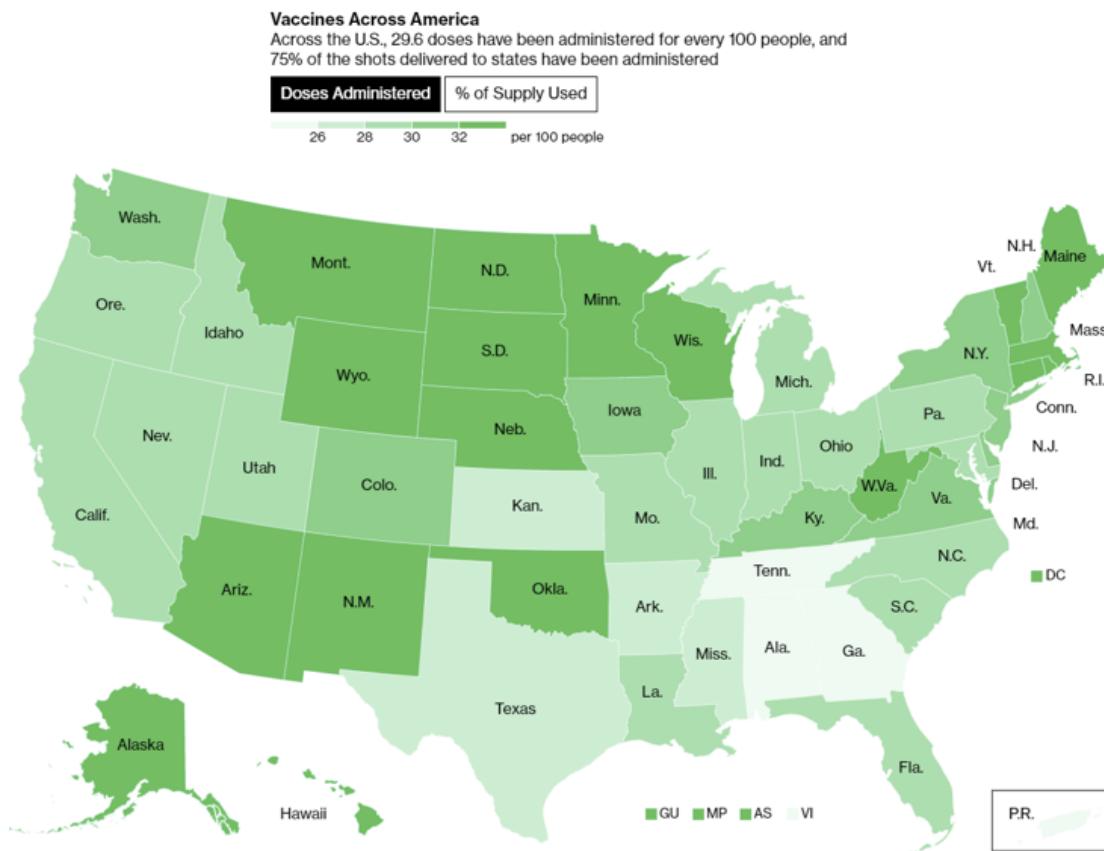
The next few weeks on this chart will be the moment of truth. If cases don't pick up by the end of March, they're likely not going to pick up at all from the current wave of variants, and vaccinations will have enough time to dominate. If cases do pick up, it's going to be very difficult to pivot quickly.

Vaccinations

(Data here is as of March 12 rather than March 11, so it's 8 days after last time.)

The biggest vaccination campaign in history is underway. More than **334 million doses** have been administered across 121 countries, according to data collected by Bloomberg. The latest rate was roughly **8.41 million doses a day**.

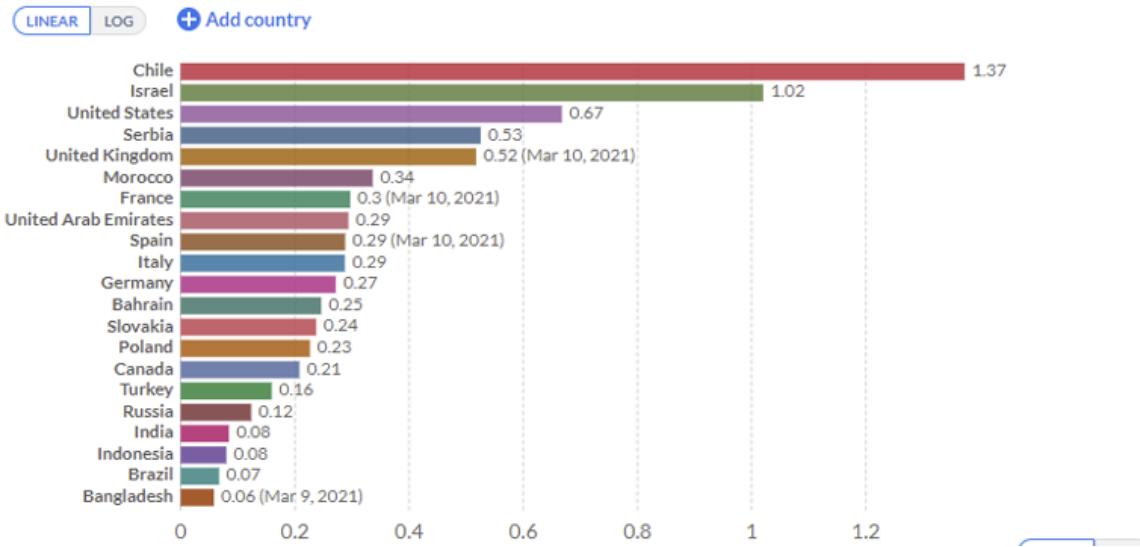
In the U.S., more Americans have received at least one dose than have tested positive for the virus since the pandemic began. So far, **98.2 million doses** have been given. In the last week, an average of **2.23 million doses per day** were administered.



Daily COVID-19 vaccine doses administered per 100 people, Mar 11, 2021

Shown is the rolling 7-day average per 100 people in the total population. This is counted as a single dose, and may not equal the total number of people vaccinated, depending on the specific dose regime (e.g. people receive multiple doses).

Our World
in Data



Not only is it clear we can sustain and further increase this pace of vaccinations, we are building up an increasing surplus of vaccine doses, and getting appointments is becoming steadily easier in most places.

We had (and continue to have) a ton of unforced errors along the way that caused (and continue to cause) massive delays, but we are on a clear path to vaccinations on demand for every adult within a few months, [and yesterday Biden made that official](#) (WaPo). Every state has been directed to make the vaccines available to everyone over the age of 18 no later than May 1, and [Alaska has already gone first and opened up vaccinations to all adults](#).

Biden delivered the announcement and the rest of his speech well, highlighting that while not everyone will be able to *get vaccinated* on May 1 or that soon after May 1, at least everyone can *get in line* on May 1, and emphasizing the need for basic safety measures for now. Needless to say, there was no discussion of cost/benefit, or why we did something very different from this earlier.

Meanwhile, Biden continues to double down on underpromising to maximize the chances of being able to claim overdelivery on all fronts. Meeting one's 100-day vaccination goal in 60 days without anything unexpected happening that much impacted the pace is more of a sign that you set a very low bar than it is a sign that you went above and beyond.

It is also important to note that we have a strange two-tiered prioritization system. If you are *actually high priority*, mostly via being elderly, you can get a vaccine appointment (in at least many places) at pharmacies without competing against people who checked a box saying they once smoked a few cigarettes. If you are *technically eligible*, you can use one of the less convenient, harder to book vaccination sites, or go overnight for Johnson & Johnson.

Compared to most plausible alternatives, this is all actually pretty great. We're not allocating by price in dollars, but we're allocating by price *at least a little*. Rather than be obsessed with exactly what order people get shots in, we make it (relatively) easy to get a shot, and get it safely, if you're at high risk, and charge a fee in annoyance to those not at as high a risk who want a relatively early shot. So those who are actually high-risk in a less legible way, or who highly value the shot, can mostly get one, and those who are mostly indifferent can wait

while others pave the way. Best of all, the annoyance of going to a worse vaccination site is a built-in cost rather than a wasteful tax, so it's even efficient. Bravo, I suppose.

That doesn't mean the system doesn't sometimes fail people when they need it most. [It absolutely does](#):



Mason 🚶‍♂️🌐 @webdevMason · 11h

My >90 year old grandfather still hasn't been able to get a vaccine in Oregon. He's apparently had two appointments but they were out of vaccine by the time he arrived to both. Now he's in a lottery. This is ridiculous.

12

14

151



Mason 🚶‍♂️🌐 @webdevMason · 11h

He's close to the southern border of Washington state. If anyone has a lead on vaccines in northern Oregon or southern Washington, please let me know. DMs are open.

I do think that particular case is mostly extraordinarily poor luck, but it still happens. Presumably in this particular case help is already on the way, but likely still worth DMing her if you have a lead.

Then again, [remember it could always be worse](#), if we grade on a curve we're killing it here in the good old USA:



Alex Tabarrok
@ATabarrok

...

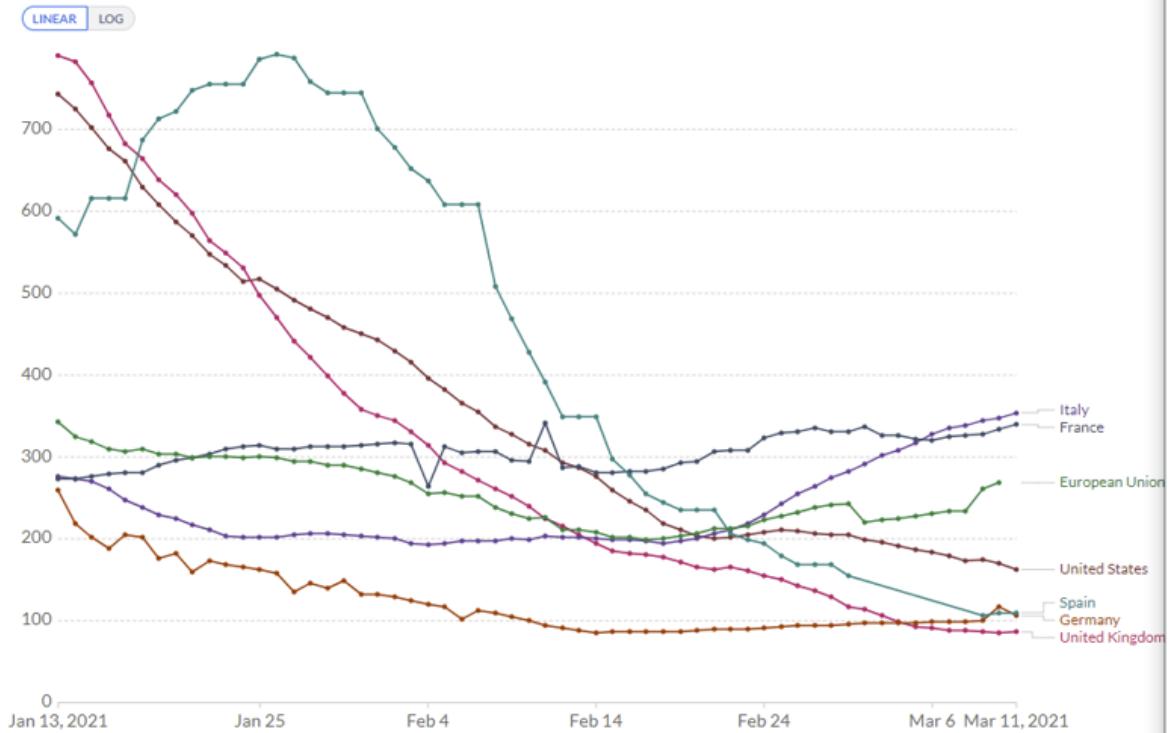
It's stunning how deep and widespread the rot is.
"Recent data showed about 40 percent of doses procured by the E.U. and distributed to member states were languishing in storage, partly due to poor logistics."

Europe

Daily new confirmed COVID-19 cases per million people

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

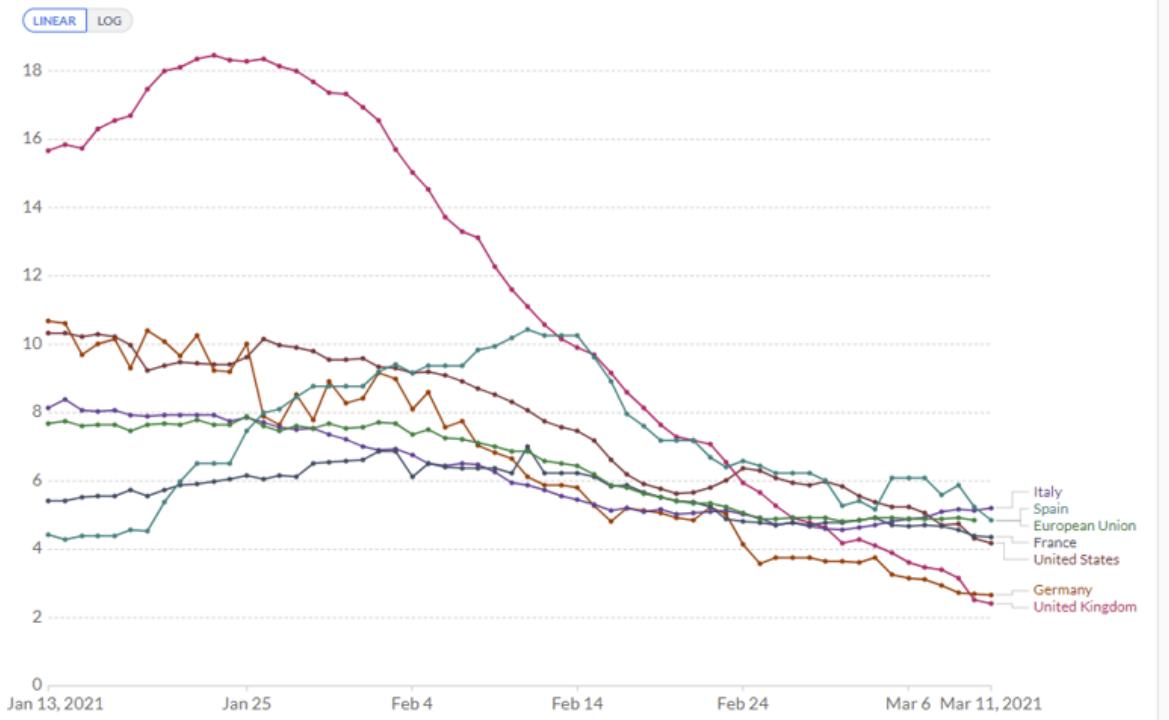
Our World
in Data



Daily new confirmed COVID-19 deaths per million people

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

Our World
in Data



European lockdown strategies continue to have stabilized things for now but not to have improved matters much, and there are signs things are slowly getting worse rather than better. The vaccine efforts are a huge fiasco across the European Union, and should be seen as a challenge to the heart of the entire European project.

The English Strain

Oh no. New strain versus old fatality numbers [from this preprint](#):

	non-VOC % (95% CI)	VOC % (95% CI)
No Comorbidities		
Female: 0-<65	0.05 (0.03-0.06)	0.07 (0.06-0.09)
65-<75	0.45 (0.30-0.59)	0.72 (0.50-0.95)
75-<85	1.08 (0.71-1.45)	1.73 (1.15-2.31)
85+	2.36 (1.47-3.25)	3.75 (2.34-5.16)
Male: 0-<65	0.09 (0.07-0.11)	0.14 (0.11-0.17)
65-<75	0.85 (0.59-1.12)	1.37 (0.96-1.77)
75-<85	2.03 (1.35-2.71)	3.24 (2.19-4.30)
85+	4.38 (2.72-6.03)	6.87 (4.33-9.42)
1 Comorbidity		
Female: 0-<65	0.11 (0.08-0.15)	0.18 (0.13-0.24)
65-<75	1.09 (0.78-1.41)	1.75 (1.25-2.25)
75-<85	2.60 (1.84-3.35)	4.13 (2.94-5.32)
85+	5.54 (3.77-7.31)	8.64 (5.91-11.38)
Male: 0-<65	0.22 (0.15-0.28)	0.35 (0.25-0.45)
65-<75	2.06 (1.51-2.62)	3.29 (2.44-4.14)
75-<85	4.81 (3.48-6.14)	7.54 (5.52-9.55)
85+	9.94 (6.87-13.01)	15.10 (10.63-19.58)
2+ Comorbidities		
Female: 0-<65	0.21 (0.14-0.28)	0.34 (0.22-0.45)
65-<75	1.99 (1.41-2.57)	3.18 (2.27-4.09)
75-<85	4.66 (3.45-5.87)	7.31 (5.42-9.20)
85+	9.65 (7.01-12.29)	14.68 (10.73-18.63)
Male: 0-<65	0.40 (0.27-0.52)	0.64 (0.44-0.84)
65-<75	3.72 (2.74-4.69)	5.87 (4.38-7.35)
75-<85	8.44 (6.44-10.44)	12.93 (9.99-15.87)
85+	16.65 (12.42-20.88)	24.34 (18.55-30.13)

[The moment will soon be here](#), and case numbers are already substantially higher than they would be otherwise.



Mark D. Levine @MarkLevineNYC · Mar 10

...

Virus is still spreading at an extraordinary rate in NYC. ~4,000 cases/day, despite steady progress on vaccination.

One big reason: variants

B.1.117 (UK) & B.1.526 (NYC) now together make up 51% of new cases here, up from 31% last week.

We still need to take this seriously.

226

3.2K

7.9K

The moment of truth is fast approaching. Within a few weeks, new variants will be a majority of new Covid cases in the United States. Very soon after that, they will account for most cases.

It is clear that the number of people with the variants is continuing to rise, as the overall number of infected people is only falling slowly. It would be very surprising if the number of cases doesn't rise before it starts dropping again - we are indeed almost certainly at least somewhat f***ed, and the resulting death rate will reflect the higher death rate from the English strain.

The question is how bad things will get. There are naively plausible mathematical models where we are rapidly vaccinating enough people to make up for the shift to the new strains. [For example](#), there are [these CDC projections](#) which even underestimated the rate of vaccinations:



Nate Silver ✅ @NateSilver538 · Mar 4

...

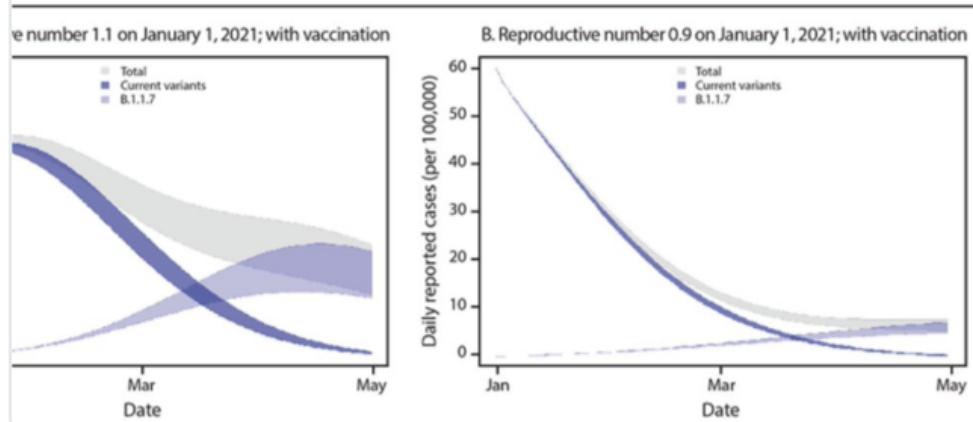
FWIW, the CDC's forecast from January of what happens when you have rising levels of B.1.1.7 on the one hand, but vaccination happening on the other hand, has been rather accurate so far (see the graph on the right which pretty closely matches real-world data).



Nate Silver ✅ @NateSilver538 · Jan 25

The CDC's forecasting agrees with this. The graph on the right shows what happens if we start out with an R_t of ~0.9 and then start to vaccinate people. The new strain slows down progress a bit and makes it take longer to get to ~0, but not so bad overall.
cdc.gov/mmwr/volumes/7...

Estimated case incidence trajectories* of current SARS-CoV-2 variants and the B.1.1.7 variant,[†] unity vaccination[§] and initial $R_t = 1.1$ (A) or initial $R_t = 0.9$ (B) for current variants — United April 2021



The problem with that model is that it fails to include a control system, and the control system is going to spend a while making things worse rather than better.

Open Sesame

[You can say things like this all you want](#), but all it will likely do is backfire because it will be seen as a completely unrealistic and unreasonable demand:



CNN Breaking News ✅ @cnnbrk · Mar 4

...

The United States shouldn't ease restrictions in place to prevent Covid-19 before the number of new coronavirus cases falls below 10,000 daily, "and maybe even considerably less than that," Dr. Fauci says

That's pretty unreasonable! When those who always make demands in a direction make completely unreasonable demands – no loosening of restrictions of any kind for a very long time – the response is to go 'yeah, that's public health experts for you' and that's that.

That's what happened.

[Connecticut fully reopens.](#)

Not to be outdone or even matched, [here's Texas](#):

 **BNO Newsroom**  @BNODesk · Mar 10 ...

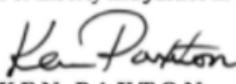
NEW: Texas Attorney General Ken Paxton threatens to sue Austin and Travis County if mask mandates and business-operating restrictions are not lifted within 3 hours

The decision to require masks or otherwise impose COVID-19-related operating limits is expressly reserved to private businesses on their own premises. It does not rest with jurisdictions like the City of Austin or Travis County or their local health authorities. Nor do they have the authority to threaten fines for non-compliance.

We have already taken you to court under similar circumstances. You lost. If you continue to flout the law in this manner, we'll take you to court again and you will lose again.

To that end, you and your local health authorities have until 6:00 p.m. today to rescind any local mask mandates or business-operating restrictions, retract any related public statements, and come into full compliance with GA-34. Otherwise, on behalf of the State of Texas, I will sue you.

For Liberty and Justice in Texas,


KEN PAXTON
Attorney General of Texas

cc: Dr. Mark Escott, Austin Public Health

🗨 78 ⬇️ 276 ❤️ 459 ⤴

Everywhere, we see states lifting restrictions in response to the progress we've made. All the vaccinated people that will start acting reasonably, especially now that the CDC has offered guidelines, will likely weaken the ability to enforce norms on the unvaccinated. We've already seen our rate of progress dramatically slowed before the impact of a number of rules being loosened, and it's not clear that those loosenings can be easily reversed, or if they can be reversed before things get back to alarming measured (and thus delayed) levels first, and then are implemented with a delay.

If public health advocates had wanted a different result, they could have offered a reasonable policy backed by object-level logic. I'd suggest something like this: The new strains are taking over. We have to survive one last spike due to the new strains. If a month or more from now, cases are falling and are at or below current levels, you can start loosening restrictions then. That gives people hope and is clearly reasonable, while covering the issues we need to actually worry about. Yes, they might prefer waiting longer than that, but it ain't gonna happen.

In Soviet America, Vaccines Still Work But Can You?

What can you safely do if you are vaccinated?

For a long time, Very Serious People have told us that the answer is exactly the same as what unvaccinated people can safely do, all but saying "[If I were you I'd lock my doors and](#)

[windows and never ever ever leave my house again."](#)

This was neither a realistic ask nor a way to get people excited about getting vaccinated. It's the same problem as telling states they can never open, except even more obviously untenable. Weeks went by, and the CDC issued no guidance on what vaccinated people could do. Thus, [this was the situation as late as March 5](#):



zeynep tufekci

@zeynep

...

This. The more CDC waits, the less influence they have on how spring will unfold. (And messaging is like fighting exponentials: early framing/influence is crucial —can't just pivot late and get the same bang).



Leana Wen, M.D. @DrLeanaWen · Mar 5

BREAKING: CDC Director says that they are STILL not releasing guidance on what vaccinated people can do.

I understand being careful, but the longer they wait, the more people will take matters into their own hands & render eventual CDC advice irrelevant.
[washingtonpost.com/opinions/2021/...](http://washingtonpost.com/opinions/2021/)

2:11 PM · Mar 5, 2021 · Twitter for iPhone



Youyang Gu @youyanggu · Mar 5

...

We have no guidance from the highest levels of public health on how to safely reopen our society. Their current strategy is just: "not yet". The CDC still won't even announce that vaccinated people can gather together.

And we wonder why states are going their own separate ways.



136



460



2.9K



Then, finally, we got guidelines! Real ones that made a non-zero amount of sense! Nice.

That doesn't mean we got there for the right reasons, [and I definitely agree with this take](#).



PoliMath @politicalmath · Mar 8

It has become abundantly clear that the CDC recommendations trail the political considerations

...

If the CDC is "the science" we're not "following the science" but the science is following what the political realities are

18

144

775



PoliMath @politicalmath · Mar 8

yes, yes, I know, astronaut_with_a_gun.gif

...

1

4

118



And also, while we're on PoliMath, [this take as well, even more strongly](#):



PoliMath
@politicalmath

...

Replying to @politicalmath

If I was in charge of Google search, I'd specifically hijack that exact search phrase and redirect people to a page that just says

YES

This is the first search result you get instead

The screenshot shows a web browser window with the AARP website. The top navigation bar includes links for 'Join', 'Renew', 'Help', 'Member Benefits', 'AARP Rewards', 'Register | Login', and a search icon. Below the navigation, a 'HEALTH' category is selected. The main headline reads 'Conditions & Treatments'. The specific article title is 'Can You Hug Your Grandkids After Receiving the COVID-19 Vaccine?'. Below the title, there are social sharing icons for Facebook, Twitter, LinkedIn, Email, and Print. A subtitle states 'Experts say different things, so consider the facts when deciding if it's safe to snuggle'. At the bottom of the article preview, it says 'by Christina Lanzito, AARP, February 4, 2021 | Comments: 12'.

1:23 PM · Mar 9, 2021 · Twitter Web App

Then again, if they did the other thing they would never be following the science, they'd be (Following Science™), which is what they do most of the time and what leads to telling everyone that they should build their lives around minimal symbolic improvements in protection against infectious diseases rather than living life. I'd much rather this follow the political realities if it can be combined, as it was here, with an attempt to get the priority order of interventions into a reasonable state. Also, if you are (Following Science™) and make impossible demands, even the politicians ignore you, as we're seeing with the reopenings.

As a reminder, the key for us to remember is that the political considerations center around blame avoidance on a two week time horizon, so if we want to get sensible policy, what we need to do is create sufficient expectation of back propagation of consequences that blame can be inflicted for bad decisions (or lack of good decisions) within that two week window. Then we've got something.

Due to the intensity of the current blame avoidance and conform-to-authority pressures, people who would usually have treated CDC guidelines as an upper bound beyond which you

get diagnosed with obsessive-compulsive disorder are now taking them literally, [for example](#):

Chana @ChanaMessinger · Mar 10
There are a lot of relevant dynamics, so I'm not as mad as I could be, but an administrator just told us that CDC regulations said that vaccinated people can be together maskless in homes, so we don't know about offices, and I think the only answer is to find it darkly hilarious.

7 replies 4 retweets 49 likes ⬆️

Chana @ChanaMessinger · Mar 10
What if we *didn't* use different reasoning for this than we would for successfully buying milk from a grocery store.

2 replies 1 retweet 21 likes ⬆️

Chana @ChanaMessinger · Mar 10
Do we think offices have dark transmission magic

2 replies 0 retweets 22 likes ⬆️

Chana @ChanaMessinger · Mar 10
The cDC hasn't said they don't, so....

...
...

Yeah, that's slightly unfair because homes have some self-limiting dynamics and offices can get very large with lots of people, but the ones who are thinking that only authorities can determine who does and does not have dark transmission magic are the problem here. Or rather, the dynamics training and forcing people into the posture that they must look to the relevant authorities for currently believed locations of dark transmission magic one would be blameworthy for not avoiding.

Also [Robin Hanson points out](#) that the authorities continue to treat people who have previously been infected as if they aren't immune and this in no way counts, presumably because "no evidence."



Robin Hanson ✅ @robinhanson · Mar 8

...

Why do these keep saying nothing about people who've been infected and recovered?



CDC Says Fully Vaccinated People Can Gather in Small Groups Without...
New guidelines direct fully inoculated people to wear masks and observe social distancing in public.

-wsj.com

23

6

108



Robin Hanson ✅ @robinhanson · Mar 9

...

Yes, there might be good reasons to treat the recovered differently. But the total lack of discussion of the recovered in such articles suggests that authorities and elite media readers don't care much about the infected, relative to the vaccinated.

3



26



([Your periodic reminder that the WHO quietly changed its definition of herd immunity to exclude the previously infected.](#))

I don't share his 'they don't much care about the infected' conclusion, or rather I find it imprecise. They mostly don't care about *anyone*, so why should the infected be an exception? Also, a lot of this is [burning down the village because it isn't legible](#). They're likely thinking a lot of people *think* they have had it when they didn't, and they don't want to open up the can of worms that involves, and they want everyone to get vaccinated and to provide as many incentives towards that as possible, so better to pretend the whole thing doesn't exist.

CDC Guidelines “Key Points”

[Here is the CDC brief on the new guidelines.](#)

Key Points

- COVID-19 vaccines currently authorized in the United States are effective against COVID-19, including severe disease.
- Preliminary evidence suggests that the currently authorized COVID-19 vaccines may provide some protection against a variety of strains, including B.1.1.7 (originally identified in the United Kingdom). However, reduced antibody neutralization and efficacy have been observed for the B.1.351 strain (originally identified in South Africa).
- A growing body of evidence suggests that fully vaccinated people are less likely to have asymptomatic infection and potentially less likely to transmit SARS-CoV-2 to others. However, further investigation is ongoing.
- Modeling studies suggest that preventive measures such as mask use and social distancing will continue to be important during vaccine implementation. However, there are ways to take a balanced approach by allowing vaccinated people to resume some lower-risk activities.
- Taking steps towards relaxing certain measures for vaccinated persons may help improve COVID-19 vaccine acceptance and uptake.
- The risks of SARS-CoV-2 infection in fully vaccinated people cannot be completely eliminated as long as there is continued community transmission of the virus. Vaccinated people could potentially still get COVID-19 and spread it to others. However, the benefits of relaxing some measures such as quarantine requirements and reducing social isolation may outweigh the residual risk of fully vaccinated people becoming ill with COVID-19 or transmitting the virus to others.
- [Guidance for fully vaccinated people](#) is available and will continue to be updated as more information becomes available.

So before we move on to the guideline details let's look at these Key Points. And remember, this is their introductory explanation everyone is praising, for what everyone says are the *pretty good, reasonable guidelines*.

First one is good.

Then we learn that the vaccines *may provide some* protection against a variety of strains. However, reduced efficacy has been observed for the B.1.351 strain.

Saying that something *may provide some* protection is saying almost exactly nothing. It's saying that we haven't proven that it doesn't provide any protection. It's the kind of language one wants to use if one wants to avoid blame for claiming something worked, while also avoiding blame for not saying something worked, and also give the impression it likely doesn't work.

So not only does this offer basically no confidence that the vaccines work against B.1.1.7, where we know they flat out fully work, it then says that one should be even more skeptical than the level of "*may offer some* protection" for this additional strain.

Next, we see that 'a growing body of evidence *suggests*' that fully vaccinated people are 'less likely to have asymptomatic infection.' But then it warns us the investigation is ongoing.

So we have something *suggestive* that they might be *less* likely, but who knows, these things are tricky, and no suggestion of things like ‘dramatically less’ or ‘prevents almost all’ or anything like that. Are we *trying* to prevent vaccinations here?

Next we are told that ‘modeling studies’ tell us that masks continue to be important, but that they generously will allow resumption of ‘some low-risk activities.’ I think this could be better summarized as ‘f*** you’ and also it seems modeling can be used to require precautions but a completely different standard of evidence applies to claims of prevention. It’s almost like it’s all about something else entirely.

Then they say ‘Taking steps towards relaxing certain measures for vaccinated persons may help improve Covid-19 vaccine acceptance and uptake.’ No s***, sherlock. Thank you for pointing this out, took you long enough. Also would help if you told people vaccines actually, what’s the word for it, *worked*.

The next line essentially says “ordinarily we’d tell you to lock your doors and windows and never ever ever leave your house again and actually that’s mostly what our guidelines say elsewhere if you look carefully, but we’ve driven half the population crazy so maybe we can reach a little bit of compromise this one time.” But it wants us to know that if there *wasn’t* a particular medical issue called ‘social isolation’ they wouldn’t let us meddling kids get away with being in the same room together.

As a side note: [The CDC guidelines for gyms call for “consistent and correct mask use.”](#) I have at various times used gyms, but it would never occur to me to use them during a pandemic until after I’d been vaccinated. The whole point of going is to *improve* your health, and there are plenty of other options. Also, the whole mask issue, which is going to interfere with exercising properly. It’s a small cost in many contexts, but not in this one. If vaccinated people still have to wear masks at the gym no matter the level of distancing, this seems a lot like a soft ban on gyms extended indefinitely. Perhaps we simply can’t hope to enforce a ‘vaccinated people can unmask’ norm in any public space properly, and we can mostly live with gyms being (more than usual levels of) terrible for another few months. The few true gym rats can get home equipment and/or are healthy enough that they can deal with the masks, I suppose.

[Or \(giant spoiler for John Wick 3 which you should definitely see\), a shorter summary of our overall situation in video form.](#)

All right, fine, yes, that’s not the part people are focused on and no one reads such words as if they mean things. I get it. Let’s see [the actual guidelines for vaccinated people.](#) The part that actually matters.

CDC Guidelines For Fully Vaccinated People

Key Points

This is the first set of public health recommendations for fully vaccinated people. This guidance will be updated and expanded based on the level of community spread of SARS-CoV-2, the proportion of the population that is fully vaccinated, and the rapidly evolving science on COVID-19 vaccines.

For the purposes of this guidance, people are considered fully vaccinated for COVID-19 ≥ 2 weeks after they have received the second dose in a 2-dose series (Pfizer-BioNTech or Moderna), or ≥ 2 weeks after they have received a single-dose vaccine (Johnson and Johnson (J&J)/Janssen).[†]

The following recommendations apply to non-healthcare settings. For related information for healthcare settings, visit [Updated Healthcare Infection Prevention and Control Recommendations in Response to COVID-19 Vaccination](#).

Fully vaccinated people can:

- Visit with other fully vaccinated people indoors without wearing masks or physical distancing
- Visit with unvaccinated people from a single household who are at low risk for severe COVID-19 disease indoors without wearing masks or physical distancing
- Refrain from quarantine and testing following a known exposure if asymptomatic

For now, fully vaccinated people should continue to:

- Take precautions in public like wearing a well-fitted mask and physical distancing
- Wear masks, practice physical distancing, and adhere to other prevention measures when visiting with unvaccinated people who are at [increased risk for severe COVID-19](#) disease or who have an unvaccinated household member who is at increased risk for severe COVID-19 disease
- Wear masks, maintain physical distance, and practice other prevention measures when visiting with unvaccinated people from multiple households
- Avoid medium- and large-sized in-person gatherings
- Get tested if experiencing [COVID-19 symptoms](#)
- Follow guidance issued by individual employers
- Follow CDC and health department travel requirements and recommendations

This is the part that *matters* so let's see the details.

The first line is that if *everyone* involved is vaccinated fully, you can do whatever you want, at least for small size gatherings. Good. Excellent. Some common sense. Yes.

The second is the principle that, essentially, a vaccinated person *is not a person with regard to gatherings* so long as everyone exposed is low-risk.

Thus, you get *one household with unvaccinated people*, since that's happening anyway, and if no one is high risk you can add any number of vaccinated people to the mix up to the limit of a 'small gathering'. Good. Excellent. Some common sense. Yes.

That approval assumes a common sense evaluation of what 'high risk' and 'low risk' mean. It's one thing to be cautious around the truly vulnerable, it's another to look at the technical 'list of high risk conditions.' The orders of magnitude in no way match. If we treat 'low-risk' as basically 'under the age of 65' I think this is conservative but at least somewhat sane.

Note that this guideline is contradicted by the guideline that one must wear masks when in the presence of someone *whose household includes a high risk member*. That's another degree of separation, and increases the effective annoyance level substantially if it trumps the permissive rule. It makes sense if one appreciates how the risk multiplications work, and you adjust the barrier for 'high risk' accordingly.

The third line is that vaccinated people need not quarantine. Again, yes. I could see asking them to *act like unvaccinated persons* during what would otherwise be a quarantine period, or otherwise use higher precaution levels, but for guideline purposes telling them to ignore it is likely even better. People with common sense will scale back on exposing others anyway if it looks like they took a big risk.

Then there's the whole 'you still have to follow every other rule same as everyone else' clauses. Still getting tested makes sense, although presumably the bar for what counts as symptoms would go up. Following employer guidelines and CDC recommendations is something you gotta say.

Avoiding medium and large size gatherings seems overly broad, depending on what counts as medium versus small. If the concern is that medium gatherings of vaccinated people are actually risky, I think that's mostly silly. If the concern is that people who aren't vaccinated will come anyway, or this will normalize larger gatherings and we want to hold off on that, those reasons seem reasonable. Given how vague medium is, I'll allow it.

Wearing masks when visiting with multiple other households is rather hilarious if you break down what is happening. It's norm enforcement with a side of punishment. The CDC does not want multiple households of unvaccinated people meeting up, for obvious reasons, and wants to at least ensure mask compliance, so it's not about to give anyone vaccinated permission to take off their masks in such a room. And I get that. Once anyone at a gathering takes their mask off, there's a strong tendency for everyone else to take theirs off as well (or to tell them to put it back on). If everyone isn't in it together, no one wants to be the schmuck going through the annoyance of wearing the mask.

Overall, these guidelines do seem reasonable, as a compromise between what makes physical sense and what preserve necessary norms of behavior, and as a compromise between encouraging vaccination versus letting risk get too out of hand. That doesn't mean they need to be followed to the letter, but we could have done a lot worse, and this is far better than no guidelines at all, and far better than the previous FUD of 'act the same as before.'

So yeah, all in all, I'll take it.

Vaccines Only Work If You Use Them

[AstraZeneca remains unapproved.](#)



Ezra Klein ✅ @ezraklein · Mar 8

Do any public health experts think these results are wrong, and Scotland made a terrible error approving the AZ vaccine?

...

And if not, are we making a terrible error with every subsequent day we fail to approve it?

It seems to me you need to believe one of these things is true.



Isaac Bogoch ✅ @BogochIsaac · Mar 1

In Scotland, the AstraZeneca vaccine reduced the risk of #COVID19-related hospitalization by **94%** after the first of 2 doses.

This is data from 490000 people.

Yet another reminder that the first vaccine available is the best vaccine.

bit.ly/3q4ysd0

[Show this thread](#)

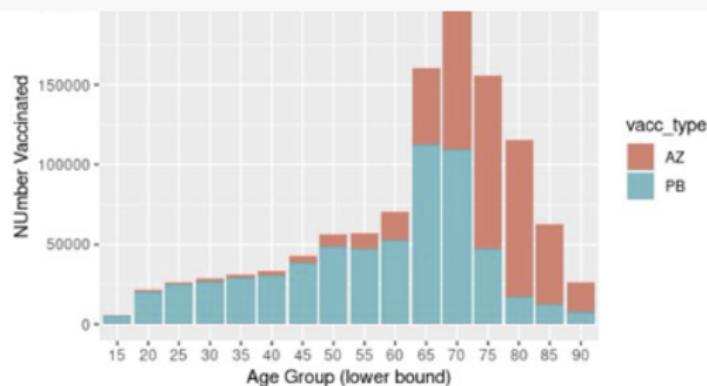


Figure 2: Vaccine uptake by age and vaccine type (AZ: Oxford-AstraZeneca. PB: Pfizer-BioNTech).

63

271

1.4K

↑



Ezra Klein ✅ @ezraklein · Mar 8

There were particular fears that AZ may be ineffective in the elderly, but a lot of elderly folks getting it here, and no evidence of reduced effectiveness.

...

At some point the mass of real world data coming out of Europe should influence us, right?

32

33

518

↑

Novavax remains unapproved, [and has new results \(press release\)](#):



Mac n' Chise 🧬🦠🧫 @sailorrooscout · 16h

Novavax's updated trial results are in:

...

- 100% protection against severe disease
- Final analysis in U.K. trial confirms 96% efficacy against original strain of COVID-19
- Efficacy against variants confirmed in U.K. and South Africa

[ir.novavax.com/news-releases/...](https://ir.novavax.com/news-releases/)

32

362

1.1K



Mac n' Chise 🧬🦠🧫 @sailorrooscout · 16h

96.4% against mild, moderate and severe disease. 100% protection against severe disease, including all hospitalization and death. 55.4% among the HIV- negative trial participants in a region where the vast majority of strains are B.1.351. Before you scoff- this is stellar in RWD.

...



Novavax Confirms High Levels of NVX-CoV2373 Vaccine Efficacy Against Original and Variant COVID-19 Strains in United Kingdom and South Africa Trials

DATA FACTSHEET

CONCLUSIONS

- 100% protection against severe disease, including all hospitalization and death
- United Kingdom: 96.4% efficacy against original COVID-19, 86.3% efficacy against predominant variant (post-hoc)
- South Africa: 55.4% efficacy against predominant B.1.351 escape variant in HIV-negative participants

UNITED KINGDOM PHASE 3 TRIAL

Who: ~15,000 adults 18-84 years of age, including 27% over age 65.

Primary endpoint: PCR-confirmed symptomatic (mild, moderate or severe) COVID-19 with onset ≥ 7 days after the 2nd dose in serologically negative (to SARS-CoV-2) adults.

Results: See Table 1 for trial data.

- 106 cases were observed: 10 in the vaccine group and 96 in the placebo group.
- 5 severe cases were observed, all in the placebo group (1 hospitalization). Four of the 5 severe cases were attributed to the B.1.1.7 variant.
- The study met its primary endpoint with 89.7% overall vaccine efficacy (95% CI: 80.2, 94.6).
- 14 days after dose 1, vaccine efficacy was 83.4% (95% CI: 73.6, 89.5).
- In volunteers 65 years of age and older, 10 cases of COVID-19 were observed, with 90% of those cases occurring in the placebo group.

FINAL ANALYSIS		
	Vaccine n=7,020	Placebo n=7,020
Total	10	96
Mild	1	28
Moderate	9	63
Severe	0	5
Vaccine Efficacy Original COVID-19	96.4%	95% CI: 73.8, 99.5
Vaccine Efficacy B.1.1.7 variant	86.3%	95% CI: 71.3, 93.5

Table 1. Final analysis of United Kingdom Phase 3 Trial.

SOUTH AFRICA PHASE 2B TRIAL

Who: ~4,400 adults 18-65 years of age, including 245 HIV-positive participants.

Primary endpoint: PCR-confirmed mild, moderate, or severe COVID-19 illness occurring ≥ 7 days after the 2nd dose in serologically negative (to SARS-CoV-2) adults.

Results: See Table 2 for trial data.

- 147 cases were observed: 51 in the vaccine group and 96 in the placebo group.
- 5 severe cases were observed, all in the placebo group (5 hospitalizations, 2 resulting in death). The vast majority of cases circulating during the efficacy analysis were due to the B.1.351 variant circulating in South Africa.
- 14 days after dose 1, overall vaccine efficacy was 42.7% (95% CI: 25.0, 56.3). In HIV-negative participants 14 days after dose 1, vaccine efficacy was 47.4% (95% CI: 29.9, 60.6).

COMPLETE ANALYSIS		
	Vaccine n=1,408	Placebo n=1,362
Total	51	96
Severe	0	5
Vaccine Efficacy Overall	48.6%	95% CI: 28.4, 63.1
Vaccine Efficacy HIV-negative	55.4%	95% CI: 35.9, 68.9

Table 2. Complete analysis of South Africa Phase 2B Trial.

Johnson & Johnson did get approved but after several weeks of pointless delay and with still essentially no plan, after all the complaining about the previous administration's lack of planning, [so here's what we got there](#):



Walid Gellad, MD MPH ✅ @walidgellad · Mar 8

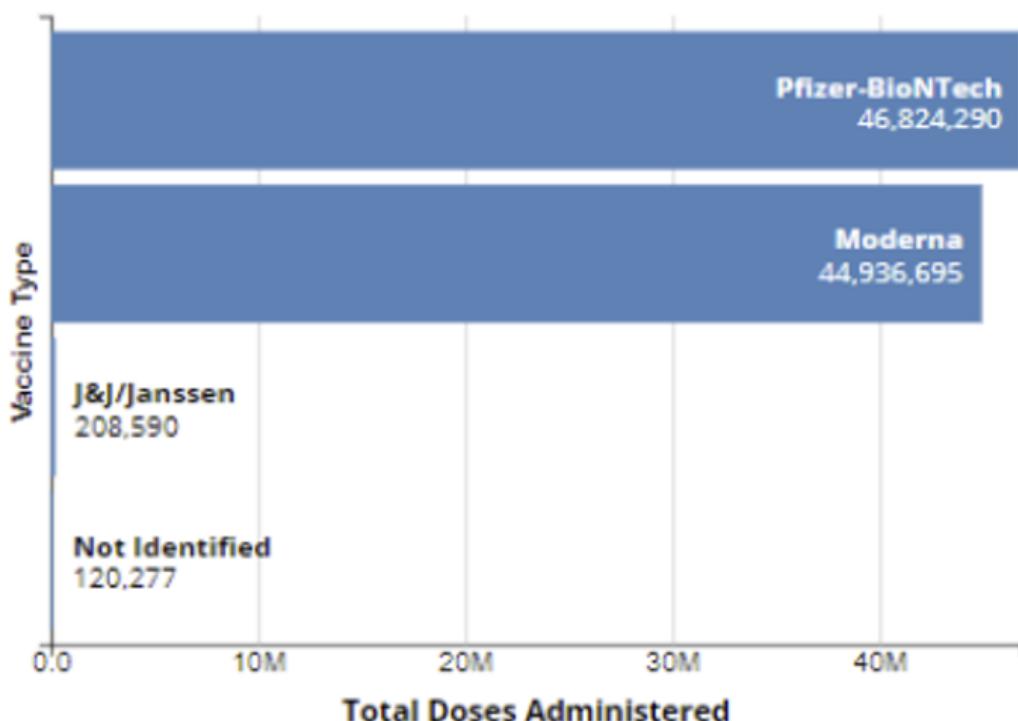
...

These J&J vaccine numbers are disappointing.

Less than 10% administered after a week? Where is the urgency?

Some states are reserving them for teachers and haven't started yet. Where are the others?

U.S. COVID-19 Vaccine Administration by Vaccine Type



[Here's a Bloomberg article \(behind a paywall\) describing what happened.](#)

There's nothing much left to say at this point, other than in a real sense #YouHadOneJob applies to all of this, and if you can't do it, you're a failure, period.

On another vaccine policy front, MR gives your periodic reminder that dose stretching (also known as vaccinating more people) reduces risk, [and this includes mutation risk](#) for obvious 'shut up and multiply' reasons.

On Lying

I am in general strongly against lying. Even when dealing with people or institutions who lie to you, who are mostly against you, who are not to be trusted, I still believe it is *usually* best to not lie.

There are limits and exceptions. Some people and systems flat out refuse to accept not lying, and some of those aren't things one can reasonably avoid, often gating vital resources

behind lying, or otherwise imposing very large costs to not lying. They train people to lie and reward liars, and punish those unwilling to lie or who endure costs when forced to tell lies, and generate norms that lying is The Way Things Are Done, and is justified and to be expected.

If one wishes to cultivate virtues like honesty, justice and honor, what is one to do? Where do we draw the line?

There's no clear answer, but I think [this is clearly one of those cases](#):



Ana Mardoll @AnaMardoll · Mar 4

...

On a covid paper at the doctor's office:

"Have you experienced any of the following symptoms in the past 48 hours?" and one of the symptoms is "fatigue".

Have I experienced fatigue in the past two days. No. I am a magical pixie forest-child who never sleeps. 😊

33

52

653



Ana Mardoll @AnaMardoll · Mar 4

...

Other symptoms I have experienced but have to lie about:

Diarrhea

Congestion / Runny Nose

Muscle Aches (this is at my PAIN DOCTOR, by the way)

8

6

188



Ana Mardoll @AnaMardoll · Mar 4

...

I really hate that the forever-covid Republicans have signed us up for means that these perfectly normal symptoms of chronic disease are now this weird shameful stigma you have to lie about or be denied care.



Ana Mardoll @AnaMardoll · Mar 4

LITERALLY, the page says that if I answer yes to any of the symptoms I have to leave the premises immediately. Do not pass Go, do not collect the pain meds that help me with those Muscle Aches and cause Fatigue.

11

6

174



Ana Mardoll @AnaMardoll · Mar 4

This is a CDC .gov checklist, too. Not something my doctor dreamed up.

8

3

125



Ana Mardoll @AnaMardoll · Mar 4

Every time I point this out, I get a million people telling me "I usually read the question to mean NEW/UNEXPLAINED symptoms" and sure ok great.

But that's your headcanon, not the language as-written. As-written, chronic pain patients have to LIE just to get in the door.

11

9

183



I understand what the people designing the checklist were thinking. The first half of their thinking, that we need a checklist of questions to see if anyone has symptoms, makes perfect sense. Good thinking there. The problem is the other half where they implicitly assume that everybody knows that words do not have meaning and that everyone knows to lie about the questions when it would be pragmatic to lie.

There are several problems with this approach.

One is that once people realize 'oh, clearly they don't think their words to be taken literally' then everyone makes their own determination of what things to mention and what things are none of anyone's damn business, perhaps because they want to not be shown the door out of the building. Then there's that to normalize that anywhere is to normalize it anywhere. We're teaching that words are not supposed to have meaning, that we shouldn't put necessary qualifiers on statements.

Then again, that could all be wrong. I am not at all convinced that the head cannon that it's 'new/unexplained' symptoms is actually the intention at all. If you give people that kind of wiggle room, a lot of them will think 'oh, sure, I can explain that' and pretend everything's fine when everything is very much not fine, and also there's constant pressure on everyone to not be socially awkward, so you kind of need hard and fast rules to avoid disaster. Which people will then of course lie about, once they realize how this works.

Presumably the solution is to ask the question, then if someone says yes to check if it's chronic or otherwise explained before escorting them automatically out of the building, or at least to say 'non-chronic' or 'new' or something. I would actually want to avoid 'explained' here and at most let a follow-up determine what counts as explained.

If You Aren't At High Risk, Should You Get The Vaccine Yet?

The trickier moral dilemma is the vaccine.

If you are at actually high risk and are eligible for the shot, yes, 100%, you should absolutely get the vaccine as soon as possible.

The questions worth asking are, should you be willing to lie to get the vaccine? Should you get it while there are others who are high risk, even if you can get it without lying?

I'll take the second question first. **If you are legally eligible and can get the vaccine without lying, I say yes, 100%, you should absolutely get the vaccine as soon as possible.**

This is rather overdetermined.

Authorities explicitly want you to do this, and I want you to do this as well, because the most important thing is getting shots into arms and not letting shots sit on shelves, and we've set up alternative methods to help the most vulnerable via pharmacies and also the best way to protect most of the remaining most vulnerable is to get as many people as possible vaccinated.

On the margin, if you don't book an appointment, either the appointment and shot you decline will go unfilled, or it will probably go to someone else who is 'high risk' according to some list but unlikely to be actually high risk, or someone who is lying. In many jurisdictions all you have to do is say you are somehow eligible. That's it. No one is verifying anyone's claims.

If there's someone in contact with you refusing to claim a shot they are eligible for on these grounds, and this is exposing you or those around you to covid risk, I think it is correct to be rather upset about this. It's not a reasonable concern.

The real question and least convenient world is, suppose (because this is the case in at least many places) that the law is inefficient, unjust and unenforced and none of that is an accident. For example, suppose there's a giant list of 'high risk conditions' that qualify people, including ever having smoked a few cigarettes, or having a 'developmental disorder' which explicitly includes your motherf***ing Tourettes (which [can also get you medical weed](#)). And it's clear that they never actually ask for any kind of verification - in Washington DC they literally just ask 'do you have one of these 20 things?' and [all you have to do is say yes. Press X to not die.](#)

It is valuable and important to cultivate the virtue of not lying, but at some point this isn't even lying anymore because [you are dealing with the words of actors rather than scribes](#) and the actual meaning of your words is the pure and truthful 'I want to get vaccinated.'

How meaningfully different is all this from a box that says "I want this vaccine"? How meaningfully different is this box from the box that says 'I have carefully reviewed the 40 page user agreement?'

Did you pack your own bags?

[Then there's the question of Prizer's CEO, who it seems is not vaccinated, and this forced him to postpone a trip to Israel.](#) Presumably the PR department decided that it would be a bad look to 'skip the line.' My suspicion is he's actually vaccinated but is pretending he isn't and he can't tell the Israelis that.

Not only should he be vaccinated, *he should have been the first person vaccinated*. That's basic Skin In The Game 101. The person in charge of making the vaccine takes the vaccine. Instead, we're so concerned about perceptions of 'line jumping' that the person who *literally led the vaccine development effort* doesn't feel entitled to publicly claim a dose for himself,

let alone feel under his proper *obligation* to take a dose (that would also benefit him, but the point is that he proves that he believes this and we can know that.)

Similarly, a better vaccine approval system (for the first vaccine, anyway) might be that everyone at the FDA decides secretly when to get themselves and their families vaccinated, entirely up to them, and when enough of them decide to do it, the vaccine is approved. You can have any meetings you want, but they don't count for anything. Ideally you'd hold some people out-of-sample so you could do this for other vaccine candidates later.

Who Wants the Vaccine?

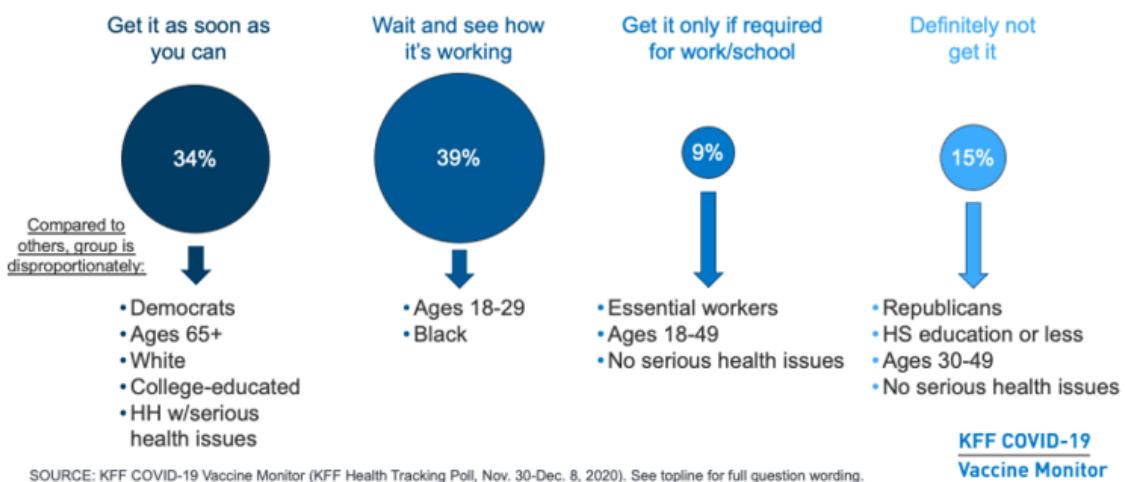
The above section assumes that if you're reading this, you're aware that vaccines are the greatest thing and the only question is how to get one.

Alas, this is far from a universal perspective. [Here's current survey data, and some more:](#)

Figure 12

Profile Of Groups By Vaccine Enthusiasm

When a vaccine for COVID-19 is approved and widely available to anyone who wants it, do you think you will...?





Ryan Struyk ✅ @ryanstruyk · 18h

...

Americans who will choose *not* to be vaccinated when one is available via new NPR/PBS/Marist poll:

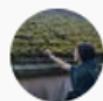
47% Trump voters
41% Republicans
38% White evangelicals
37% Latino
34% independents
34% White non-college
28% White
25% Black
18% White college
11% Democrats
10% Biden voters

The six point gap between Republicans and Trump voters makes the role of tribal identity here very clear.

I wonder several things. First, I wonder how much the 'wait and see' category is mostly 'I don't want to have to think ever and this is an excuse not to for now' because most of those people aren't yet eligible. They know they don't want to think enough to look over a list of conditions and figure out how to get an appointment, so why think about something now when they can at worst have to think about it later, and hopefully by then the answer is obvious?

The group that says 'get only if required' feels like it has to be way bigger than that. Are the majority of people who aren't in the first two categories really going to *give up their job or place in school* rather than get the shot? That's a super strong preference to not take the vaccine. I am super excited for the vaccine but would *I quit my job or drop out of school* to avoid it under a normal person's life circumstances? No, I would not. I'm guessing most of this is all talk.

There's also the question of blame and social pressure, because it seems (standard warning about anecdotes) [like there's a lot of this](#):



Marie 🌿 @mesolude · Mar 5

...

Uber driver isn't planning to get vaccine until uber / lyft force him to as a requirement to use the platform

He's worried about it, lots of people in his life telling him not to do it

1

1

9

↑

Mostly people deciding about vaccines aren't basing their decision on physical world models and a study of immunology and statistical findings. They're responding to various forms of social pressure and information cascades and blame dynamics. For various reasons there's a bunch of 'vaccination bad' social pressure in many places, so my model says that such folks will mostly reverse when the forces going the other way are sufficiently strong that the social pressure has to back off, and such folks aren't realizing that the pressure from other sources will relieve the social pressure when the time comes. If you can't (hold a job / go to school /

travel / go to a restaurant / etc) without vaccination, they're gonna fold and they know it, but for now their social rewards are for putting up a front that they're not going to get vaccinated, so that carries over to the survey because there's no incentive to be super-honest and self-aware.

I continue to be rather baffled by the whole 'wait and see' attitude, unless it's another case of avoiding blame and social pressure, in this case by avoiding taking a position, and doesn't cash out to anything at all. What are we waiting to see, at this point? We've given about 100 million doses in the United States and everything's going great so literally what do these people worry they are going to see? [Is it basically this?](#)

In "Bart to the Future", Bart sees his own future. Ned is seen blind, revealing that he received laser eye surgery, which was great at the beginning, but after the 10-year-point, your eyes fall out. He often bails Bart out with money as a way of thanking him for not "outing" Rod and Todd.

Seems like it would have to be, except without as plausible a mechanism as that one.

So I guess that means it's time to address the whole blood clot thing...

AstraZeneca and Blood Clots

Even by 2020-2021 pandemic standards this one seems beyond ridiculous.

Several places, including Denmark and Thailand, have suspended use of the AstraZeneca vaccine due to concerns about blood clots. There have been voices of panic, warning that if this turned out to be a real thing that it could destroy public confidence in all vaccines, perhaps for all time.

I did not have to look to know this was almost certainly Obvious Nonsense, because it looked nothing like what the response would be if it *wasn't* Obvious Nonsense.

And also because blood clots on this scale are *somewhat* more plausible an issue than being struck by lightning more often, but not *that much more* plausible.

And also because math. If there was a blood clot issue big enough to make the AstraZeneca vaccine potentially not worth using, either it comes along only after a several month delay and then suddenly happens to tons of people, or we somehow missed it for several months under mass vaccination drives. Thus, until I got back home I didn't bother checking for data.

Then, of course, [I saw this from the BBC](#):

Oxford-AstraZeneca: EU says 'no indication' vaccine linked to clots

It said the number of cases in vaccinated people was no higher than in the general population.

The statement came after a number of countries, including Denmark and Norway, suspended the use of the jab.

The suspension followed reports that a small number of people had developed clots after receiving the vaccine.

There were also reports that a 50-year-old man had died in Italy after developing deep vein thrombosis (DVT) following a dose of the jab.

"There is currently no indication that vaccination has caused these conditions, which are not listed as side effects with this vaccine," the European Medicines Agency (EMA) said on Thursday.

"The vaccine's benefits continue to outweigh its risks and the vaccine can continue to be administered while investigation of cases of thromboembolic events is ongoing," it added.

It said there had been 30 cases of "thromboembolic events" among the five million Europeans who have received the jab.

Here's the BMJ [confirming that investigations show zero signs of a statistical effect of any kind](#). Even if somehow there was some small effect, the chance of this being big enough to justify not using the vaccine is very, very close to actual zero.

[I love this post by Nate Silver...](#)



Nate Silver
@NateSilver538

...

Several EU countries are temporarily suspending use of the AstraZeneca vaccine (see story below). So it's worth pointing out that the EU's slow/conservative vaccine rollout is starting to have real consequences.



Reuters @Reuters · Mar 11

Denmark, Norway temporarily suspend AstraZeneca COVID shots after blood clot reports reut.rs/2PV4qvO



11:25 AM · Mar 11, 2021 · Twitter Web App

...because, while there are a lot of previously existing real consequences and they're really terrible (and thus that part of it is really weird), the implicit assumption here is that none of this has anything to do with an actual problem with the vaccine *and that no one even thinks anyone else actually doubts seriously this*. There's *common knowledge* that the concerns are stupid.

The amount of damage done to vaccine efforts due to the suspensions and beyond catastrophic messaging and failure to do any math at all, both by suspending them directly and by making everyone around the world have one more reason to worry, is rather large here.

I wonder how many people didn't get laser eye surgery due to the throwaway Ned Flanders joke on The Simpsons. I'm guessing more than one might think. I know that I was for a time

under some social pressure to get it, but I didn't want to, and the throwaway joke gave me some ammo to push back. Then compare that to this.

We Have Established as Common Knowledge That Andrew Cuomo Is The Worst

Andrew Cuomo has been The Worst for a long, long time. I have it on very good authority that he got his start cruelly bullying his brother Chris in early childhood, kept going from there, and Mario Cuomo should be viewed in many of the same ways we might think of Marcus Aurelius - you can be a great leader while you're alive, but all is lost if you botch the line of succession.

For those following the actual underlying scandal and cover-up, there's this: [Cuomo administration altered a Covid report, intentionally omitting the true magnitude of Covid's impact on nursing homes. \(HT/Source, administration response that their cover-up was sufficiently badly implemented that it was legitimate\).](#)

[Oh, and then there's this:](#)



Batya Ungar-Sargon ✅ @bungarsargon · Mar 9

Oh, man. It wasn't just seniors: Cuomo ordered homes for people with developmental disabilities to accept covid patients — and never rescinded the order. 552 died. Alone. Can you even imagine what their final moments were like?



Cuomo admin ordered homes for disabled to accept coronavirus patie...

New York Gov. Andrew Cuomo is already under fire for his administration's handling of nursing home death numbers

foxnews.com

167

1.6K

2.4K



Batya Ungar-Sargon ✅ @bungarsargon · Mar 9

Cuomo made a decision to send seniors and the developmentally disabled back to group and nursing homes so that hospital beds would be available for those he deemed more deserving. It's sickening. The sanctity of human life is infinite. That means ALL human life.

18

141

503



Batya Ungar-Sargon ✅ @bungarsargon · Mar 9

Instead of protecting our most vulnerable, Cuomo signed their death warrants. I just can't get over it. THIS is the story, THIS is why he should resign.

22

158

580



Batya Ungar-Sargon ✅ @bungarsargon · Mar 9

And yes, I'll say it: Demanding he resign over inappropriate workplace behavior is at least to some extent a denial of his true crime, which is sending our most vulnerable, our most defenseless - seniors and the developmentally disabled - to their deaths.

...

It's worth noting also that the inappropriate workplace behavior is looking worse and worse, as there's more and more accusers and the things he is accused of get worse and worse. By any reasonable standard *yes he should definitely go down for the sexual harassment* if it

wasn't for the fact that *he very much needs far more to go down for the directly caused mass deaths and cover up of the mass deaths* and it's hard to truly go down for multiple things at once.

[So basically](#) this (and you gotta love the photo):

The Babylon Bee @TheBabylonBee · Mar 10
10,000th Victim Comes Forward To Accuse Cuomo Of Inappropriately Killing Her Grandma



10,000th Victim Comes Forward To Accuse Cuomo Of Inappropriately ...
NEW YORK, NY - Yet another victim has come forward to accuse Governor Cuomo of inappropriately killing her grandma, sources ...
babylonbee.com

At first it was plausible to claim the sexual harassment accusations were some combination of false or not that serious. That no longer seems plausible.

What seems clear is that, if it wasn't for the nursing home situation, Cuomo wouldn't be facing any of these accusations, would likely never have faced them, and he would have continued to get away with lots of sexual harassment.

All the revelations are entirely unsurprising. What would have been surprising would have been if Cuomo *wasn't* engaging in and getting away with lots of sexual harassment, because the prior on such behaviors by men with power is rather high, and Cuomo is a mean petty tyrant and bully and a liar – remember, he's the worst – so to have his behavior in this other realm suddenly be appropriate and reasonable would have required an explanation. Then, once his grip on power had sufficiently slipped and things turned against him, things turned against him and others amplified and encouraged reports of his actions rather than discouraging and suppressing them, and others (including the victims) suddenly interpreted his actions as unwelcome and offensive rather than maximally permissible.

And now essentially every Democratic politician in New York is lining up to call for his resignation.

I do find the dynamics here interesting in a broader sense, but this is already dangerously deep into 'there be dragons' territory and I'm only going here because it's Cuomo, so let's

move any further discussion to a different venue more appropriate to such issues.

And Yet, No, Technically We Are Incorrect, Eric Topol Is Actually The Worst

Beware scope insensitivity! For it seems likely [Eric Topol did this](#):



Eric Topol @EricTopol · Oct 10, 2020

We were on a path for a vaccine emergency authorization (EUA) before November 3rd. Thanks to the FDA, Trump's plan was disrupted. That won't happen.
First real sign of the independence of FDA since the pandemic started. And that's important.

...



Medscape @Medscape · Oct 10, 2020

EXCLUSIVE: @EricTopol and @SteveFDA meet, following that "very tough letter," for a valuable conversation on the state of COVID-19 in the U.S. ms.spr.ly/6018TIB0a



35



241



934



[No, seriously, it looks like he did that. Thread](#). See the MIT link for more details.

If this actually made a difference, the amount of blood on this man's hands is staggering. Not history's greatest villain, but we should not be confident he doesn't make the list.

From The Drunkwriting Files of Polimath: A Short Dramatic Scene Within the CDC

A Short Scene Deep Within the CDC

Expert 1: ok, so physical distancing is never a bad thing. How much should we say they need?

CDC Director Rochelle Walensky: Well, 3 feet is probably ok. That's the advice I've been giving to schools so far and it's in line with the WHO recommendations.

Expert 2: Well... 3 feet is ok, but 6 feet is better.

Expert 3: I mean... obviously. Six feet is better than three feet in the same way that 100 feet is better than 6 feet since no one is going to catch a pathogen from 100 feet away. Should we make it 100 feet?

Director Walensky: That seems excessive.

Expert 1: Yeah, one student per classroom is probably overkill. How big is a normal classroom?

Expert 2: I don't know... 25 kids to a classroom right? So 25 kids spaced 6 feet apart, what is that?

Expert 3: No, I've got this one, ok... how big is a classroom?

Expert 1: Oh man, I don't know. 50 feet? That sounds right. 50 feet by 50 feet.

Expert 2 (thinking back to high school): yeah, that's probably right.

Expert 3: ok, so 50 squared is 2500. How much space does each student need?
Rochelle?

Director Walensky: I am 100% not participating in this imaginary satirical discussion.

Expert 3 (distainfully): You're no fun.

Director Walensky: I am plenty of fun.

Expert 2: Focus, people! We're writing national policy! So each kid needs 6 feet of space... assuming every child is a massless singular point in space, they require 113 square feet of space.

Expert 1: Well, we can round that down to 100, so we can fit the average class size of 25 students into an average classroom of 2500 square feet!

Expert 3: Math saves the day! 6 feet distance it is! Well done everyone, let's order tacos. Dr. Walensky, do you like tacos?

Director Walensky: I have no opinions about tacos.

Expert 3: Look, no one is going to take a person who has no opinions about tacos seriously.

Director Walensky: Tacos are fine.

Expert 3: Yes. that's the kind of decisive action we at the CDC provide!

Seems legit. Tacos are great.

In Other News

[France is going to start vaccinating on weekends.](#)

[Report from WSJ that P.1, the new Brazilian variant, is really dangerous \(WSJ link\).](#)



Randall Parker #ApproveMoreVaccines @futurepundit · Mar 6

...

P.1 is bad: The new variant, known as P.1, is 1.4 to 2.2 times more contagious than versions of the virus previously found in Brazil, and 25% to 61% more capable of reinfecting people who had been infected by an earlier strain, according to a study released

WSJ The Wall Street Journal ✅ @WSJ · Mar 6

"The virus is behaving differently. It's really aggressive." Doctors are sounding the alarm over a new variant causing patients to become seriously ill faster in Brazil and Peru. Cases include people in their 30s and 40s with no underlying health problems. on.wsj.com/2O1Y5c7

I haven't seen talk elsewhere and haven't had an opportunity to follow up on this. These numbers even if accurate are definitely something vaccination can overcome if we have enough time, which we likely do if this hasn't arrived here yet. The highlighting of 'cases' include people in their 30s and 40s with no underlying conditions' highlights that this is clueless journalism where wet ground causes rain, so I'll hold off on updating much until more information is available.

[Goodbye what-to-do-now thread from Covid Tracking Project](#), alas not that helpful. Links to [data summary](#) and a [guide to federal resources](#).

[New nature study on long Covid \(paper\)](#). As he says, it's not great, but the alternative data points seem even worse.



Prof Francois Balloux ✅

@BallouxFrancois

...

This study on longcovid is not perfect as it relies on self-reported symptoms, by self-selected users. That said, it may be the best we have and the results feel plausible.

- 13.3% symptoms lasting ≥ 28 days,
- 4.5% for ≥ 8 weeks
- 2.3% for ≥ 12 weeks

[Alex Tabarrok reports that the condition of his students is increasingly dire.](#)



Alex Tabarrok
@ATabarrok

...

I have over 300 students and judging by that sample many people are reaching the breaking point. It's worse now than last semester. The pandemic is amplifying all the other problems students may have--family, work, health. We cannot end this thing soon enough.

2:37 PM · Mar 7, 2021 · Twitter Web App

[California variant seems unlikely to be important.](#)

[CDC study on Covid and obesity.](#)

[FDA decides its agents not flying is more important than doing drug company inspections, then says they had no choice and that the backlog of drug approvals is 'due to pandemic.'](#)

[Australian doctors are uncertain in what ways they can legally promote vaccinations due to anti-drug-advertising laws. MR chimes in and reminds us that England banned mask advertising.](#)

[A modest proposal on price gouging.](#)

[CDC still, today, discouraging use of N95 masks because of supply concerns.](#) Delenda est.

Our vaccine messaging is so terrible that [the mayor of Detroit turned down an allocation of J&J vaccine](#) doses, so his city's residents can 'get the best.' As far as I know, he remains the mayor.

[Twitter thread of examples of 'public health experts' calling for kids to return to school.](#)

[If you're finally going to vaccinate around the clock, why not give it an '80s theme?](#)

[Doing a randomized oncology trial means overcoming 50+ people with veto power over several steps.](#) There is indeed likely someone you forgot to ask.

[Teachers refusing to return to 'unsafe' in-person schooling warned by their union not to post social media pictures of themselves on spring break.](#)

From LessWrong: [A tool called MetaForecast that is Exactly What It Says On the Tin.](#) I don't find such things that useful right now but this seems much better than trying to find the data elsewhere.

[Facebook is censoring doctors writing in the Wall Street Journal on grounds of 'misleading information.'](#) As Gu points out in his thread, this sets a highly dangerous precedent and the procedure being used to decide what to censor makes no sense and is fully arbitrary. It would be entirely unsurprising if links to this column were to be censored by Facebook. Please do not rely on them as a source of news, or ideally for anything at all.

Next week I plan to return to the Thursday cycle of posting.

Coherence arguments imply a force for goal-directed behavior

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

[Epistemic status: my current view, but I haven't read all the stuff on this topic even in the LessWrong community, let alone more broadly.]

There is a line of thought that says that advanced AI will tend to be 'goal-directed'—that is, consistently doing whatever makes certain favored outcomes more likely—and that this is to do with the 'coherence arguments'. [Rohin Shah](#), and probably others¹, have argued against this. I want to argue against them.

The old argument for coherence implying (worrisome) goal-directedness

I'd reconstruct the original argument that Rohin is arguing against as something like this (making no claim about my own beliefs here):

1. '**Whatever things you care about, you are best off assigning consistent numerical values to them and maximizing the expected sum of those values'**
‘Coherence arguments²’ mean that if you don’t maximize ‘expected utility’ (EU)—that is, if you don’t make every choice in accordance with what gets the highest average score, given consistent preferability scores that you assign to all outcomes—then you will make strictly worse choices by your own lights than if you followed some alternate EU-maximizing strategy (at least in some situations, though they may not arise). For instance, you’ll be vulnerable to ‘[money-pumping](#)’—being predictably parted from your money for nothing.³
2. '**Advanced AI will tend to do better things instead of worse things, by its own lights'**
Advanced AI will tend to avoid options that are predictably strictly worse by its own lights, due to being highly optimized for making good choices (by some combination of external processes that produced it, its own efforts, and the selection pressure acting on its existence).
3. '**Therefore advanced AI will maximize EU, roughly'**
Advanced AI will tend to be fairly coherent, at least to a level of approximation where becoming more coherent isn’t worth the cost.⁴ Which will probably be fairly coherent (e.g. close enough to coherent that [humans can’t anticipate the inconsistencies](#)).
4. '**Maximizing EU is pretty much the same as being goal-directed'**
To maximize expected utility is to pursue the goal of that which you have assigned higher utility to.⁵

And since the point of all this is to argue that advanced AI might be hard to deal with, note that we can get to that conclusion with:

1. ‘Highly intelligent goal-directed agents are dangerous’

If AI systems exist that very competently pursue goals, they will likely be better than us at attaining their goals, and therefore to the extent there is a risk of mismatch between their goals and ours, we face a serious risk.

Rohin’s counterargument

Rohin’s counterargument begins with an observation made by others before: any behavior is consistent with maximizing expected utility, given *some* utility function. For instance, a creature just twitching around on the ground may have the utility function that returns 1 if the agent does whatever it in fact does in each situation (where ‘situation’ means, ‘entire history of the world so far’), and 0 otherwise. This is a creature that just wants to make the right twitch in each detailed, history-indexed situation, with no regard for further consequences. Alternately the twitching agent might care about outcomes, but just happen to want the particular holistic unfolding of the universe that is occurring, including this particular series of twitches. Or it could be indifferent between all outcomes.

The basic point is that rationality doesn’t say what ‘things’ you can want. And in particular, it doesn’t say that you have to care about particular atomic units that larger situations can be broken down into. If I try to call you out for first spending money to get to Paris, then spending money to get back from Paris, there is nothing to say you can’t just have wanted to go to Paris for a bit and then to come home. In fact, this is a common human situation. ‘Aha, I money pumped you!’ says the airline, but you aren’t worried. The twitching agent might always be like this—a creature of more refined tastes, who cares about whole delicate histories and relationships, rather than just summing up modular momentarily-defined successes. And given this freedom, any behavior might conceivably be what a creature wants.

Then I would put the full argument, as I understand it, like this:

1. Any observable sequence of behavior is consistent with the entity doing EU maximization (see observation above)
2. Doing EU maximization doesn’t imply anything about what behavior we might observe (from 1)
3. In particular, knowing that a creature is an EU maximizer doesn’t imply that it will behave in a ‘goal-directed’ way, assuming that *that* concept doesn’t apply to all behavior. (from 2)

Is this just some disagreement about the meaning of the word ‘goal-directed’? No, because we can get back to a major difference in physical expectations by adding:

1. Not all behavior in a creature implicates dire risk to humanity, so any concept of goal-directedness that is consistent with any behavior—and so might be implied by the coherence arguments—cannot imply AI risk.

So where the original argument says that the coherence arguments plus some other assumptions imply danger from AI, this counterargument says that they do not.

(There is also at least some variety in the meaning of ‘goal-directed’. I’ll use goal-directed_{Rohin} to refer to what I think is Rohin’s preferred usage: roughly, that which seems intuitively goal directed to us, e.g. behaving similarly across situations, and accruing resources, and not flopping around in possible pursuit of some exact history of personal floppage, or peaceably preferring to always take the option labeled ‘A’.⁶)

My counter-counterarguments

What's wrong with Rohin's counterargument? It sounded tight.

In brief, I see two problems:

1. The whole argument is in terms of logical implication. But what seems to matter is changes in probability. Coherence doesn't need to rule out any behavior to matter, it just has to change the probabilities of behaviors. Understood in terms of probability, argument 2 is a false inference: just because any sequence of behavior is consistent with EU maximization doesn't mean that EU maximization says nothing about what behavior we will see, probabilistically. All it says is that the probability of a behavioral sequence is never reduced to zero by considerations of coherence alone, which is hardly saying anything.

You might then think that a probabilistic version still applies: since every entity appears to be in good standing with the coherence arguments, the arguments don't exert any force, probabilistically, on what entities we might see. But:

1. An outside observer being able to rationalize a sequence of observed behavior as coherent doesn't mean that the behavior is actually coherent. Coherence arguments constrain combinations of external behavior and internal features —'preferences'⁷ and beliefs. So whether an actor is coherent depends on what preferences and beliefs it actually has. And if it isn't coherent in light of these, then coherence pressures will apply, whether or not its behavior *looks* coherent. And in many cases, revision of preferences due to coherence pressures will end up affecting external behavior. So 2) is not only not a sound inference from 1), but actually a wrong conclusion: if a system moves toward EU maximization, that does imply things about the behavior that we will observe (probabilistically).

Perhaps Rohin only meant to argue about whether it is *logically possible* to be coherent and not goal-directed-seeming, for the purpose of arguing that humanity can construct creatures in that perhaps-unlikely-in-nature corner of mindsphere, if we try hard. In which case, I agree that it is logically possible. But I think his argument is often taken to be relevant more broadly, to questions of whether advanced AI will tend to be goal-directed, or to be goal-directed in places where they were not intended to be.

I take 1) to be fairly clear. I'll lay out 2) in more detail.

My counter-counterarguments in more detail

How might coherence arguments affect creatures?

Let us step back.

How would coherence arguments affect an AI system—or anyone—anyway? They're not going to fly in from the platonic realm and reshape irrational creatures.

The main routes, as I see it, are via implying:

1. incentives for the agent itself to reform incoherent preferences
2. incentives for the processes giving rise to the agent (explicit design, or selection procedures directed at success) to make them more coherent

3. some advantage for coherent agents in competition with incoherent agents

To be clear, the agent, the makers, or the world are not necessarily thinking about the arguments here—the arguments correspond to incentives in the world, which these parties are responding to. So I'll often talk about ‘incentives for coherence’ or ‘forces for coherence’ rather than ‘coherence arguments’.

I'll talk more about 1 for simplicity, expecting 2 and 3 to be similar, though I haven't thought them through.

Looking coherent isn't enough: if you aren't coherent inside, coherence forces apply

If self-adjustment is the mechanism for the coherence, this doesn't depend on what a sequence of actions looks like from the outside, but from what it looks like from the inside.

Consider the aforementioned creature just twitching sporadically on the ground. Let's call it Alex.

As noted earlier, there is a utility function under which Alex is maximizing expected utility: the one that assigns utility 1 to however Alex in fact acts in every specific history, and utility 0 to anything else.

But from the inside, this creature you excuse as ‘maybe just wanting that series of twitches’ has—let us suppose—actual preferences and beliefs. And if its preferences do not in fact prioritize this elaborate sequence of twitching in an unconflicted way, and it has the self-awareness and means to make corrections, then it will make corrections⁸. And having done so, its behavior will change.

Thus excusable-as-coherent Alex is still moved by coherence arguments, even while the arguments have no complaints about its behavior *per se*.

For a more realistic example: suppose Assistant-Bot is observed making this sequence of actions:

- Offers to buy gym membership for \$5/week
- Consents to upgrade to gym-pro membership for \$7/week, which is like gym membership but with added morning classes
- Takes discounted ‘off-time’ deal, saving \$1 per week for only using gym in evenings

This is consistent with coherence: Assistant-Bot might prefer that exact sequence of actions over all others, or might prefer incurring gym costs with a larger sum of prime factors, or might prefer talking to Gym-sales-bot over ending the conversation, or prefer agreeing to things.

But suppose that *in fact*, in terms of the structure of the internal motivations producing this behavior, Assistant-Bot just prefers you to have a gym membership, and prefers you to have a better membership, and prefers you to have money, but is treating these preferences with inconsistent levels of strength in the different comparisons. Then there appears to be a coherence-related force for Assistant-Bot to change. One way that that could look is that since Assistant-Bot’s overall behavioral policy currently entails giving away money for nothing, and also Assistant-Bot prefers money over

nothing, that preference gives Assistant-Bot reason to alter its current overall policy, to avert the ongoing exchange of money for nothing.⁹ And if its behavioral policy is arising from something like preferences, then the natural way to alter it is via altering those preferences, and in particular, altering them in the direction of coherence.

One issue with this line of thought is that it's not obvious in what sense there is anything inside a creature that corresponds to 'preferences'. Often when people posit preferences, the preferences are defined in terms of behavior. Does it make sense to discuss different possible 'internal' preferences, distinct from behavior? I find it helpful to consider the behavior and 'preferences' of groups:

Suppose two cars are parked in driveways, each containing a couple. One couple are just enjoying hanging out in the car. The other couple are dealing with a conflict: one wants to climb a mountain together, and the other wants to swim in the sea together, and they aren't moving because neither is willing to let the outing proceed as the other wants. 'Behaviorally', both cars are the same: stopped. But their internal parts (the partners) are importantly different. And in the long run, we expect different behavior: the car with the unconflicted couple will probably stay where it is, and the conflicted car will (hopefully) eventually resolve the conflict and drive off.

I think here it makes sense to talk about internal parts, separate from behavior, and real. And similarly in the single agent case: there are physical mechanisms producing the behavior, which can have different characteristics, and which in particular can be 'in conflict'—in a way that motivates change—or not. I think it is also worth observing that humans find their preferences 'in conflict' and try to resolve them, which suggests that they at least are better understood in terms of both behavior and underlying preferences that are separate from it.

So we have: even if you can excuse any seizing as consistent with coherence, coherence incentives still exert a force on creatures that are *in fact* incoherent, given their real internal state (or would be incoherent if created). At least if they or their creator have machinery for noticing their incoherence, caring about it, and making changes.

Or put another way, coherence doesn't exclude overt behaviors alone, but does exclude combinations of preferences, and preferences beget behaviors. This changes how specific creatures behave, even if it doesn't entirely rule out any behavior ever being correct for some creature, somewhere.

That is, the coherence theorems may change what behavior is *likely* to appear amongst creatures with preferences.

Reform for coherence probably makes a thing more goal-directed Rohin

Ok, but moving toward coherence might sound totally innocuous, since, per Rohin's argument, coherence includes all sorts of things, such as absolutely any sequence of behavior.

But the relevant question is again whether a coherence-increasing reform process is likely to result in some kinds of behavior over others, probabilistically.

This is partly a practical question—what kind of reform process is it? Where a creature ends up depends not just on what it incoherently 'prefers', but on what kinds of things

its so-called ‘preferences’ are at all¹⁰, and what mechanisms detect problems, and how problems are resolved.

My guess is that there are also things we can say in general. It’s too big a topic to investigate properly here, but some initially plausible hypotheses about a wide range of coherence-reform processes:

- 1. Coherence-reformed entities will tend to end up looking similar to their starting point but less conflicted**

For instance, if a creature starts out being indifferent to buying red balls when they cost between ten and fifteen blue balls, it is more likely to end up treating red balls as exactly 12x the value of blue balls than it is to end up very much wanting the sequence where it takes the blue ball option, then the red ball option, then blue, red, red, blue, red. Or wanting red squares. Or wanting to ride a dolphin.

(I agree that if a creature starts out valuing Tuesday-red balls at fifteen blue balls and yet all other red balls at ten blue balls, then it faces no obvious pressure from within to become ‘coherent’, since it is not incoherent.)

- 2. More coherent strategies are systematically less wasteful, and waste inhibits goal-direction_{Rohin}, which means more coherent strategies are more forcefully goal-directed_{Rohin} on average**

In general, if you are sometimes a force for A and sometimes a force against A, then you are not moving the world with respect to A as forcefully as you would be if you picked one or the other. Two people intermittently changing who is in the driving seat, who want to go to different places, will not cover distance in any direction as effectively as either one of them. A company that cycles through three CEOs with different evaluations of everything will—even if they don’t actively scheme to thwart one another—tend to waste a lot of effort bringing in and out different policies and efforts (e.g. one week trying to expand into textiles, the next week trying to cut everything not involved in the central business).

- 3. Combining points 1 and 2 above, as entities become more coherent, they generally become more goal-directed_{Rohin}.** As opposed to, for instance, becoming more goal-directed_{Rohin} on average, but individual agents being about as likely to become worse as better as they are reformed. Consider: a creature that values red balls at 12x blue balls is very similar to one that values them inconsistently, except a little less wasteful. So it is probably similar but more goal-directed_{Rohin}. Whereas it’s fairly unclear how goal-directed_{Rohin} a creature that wants to ride a dolphin is compared to one that wanted red balls inconsistently much. In a world with lots of balls and no possible access to dolphins, it might be much less goal-directed_{Rohin}, in spite of its greater coherence.

- 4. Coherence-increasing processes rarely lead to non-goal-directed_{Rohin} agents—like the one that twitches on the ground**

In the abstract, few starting points and coherence-motivated reform processes will lead to an agent with the goal of carrying out a specific convoluted moment-indexed policy without regard for consequence, like Rohin’s twitching agent, or to valuing the sequence of history-action pairs that will happen anyway, or to being indifferent to everything. And these outcomes will be even less likely in practice, where AI systems with anything like preferences probably start out caring about much more normal things, such as money and points and clicks, so will probably land at a more consistent and shrewd version of that, if 1 is true. (Which is not to

say that you couldn't intentionally create such a creature.)

These hypotheses suggest to me that the changes in behavior brought about by coherence forces favor moving toward goal-directedness_{Rohin}, and therefore at least weakly toward risk.

Does this mean advanced AI will be goal-directed_{Rohin}?

Together, this does not imply that advanced AI will tend to be goal-directed_{Rohin}. We don't know how strong such forces are. Evidently not so strong that humans¹¹, or our other artifacts, are whipped into coherence in mere hundreds of thousands of years¹². If a creature doesn't have anything like preferences (beyond a tendency to behave certain ways), then coherence arguments don't obviously even apply to it (though discrepancies between the creature's behavior and its makers' preferences probably produce an analogous force¹³ and competitive pressures probably produce a similar force for coherence in valuing resources instrumental to survival). Coherence arguments mark out an aspect of the incentive landscape, but to say that there is an incentive for something, all things equal, is not to say that it will happen.

In sum

- 1) Even though any behavior could be coherent in principle, if it is not coherent in combination with an entity's internal state, then coherence arguments point to a real force for different (more coherent) behavior.
 - 2) My guess is that this force for coherent behavior is also a force for goal-directed behavior. This isn't clear, but seems likely, and also isn't undermined by Rohin's argument, as seems commonly believed.
- .



Two dogs attached to the same leash are pulling in different directions. [Etching by J. Eyt, 1642](#)

Direct effects matter!

This is a linkpost for <https://aaronbergman.substack.com/p/direct-effects-matter>

Note: this is from my personal blog [here](#).

A strange phenomenon plagues public discourse. Subtle and largely detached from the culture war, it often manages to evade detection. Can you spot it in each of the following arguments or discussion points?

1. Drugs like cocaine and heroin are bad. They are ruinous to users' health, and use imposes a large negative externality on society.
2. We should decriminalize drugs. There is no reason to tear families apart and ruin lives for the sake of regulating consciousness.
3. Most American adults should exercise more. Exercise has a plethora of mental and physical health benefits, after all, and most adults are sedentary.
4. This expansion of unemployment insurance is bad. It creates distortionary incentives and will reduce workforce participation.

All four arguments completely ignore the first-order, direct effects of the practice in question!

Drug use *and* the war on drugs might impose serious costs on society, but we can't forget that drugs themselves are fun to use! Exercise may be very important to personal wellbeing, but don't ignore that (for some people, some of the time) exercise sucks! And we can talk about the second-order consequences of any sort of social welfare program, but let's not neglect that the money *directly* improves poor people's lives!

Saying this out loud sounds trivial, almost stupid. But I think "first-order effect neglect," as I'll call it, is a serious issue with public discourse.

When was the last time...

When was the last time that you heard a debate over the merits of drug legalization consider the stupidly-obvious fact that drugs make people happy? I don't know if I ever have! Consider this list from the [first result](#) of my Google search for "pros and cons drug decriminalization."

Now that Portugal's decriminalization process is over a decade old, there are several long-term benefits that have been recognized, including the following:

- Substance abuse and addiction rates have been cut in half since decriminalization
- Addiction treatment and rehabilitation is less expensive than incarceration
- Individuals with substance abuse problems are much more likely to find recovery in rehab than in jail
- People completing treatment can become productive members of society much more easily than convicted felons
- Violence related to drug trafficking is greatly reduced
- Courts are freed up for other important work
- The rebellious, countercultural essence of drug use is changed when society sees it as a disease and not a crime²

These are all valid and important points to consider, but nowhere is a mention of the most direct and obvious effect of drug use: if 10% more people smoke weed because of decriminalization, that's a lot in chemical-induced pleasure!

Some public discussions have only a mild case of the disease. Consider Biden's COVID relief package, as well as the larger discourse about public assistance in America. I'm sure there are better examples, but I managed to find this tweet pretty quickly.



justine says join your union
@kvetchings

...

Turns out the solution to reducing poverty was just giving people money the whole time

5:33 PM · Mar 10, 2021 · Twitter for iPhone

59 Retweets 4 Quote Tweets 489 Likes



As an observer of the online liberal-technocrat-effective altruism nexus, I see points like this made on Twitter or elsewhere from time to time. But—as the tweet's sarcastic tone indicates

—the Discourse too often neglects the simple, obvious, and direct positive impacts of giving people money.

Am I being unfair? Take a look at this list from the [first result](#) of my Google search for “pros and cons of basic income”

✓ Pros

- Workers could afford to wait for a better job or better wages
- People would have the freedom to return to school or stay home to care for a relative
- May help remove the "poverty trap" from traditional welfare programs
- Citizens could have simple, straightforward financial assistance that minimizes bureaucracy
- The government would spend less to administer the program than with traditional welfare
- Young couples would have more money to start families in countries with low birth rates
- The payments could help stabilize the economy during recessionary periods

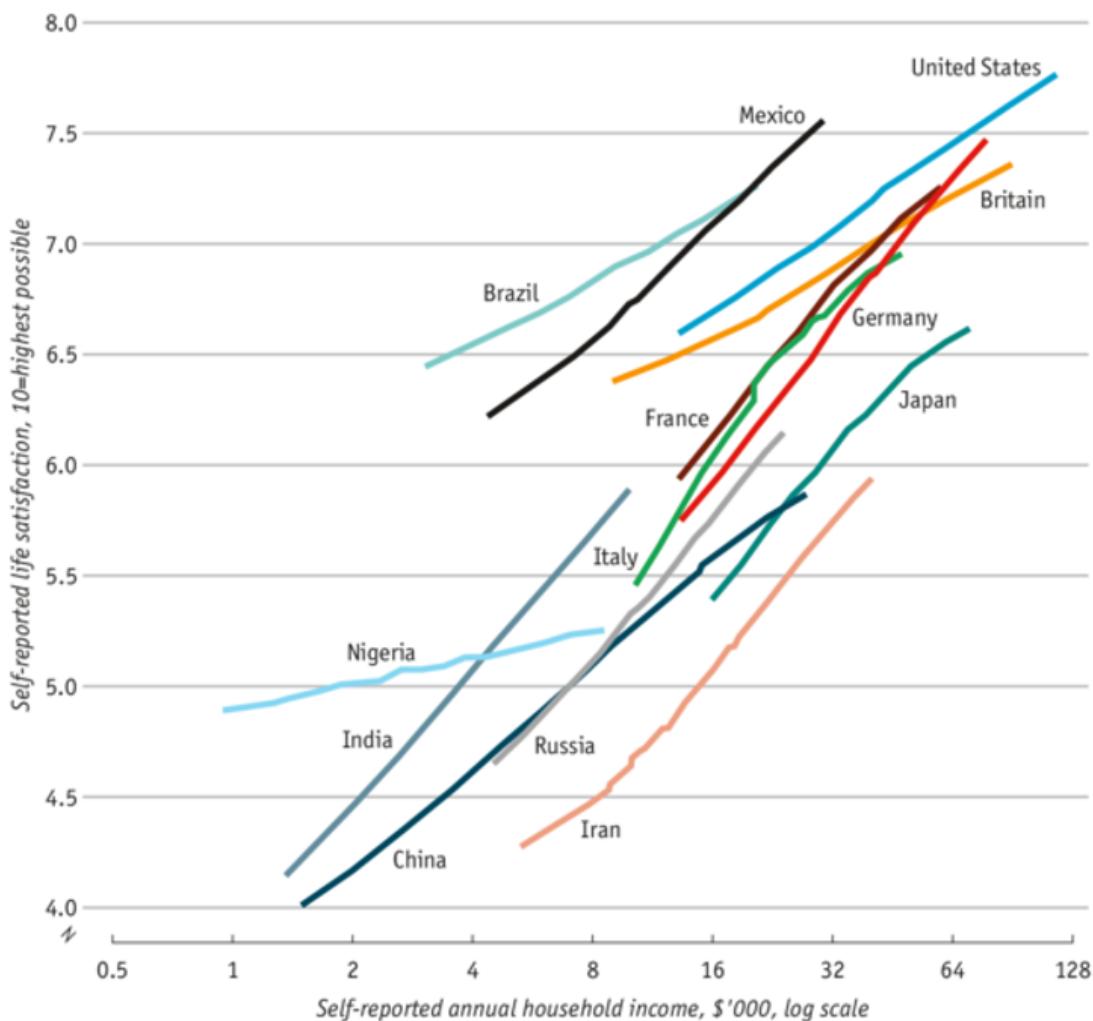
✗ Cons

- Inflation could be triggered because of the increase in demand for goods and services
- There won't be an increased standard of living in the long run because of inflated prices
- A reduced program with smaller payments won't make a real difference to poverty-stricken families
- Free income may not incentivize people to get jobs, and could make work seem optional
- Free income could perpetuate the falling labor force participation rate
- There are many opposed to handouts for the unemployed

At the risk of beating a dead horse, there is no mention of the fact that money makes people better off, so giving them money is good. This is *sort of* gestured at in the second benefit listed, “people would have the freedom to return to school or stay home to care for a relative,” but even this point associates the benefit with doing something economically productive (education) or helping someone *other* than the recipient (caring for a relative). Nowhere on the list does it say “people can now buy things they like.”

Life satisfaction and income

2012 or latest



Source: "Subjective Well-Being and Income: Is There Any Evidence of Satiation?",
by Betsey Stevenson and Justin Wolfers. NBER Working Paper 18992. April 2013

Economist.com/graphicdetail

What's going on?

It is important to note that “first-order effect neglect” of some policy or idea is by no means limited to that idea’s opponents. In other words, it can’t be explained solely by ideological incentives. So, what are some other possible explanations?

1. Fish in water (or humans in air)

For one thing, first order effects might seem so obvious that explicit recognition would appear awkward or forced. The “water we swim in” analogy isn’t perfect, since fish (presumably) don’t have any concept of water.

Humans, though, have an understanding of air. We implicitly know that we’re surrounded by a gaseous mix of some sort even when we are not explicitly considering the fact.

Nonetheless, it requires some sort of unusual stimulus to regard the air as a thing in and of itself. Ask someone to enumerate the things in a room, and “air” will likely fail to make the list. Even when the air is a vehicle for some physical sensation, such as cold, we still may think of “cold” itself as the relevant object itself instead of as a property of the air.

I think this likely explains a lot of the problem. First-order effects are the air we walk in. *Of course* drugs feel good. *Of course* exercise is hard. You don’t need to *tell me that*.

But banality can beget its own destruction. When everyone knows (or simply believes) that everyone else knows something (who in turn know that everyone else knows, *ad infinitum*), there is no direct reason for a person to state the fact. Humans, though, aren’t omniscient agents with perfect memories. If some fact remains unstated for too long (because of its mutually-understood “obviousness,”) it can drift out of some people’s—or everyone’s—conscious consideration.

2. Signalling

No blog post would be complete without a more cynical explanation. I don’t think this one is entirely disentangled from the former - rather, this might be the mechanism by which the “humans in air” phenomenon occurs. There are couple different points here:

Coming up with indirect, non-obvious effects of some policy or idea signals intelligence or wit.

Explicitly stating obvious, first-order effects might signal social ineptitude and/or lack of wit.

The first point is pretty self-explanatory; you don’t get smart points for saying “money lets people buy things,” but you *do* get smart points for saying “a basic income decreases downside risk associated with entrepreneurship, so we should expect it to boost socially-productive innovation and risk taking.”

The second is more subtle, For one thing, stating the obvious might indicate that you *don’t* know the direct effect is “common knowledge.” Think of a math major working on homework with a friend who says out loud “since $3x=15$, we can divide both sides by 3 which means that $x=5$.” Sounds kinda dumb, right? Math majors are *supposed* to know that the mechanics of this operation are trivial and obvious.

This itself can have two implications

A lack of *social* awareness about what is common knowledge (that ‘everyone knows we can divide both sides by 3’)

A lack of direct knowledge/intelligence/wit (that the operation in question is trivial to the speaker himself).

Countersignaling

My favorite part of our signaling-centric-psychology is [countersignaling](#).

If *not* stating the obvious signals some amount of wit, stating the obvious might come to signal that a person is so self-assured of his social status and intelligence that he isn’t worried about coming across as dumb or inept. Further, at some point the “obvious” first order effects stop being obvious, so stating them is a sort of direct signal of independent thinking.



Donald Trump signaling and Warren Buffet countersignaling with their houses.

That's what I'm taking advantage of with this blog post! At least consciously, I am intending to do more of the "obvious' things aren't obvious anymore" thing instead of the "I'm so secure that I can state something that is genuinely obvious to everyone" countersignal, but there might be some of that going on as well.

3. Domains of respectability

A third explanation involves the fact that some types of effects or values are considered more important/legitimate or higher status than others. The pleasure people get from heroin is regarded as lower and less valuable than the social benefits associated with reducing mass-incarceration, regardless of which is larger in an absolute utilitarian sense.

I'm not quite sure why, but it seems that simple, first-order effects are often indeed lower status than higher-order effects. Perhaps it has something to do with direct effects often

affecting people at the individual level, whereas secondary effects are more likely to affect communities and societies.

Once again, I am unsure how entangled this is with the last two explanations; it doesn't seem entirely distinct. Focusing on high status second-order concerns might be a signal that one has lofty, respectable values, but it also might just be a more directly effective argument in favor of one's point.

Conclusion

I've relied heavily on the examples of drug use, exercise, and public cash assistance, but I can think of more.

D.C. statehood being considered solely in terms of its effect on electoral politics rather than in terms of whether its citizens should have political representation, a point made by countersignaling-extraordinaire Ezra Klein.



Ezra Klein  @ezraklein · Sep 21, 2020

Statehood for DC and Puerto Rico isn't a punishment Democrats could mete out against Republicans. Statehood for DC and Puerto Rico is the right thing to do because US citizens deserve political representation.

vox.com/21448513/democ...

Furious Democrats are considering total war — profound changes to two branches of government, and even adding stars to the flag — if Republicans jam through a Supreme Court nominee then lose control of the Senate.

On the table: Adding Supreme Court justices ... eliminating the Senate's 60-vote threshold to end filibusters ... and statehood for D.C. and Puerto Rico. "If he holds a vote in 2020, we pack the court in 2021," Rep. Joe Kennedy III (D-Mass.) [tweeted](#).

238 2.6K 12.1K

2. Universal healthcare being considered in terms of its effect on total healthcare spending, aggregate public health, or labor market fluidity instead of whether it would directly improve people's lives ([proof](#)).

3. Neglecting the intrinsic value of hobbies/activities of any kind, as with this poster:

THE Benefits Of Tennis

Physical

LONGEVITY: A Harvard study of 10,000 people over 20 years said play three hours of tennis a week and you will cut your risk of death in half from any cause.

CALORIES AND WEIGHT: A Mayo Clinic says an hour of singles can burn 580 to 870 calories.

HEART: A Cleveland Clinic says tennis is an "ideal sport for a healthy heart." A Johns Hopkins study showed middle-aged men who played tennis, more than any other activity, had a significantly lower incidence of cardiovascular disease as they aged.

MUSCLES: Tennis develops and strengthens muscles.

BONE STRENGTH AND DENSITY: Weight-bearing activities are important for bone health, according to the NIH.

FLEXIBILITY: Out of 60 sports, ESPN ranked tennis in the top 12 for flexibility.

BALANCE, COORDINATION & REACTION TIME: Moving and adjusting to hit the ball improves coordination.

Social

POSITIVE PERSONALITY AND FITNESS DEVELOPMENT:

Tennis out-performs all other sports in developing positive personality characteristics, according to Dr. Jim Gavin, an author at Concordia University.

SOCIAL SKILLS: Tennis develops social skills for people of all ages and abilities.

IMPORTANT SOCIETAL VALUES: Tennis teaches the values of fair play and sportsmanship.



Mental

CONTINUING DEVELOPMENT OF THE BRAIN: A University of Illinois study finds that by requiring alertness and tactical thinking, tennis generates new connections between nerves in the brain and promotes a lifetime of continuing development.

CRITICAL THINKING AND PROBLEM SOLVING:

These activities are a major part of tennis and they keep your brain active.

HARD WORK AND SELF-DISCIPLINE: Tennis reinforces the value of hard work.

MANAGING MISTAKES: Properly managing mistakes is a trait that is critical in all aspects of life.

MANAGING ADVERSITY: Managing adversity is a trait that is critical in building leadership skills.

MANAGING AND REDUCING STRESS:

The physical, mental and emotional challenges of tennis help increase your overall capacity to deal with stress.

FAMILY: Tennis is a great activity for the whole family.

COMMUNICATION SKILLS: Tennis develops teamwork and communication skills, whether through doubles, or through league or school teams.

SELF-IMAGE: Tennis players scored higher in vigor, optimism and self-esteem, while scoring lower in depression, anger, confusion, anxiety and tension, than other athletes or non-athletes, according to a study by Southern Connecticut State University.

#TennisBenefitsLife



Is there any hope that we can elevate the salience of direct effects? I think so!

We're already seeing some improvements in the "give people cash" discourse as I described above. As folks realize that "obvious," "common knowledge" things are no longer so obvious,

emphasizing direct effects will become more directly appealing as an argumentative device
and will come to signal wit instead of ineptitude.

So, take advantage of countersignaling and make some stupidly-obvious points that aren't getting made!

Covid 3/25: Own Goals

AstraZeneca has made quite the mess of things. First they screwed up their initial studies in ways that kind of boggle the mind. Then, with the studies designed to repair trust, fix the problem and allow approval, they report incomplete results in order to make themselves look better, even though inevitably they were caught doing this within a day – it's pretty inevitable that you'll be caught when you do something in public that someone already warned you not to do, especially when that someone is also the regulatory authority. Oops.

In addition to AZ's own goals, health officials continue to score additional own goals around the whole issue of blood clots (that don't exist, and wouldn't matter even if they did). Most but not all places have resumed vaccinations, but trust in the vaccine, and plausibly in all vaccines, is permanently damaged.

Those developments are infuriating, and also enlightening as to how the system of the world functions these days, but the main event remains the race between new strains and vaccinations.

In America the race is plausibly close. Cases are rising, and likely will continue to rise for several more weeks, especially if vaccination rates continue to stagnate. But that acceleration should start soon, and at an additional 3% protection per week that grows and compounds, the vaccinations won't take that long to turn the tide even if they don't accelerate much.

In Europe the race is not so close. Vaccinations are running far slower, with no short term hope for things to get much better. The recent own goals only made a bad situation worse, and in many European countries things are looking quite bad. Lockdowns are once again the order of the day in many places, most notably Germany, and yet the situation is getting rapidly worse, in some places reaching crisis proportions.

I spend a lot of focus on everything that's wrong with our vaccine efforts, but as with many such things it's equally vital to remember that things could be, and in almost all other places are, so much worse. We failed versus the standards of a fully functional civilization, but compared to the rest of Earth 2020, we passed with flying colors. We need to remember both results.

Let's run the numbers.

The Numbers

Predictions

Last week's prediction (WaPo numbers): Positivity rate of 4.3% (up 0.2%), deaths decline by 8%.

Result:

In the past week in the U.S....

New daily reported **cases rose 4.3% ↑**

New daily reported **deaths fell 11.7% ↓**

Covid-related **hospitalizations fell 4.1% ↓** [Read more](#)

Among reported tests, **the positivity rate was 4.6%**.

The **number of tests reported fell 29.9% ↓** from the previous week. [Read more](#)

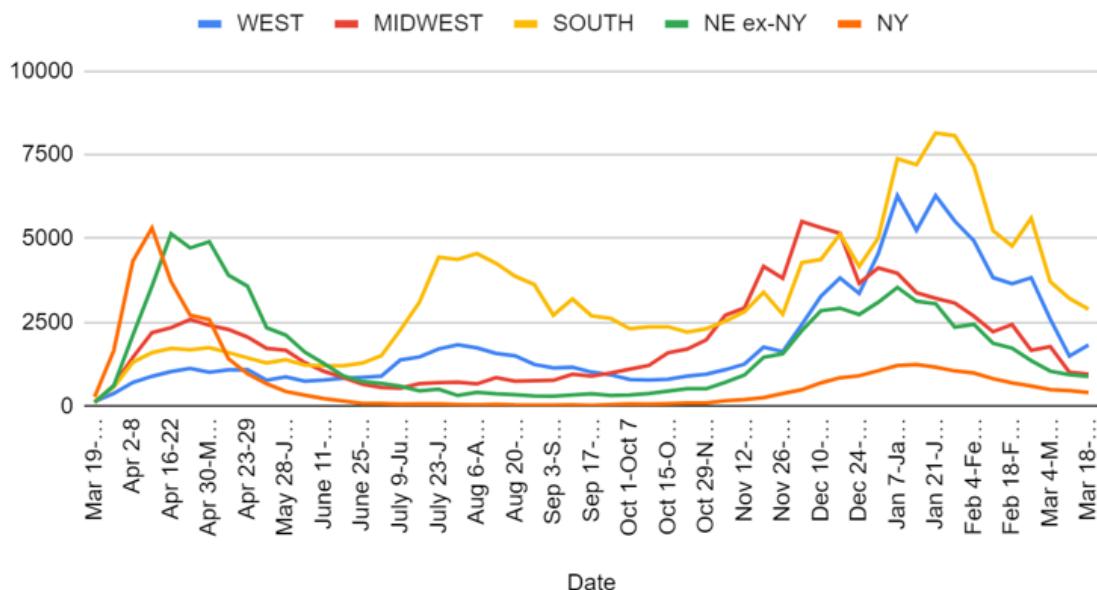
Pretty close. It's worth noting that these changes don't match the Wikipedia data, which only has a 2% or so drop in deaths. Based on regional changes I'd be inclined to believe the WaPo decline is more accurate.

The positivity rate was higher than I expected, and I presume that's mostly because the number of tests fell by almost 30%. That's a very scary drop to see while we're seeing a rise in cases in the same data set. It's much bigger than vaccinations could explain. What's going on there? I certainly hope we don't see another big drop until the need goes away.

Prediction for next week: Positivity rate of 4.9% (up 0.3%) and deaths fall by 7%. Deaths should continue to decline due to lag, but that effect should shrink. Vaccinations continue to help, but in the short term I'd expect new strains to have the bigger impact, so the short term situation should get somewhat worse.

Deaths

Deaths by Region



Date

WEST MIDWEST SOUTH NORTHEAST TOTAL

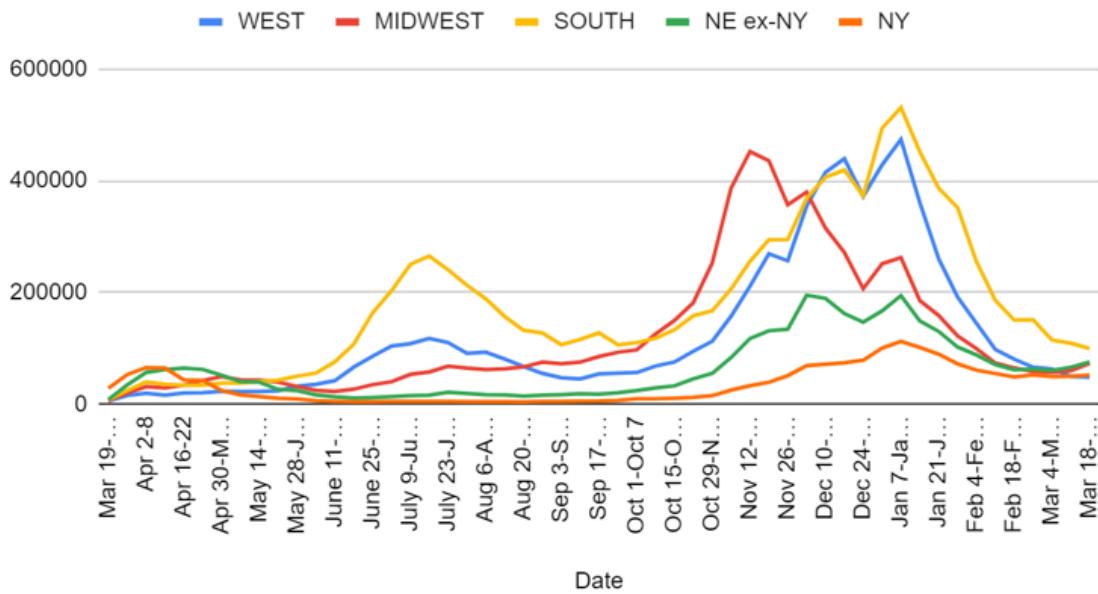
Feb 4-Feb 10	4937	2687	7165	3429	18218
Feb 11-Feb 17	3837	2221	5239	2700	13997
Feb 18-Feb 24	3652	2433	4782	2427	13294
Feb 25-Mar 3	3834	1669	5610	1958	13071
Mar 4-Mar 10	2595	1775	3714	1539	9623
Mar 11-Mar 17	1492	1010	3217	1402	7121
Mar 18-Mar 24	1823	957	2895	1294	6969

The change in the West seems like it has to be timeshifted data messing things up, with the declines in the other regions closely matching the WaPo declines. Deaths are on a solid trajectory downwards.

Positive Tests

Remember that the number of tests went way down this week, so this is a much bigger deal than it appears to be.

Positive Tests by Region

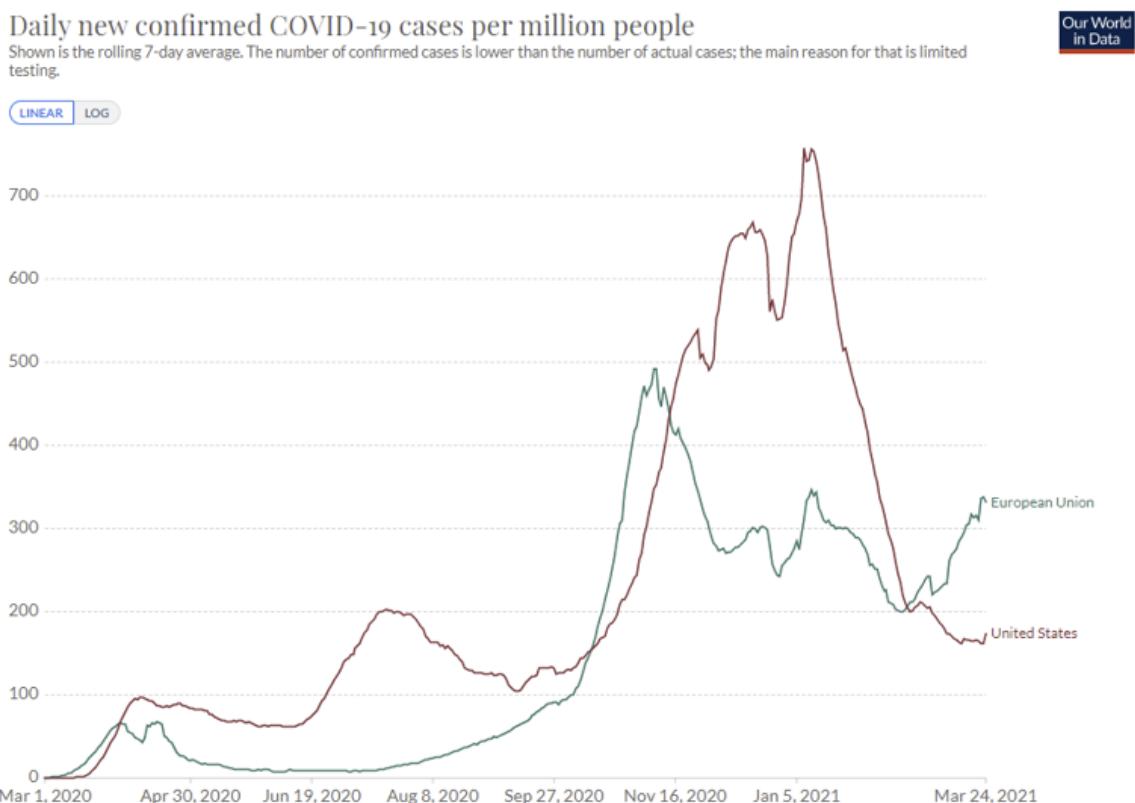


Date	WEST	MIDWEST	SOUTH	NORTHEAST
Feb 11-Feb 17	97,894	73,713	185,765	125,773
Feb 18-Feb 24	80,625	64,857	150,493	110,339
Feb 25-Mar 3	66,151	58,295	151,253	115,426
Mar 4-Mar 10	62,935	57,262	114,830	109,916
Mar 11-Mar 17	49,696	59,881	109,141	115,893
Mar 18-Mar 24	47,921	72,810	99,568	127,421

A number of states saw 20%+ rises in cases, including Michigan, Pennsylvania and Illinois.

It's also worth remembering, if you are unvaccinated, that these tallies include an increasing number of vaccinated people. Every week another 3% or so of the adult population gets vaccinated for what Bloomberg currently calls 20% overall vaccine coverage, and most of the infections are coming out of the remaining unvaccinated population. That's an extra 25% effective risk above the numbers for anyone still fully vulnerable, compared to numbers from last year, and that gap will continue to rise. Take action accordingly.

In Europe, the situation is rapidly getting worse despite strong lockdown efforts in many places. Rather than go with the previous hard-to-read multi-country graph, I'll show only the European Union:

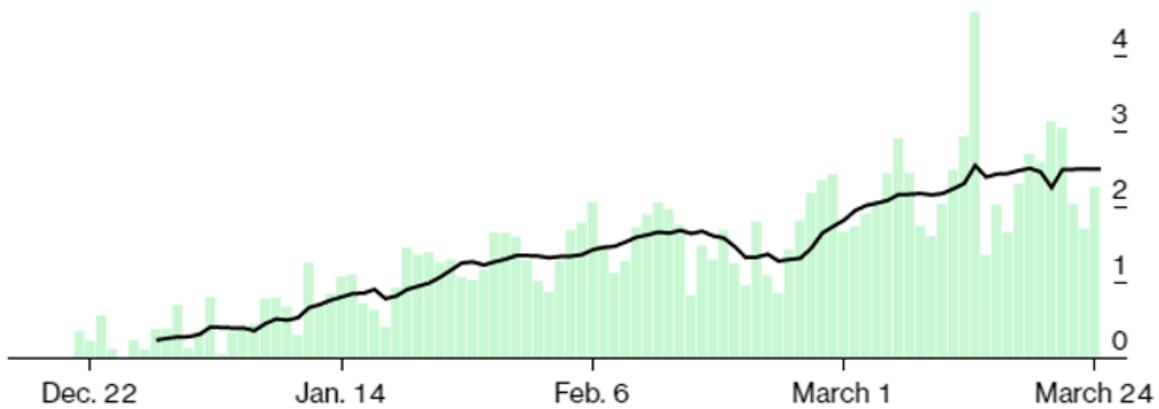


Vaccinations

In the U.S., more Americans have received at least one dose than have tested positive for the virus since the pandemic began. So far, **130 million doses** have been given. In the last week, an average of **2.49 million doses per day** were administered.

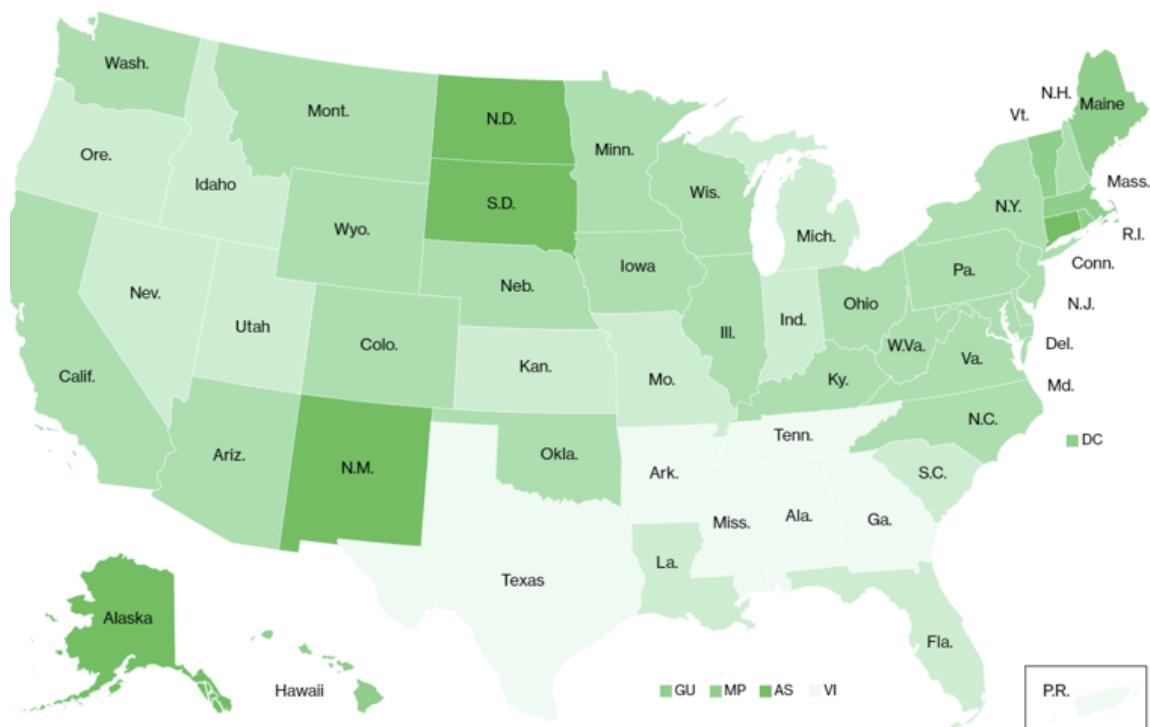
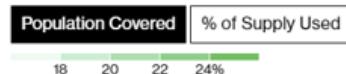
↗ Average daily rate estimate

Doses administered: 5M



Vaccines Across America

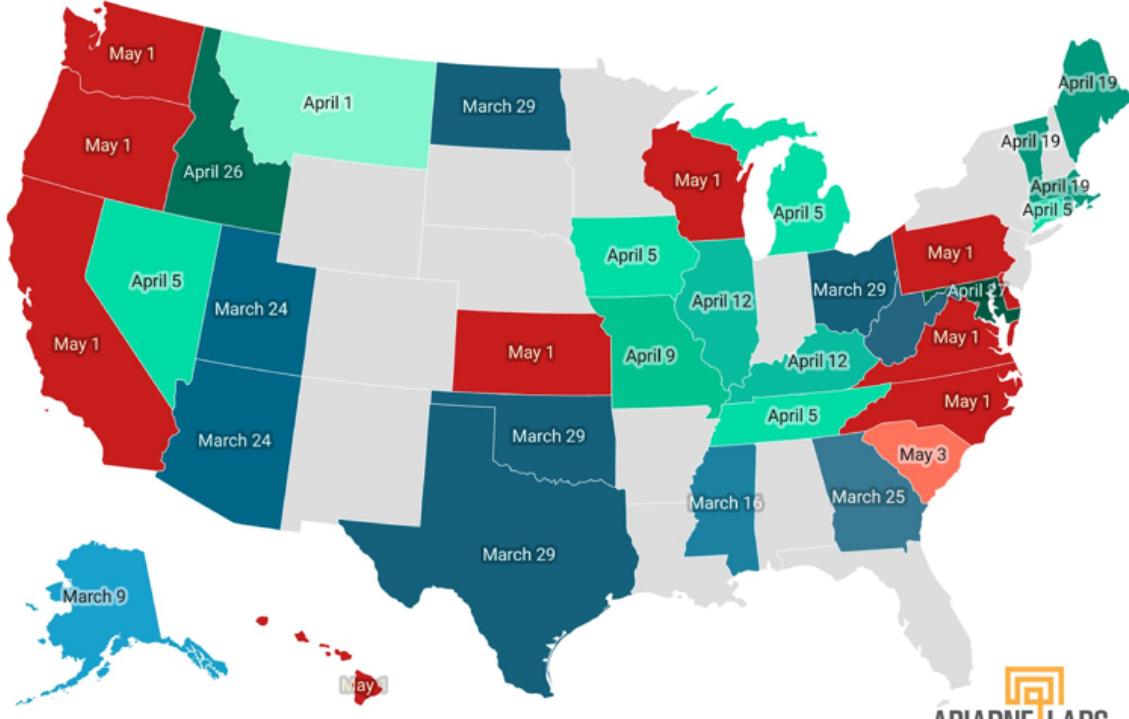
Across the U.S., enough doses have been administered to cover 20% of the population, and 77% of the delivered shots have been used



When can anyone who wants a vaccine fail to load a crashing website where they will eventually be able to book an appointment? Here's a handy guide.

States Opening Eligibility to All Adults

March 9 March 16 March 22 March 24 March 25 March 29 April 1 April 5 April 9
April 12 April 19 April 26 April 27 May 1 May 3



That means New York and Florida among other states so far have not heeded Biden's call for universal May 1 eligibility. My guess is that they'll do it anyway, but we'll see.

On a similar note, I'm super late to the party but in case you also missed it, this is great and all too accurate: [SNL's So You Think You Can Get The Vaccine](#).

[Here's a piece on vaccine access and the problems people have in practice trying to get an appointment](#), by journalists who booked thousands of appointments and wrote up their experiences. Many strong takeaways, including several important practical takeaways, including that mostly the exact time of your appointment does not matter at the big sites so long as you're there the same day. This is the way.

It does seem like appointment access is steadily improving. We have [anecdotes like this](#):



Matt Haney 
@MattHaneySF

...

A TON OF available appointments at Moscone right now.

There haven't been many in a while, so this good and needed.

Please sign up and spread the word: myturn.ca.gov

7:52 PM · Mar 24, 2021 · Twitter Web App

Also thought I'd signal boost this even though it's likely largely out of date:



Kelsey Piper @KelseyTuoc · Mar 20

...

If you're in Berkeley/Oakland and eligible for the vaccine, there are 1000s of slots available at the Berkeley mass vaccination site:
[curative.com/sites/24548#9/...](http://curative.com/sites/24548#9/) This site isn't bookable through MyTurn so MyTurn is saying there are no vaccines in your area while these sit open.

If you want the Johnson & Johnson shot, that's going less well. It seems that when officials said explicitly they have no plan, [that did not bode so well](#).



Walid Gellad, MD MPH  @walidgellad · Mar 22

...

Important story.

Although overall vaccinations are accelerating, the rollout of J&J has been a failure.

(A failure assuming speed is of the essence, and states/fed had months to plan).

Three weeks in, still >1million of the initial 3.9million shots have not been put in arms.

That's on top of J&J failing to produce that much supply, which is also not going well.

One source of that seems to be that a contract manufacturer Catalent they used *needed its own emergency use authorization* [which they only got this week](#). That was holding back millions of doses, and the delay plausibly killed thousands. And that saga isn't even over yet:

Emergent is already sending millions of doses to Catalent, the people said. But those shots cannot be used until Emergent receives its own FDA authorization. Catalent has not yet responded to questions about which company made the active ingredient for the doses it has begun shipping out. But a person familiar with the matter said J&J is currently using drug substance flown in from the Netherlands, not substance produced by Emergent, to manufacture its vaccine.

The whole debacle makes it clear that the FDA is causing far more delays in vaccine manufacturing and delivery with its red tape than one would guess only looking at the topline delays in authorizations.

Also, if you have gotten vaccinated, [don't forget to claim your free Krispy Kreme donuts](#). One free glazed any time you show your card, not even a one time deal. Being obese and thus at high risk for Covid-19 no longer matters to you, so it's doubly good! Remember, the key is not eating several additional donuts while you're there.

AstraZeneca Vaccinations Resume In Europe

The pause is over and vaccinations have resumed, after the EMA issued its official decision shortly after last week's post went up. This was a supremely overdetermined decision from a cost/benefit standpoint.

[Here's Kai's summary of the statement](#), which somehow manages to call for everyone to still be concerned and calls for us to Raise Awareness of a phantom danger, which will have the primary effect of making people not get vaccinated (and almost no secondary effects).



Kai Kupferschmidt ✅ @kakape · Mar 18

"The committee has come to a clear scientific conclusion: This is a safe and effective vaccine. Its benefits in protecting people from #COVID19 with the associated risks of death and hospitalization outweigh the possible risks", says Emer Cooke at [@EMA_News](#) press conference.

39

850

1.3K



Kai Kupferschmidt ✅ @kakape · Mar 18

"The committee also concluded that the vaccine is not associated with an increase in the overall risk of thrombo embolic events or blood clots", says Cooke.

5

108

235



Kai Kupferschmidt ✅ @kakape · Mar 18

"During the investigation and review we began to see a small number of cases of rare and unusual but very serious clotting disorders, and this then triggered a more focused review based on the evidence available", says Cooke.

3

36

166



Kai Kupferschmidt ✅ @kakape · Mar 18

"After days of in-depth analysis of lab results, clinical reports, autopsy reports and other information from the clinical trials, we still cannot rule out definitively a link between these cases and the vaccine", says Cooke.

8

88

217



Kai Kupferschmidt ✅ @kakape · Mar 18

"What the committee has therefore recommended is to raise awareness of these possible risks making sure that they're included in the product information, drawing attention to these possible rare conditions", says Cooke.

2

59

191



Kai Kupferschmidt ✅ @kakape · Mar 18

"We're also launching additional investigations to understand more about these rare cases, and we're conducting targeted observational studies", says Cooke.

As expected, elites that backed halting vaccinations in order to show how Very Serious they were got the Official Word right away and pivoted to explaining how important it was to be cautious for no reason, and how it's now important to resume giving life saving medicine

once the Official Word came down that the exact same information we had last week still contains no case whatsoever for halting vaccinations.

There seem to be two stories being told at once by the Very Serious. One is that the pause was necessary to assure people we were being Very Serious, and took potential risks seriously; if we didn't signal we were willing to move the trolley to the track with thousands of people on it, in order to maybe save the handful of people who *might* be on the other track, who would trust vaccines? The other case is that we couldn't possibly know anything until the Proper Authorities at the EMA say it *in the proper form*, so once safety had been thrown into official doubt, we needed to wait for Official Word to be sure we weren't doing something awful.

I'm glad the full pause is over. I am highly confident that this did far more damage to vaccine confidence than would have been done without a suspension, and I am also highly confident that elites will continue to insist that this is not the case no matter what happens, and that they will keep patting themselves on the back for their actions.

[I strongly endorse this strong prior:](#)



Rob Bensinger @robbensinger · 10h

Pretending that things are dangerous is not a good way to make people trust them!

...

[Here's some concrete data](#) on what's happened to vaccine hesitancy in Europe. I don't think [Nate's comment](#) is entirely fair, in the sense that we'd expect increasing hesitancy right now under any scenario, but still, magnitude here is large.



Nate Silver ✅ @NateSilver538 · Mar 22

One piece of evidence that when public agencies are irrationally risk-averse about vaccines in the hopes of averting vaccine hesitancy, it winds up only increasing vaccine hesitancy.

...



YouGov ✅
@YouGov

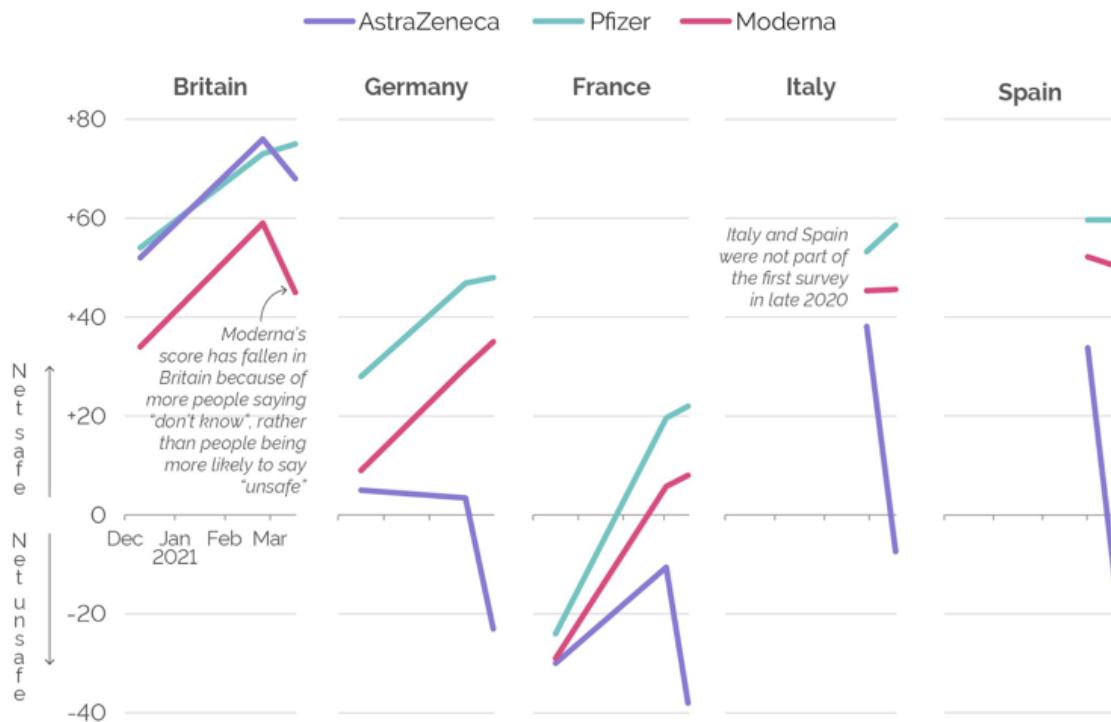
...

European confidence in AstraZeneca vaccine safety sinks after blood clot scare (changes in last 2/3 weeks)

- 🇫🇷 safe 23% (-10) / unsafe 61% (+18)
- 🇮🇪 safe 36% (-18) / unsafe 43% (+27)
- 🇩🇪 safe 32% (-11) / unsafe 55% (+15)
- 🇪🇸 safe 38% (-21) / unsafe 52% (+27)

Europeans now see AstraZeneca vaccine as unsafe, following blood clots scare

How safe, or unsafe, do you think the Pfizer-BioNTech / Oxford-AstraZeneca / Moderna vaccine is?
Figures shown are NET scores



YouGov

Latest data: 15-18 March 2021

Australia is trying the plausible-sounding gambit of [telling people who have explicit blood clot issues to avoid the AstraZeneca vaccine](#). Which has essentially no first-order object-level consequences, because such conditions are super rare:

"It's just precautionary, but these conditions are incredibly rare."

Professor Cheng said he didn't know how many people this advice would apply to, but "it wouldn't be many".

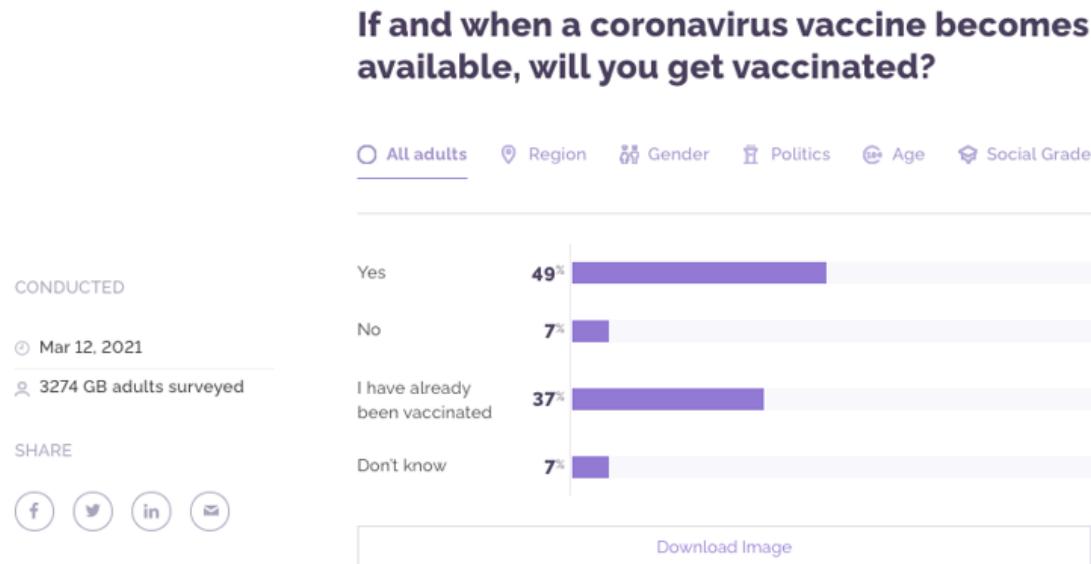
"I would imagine it could be a couple of hundred people, but I don't know for sure."

That's unlikely to kill anyone *directly* on net given Australia's situation, but what it does is send the clear message to everyone in Australia that the reasonable authority figures think

there's something wrong with the AstraZeneca vaccine.

Thus, I expect this approach, while far less bad than what Europe did, to still backfire.

[Whereas here's how it's going in the United Kingdom](#), which ignored the whole thing:



Lest we think this is fully over, [the madness will continue until morale improves](#) (WaPo), and I do not expect improved morale:

In a sign of more potential difficulties ahead, French Health Minister Olivier Véran said the country's health advisory body was now recommending AstraZeneca vaccinations only for people age 55 or older. Just weeks ago, the same authority lifted a recommendation that only adults under 65 should receive the AstraZeneca shot.

French officials cited the European Medicines Agency assessment, which identified younger women as potentially at increased risk for rare blood clots in the days after vaccination.

I can't blame the people for getting confused and skeptical when they get these kinds of messages. Model what such authorities are thinking, and what they care about, based on what they say and do.

Plus things are still halted in some places:

Still, Denmark, [Sweden](#) and [Norway](#) said they were not yet ready to resume their AstraZeneca inoculations. “We need time to get to the bottom of this,” [Soren Brostrom](#), director of the Danish Health Authority, said Friday.

Perhaps the most interesting part of the whole situation is the question of what is the ‘default’ action, which thus counts as inaction and is not blameworthy, versus what would constitute a non-default action, and thus be fully blameworthy for incurring even the theoretical risk of loss while getting zero credit for any gains? In other words, which track is the train already on when we [decide whether to flip the switch?](#)



Charles Kenny @charlesjkenny · Mar 17

Nobody is arguing continuing AZ shots in Europe will increase deaths. It *might* cause a few deaths while saving many. Stopping the program is pulling the lever to send the train down the tracks with lots of people tied to them because there *might* be a person on the other track

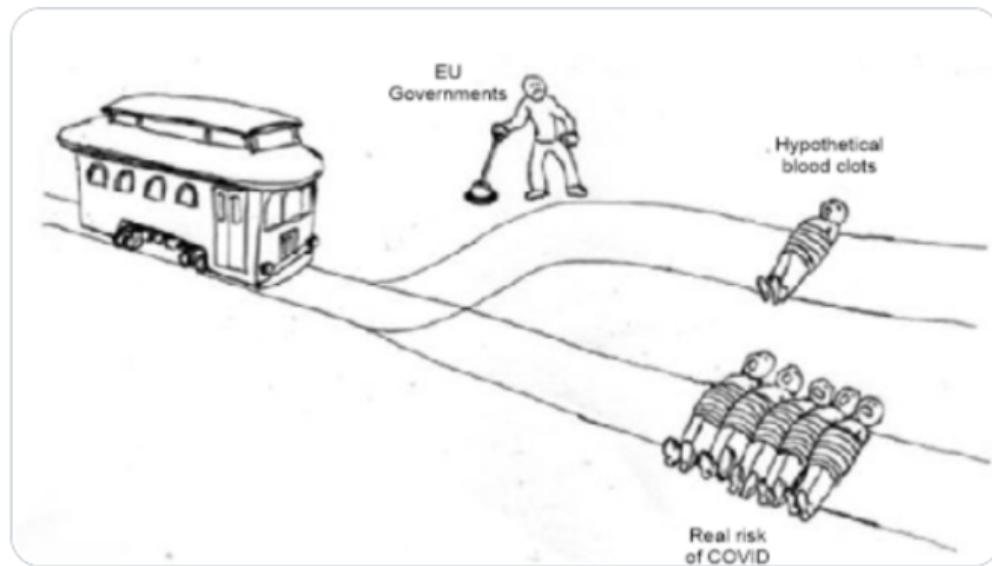
1

7

46



Alex Tabarrok @ATabarrok · Mar 17



If we could get the ‘stopping AZ is flipping the switch’ intuition to stick, then no one would have flipped the switch. Somehow, elite and ‘ethicist’ wisdom got so internalized that halting vaccinations over a phantom in the wind effectively became the default action.

Meanwhile, if you’re European and still aren’t sure what happened, [Johan Norberg provided a handy FAQ](#) during the suspension, which would now need to be slightly modified since vaccinations have resumed, but still seems worth sharing in full:

WHERE ARE EUROPE'S VACCINES? A GUIDE

Why haven't we been vaccinated already?

A: Because the European Commission procured vaccines centrally and wanted to show it could get better terms and lower prices, and that took time. UK signed with AstraZeneca three months earlier than the EU, for example, so it had another three months to fix glitches in production.

Did the EU get lower prices?

A: Yes – and more deaths and longer lockdowns that cost much more than costlier vaccines would. And there were glitches at AstraZeneca's factory in Seneffe, Belgium.

But doesn't AstraZeneca have another vaccine factory in Halix in the Netherlands?

Yes, but that factory does not have regulatory approval to supply Europe.

Gosh. But couldn't their US factories sell to Europe?

Yes, AstraZeneca has 30 million unused doses in a warehouse in Ohio that it wants to send to Europe, but the US does not allow it.

Fair enough, I guess President Biden wants the doses for Americans.

No, he bans Americans from using them, and the US regulatory process will take at least a month.

Oh dear, ok, just roll out the few vaccines we have in Europe then.

Sorry, European governments are suspending it now, since some who got the jab developed blood clots.

More people than would have developed blood clots in a similar population anyway?

No, fewer, actually, so there is "no indication" vaccines caused it, says the European Medicines Agency. But that's the precautionary principle for you. We don't accept any risks.

Because this potential side effect would be worse than Covid-19?

Oh no, we are talking of fewer than ten deaths from blood clots, while delaying Covid vaccination by just a week will probably kill thousands.

So we have to accept certain disease and death because you worry about a minimal risk of something that would in any case be much less dangerous?

Yes, we just want to keep you safe. You are welcome.

We also have [MR's take on Europe's vaccine efforts](#).

[And this is offered without further comment](#), beyond my congratulations:



Jason Groves @JasonGroves1 · Mar 19

...

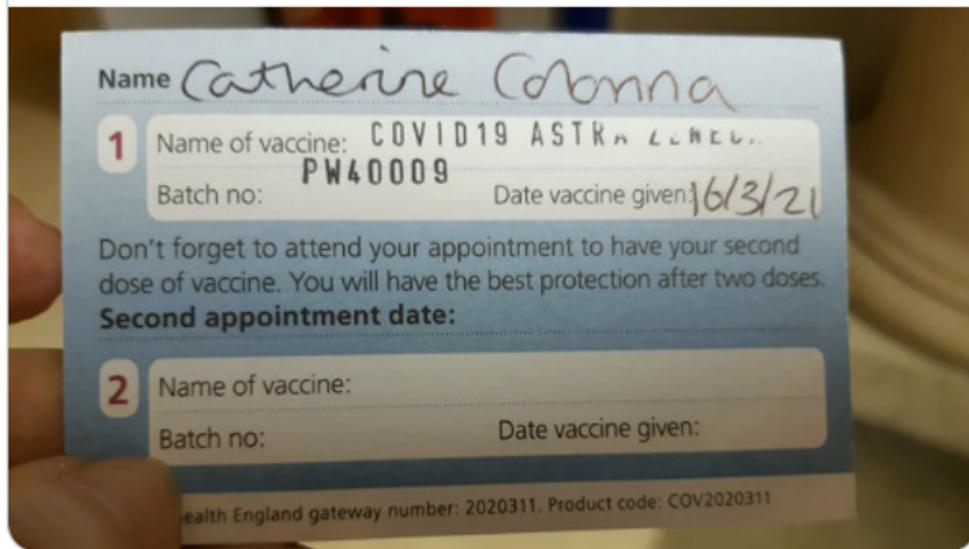
Looks like the French ambassador had the AZ jab three days ago, when its use was still suspended in France. Good for her



Catherine Colonna ✅ @AmbColonna · Mar 19

France government official

Done. Safely.



2

20

64



I Will Not Invoke [The Copenhagen Interpretation of Ethics](#)

With 30-60 million doses in reserve, and no intention of approving or using the AstraZeneca vaccine until after we have enough vaccines without it (we're getting closer than when I wrote this section, but not closer enough, see below), [we generously agreed to send 4 million of them overseas](#).



Jeff Mason ✅ @jeffmason1 · Mar 18

...

SCOOP: The U.S. plans to send 2.5 million doses of AstraZeneca COVID-19 vaccine to Mexico and 1.5 million doses to Canada, administration officials tell me. The loan does not affect [@POTUS](#) goal to have enough vaccine for all U.S. adults by the end of May.

186

1.3K

4.4K





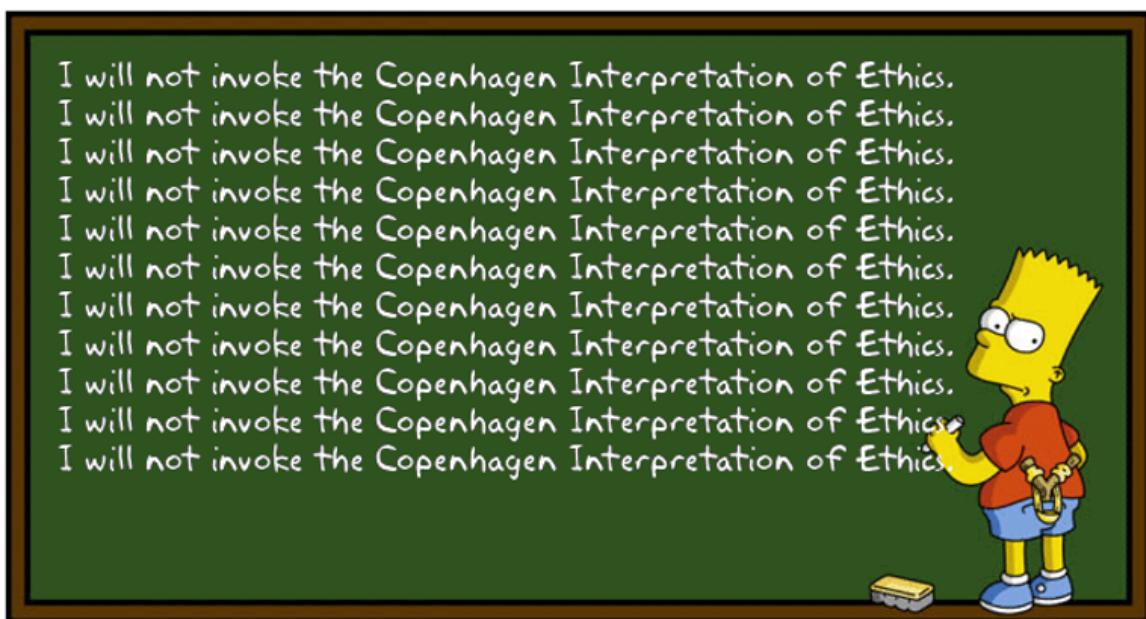
Steven Portnoy
@stevenportnoy

...

"If we have a surplus [of vaccines], we're going to share it with the rest of the world," Biden tells TV pooler @nancycordes.

3:53 PM · Mar 10, 2021 · TweetDeck

I am glad to see that we saw a problem, and in response we did something good and helpful. Excellent. Nothing wrong here. [Woo-hoo](#).



AstraZeneca Vaccine Once Again Found Safe and Effective?

It was looking to be very good news.

AstraZeneca's American trial results were in ([WaPo](#)). Sample size of 32,449. The headline numbers are 79% effectiveness against symptomatic Covid-19 and 100% (once again!) against severe disease and death, although that was only 5 cases versus 0 cases so it doesn't mean much.

Part of the justification for all this delay was the need to check various subgroups, both age and ethnicity. The study made extra effort to 'look like America' in these ways, and found essentially no variation between groups. This is of course rather silly, since if you *actually* cared mostly about such groups you'd want to *overrepresent* them, but theater, [like war](#), [never changes](#).

The drug company said, “vaccine efficacy was consistent across ethnicity and age. Notably, in participants aged 65 years and over, vaccine efficacy was 80%.”

As one would expect, blood clotting was examined, and they found nothing:

An independent monitoring board combed through the data to look for any cases of blood clotting events similar to those that caused the vaccination effort to be suspended in many European countries. While vaccination was started again after a pause, it undermined confidence in the vaccine. The independent board found no suggestion that the vaccine carried an increased risk of clotting.

We've managed to stall out the process long enough that we likely won't need the vaccine once it gets approved, even if nothing unexpected goes wrong from here:

The company said in its release it will apply for emergency authorization from the Food and Drug Administration “in the coming weeks,” but a question that has emerged recently is whether the United States, which President Biden has promised will have enough vaccine for all adults by the end of May, will need the vaccine. Fauci said it was too soon to know how it would fit into the U.S. vaccine portfolio

There were already real messaging issues with the AstraZeneca vaccine, as it is less effective than Moderna/Pfizer, and now the Europeans have muddied the waters further, to the point where introducing AstraZeneca in America might end up doing more harm than good via confusion and increasing overall vaccine reluctance. I'd still approve and use it if I could do so *now*, but if we're going to wait weeks for an application then weeks for a review, we could plausibly be far enough along that it makes sense to ship all our AZ doses overseas anyway.

If it gets to be May 1, and we make everyone suddenly eligible while also approving AZ and making available 60 million doses, it's going to be an interesting dilemma for unvaccinated people who can't find an appointment for a better vaccine right away. With a vaccine surplus likely imminent at that point, the advice ‘get any vaccine you can as soon as you can’ no longer would obviously apply even to those who know the ‘risks’ involved are pure phantoms.

So, about that ‘nothing goes wrong from here’ line, [there's one little problem](#).

NIAID Statement on AstraZeneca Vaccine



Late Monday, the Data and Safety Monitoring Board (DSMB) notified NIAID, BARDA, and AstraZeneca that it was concerned by information released by AstraZeneca on initial data from its COVID-19 vaccine clinical trial. The DSMB expressed concern that AstraZeneca may have included outdated information from that trial, which may have provided an incomplete view of the efficacy data. We urge the company to work with the DSMB to review the efficacy data and ensure the most accurate, up-to-date efficacy data be made public as quickly as possible.

My consistent line on companies releasing their trial data has been that one can assume it's right because if it isn't right then officials will quickly figure this out and it would be a disaster within the two week blame horizon, so there's no reason for them not to get it right.

AstraZeneca has decided to test that theory on multiple fronts. The part where we can trust the initial release was decisively disproven. The part where any deception would be quickly spotted seems strongly confirmed.

The mystery is how in the world AstraZeneca thought they would ever get away with this, especially on top of both the botched initial studies and the recent ongoing (completely unfounded and undeserved) doubts involving blood clots.

Or, alternatively, did Oxford really find a pharmaceutical company so incompetent that they did this by mistake, on top of giving an entire trial segment the wrong dose of vaccine the first time around? These are some rather epic screwups.

The safety board *outright told AZ* that they should be using a lower safety estimate, and AZ went ahead and used the higher number anyway ([from WaPo](#)).

The letter goes on to explain that while the company announced its vaccine was 79 percent effective on Monday, the panel had been meeting with the company through February and March and had seen data showing the vaccine may be 69 to 74 percent effective, and had “strongly recommended” that information should be included in the news release.

There was a known actual 0% chance this would work.

[The widespread response seems to similarly be disbelief that AZ could be this stupid.](#)



Nate Silver @NateSilver538 · 21h

I'm just sort of gobsmacked at why AstraZeneca thought this was a good idea. (I don't think I've ever used the word gobsmacked before but that's just how gobsmacked I am.)

...



Helen Branswell ✅
@HelenBranswell

...

I am rarely speechless. This turn of events has rendered me speechless.
What a debacle.



Helen Branswell ✅ @HelenBranswell · 21h

...

1. Turning this into a thread about the extraordinary public rebuke of AstraZeneca today by the DSMB* of its US clinical trial.
(*data & safety monitoring board, independent experts who oversee the trial & monitor the data)



Helen Branswell ✅ @HelenBranswell · 21h

...

2. The [@washingtonpost](#) got or saw a copy of the letter. It's clear the DSMB told AZ the vaccine efficacy estimate they should be using was between 69% - 74%. The company used data from an interim analysis that pointed to 79%, didn't mention the more recent data.

3. This quote from the Post piece is a killer. Releasing the 79% figure — while knowing it was out-of-date — was like telling your mother you got an A in a course when really you got an A on one test, but a C overall.
OUCH!

Post story: [washingtonpost.com/world/astrazen...](https://www.washingtonpost.com/world/astrazen...)

Federal officials were taken aback by the letter from the board. One said the AstraZeneca results were the equivalent of “telling your mother you got an A in a course, when you got an A in the first quiz but a C in the overall course.” Another said the disclosure by the board would inevitably hurt the company’s credibility with U.S. regulators.

HEALTH

'I was sort of stunned': Fauci and U.S. officials say AstraZeneca released 'outdated information' from Covid-19 vaccine trial



zeynep tufekci ✅ @zeynep · Mar 23

Fuming. I supported a US trial of AstraZeneca so it could undo the *unnecessary* damage to vaccine confidence from their botched initial rollout. Yesterday, I was cautious because, again, press release, no data. Turns out AZ is botching this rollout, too.



zeynep tufekci ✅ @zeynep · Mar 23

The level of irresponsibility on display here is unforgivable—especially since their vaccine seems fine! It's delivering in the real world, and the actual trial results are excellent! But they keep screwing around with confidence with their callous hubris.



zeynep tufekci ✅ @zeynep · 23h

Can't describe how angry I think we should be at AstraZeneca right now. I hope for an investigation on the whole thing from start, including what if any role/involvement Oxford had on the botched side of things. The vaccine is fine! People in charge need to face accountability.

This was worse than telling your mother you got an A in the course when you got a C overall and an A on one test. This is you telling your mother you got a C+ *after your father already knows you got a C-, told you that you'd better not pretend you got a C-, and there's no real consequences to you based on the slightly lower grade, and your father is standing right there during the conversation.*

The rest of that WaPo article and the rest of the Stat article flesh out the details, with a heartwarming amount of scrambling to stop this debacle from sinking the vaccine's approval or the public's trust in it. My model of how this would *normally* work is that they'd be looking for ways to punish AstraZeneca for this, which is quite reasonable given that the incentives involved have to be maintained. Instead, everyone is doing their best to admit they are stunned but essentially turn around and say 'la-la-la, I can't hear you, I'm sure you said the right number, right?' for practical purposes, and hope the public doesn't notice.

I agree strongly that we need to do a bunch of [whistling in the dark](#) and get this approved anyway, even if it's a little late to make that much difference, and to paper over the potential loss of trust in the vaccine. To the extent that I wrote a paragraph after this and then deleted it rather than give ammunition where it shouldn't be given.

We still need to punish AstraZeneca the company, and hard. The default outcome is of course this:



zeynep tufekci @zeynep · Mar 23

Yes, exactly this. There will likely be no accountability for a series of terrible and completely unavoidable actions. Next time we blame people for falling for misinformation, remember the role they played in contributing to this environment of mistrust.

negativemc @mathnegative · Mar 23

Replies to @zeynep

The only punishment AZN will face is lump sum fines paid via monopoly profits and a few embarrassing but mostly just boring congressional hearings.

It's the world that suffers and bears these costs.

The worst part of it all? The vaccine is good.

And that's completely unacceptable. The fine for this should be *freaking huge*. Enough to materially move the stock price. Heads at the company need to roll. And the next time they come to us for an approval, or issue a press release, we need to remember that this happened. No reputational bailouts, please.

What happened in the end, at least so far?

It looks like they're now going with 76% effective as their new story, and so far it's still sticking. That's not that different from 79% for any regulatory purposes, so the whole thing really was a pure own goal.

Have You Tried Increasing Supply?

For a while it's been clear that the world is capable of producing far more Covid-19 vaccine doses than it is producing. A few billion dollars would have bought us vastly greater production capacity, and it still could do so in time to impact the third world greatly, which would be highly worthwhile for direct impact and also selfishly wise for us to do purely in order to limit potential mutations. As Tyler Cowen puts it, trillions are bigger than billions.

This week [we have yet more evidence that there is lots of slack in the vaccine supply](#) (WaPo) if we collectively paid what it would take to use it.

(Also, you gotta admire the chutzpah of saying vaccine manufacturers are 'defending their monopolies' when many vaccines *still* sit ready to go, safe and effective yet unapproved, the rest were delayed for needless months, and while companies are forced to accept fixed monopsony prices from governments who refuse to pay for more capacity or speed, and who pause for investigation and cripple public faith in all vaccines when there are side effects at below the rate in the general population. I mean, yeah, wow.)

In particular, there are additional pharmaceutical firms capable of manufacturing vaccine doses if given the legal right to do so, and the holders of the relevant intellectual property

are saying no.

Except, from how I read this, they're not saying no because they want to preserve their patents – Moderna in particular already said they wouldn't enforce theirs, and I doubt anyone else would either no matter what the drug companies wanted.

They're saying no because *they don't want to teach other companies how to produce vaccines*. They're also concerned about quality control, and what would happen if another manufacturer made a mistake – it would doubtless still be a net huge win on first order grounds, but faith in vaccines in general, and in each company's product in particular, is fragile.

Which all makes perfect sense from a business standpoint. Their technical expertise is their entire business, so is the trust in their brands, and it all transfers between products.

Here are some choice quotes from that piece, in which cases are made from all sides, from which I believe the overall picture becomes clear:

Abdul Muktadir, the chief executive of Bangladeshi pharmaceutical maker Incepta, has emailed executives of Moderna, Johnson & Johnson, and Novavax offering his company's help. He said he has enough capacity to fill vials for 600 million to 800 million doses of coronavirus vaccine a year to distribute throughout Asia.

He never heard back from any of them. The lack of interest has left Muktadir worried about prolonged coronavirus exposure for millions of citizens of Bangladesh and other low-income nations throughout Asia and Africa who are at the back of the [global queue](#) for shots.

“Unfortunately, only limited, exclusive and often non-transparent voluntary licensing is the preferred approach of some companies, and this is proven to be insufficient to address the needs of the current COVID-19 pandemic,” the WHO said in response to questions from The Washington Post. “The entire population and the global economy are in crisis because of that approach and vaccines nationalism.”

Pfizer, which partnered with Germany's BioNTech, a company that received German subsidies, has predicted it will get \$15 billion from sales of its vaccine, an estimate that is considered conservative. Pfizer did not accept U.S. government funding.

Drug companies are lobbying the Biden administration to block a push at the WTO by India, South Africa and about 80 other countries for a temporary waiver on patent protections for the new vaccines. The pharmaceutical industry argues that innovation as well as vaccine quality and safety depend on maintaining exclusive intellectual property rights.

Pfizer, which says it plans to produce 2 billion doses of vaccine in 2021, has begun selling its vaccine directly to countries. The company said 36 percent of its production will be reserved for middle- and low-income countries, with nonprofit pricing baked in for the poorest nations.

Moderna did not comment on the conversation but referred to the October patent pledge. “Our patent pledge stated that, while the pandemic persists, Moderna will not use its patents to block others from making a coronavirus vaccine intended to combat the pandemic. There was no mention of a commitment to transfer our know-how beyond our chosen partners,” Moderna spokesman Ray Jordan said in an email.

“WHO criticism of industry is showing a lack of understanding for the complexity of vaccine manufacturing and global supply chain and a disrespect for the daunting challenge of literally trebling global vaccine capacity for one single disease almost overnight,” Thomas Cueni, director general of the International Federation of Pharmaceutical Manufacturers and Associations, said in an email.

“COVID-19 vaccine makers have been making agreements with other vaccine makers, wherever they are in the world,” he said. “Speed is of the essence; and for these relationships to be established quickly, you need trust, as well as a total shared commitment to the quality and safety of COVID-19 vaccines produced.”

If we cared enough, we could pay off the necessary stakeholders and make all of this happen, as an extreme upper bound all the vaccine manufacturers have a combined market cap below a trillion dollars most of which is unrelated to any of this, but like paying to expand capacity, we are choosing not to do so or even seriously consider doing so. There's still time to get the rest of the world vaccinated at a much faster pace than would otherwise happen, and in doing so save a lot of people and also dramatically lower our risk of mutations.

The Safe Distance For Children Has Always Been Three Feet

That is all.

Talk To Your Kids About Covid-19

An important fact about teenagers is that they mostly don't underestimate the risks from activities like drunk driving. Instead, they mostly *greatly overestimate* those risks, then often decide they don't care and do the stupid and dangerous stuff anyway, because they're teenagers. The typical response strategy is to misleadingly scare kids even *more* so that they'll end up acting more like we want them to act.

Why should Covid-19 be any different? [In a thread that's centered on explicit bothsidesism with regard to Covid mistakes](#), David Leonhardt drops this chart ([source](#) that has a lot of good data):



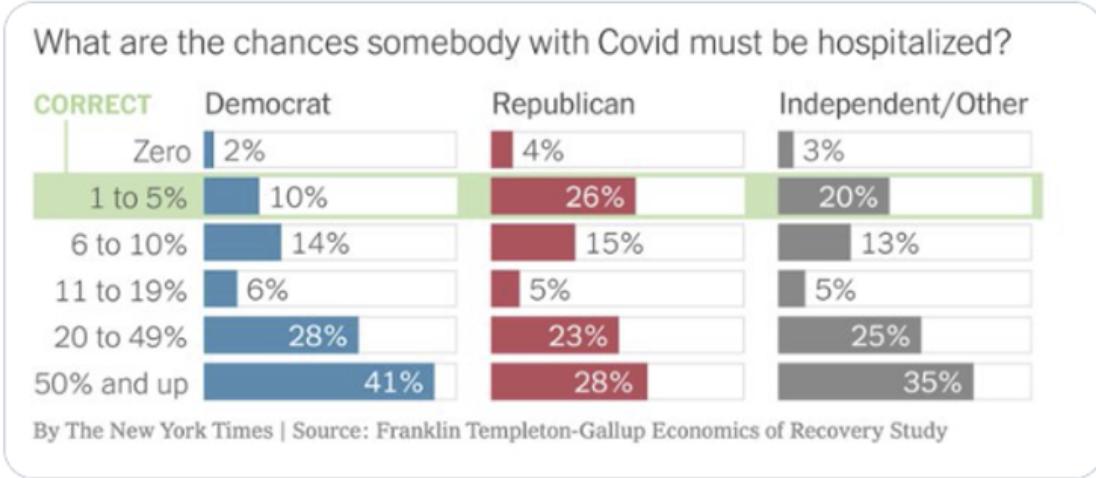
David Leonhardt

@DLeonhardt

...

Replies to @DLeonhardt

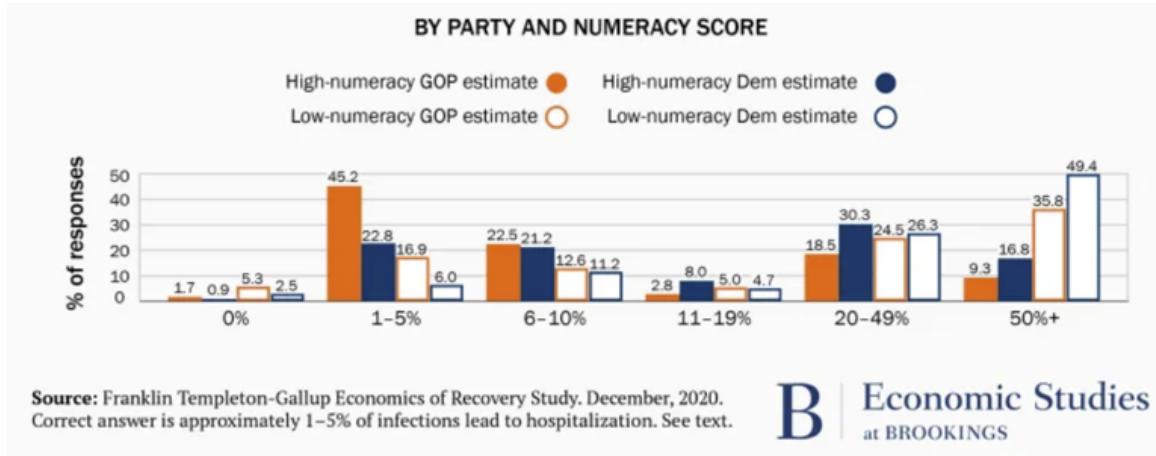
When asked to estimate how often Covid patients have to be hospitalized, Democratic voters do worse than Republican voters:



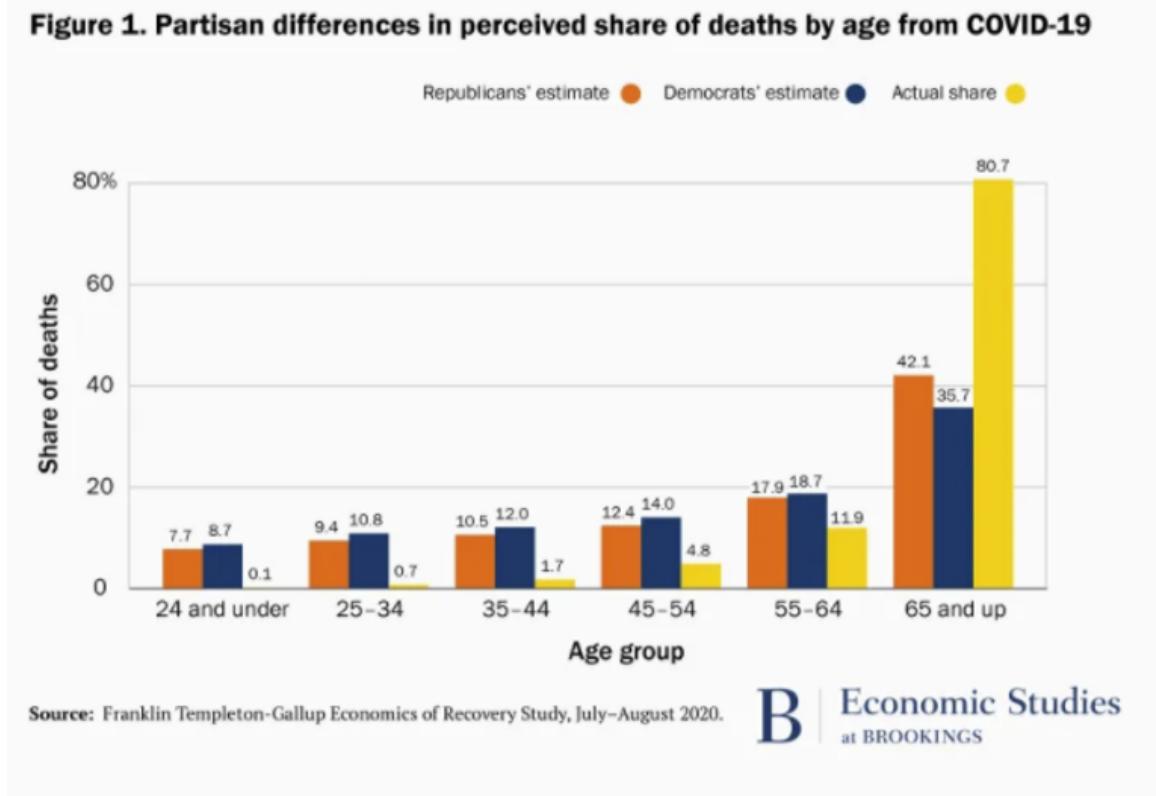
As usual, when one is tallying up points for who is doing better or worse and deserves more credit or blame, the lead gets buried. The lead is the implicit 'overall' chart that combines all three of these, and that the three charts aren't fundamentally that different.

A large majority even of Republicans still are vastly overestimating the hospitalization rate, with half of them thinking at least *one person in five* needs hospitalization.

There's a *huge* correlation with the [Berlin Numeracy Score](#) (and also sadly if you do the math not that many people scoring high on the Berlin Numeracy Score, looks like about 29% of Republicans and 27% of Democrats [if my calculations are correct](#)):



You see a similar thing here:



The small extent to which the Republican estimates here are better is much less interesting and important than the extent to which everyone's estimates are way, way off, in the direction that would lead to people being far more cautious than they would otherwise be, and much closer to each other than to the right answer.

We don't have the high vs. low numeracy chart for this one, but I'd be curious to see it.

My guess is that we can afford to mostly not treat the low-numeracy people as meaningfully having numbers assigned to things at all. When we ask them for numbers, we get garbage answers because they don't think that way at all, and you get a detached wild guess slash attempt to give the right no-context social response to a question rather than anything that's meant to represent reality.

For the high-numeracy people, the numbers likely at least mean *something*. They're still wildly wrong, in ways that are more enlightening, and they are more different by party than the low-numeracy numbers rather than the reverse, which is interesting.

Should we score this as a clear win for Fauci-style level-two thinking and communication (also known as 'lying')? That depends on people adjusting their behaviors based on their perceived risk levels, and on those adjustments being useful rather than not useful, plus a willingness to pay the costs involved in misrepresentation. I don't think those questions have clear-cut answers. I do think that some sympathy for the devil is reasonable here, in a 'if you told them the full truth and they understood it you wouldn't like their answer' kind of way. Humans often solve risk problems and collective action problems by having very false primary and intermediate values in their calculations, then doing incorrect manipulations to them that get tuned in various ways until they give reasonable answers. This is another illustration of all that.

The other question is, if we told people the truth, would they make all these errors anyway? Or to put it another way, if you listen to me say 1%, when asked for the answer write down 25%, but then tend to make decisions that make sense if the answer is 1%, and you're not ever asked for explicit calculations in an impactful sense, where's the mistake? *In what sense are you really being wrong here?*

I think that's an excellent question and that we haven't given it enough thought.

Public Broadly Supports Challenge Trials

Ethics is hard, but it's nowhere near as hard as 'ethicists' think it is. [The public knows better \(source\)](#).



Eliezer Yudkowsky @ESYudkowsky · Mar 19

...

An extreme case in point of "handwringing about the Overton Window in fact constituted the Overton Window's implementation". But hey, at least this instance only killed a couple of million people and down-geared the global economy for a year.



Robert Wiblin @robertwiblin · Mar 19

Intellectuals: "People will freak out about Human Challenge Trials because they're so ethically fraught."

Ordinary people: "WTF is wrong with you this is totally fine, please do it immediately."

One final concern that skeptics have raised is that the public will [view these trials as unethical](#), feeding distrust and delaying uptake.

Fortunately, new research suggests this concern is unfounded. In a [paper](#) published this January, a team of political scientists, bioethicists, and public health researchers surveyed a representative sample of nearly 2,200 people in the United States, as well as around 500 people in each of seven other locations: Australia, Canada, Hong Kong, New Zealand, Singapore, South Africa, and the United Kingdom. The researchers explained how challenge trials differ from standard trials and asked the participants what they thought about the ethics.

They found staggering support for challenge trials in every location. Three-quarters of respondents preferred them to standard Phase 3 trials — and among those who seemed to thoroughly understand how the trials worked, nearly 90 percent preferred them to the standard trials. In the United States, there were similar levels of support among Democrats, Republicans, and independents, as well as among vulnerable groups, like those over 65, essential workers, and racial minorities. Every demographic group thought challenge trials were more ethical and more scientifically valid — and, perhaps most important, expressed greater willingness to take a vaccine if it had been tested in a challenge trial. Sometimes, the public's views surprise the experts.

Challenge trials are super popular, and the concerns of 'ethicists' are simply wrong. Excellent.

In general, if something has 75% popular support, and those who best understand it give it 90% support, that's a pretty popular thing, and when the only thing stopping it is that it's illegal, supporting legalization seems like a very good political issue for either or both parties or any politician to take up and make their own.

I also wouldn't worry about those 'ethicists' on any level. Not only are they always wrong and destructive in everything they do, they'll fall in line on this when policy changes, and move on to trying to slow down and cripple the challenge trials the way they try to slow down and cripple every other type of experiment or other useful action. It's their job, after all.

Going forward, the agenda is clear. Human challenge trials, always, in all things, as the standard, *default* way to test new vaccines or other applicable treatments in Phase III, with proper compensation to participants.

In Other News

[Take me out to the ballgame](#), at the latest some time in May, concerts getting similar treatment, I'd prefer a higher percentage with only the vaccinated which seems better all around:



Morgan Mckay @morganfmckay · Mar 18

Also News: Starting April 1, sports venues with 1,500+ indoor capacity can reopen at 10% and sport venues with 2,500+ outdoor capacity can reopen at 20% capacity.

...



Morgan Mckay @morganfmckay · Mar 18

Replies to [@morganfmckay](#)

Attending fans will need proof of a negative COVID-19 test or proof they have been vaccinated. This requirement will be re-evaluated in May

...

[Take me out of the ballgame and keep me out long enough to fully recover](#), because Covid-19 is really nasty and you very much do not want to get it even if you know you're not going to die. Article looks primarily at the NBA and the longer-term impact of Covid on its players, especially players who push hard to return quickly. Post includes reports of many athletes where there's nothing physically wrong with them that can be identified, yet their performance levels haven't corrected themselves for extended periods. Very hard to tell from things like this how frequent or severe such effects are in general, and how scared young people especially should therefore be of the virus.

[Data on infections post-vaccination among California healthcare workers](#). I'm doing a bunch more thinking on exactly how delayed we should treat the vaccine response. The initial trials strongly suggested very good protection after about day 10, whereas the in-the-wild data we have, both here and from Israel, suggest similar longer term results but that it takes longer to happen. I think I initially mostly dismissed the Israeli data for reasons that don't hold up under scrutiny, and we should be more skeptical of how protected we are in the second and third weeks after vaccination – it's still a lot, but probably not a 90%+ 'hot damn look at this chart' level of protection, and that's likely due to real world conditions messing things up reasonably often, in ways the second shot still fixes.

Table 1. New SARS-CoV-2 Infections among Vaccinated Health Care Workers from December 16, 2020, through February 9, 2021.

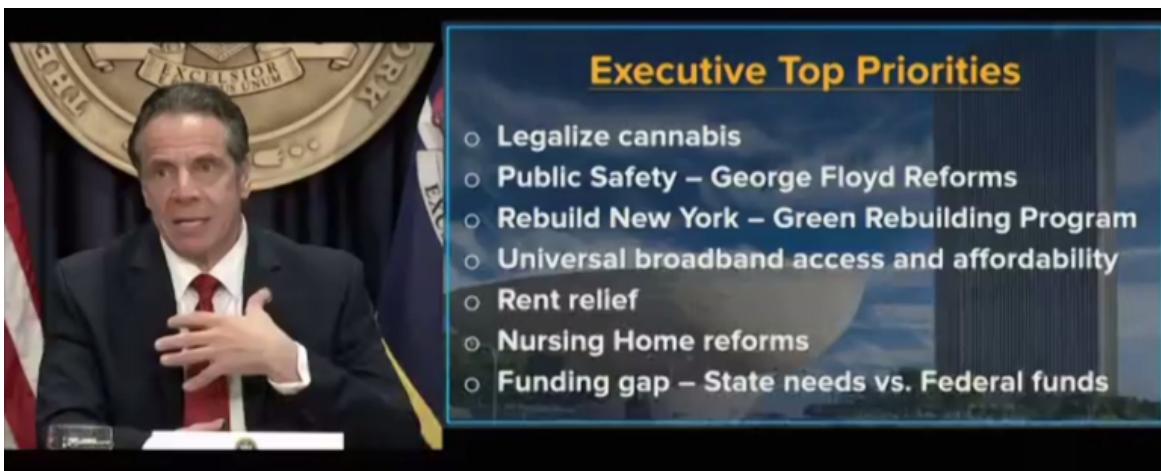
Days after Vaccination	Vaccinated Persons		
	With New Infection (N=379)	Tested (N=14,604)*	Eligible for Testing (N=36,659)†
	number	number (percent)	
Dose 1			
Days 1–7	145	5794	35,673 (97.3)
Days 8–14	125	7844	34,404 (93.8)
Days 15–21	57	7958	32,667 (89.1)
Day 22 or later, before dose 2	15	4286	32,327 (88.2)
Dose 2			
Days 1–7	22	5546	23,100 (63.0)
Days 8–14	8	4909	16,082 (43.9)
Day 15 or later	7	4167	14,990 (40.9)

* Shown are the numbers of unique health care workers who underwent testing (not the number of individual tests).

† Shown are the numbers and percentages of persons among 36,659 vaccinated health care workers who were eligible to undergo testing each week as of February 9, 2021.

[New Zealand experiences the joys of vaccine prioritization by politics and power](#), except the relatively sane one both because New Zealand is relatively sane in general, and because they don't have an ongoing pandemic so anyone not looking to leave the country can afford to wait a bit no matter how high-risk they would be.

[Cuomo Watch: New York is over Covid and has its priorities straight:](#)



(Yes, those are only *budget* priorities, but still. Also, I get it, it's been a rough time and you gotta mello out, man.)

[An analysis of a likely airborne infection](#) inside the quarantine hotel in New Zealand. I'm not as blown away as the thread author is, and in general such things are unlikely, but I have little doubt that they happen.

[Also Cuomo watch:](#)



Bill Hammond
@NYHammond

...

Cuomo: "We are lower now than we were before the holiday surge."

I don't know how he can say that.

Infection, positivity, hospitalization and mortality rates are all higher now than they were on Thanksgiving Day.

Also Cuomo watch ([WaPo](#)):

Andrew Cuomo's family members were given special access to covid testing, according to people familiar with the arrangement

Now that everyone's truly out to get him, we should expect to find lots of abuses of power, because abuse of power should come as no surprise. Cuomo I do think was unusually abusive in all senses, but the main difference between Cuomo and similar others is mostly that Cuomo is now a scapegoat and thus is getting caught and called out, rather than that such behaviors are rare.

In other allocation by politics and power news, if your diversion of vaccines to your employees includes a place with the word Trump in its title, [then maybe we'll notice](#) (WaPo). Doesn't seem like the consequences for this are going to be much of a deterrent, even as the symbolic scapegoat case, and I presume this kind of thing happened a lot and mostly wasn't caught.

Germany is imposing new restrictions. The explanation is something that [perhaps they could have seen coming](#)...



Kai Kupferschmidt  @kakape · Mar 23

...

This is kind of true, but we also already knew all of this in early March when the same group of politicians decided to ease restrictions.

We call this kind of thing „into the potatoes, out of the potatoes“ in German or maybe in future „into the pandemic, out of the pandemic“.. 



Marcel Dirsus  @marceldirus · Mar 23

Chancellor Merkel: "We basically have a new pandemic. The mutation from Great Britain has taken over...it is clearly more lethal, more contagious and contagious for longer."

Merkel [tried to go even farther this time](#), then backed down saying it wasn't practical, and accepted responsibility for the overreach, providing good evidence that there isn't much more slack left in the system in that direction.



BNO Newsroom 
@BNODesk

...

Germany will impose an ever stricter lockdown over Easter, when even grocery stores will have to close

Why did Germany previously loosen? [Pandemic fatigue](#). I'm more sympathetic to this than Kai is, in the sense that if there are increasing costs over time eventually you can't and won't keep paying them, even if the consequences are dire. The vaccination campaign in Europe isn't far enough along to give people a reasonable end point, and hoping to sustain harsh restrictions that long never seemed that likely to work.

Then there's [this from Austria](#), where similar dynamics seem to have forced the issue. At some point, all options become impossible for different reasons.



Christine Syrowatka
@chrissi_syro

...

For my non-German speaking followers: The situation in Austria now is bad, really bad.

The ICUs in Vienna reached their capacity limit. But, Vienna's mayor now decided that „We take the risk“. They do nothing and still discuss to loosen the lockdown.

We shouldn't feel superior on this, any more than Europe should have felt all that superior to us earlier. If our vaccine effort wasn't where it is, would we be any different?

I hadn't heard what was going on with Novavax for a while despite it looking like they had a proven safe and effective vaccine, [found this article from early in March saying they're on track to ship in June](#). Article in question doesn't seem to notice this is a problem.

[Paper finding cognitive function declines in face of prolonged social isolation](#). Social isolation costs are definitely being downplayed by many, with the 'cognitive decline' aspect seeming like a way to make a part of it legible so that at least part of it can be properly considered in decisions. Reminds me of other similar concerns that got ignored without quantification.

[Getting vaccinated correlates to answering polls](#) in a way that isn't easy to correct for ([poll](#)).



Nate Silver  @NateSilver538 · Mar 18 · ...
This is sort of an interesting example of response bias in polls. A new Quinnipiac poll found that 38% of adults in New York say they've "gotten a COVID-19 vaccine". But we know what the *actual* number is and it's closer to 30% (at least one dose). poll.qu.edu/new-york-state...

Results are also interesting, with a bunch of different dynamics going on here and being combined. I'll let the numbers speak for themselves, remember to adjust for who answers polls.

14. (Adults) Have you gotten a COVID-19 vaccine, or not?

		ADULTS.....						WHITE.....	
		Tot	Rep	Dem	Ind	Men	Wom	Yes	No
Yes		38%	35%	43%	36%	34%	41%	60%	40%
No		62	62	57	63	66	58	39	59
DK/NA		1	3	-	1	1	1	1	1
		AGE IN YRS.....			WHITE.....		4 YR COLL DEG		
		18-49	50-64	65+	Men	Wom	Wht	Blk	
Yes		19%	42%	77%	45%	51%	48%	29%	
No		80	58	23	54	49	51	71	
DK/NA		1	-	-	1	1	1	-	
		AREA.....		UpStat NYC		Sub			
		45%	32%	37%					
Yes		54	68	62					
No		1	1	1					

Not Covid, Not a Crime Either: Bill de Blasio calls for [sending the police in when people are spotting committing wrongthink](#).



Matthew Chayes  @chayesmatthew · Mar 18

Those who commit hateful but noncriminal conduct should be confronted by the NYPD, [@NYCMayor](#) [@BilldeBlasio](#) says:

"I assure you, if an NYPD officer calls you or shows up at your door to ask about something you did, that makes people think twice, and we need that."

1.5K

2.2K

373



Matthew Chayes  @chayesmatthew · Mar 18

"I think that has an educating impact on people; I think that has a sobering impact that we need."

101

34

79



[Twitter thread on Supreme Court rulings on Covid-19 restrictions](#), the opinions written in support of them, and how it is being treated as a gotcha when justices stop pretending they don't know things, and illustrating what happens when there's no such thing as knowledge or law but only blame avoidance.

[This Vitamin D observational study is super weird](#). It's saying that in black individuals vitamin D levels mattered, but in white individuals they didn't, *even for correlations and even for fully deficient levels of Vitamin D*. This proves way, way too much, so I notice I'm confused. We have seen lots of studies showing *correlations* of D-levels with results in ways that ethnicity doesn't screen off, and the counterarguments against it were along the lines of *of course there's a correlation* because people in poor health have low D levels and to the extent D is doing work it could be by needing to not be deficient. Yet this is claiming to blow past all of that. What's up here?

[A fine idea:](#)



Nick Reisman  @NickReisman · 2h

Madison County officials say they are establishing a "will call" list for those age 18 and older that will be contacted when there are extra doses or no shows for appointments.

4

6

38



[Retrospective called "How the West Lost Covid."](#)

Looks like there was a brief period where [Danville was handing out vaccinations to whoever wanted it](#), but the health department shut it down.

A survey on how effective 'reasonable prevention' was at preventing Covid:



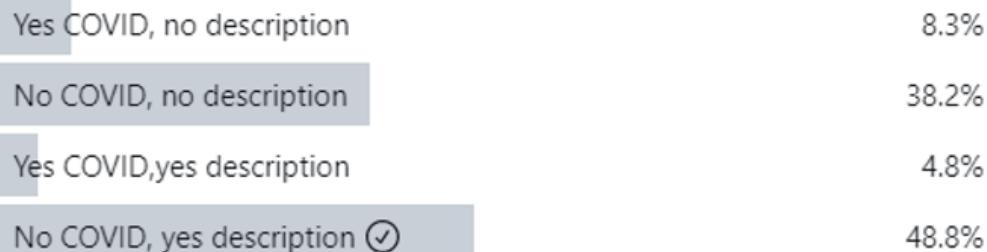
Rob Bensinger @robbensinger · Mar 21

...

1. Did you catch COVID?

2. Does this accurately describe you? 'I was avoiding seeing people in person, not going in stores, etc., though I ate some delivery meals and wasn't sanitizing packages.'

(Leave a comment if 'yes' to both.)



1,494 votes · 3 days left

My worry here is that some people who were even more paranoid than that got put in the no category, and given who follows Rob I'm guessing that could be a sizable group, but it's cool to see. This is an overall 13.1% rate of getting Covid, and 17.8% even in the "no" group, which is substantially lower than the nationwide average of about 30% (some people answering weren't American, but still that's a big gap), which supports the idea that this group is a lot more cautious than usual. The "yes" group had a 9% rate, about half of the "no" group.

Between people thinking they followed such rules when they didn't, the lizardman constant slash misclicks/misunderstandings, false beliefs about having had Covid (which aren't that rare) and the most cautious people of all being in the "no" group, that 2:1 ratio is almost certainly too low. I don't think I'm the best person to repeat this quiz (since my followers largely read this) while fixing the wording (something like ending with the etc in the description on #2, and ideally with "USA only" on it and if desired a second copy for non-Americans) but I'd encourage those with enough and diverse enough Twitter followers to give it a shot. Would also be cool to see how often different groups did this level of prevention, in addition to the goal of being able to get effectiveness.

[New analysis on how we could do the most important thing and speed up vaccination capacity.](#) Le sigh.

[Vaccination is a superpower](#), it even gives you the ability to fly, but no, it isn't as cool as being bulletproof, even if you discount required secondary powers and healing factors involved. If one was bulletproof, one could change one's behavior even more than you can when you're vaccinated. Luke Cage is a lot more likely to get shot than Alex Tabarrok or even Trish Walker, and she's not exactly doing proper social distancing.

This week I'll once again be travelling into New York City, this time to get my second vaccine shot. It'll be that much better not having to be so paranoid while I'm there. The third trip a few weeks from now will be even better.

A Retrospective Look at the Guild of Servants: Alpha Phase

Executive Summary

The Guild of Servants is an organization founded in August 2020 aimed at providing structure and community to rationalists, and others with similar ambitions, in the midst of the COVID-19 pandemic. The Alpha Phase of the project recently concluded.

The Alpha Phase was an experimental rollout and proof-of-concept involving ~35 volunteer participants, consisting of an online community component and an online course component. We have collected feedback from stakeholders in the form of interviews, and consolidated that feedback in this document.

The courses, as conducted in the Alpha, were determined to be too time-consuming and demanding, both for the instructors and the participants. It is recommended that the duration and intensity of courses be scaled down in subsequent phases. Large and complex courses such as the Character Sheet ought to be broken down into smaller discrete pieces. In this document we make further, more detailed suggestions for changing the format of the courses.

The cohort community structure, in which participants were placed into fixed cohorts of six members, was deemed a success. However, due to a natural expectation of attrition, it is recommended that the starting cohort size be increased to eight members, and that initial cohort sorting be based on mutual timezone compatibility.

The Alpha Phase was a process of learning and experimentation, and each founding Council member played many different roles. For the Beta Phase, it is recommended that the Guild leadership structure be reorganized such that each Guild Council member is assigned a specific administrative role, and that participation and contribution be compensated either monetarily or with ownership equity. It is further recommended that Guild Council members and other administrators not be expected to do both administrative and course-teaching duties at the same time.

The Alpha Phase is considered a success, based on our own stated expectations and goals in the planning. We avoided the pitfalls that we anticipated in our pre-mortem analysis. Before moving into the Beta Phase, it behooves us to recalibrate expectations toward a more ambitious goal, and specify our goals and vision as precisely as practical.

Introduction

August 2020, a group of five rationalists founded an organization to provide a formal community and development structure for rationalists. Due to the restrictions of the COVID-19 pandemic and the belief that the Future is Online, we aimed to build an entirely-online community structure that was still valuable and meaningful. We named the organization the Guild of Servants, to communicate our intent to make ourselves valuable to the world.

Approximately thirty rationalists responded to our open invitation and took part in the Alpha Phase of the Guild of Servants. We cannot sufficiently express our gratitude to these people for their patience and enthusiasm as we tested out ideas for community infrastructure and course design.

The Guild of Servants Alpha was organized around three courses which were designed and implemented by the founding Council members. These courses were:

- **Fashion**, a module which taught principles of personal aesthetics, and provided constructive feedback to participants. The purpose of the course was to encourage participants to pay attention to the ways in which they broadcast information about themselves, and to provide quick and affordable ways to improve in this area.
- **The Character Sheet**. This module aimed to guide each participant through a process of uncovering their own unique strengths and weaknesses, clarifying their personal goals and ambitions, and generally aided them in leading a strategic and examined life. It was based on Alex Hedtke's workshop material.
- **How to Spot a Con** was a course rooted in Cialdini's book **Influence** and designed to lead participants to be less likely to be exploited by social manipulation.

The courses were conducted via weekly Zoom sessions and relied heavily upon the six-person "cohort" structure. The cohort groups provided a framework for group work and group discussions, as well as a sense of camaraderie.

Courses

It was widely agreed that the courses were too demanding and time-consuming, on both the student and instructor side. A good target for the future would be less than one hour per week of total class time and less than one hour per week of homework. In future phases, students may end up taking more than one class at a time, but this should not be expected or default.

Each class should be a maximum of three sessions of one hour each. In general, classes should be short, focused and specific. Big topics should be broken out into multiple discrete classes. Something like the Character Sheet class, for example, could perhaps be expanded into three or more distinct classes.

It was recommended that all courses begin with a questionnaire to determine each student's background knowledge and competency in the course domain. If possible, in future phases, students can be "tracked" according to their starting knowledge and skill set. Additionally, setting course expectations at the beginning of each course is an indispensable step which we did not always attend to. It was difficult to calibrate course materials without an understanding of where the class attendees were coming from, and difficult for students to calibrate their expectations without explicit course objectives.

Standardization of what a guild "class" looks like would provide a better sense of cohesion both between and within classes. Wherever possible, classes should be the same length, involve the same procedures and follow the same patterns. Deviations will happen, but only with intentionality and purpose.

Multiple council members expressed that the group breakout sessions would have been greatly facilitated by the presence of teaching assistants. These teaching assistants could be guild members who previously took the class and are serving in this role under some incentive, or could simply be cohort members tapped randomly or on the basis of competency.

Some instructors reported a feeling of imposter syndrome and/or inadequacy in the role of teacher. This is a normal reaction to placing oneself in the role of “expert”, and it should be kept in mind that all good teachers are constantly revising and refining their teaching materials. The leadership should focus on keeping up the spirits of course instructors who may be experiencing difficulties.

Purely elective online courses must be regarded as a kind of “infotainment”, with heavy emphasis on the “entertainment” half of the equation. Recommend shortening/tightening classes and punching up their enjoyability level.

Fashion

This was widely agreed to be the most popular and fun class with the highest degree of engagement. It was suggested that a professional fashion expert be paid to give feedback at specific junctures to avoid groupthink and also enhance perceived credibility.

Character Sheet

All agreed that certain elements of the Character Sheet class ought to be made more central to the mission/vision of what the Guild is trying to do. More focus should be placed on taking inventory of the student’s pain points and weaknesses, and also understanding what their goals are. Something like a “skill tree” concept should be implemented only after understanding what skills the student is actually interested in obtaining. We can and should provide a “guided wayfinding” approach that locates the inductee and then leads them down a path toward their goals. In short: the Character Sheet content is important to the Guild mission, too important to fit into one class.

People didn’t seem to understand the purpose of certain portions of the Character Sheet. The middle section on strengths and weakness was remarked on as something that people got hung up on.

How to Spot a Con

Regarded as having portions that were fun and hands on and also portions that were dry.

It was suggested that this class could benefit from some social manipulation games or inter-cohort competitions, potentially with prizes. The lifestyle restrictions caused by the pandemic screened off a number of potential “field test” style activities.

Proposed actions:

- Break the Character Sheet class out into several short modules, with more focus on locating people where they are at the beginning and putting them on track.

- Consider hiring domain experts for specific parts of classes where expert-level feedback would be valuable.
- Implement inter-cohort competitions as often as possible. People find competition motivating, and we should take advantage of the cohort structure.
- Define a fixed template for what all courses should look like, structurally, with allowance for necessary deviations.

Cohorts

All Guild members were placed into cohorts of ~6 people, based on research in optimal task force size. These cohorts were the organizing unit for almost all Guild activities, including class assignments and group discussions.

The cohorts were agreed by all to be a fundamentally good idea in need of a few tweaks. A significant degree of variation between cohorts in terms of involvement and interaction was observed.

There were a small number of minor personality clashes reported within cohorts, which seem to have been addressed and rectified with minimal cognitive-emotional overhead reported by the involved parties. Some council members reported these events as being more stressful than others. Perhaps some people are just better than others at dealing with such conflicts.

This author thinks Discord and other Slack-likes, e.g. Microsoft Teams, are a cancer and would rather encourage people into having scheduled video-calls, phone calls, or practically anything other than just being passively online all day. Rationalists Don't Let Rationalists Be Online All Day. That said, I recognize the inevitable tension between this stance and the goal of creating an online community space that provides a sense of purpose and cohesion. I am not against the existence of the Guild Discord, I merely recommend that we don't actually *rely* on it.

Proposed actions:

- Increase the cohort starting size to 8 people to help with attrition and low investment.
- Implement a policy whereby cohorts are re-sorted and consolidated after the first ~3 weeks, such that no cohorts suffer from low active participation.
- Cohort members should be assigned by time zone compatibility.
- Delegate the resolution of interpersonal conflicts such that the Council does not need to get involved unless absolutely necessary.
- In general, leverage inter-cohort competition more effectively; this concept is discussed further in subsequent sections.

Leadership

It was generally agreed that the demands of simultaneous administration and teaching were untenable. Nearly every Council member reported feeling stressed and overworked. In the Beta Phase, the Guild Council will address these issues by the following strategy:

- Assigning permanent roles, such that each council member only focuses on one set of administrative tasks, and teaching/course-prep is only undertaken by

individuals with no other commitments.

- Increasing the number of Council-level administrators to spread the load out.

Even with this high degree of subjective overwork, we were generally dissatisfied with our own level of organization and logistical reliability. Specifically, we observe that students were rarely clear on where to look to find important information about courses, expectations, or scheduling. This issue will be addressed by placing one individual in charge of the student-facing organizing content, e.g. calendars and course notes.

The points system which we started out with was essentially abandoned. This was partly because of the lack of clarity in what points meant, partly due to the fact that administering and tracking points was far more time consuming than expected, and partly because points just didn't seem to be very motivating to people. The point system itself may be a good idea, but it was everyone's last priority and so its implementation floundered. The points/ranking system should either be dropped, or it should be prioritized explicitly, but it does not function effectively as a vague background element.

The Incentives Problem

The Guild of Servants Alpha actually exceeded all of our expectations, since we all predicted in our pre-mortem document that becoming overwhelmed and/or losing interest would be the most likely failure mode. Most of the Guild Council remained extremely committed and devoted large amounts of time to Guild business on an ongoing basis, and a large fraction of Guild members maintained attendance and interest through to the end.

That said, relying on volunteered time and sustained hard work for no compensation is unlikely to be a viable strategy as the organization grows. It is currently not possible, for example, to simply "assign" someone the job of improving the website, or keeping a calendar updated. The completion of all such tasks is dependent on interest level and willingness to donate time. We are all aware that we need to solve the incentives problem, but there is no broad agreement on how to do this.

A few possibilities include the following:

- Adopt a standard startup business structure, giving Council members "founder" status and allocating ownership share accordingly, and use ownership shares as an incentive for new members. Voting shares can be separated from ownership shares to give the Council more permanent control, if this is deemed wise. C-Corp startup structures are a tried-and-true framework, and a small share of a business entity that might grow very large is an empirically motivating incentive.
- Begin collecting Guild dues now, in the Beta phase, and pay out dues in return for Guild service. This is, frankly, unlikely to be very motivating at first, since the dollar amounts will be very low until membership ramps up.
- Raise capital. Seek out and secure a round of seed funding. These funds could be used to pay people fair wages for their work. Essentially we would reconsolidate the Guild as a venture-backed startup.
- Set aside the above proposals, and spend time researching monetization strategies for broadly similar organizations, then pick a monetization scheme that we agree could work and reorganize around that.

Beta Roles

We need individuals to take responsibility and ownership for critical administrative functions. Below are listed a set of roles, some of which have been assigned at the time of writing.

- Scheduling and Calendar manager. (Consult with all stakeholders to schedule meetings and courses. Consolidate all scheduling information into one, unambiguous calendar. Promptly address arising issues, such as instructors calling in sick.)
- Student advocate, or, Course design director. (Red-team the course design. Oversee and manage the structure, duration, and demands of all courses, to maintain a consistent vision and implementation across the Guild courses.)
- Technology director. (One person who oversees and makes the final call on issues relating to tech platform choices, and, possibly, handles technically difficult and repetitive tasks such as uploading YouTube videos to the channel, or administrating course surveys.)
- Cohort activities director. (Focus entirely on the cohorts. Are people having fun? Are the cohorts the right size? Are we constantly running some kind of inter-cohort competition? Can we host regular Street Epistemology sessions within cohorts? Forecasting tournaments? Good-works projects? There's a lot of room for possibility here.)
- Business manager. (Handles payments, billing, fees, invoices, approves reimbursements, officially sets prices and controls monetization streams.)
- Internal Guild Prediction Market manager. (Of course we should have one of these.)
- I am almost certainly missing some potential valuable roles; please suggest more.

Outreach

The Council is open to revisiting everything about Guild outreach and branding, including the name “Guild of Servants”. It has been agreed that we ought to hire a PR expert for a consultation, although we have not actually priced this out, nor determined where the funds would come from. (This author cautions that we ought not to expect magic; the purpose of a consultant is often to simply say obvious things with great authority.)

The twee approach (silly hats, cryptic Cyrillic letters) served to filter a certain kind of person for the Alpha. This seems to have worked well, but is unlikely to work in exactly the same way as we grow. As discussed in the next section, we shouldn't be afraid of seeming weird, as long as the weirdness is intentional.

Vision

The following is an attempt to consolidate the different but often overlapping statements of vision and mission provided by the Council:

“The mission of the Guild is to serve the Guild members by providing a structure which facilitates their growth into impressive human beings, according to their own needs and desires. A secondary but also important mission is to provide public benefit

by encouraging/incentivizing Guild members to do “good works”, improving the world and raising the sanity waterline.”

Most council members agree that we lack focus and strictly lack a solid elevator pitch. If I had to distill everyone's remarks down to something like a pitch, it might look like this:

“The Guild is an organization of people who have the same meta-goal of making themselves more rational and generally impressive, and it is innately valuable to push such people in contact with each other, providing a support structure.”

Nearly every council member expressed something like a wish for the Guild to serve as a monastery-school for rationalist-warrior-monk-Bene-Gesserit-Mentat-Freemasons. This is a good “a computer on every desktop” sort of inspirational target, but doesn't contain many gears.

Sometimes in order to define what a thing is, it is important to define what it is not. Now is the time to really focus on this. If we don't define the vision clearly and succinctly, then we will simply pursue the path of least resistance, and our failure will not be a failure of execution but rather a failure of direction; we will look up and realize that we have traveled a thousand miles on the wrong heading. The Beta Phase will be an opportunity to more carefully explore and define the Guild vision, mission and elevator pitch.

A Possible Model

This is my little corner to editorialize. I'm attempting to find a golden path that connects what we have, and where we want to be.

In my vision, the Council transitions to being administrators with defined roles. The function of the Council is to:

1. Serve and support the cohorts.
2. Facilitate and empower the teachers.

Teachers would be volunteers who come to us with a plan for a course. The Council helps them hammer out their lesson plan, their schedule, their assignments, and integrates their course with the Guild resources, calendar, website, YouTube channel, etc., including Mentat (an extensible teaching app developed by yours truly).

The course materials would exist as static YouTube videos, notes, and Mentat modules. “Class sessions” cease to be live lectures, and are instead short, scheduled check-ins with teachers/TAs to give feedback on assignments, ask questions, etc.

These course materials (or at least the YouTube videos) would be freely available to all, serving as advertising which drives/upsells people into paying for Guild membership in order to get the full benefit.

The Guild thus becomes an ever-growing library of curated, well-designed and thoroughly tested course content, and a superstructure of cohorts who are supported in their own ambitions, and in taking the courses. In this way, one can be “in the Guild” and take no courses at all, instead simply participating in the cohort activities (competitions, projects, etc.).

The business funnel consists of free YouTube videos that aim to convert people into paying Guild members, and then provide ongoing valuable support and other services to those Guild members to retain and develop them.

Conclusion

The Guild of Servants Alpha Phase is a strict success based on initial stated ambitions and expectations. The primary noted points of failure were in administrative inadequacy, and excessively high expectations for the courses. Other concerns include expected difficulties in scaling the current design up to serve more Guild members, and a lack of specificity in monetization and incentivization going forward. All things considered, we are highly optimistic that the Guild can be sustained and grown if appropriate steps are taken.

The Guild of Servants Beta Phase is set to begin in the summer of 2021, and will include a greater course variety and expanded cohort activities. We plan to scale up for many more participants this time, as well. **Please contact us at guildofservants@gmail.com if you're interested.**

Thanks to Alex Hettke, David Youssef, Errol Highborg and Gray Morrow for their service to the Guild, and thanks to each and every participant in the Alpha Phase for their patience and enthusiasm. Thanks to Steven Zuber for his feedback on this document.

Thanks for reading,

Matt Freeman

How You Can Gain Self Control Without "Self-Control"

Many people can't understand how Ty does it.

Consider a typical weekday for Ty. That means waking up at 6:30 am and getting ready to run. It's freezing, and Ty's lips turn blue, but that doesn't stop Ty from going ten miles. After that, it's time for a 12-hour workday, including two three-hour stretches with no break. Ty's meals are a salad, some hummus, some cheese, fruit, and eight no-sugar protein bars. At 9 pm, Ty decides to hit the climbing gym for an hour and finally wraps up the day with another 90 minutes of work. Ty gets in bed at 12:30 am for six hours of sleep.

This schedule, as far as Ty is concerned, is easy and unremarkable. How could this schedule possibly feel easy?

You, like most people that know Ty, probably assume Ty has remarkable self-control. But the interesting thing - that Ty will readily admit - is that Ty really doesn't. At least not the sort of "self-control" you're likely thinking about.

To understand how Ty lives such a self-controlled life, let's peel back the layers of what self-controlled behavior actually is. Rather than the typical approach of treating it as a black box, we're going to rip it apart to explore which pieces are easier to change and which ones you may be stuck with for life. By adjusting these easier-to-change parts, you may be able to increase how much control you have in your life, even if you can't change your "self-control" (as some would think of it). If you can achieve this, it may well translate into making healthier choices and achieving more of your life goals.

In this article we'll be exploring:

1. Nine traits of self-controlled behavior (what self-control looks like when we peel it apart)
2. How these traits differ from what most people think of as "self-control"
3. Whether self-control is genetic
4. Twelve strategies for gaining more control in your life
5. A step-by-step process for applying these strategies
6. The strange case of Ego Depletion (where, after hundreds of studies, scientists still disagree about whether self-control can get drained like a battery)

Before We Get To Self-Control: What's This Thought Saver thing?

In a moment you'll notice that something is different about this article. It has [Thought Saver](#) widgets embedded in it, designed to help you absorb the content as you go! Just click on each of those widgets to quiz yourself on what you've learned so far. You can also [click here](#) if you want to try out the free Thought Saver tool for yourself, or get all the flashcards for this article. We created Thought Saver to help make it faster and easier to learn and remember important ideas (based on a homegrown approach I

personally have been using for years that I LOVE). Thought Saver is still in beta though - [we'd greatly appreciate your feedback](#) regarding how to make it better! The collaboration for this post came about because the LessWrong team said they were thinking of experimenting with memory aids directly in articles. So we joined forces in order to do this fun experiment - please let us know what you think of it!

One more thing: to my knowledge the idea of embedding quizzes right into an essay began with the incredible introduction to quantum computation (called [Quantum Country](#)) written by the extremely brilliant [Michael Nielsen](#) and [Andy Matuschak](#). Note that Thought Saver isn't focussed on making spaced repetition-based essays - if that's your interest, definitely check out [Andy's work](#), as he's doing cool stuff in that space.

Now let's get going and explore self-control!

Nine Traits of Self-Controlled Behavior

Let's define "self-controlled behavior" to be taking the actions that are better for you in the long-term over those that would feel more desirable in the short-term.

There are quite a number of traits that can cause someone to act in a self-controlled manner. To make them easier to understand, I've divided them into four broad categories, each of which encompasses multiple traits:

I. Classic Self-Control

1. Awareness of temptation: at the movies, many people mindlessly shovel overly-salted popcorn into their mouth until they're surprised to find the bag empty - but Ally periodically notices how much popcorn she's been eating, is aware of her urges to eat more, and is therefore capable of making a choice regarding these urges.

2. Tendency to override temptation: whenever Todd passes the candy bowl at the reception desk, he overrides his urge to eat a piece of candy, walking on by.

II. Helpful preferences

3. High motivation: Asha used to find it really difficult to get herself to exercise, never managing to do it consistently. That is until her doctor told her that if she doesn't change her behavior, she's predicted to die ten years younger than she otherwise would. This terrified her since she wants to be there for her grandchildren. Now she exercises six days per week, without fail. Fear of death motivated Asha to change her behavior.

4. Delayed gratification: Kim has a big exam next week and really doesn't want to study. But she has always been the sort of person that makes short-term sacrifices for long-term benefit. So she studies tonight, knowing she'll be happy about it tomorrow. Her discount rate on future value (relative to the present) is lower than most people's.

5. Lack of unhealthy desires: Sue doesn't like the taste of sweets and finds alcohol unappealing. It takes her no effort to resist eating cookies. She thinks salads taste much better and rarely experiences the temptation of eating unhealthy foods in the first place.

III. Pain tolerance

6. High pain tolerance: Dan has the goal of finishing a marathon, so he runs five days per week. Sure, it's a lot of running and cold as heck outside where he lives. But the cold and burning in his muscles just don't bother him much. Others would suffer far more from those same experiences.

7. Nonchalance toward future suffering: Amilia signs up for Tough Mudder races, even though she finds them excruciatingly unpleasant. She knows how painful the race will be, yet that doesn't deter her from repeatedly putting herself in that situation. She's not one to avoid something she finds valuable just because it's painful - the thought of future pain just doesn't deter her much. Similarly, she doesn't put off painful dental work or hesitate to leap into a freezing cold pool (even though she knows it will feel awful for the first minute).

IV. Momentum

8. Energy: Donny works 90 hours per week. While his schedule would grind most people to an exhausted pulp, he's famous for his endless energy. On his tenth hour of working each day, he feels as peppy as most people do on hour number two. And even after all this work, Donny can easily avoid making short-sighted decisions (like eating pizza for dinner every night) since he doesn't experience the exhaustion that might lead others to make such choices.

9. Flow: If she were to pay close attention, Rita would notice that her sprinting routine feels like having your leg muscles ripped apart while constantly being on the verge of puking. But the reality is that Rita gets into such a deep flow state while sprinting that she isn't aware of the passage of time, let alone how awful it all feels.

Exercise 1: before reading on, in order to better understand yourself, I recommend taking a minute right now to think about which of the nine traits of self-control you are high in, vs. about average in, vs. low in:

- *Classic Self-Control*
 - Awareness of temptation
 - Tendency to override temptation
- *Helpful preferences*
 - Motivation
 - Delayed gratification
 - Lack of unhealthy desires
- *Pain tolerance*
 - Pain tolerance
 - Nonchalance toward future suffering
- *Momentum*
 - Energy
 - Flow

Of course, it's possible for these traits to be domain-specific (e.g., you may lack unhealthy desires around food, but have an unhealthy desire to play video games until your eyes hurt). But that being said, it can still be useful to consider which we are generally high or low in.



Classic Self-Control

We can think of the first two items in the list above as the components of what people usually mean when they say "self-control" - the awareness and restraining of one's impulses. For our purposes, to differentiate it from the other traits, we'll call the combination of these two traits "classic self-control." But as we can see, these represent just a fraction of the traits that lead to self-controlled behavior.

When you experience the temptation or urge to do something that's bad for you in the long-term or to not do something that's good for you in the long-term, classic self-control works like this:

1. you experience the urge
2. you notice it
3. you override it
4. you avoid the urge

But there are so many other ways a person can act in a self-controlled manner beyond being unusually aware of temptations or being unusually likely to override a temptation after noticing it.

As we saw above, someone may simply not feel the temptation in the first place, or they may experience less suffering from avoiding the temptation. They may be less motivated to prevent such suffering, or they may value their future well-being more than other people do. They may have high motivation to avoid the temptation or have unusually high levels of energy that prevent exhaustion or easily get into flow states that prevent "unpleasant" tasks from being experienced as unpleasant.

Understanding Ty

We're now ready to understand how Ty maintains the unusually disciplined lifestyle described above without actually having an above-average level of classic self-control.

This is best explained by simply quoting Ty on related topics:

"The worst part of running is...well...not running."

"Marathons are not tiring."

"I don't care about pain or discomfort; I just care about running."

"I was coming up with mantras today while running:

Suffer more.

Push harder.

Go for glory.

Suffering is glorious.

The last one is my favorite."

"I'm getting bored of boxing. It's too short. Forty-five minutes is not a workout; it's a warmup."

Two days after completing a marathon: "I'm so lazy, I shouldn't feel this good after a race."

"When I was reading [this self-control post], it struck me, perhaps for the first time, that some people don't like hard exercise."

Ty doesn't have high classic self-control (awareness of temptation + tendency to override temptation) or delayed gratification. Ty sometimes eats half a jar of peanut butter in one sitting and procrastinates on getting out the door for runs. Ty's been working on creating a stretching habit for years with little success.

On the other hand, Ty has a very unusual lack of unhealthy desires (peanut butter consumption aside), an extremely high pain tolerance, low suffering avoidance, very high motivation in work and exercise, high energy, and a tendency to get into immersive flow states.

Whenever Ty doesn't feel like running, Ty skips it. Ty loves salads and doesn't especially like sweets. Ty loves working hard. Ty isn't averse to suffering during exercise. Ty doesn't get tired easily, even on little sleep. Ty gets absorbed by work activities and can often stick with them for hours without distraction.

The Heritability of Self-Control

Why do some people have better self-control than others?

It seems that at least some (perhaps all) of the traits listed above that can lead to self-controlled behavior are at least moderately genetically determined. By many accounts, personality traits (such as those measured by the Big Five) [have heritability in the order of 40%](#). When researchers have subdivided conscientiousness (the Big Five trait most intuitively related to self-control), its individual facets (such as "Deliberation" and "Dutifulness") also [appear to have a genetic component](#). [Luciano and colleagues](#) found that all conscientiousness facets were influenced by genes, with "broad sense heritabilities" ranging from 18% to 49%. Even "nuances" of personality (which measure aspects of personality that are more fine-grained than facets) appear to be partially heritable (see [1](#) and [2](#)). A meta-analysis attempting to calculate the heritability of self-control specifically [puts it around 60%](#).

Fortunately, there is still recourse if we are born without much classic self-control. A common mistake is to assume that moderate heritability implies that a trait cannot be changed. To see why this is silly, consider the fact that muscle mass has significant heritability but can obviously be modified by weight lifting. People often think of heritability as a property of a trait itself, but really it is a property of a trait paired with a particular environment. As better strategies are developed for modifying that trait, or if such strategies are adopted more widely, the heritability can drop. For instance, if a new type of highly effective strength training became very popular, the measured heritability of muscle mass would fall as more people grew more muscular!

Importantly, even if a trait can't be changed by any known interventions, it may still be possible to change important outcomes related to that trait. Imagine a person with relatively poor working memory who can keep - at most - a five-digit number in their mind. This person can learn to use a piece of scratch paper to augment their working memory when they need to remember numbers (by simply writing down the digits and reading them back as needed). Not only does this work, but it works so well that the piece of scratch paper allows them to do better at this task than just about anyone who has to rely on memory alone.

Returning to self-control, this suggests that even if traits related to self-control are partially heritable (which seems likely), we still may be able to alter these traits. And even if we can't alter these traits, we may be able to apply strategies that achieve similar outcomes to what we'd achieve if we could alter them.

In the next section, I consider strategies you may be able to apply in your own life in order to act with more self-control.



Twelve Simple Strategies for Gaining More Control

Below is a list of simple strategies that you may be able to apply to live with more self-control, whether or not you are a person who tends to have a lot of it naturally. I've divided them into categories to make them easier to digest.

I. Preparation

1. Sidestepping temptation: Sally is a sucker for ice cream. When it's in the freezer, she eats the whole container in one go. But when she's at the store, it's easy enough not to buy any. So she doesn't buy ice cream anymore, and she asks her husband not to buy it either.

2. Total elimination: Kenneth used to have a problem with alcohol. He finally quit by going cold turkey. He doesn't even allow himself one beer, because he knows that if he does have one, he's going to be extremely tempted to have more.

3. Avoiding impaired decision-making: Don used to do his weekly food shopping after work, which meant he would do it when hungry and exhausted. Now he does it

on Saturday after lunch, when he is fresh and well-rested, so these days he finds that he naturally wants to buy healthier foods when he's at the grocery store.

4. Attention triggers: Jupiter took a [Center for Applied Rationality](#) workshop, where she learned to design Trigger Action Plans (TAPs) for paying attention. These are plans of the form "if this happens, then I'll execute this action plan" - for example, "when I'm finding my daily writing difficult, and I'm thinking about quitting, I'll notice that and try to figure out what's going wrong (rather than mindlessly checking Twitter to avoid writing)." By rehearsing this TAP, she was able to use it in real life. This helped her to make her writing sessions longer. Note: if you want to design your own implementation intention, and then apply techniques to help make it stick, you may find our [Program Yourself](#) tool useful.

II. Desire

5. Making goals more desirable: Killroy used to force himself to run daily. This was hard, and he often avoided it. But then he tried swimming and realized that he enjoys it much more, and finds it easier to get himself to do. By swapping running for swimming (which he finds a lot more fun), he made healthy exercise much easier.

6. Associations and framing: Joel used to love cake and ate it all the time. But now, when he thinks about eating it, he immediately imagines the regret he'll feel five minutes afterward - and the cake has a more negative association now. And when he thinks about going to the gym, he imagines how buff he's going to be in 12 months (from all that lifting) and gets excited. His long-term goals tend to jump to mind when he's considering doing something he may regret.

7. Temptation bundling: Marija hates going on the treadmill, but loves playing video games on her phone. By only allowing herself to play those games when on the treadmill, she finds that she now looks forward to the treadmill.

8. Mindfulness of desire: Philip sometimes has a strong desire to watch TV before bed, even though it interferes with his sleep. Fortunately, from his meditation practice, he's developed the skill of viewing his thoughts and feelings from an outside perspective, and observing them without being sucked into them. He's learned that he can observe a desire, and then just watch it without acting on it. So he notes the desire to watch TV and then lets it drift away, imagining that the desire is a leaf drifting away slowly on a stream. This is one of the skills we teach in our app for anxiety, [Mind Ease](#) (there we call it "defusion" because you are un-fusing from a thought).

III. Automaticity

9. Routines and habits: Jimmy had set a goal for himself of doing push-ups the moment he gets home from work. The first three weeks were a struggle. But now Jimmy hops down onto the ground for push-ups as soon as he's home (sometimes before he even realizes what he's doing). Doing push-ups went from something Jimmy would occasionally think about doing but usually avoid, to a routine that he would think of upon entering his home (and then usually do by default), to a habit that has become automatic. Note: if you want to form a new daily habit, you may find our free habit formation tool, [Daily Ritual](#), useful for that purpose. It can be especially useful to start with a simple morning habit that you can then attach more pieces to overtime. Your morning habit could even be a "meta habit," such as "when I wake up I do the things from this list I keep," then you can keep adjusting what's on the list over time.

10. Plunging ahead: Teddy finds it really hard to sit down and write for an hour, and he typically avoids it, even though he aspires to be a great writer. However, he finds it easy to commit to a five-minute writing session. Once he has started, he finds that he usually gets so engrossed that he can write for twenty or thirty minutes before he even realizes how long he's been doing it. When even five minutes sounds difficult or unpleasant, he sets the simplest possible goal: just opening up the document he's working on. Even this tiny action is usually enough to get the process started.

IV. Incentives

11. Altering costs: Tom has figured out a way to drink less coffee. By keeping the coffee maker in his wife's office, she sees him each time he gets one. Since he fears her raised eyebrow with regard to his excessive caffeine intake, this makes it no longer worth it to get more coffee after the first two cups.

12. Accountability: like most of us, Roxana wants other people to think highly of her, to not be judged by the people she cares about, and to keep the commitments that she's made. She leverages those social motivations to get herself to engage in healthy behaviors by spending time around people that care a great deal about health, and who live in an exceptionally healthy manner. This influences her to eat more healthily and to exercise more. Another way that Roxana uses social influence to her advantage is that when she has an important project she keeps putting off, she'll schedule a co-working session with a friend or colleague and pre-commit to working on that project during that time. Note that some types of accountability could be viewed as a specific form of "altering costs." (Incidentally, I was finding it difficult to get myself to sit down to write this essay, and so I used this very method - thanks Clare for holding me accountable!)

How to Gain More Self Control Without Requiring "Self-Control"

So far we've considered traits of self-control, as well as strategies that can help us act in a more self-controlled way. But how can we put these ideas into action?

If you're interested in gaining more self-control in your life (whether or not you have won the genetic self-control lottery), I recommend the following procedure:

Step 1: Pick an area of your life in which you'd like to act with more self-control in (e.g., "exercise") and a concrete goal for that domain of your life (e.g., "exercise for at least 45 minutes per day at least three days per week").

Step 2: Review the list of self-control strategies (above), and pick out the two that you think are most likely to help you act with more self-control as you work toward achieving that concrete goal. Use your understanding of yourself (especially what has worked well for you in the past) to guide your selection. If you feel stuck, ask a friend to help you choose.

Step 3: Think up (or better yet, write down) the first small steps to put the two strategies you selected in the previous step into action. Take those steps immediately, or if that's not possible, choose a time when you will do them in the near future.

Step 4: Create a reminder of your intention to apply these two strategies (e.g., by sending an email to yourself, or adding an entry in your calendar, or writing a note to yourself and putting it on your desk).

Exercise 2: rather than simply reading these sentences, I highly recommend you stop right now and actually follow the steps above. Reading without taking action on what you've read is like cooking with imaginary ingredients. You don't get the delicious meal at the end. By following the steps above, you can greatly raise the chance that you put what you learned from this essay into action.



Rethinking Ego Depletion

An interesting debate in the academic literature is whether self-control is analogous to the energy in a battery, getting "used up" as we deploy it. If true, that means that after exerting a lot of self-control, it temporarily becomes harder for us to

subsequently resist further temptation (until we "recharge" ourselves back to full capacity). Bizarrely, after hundreds of studies, this debate continues. What's going on here? How on earth could this not be settled after so many studies? I have a theory.

But before I explain it, let's take a closer look at the claim being made. As [Wikipedia](#) puts it:

"Roy Baumeister and his colleagues proposed a model that described self-control like a muscle, which can become both strengthened and fatigued. The researchers proposed that initial use of the "muscle" of self-control could cause a decrease in strength, or ego depletion, for subsequent tasks. Later experimental findings showed support for this muscle model of self-control and ego depletion...They showed that people who initially resisted the temptation of chocolates were subsequently less able to persist on a difficult and frustrating puzzle task. They attributed this effect to ego depletion, which resulted from the prior resisting of a tempting treat. Additionally, it was demonstrated that when people voluntarily gave a speech that included beliefs contrary to their own, they were also less able to persist on the difficult puzzle, indicating a state of ego depletion...When the energy for mental activity is low, self-control is typically impaired, which would be considered a state of ego depletion...there is currently no direct measure of ego depletion, and studies mainly observe it by measuring how long people persist at a second task after performing a self-control task."

The existence of ego depletion effects has become increasingly controversial in recent years. [Hagger and colleagues](#) noted in their recent paper that one meta-analysis of ego-depletion experiments found a medium-sized effect but that a subsequent meta-analysis contradicted these results.

In response to these contradictions in the literature, Hagger et al. carried out a huge randomized controlled trial, conducted as a 23 laboratory collaboration (with 2141 total study participants) to measure the size of the ego-depletion effect. Overall, they failed to find an ego depletion effect (the effect size was $d=0.04$, 95% CI [-0.07, 0.15]). They derived this by having control participants identify words with the letter 'e', while another group (whose self-control was to be depleted) did the same thing but had to withhold their response if the 'e' was next to a vowel. The depletion version "was considered to be more demanding, and to require greater self-control, than the no-depletion version because participants had to inhibit the tendency to respond to any 'e' and instead apply the more restrictive rules." On a survey given immediately afterward, this task caused the depletion group (compared to the control group) to report more effort, difficulty, and frustration, but not fatigue.

Next, the researchers attempted to measure whether ego depletion had actually occurred by showing participants three digits (such as 212), for which they had to indicate with their keyboard the identity of the digit that differs from the other ones (not its position - so for the example 212 they had to click on the 1st key indicating the number one, not the 2nd key indicating that the odd one out is in the second position). Word size was used to further throw off participants (by making the wrong digit a bit bigger or smaller than the others). Each participant had 100 rounds of this task (with 100 additional rounds of easier ones mixed in).

If ego depletion is real (the theory goes), we would expect participants in the depleting letter 'e' task to have slower reaction times and higher reaction time variabilities on the difficult rounds of the second task compared to those who had the simple version of the letter 'e' task. But they didn't.

So maybe ego depletion doesn't exist? But what to make of the alleged "300 independent studies [that] have replicated this effect during the 15 years since it was first reported"? Are all of these just false positives? Perhaps this is an egregious example of the replication crisis in psychology? But what then to make of this recent analysis of all the ego depletion meta-analyses claiming all the meta-analyses DO find an ego depletion effect, except in cases where they use specific (possibly flawed) bias correction techniques ([see table](#))?

Here's my simple proposed answer to this question (epistemic status: speculative and controversial) based on the breakdown of self-control above:

MAYBE IT'S JUST BECAUSE SELF-CONTROL IS NOT ONE THING!!!

Of course, some psychologists have worked to carefully unpack what self-control means. But insofar as I'm correct about this claim above, it seems the significance of this point may be greatly underappreciated.

Imagine a scientific inquiry into the question of whether "banning shoelaces causes people to trip more". Some studies find that banning shoelaces in the lab does indeed lead to more tripping. But in other studies, banning shoelaces has no effect. A conundrum. A meta-analysis concludes that tripping is associated with banned shoelaces, but not as reliably as people thought. Another meta-analysis says that some of those studies were really bad and should be thrown away. Furthermore, if you add back some unpublished studies and correct for pro-shoelace biased research, the relationship between shoelace bans and tripping can't be distinguished from noise. But really, the answer to whether banning shoelaces causes tripping depends on whether the population in question is wearing tennis shoes, loafers, or flip-flops. *You're not going to get a satisfying answer until you consider different types of shoes!*

So here's my proposal regarding ego depletion: some aspects of self-control get "depleted", and others don't. While it is often useful and convenient to lump all aspects of self-control together (e.g., to point out to friend A that friend B has a lot of self-control, which may help friend A better understand friend B) when we are asking nuanced questions about what the properties of self-control are, we have to get more specific. My prediction, therefore, is that different tasks intended to produce "ego depletion" impact different parts of the self-control cluster concept (and some such tasks meaningfully impact *none* of the parts that have the potential to be depleted, which could explain the null findings that sometimes occur).

Let's quickly review some aspects of self-control that may be prone to "depletion" (taken here to broadly mean that they may drop in certain circumstances in a way that could lead to worse "performance" later).

1. Depletion of Motivation

As noted above, how motivated you are clearly matters for self-control. If you thought you were going to die immediately after eating your favorite unhealthy food, I'm certain you would stop eating it.

Now, suppose that you enrolled in a psychology study to contribute to advancing science or to get some class credit. If the study required you to do a highly frustrating task for long enough, you may have diminished motivation to try hard on the next (potentially even more frustrating) part of that stupid experiment. Is it any wonder

that you might then give up on an impossible puzzle task sooner than those who were given a less frustrating task at the beginning? Screw those researchers. In the failed replication attempt mentioned above, motivation was not measured (so we don't know whether the difficult 'e' finding task reduced motivation). In any event, a meaningless 'e' finding task may not be an ideal simulation of real-world scenarios.

Remember all those simple self-control strategies we talked about, that you can apply even if you aren't strong in self-control-related traits? Another potentially interesting form of motivation-related depletion might result from an ego-depletion task causing people to lose motivation to apply their most effective self-control strategies! For instance, at the beginning of the study, you might have been deploying your best weapons against scarfing down the cake they gave you, but, by the end of the study, you don't feel motivated to use those strategies anymore.

2. Depletion of Energy

It will come as a great shock to you (if you were not born on Earth and this is your first encounter with human culture) that some things make people tired, and that once people are tired, they perform worse at tasks and often want to take a break, and that after such a break they may well feel recharged. As I think we have all experienced, there are some things that are physically tiring (like carrying heavy boxes around) and others that are mentally tiring (like doing hard math homework, or trying to fill out complicated, ambiguous forms that require cross-referencing information from many documents at once). Energy itself can probably be subdivided into at least four different parts (sleepiness, mental tiredness, physical tiredness, and slowness, as we did [here](#)), though for our purposes that's probably more subdivision than we need.

Now, imagine that you enroll in a psychology study that requires you to maintain intense concentration, and you start to feel mentally tired. Is it really a surprise that you would then perform worse, or give up sooner, on a subsequent task, than if the original task had been less tiring? It would really surprise me if no such effect existed. It's worth noting that the big failed replication attempt mentioned above found that the ego-depletion task did not actually induce fatigue! It's possible the results would have been different had the task-induced fatigue, though we don't know.

Interestingly, there's [a theory](#) proposed in the academic literature (that I learned about after writing the first draft of this essay) that the ego depletion effect, rather than being connected to self-control specifically, is better thought of as being due to "transient cognitive fatigue."

3. Depletion of Desire to Delay Gratification

If you're like me, you sometimes allow yourself to indulge more than usual after you've worked really hard or if you've done a really good job. For instance, if I've worked out really hard, I'm less likely to feel bad about indulging in an unhealthy meal, even though it is just as unhealthy after a big workout. I suspect that, at least for me, this is because I have a barometer of "how good a job I'm doing at life," and I try to keep it above a certain level. Similarly, I notice that I allow myself to indulge more when I'm going through a challenging period (e.g., if I had a really bad day, I cut myself some slack in terms of keeping up healthy habits). I think a lot of people do something similar. In other words, the extent to which we delay gratification isn't a

fixed thing (even though some people have a much higher average tendency to do so than others) - we allow ourselves to do more or less of it based on various parameters. It's not too hard to believe that after contributing to science by doing that really annoying task over and over again, you're ready to let yourself eat those goddamn jelly beans they offer.

Note that I'm not claiming that the three self-control-related concepts above are the only ones that have the potential to "deplete" - but they are the ones I currently see as being the most likely to do so. It wouldn't surprise me though if, for instance, some difficult mental tasks made us feel hungry faster than others (though I doubt the 'e' finding task mentioned above would do so). Obviously, a high degree of hunger can cause us to perform less well on some subsequent tasks, and can cause us to give into food urges more readily. This potential connection between hunger and self-control is not so different from the hotly debated [link between self-control and glucose](#).

On the other hand, there are the other trait-like aspects of self-control that seem far less likely to deplete. For instance, it's not as clear how one's "awareness of temptation" would get depleted, or "lack of unhealthy desires," or "tendency to get into flow states."



Some people may want to say that things like "energy" and "motivation" are (by definition) not part of self-control or (by definition) not related to ego depletion. I don't have a horse in that semantic debate, except to say that we should try to be really clear regarding what we're talking about. I think that a lack of such clarity (even though some researchers have made efforts to clarify terms) has substantially contributed to the strange situation we find ourselves in - hundreds of studies have taken place yet there still being a debate over whether the phenomenon being studied exists at all! As [Wikipedia explains](#):

"A 2010 meta-analysis of 198 independent tests [of ego depletion] found the effect significant with a moderate effect size ($d = .6$). Even after accounting for possible unpublished failed studies, the analysis concluded that it is extremely unlikely that the effect doesn't exist. In 2015, [a meta-analysis of over 100 studies](#) by Carter and McCullough argued that the 2010 meta-analysis failed to take publication bias into account. They showed statistical evidence for publication bias. When they statistically controlled for publication bias, the effect [was not significantly different from zero]...In response, Cunningham and Baumeister argued that Carter and McCullough analysis contained errors in its data collection and in the various analyses used."

Some conclude from this research that ego depletion simply isn't real. That possibility should be taken seriously.

On the other hand, I bet you have abundant first-hand evidence that:

- some tasks tire people out, causing them to perform less well on subsequent tasks (e.g., "I've been working at this for 5 hours - now I'm totally exhausted and barely making progress")
- some tasks demotivate people, causing them to try less hard on subsequent tasks (e.g., "I clearly suck at math - I'm going to fail this test even if I try my hardest.")
- when people have worked hard they relax their self-control, allowing themselves to splurge (e.g., "I worked my arse off today on that assignment, I deserve a night of Netflix with a chocolate fudge sundae.")

We could deny that any of the above are included in what we mean by "ego depletion." If we strip away enough of these meanings then, well, there is nothing left, and surely ego depletion won't be a thing anymore. This is why, again, it's critical to get clear on what we're really talking about. If we are asking people to do something like finding all instances of the letter 'e' in a text (with some other weird constraints to make the task more difficult), what **precisely** are we trying to deplete?

If what I've claimed is true (that self-control has different parts in an important sense) then it becomes weirdly difficult to interpret meta-analyses. If some aspects of self-control deplete (including pretty boring stuff like energy and motivation), and others don't (perhaps things like the tendency to notice when you're experiencing a temptation), then averaging across many studies (even if done perfectly) just gives information about the average level of self-control depletion across the many different protocols studied. In other words, you're averaging across some situations where we would expect depletion effects and others where we wouldn't, and the result is based on how many studies there are of each. Ideally, a meta-analysis handles this by looking for heterogeneity in studies, but this is a lot easier said than done (and is even tougher to do without a strong theory of what the important sorts of heterogeneity are

- because you ultimately have to decide which studies to combine and which to leave separate).

In Summary

So, what to conclude? Well, I think that "self-control" refers to many different traits. And, due to the considerations I've outlined above, I think that "self-control" sometimes gets depleted, and it sometimes doesn't. Whether it does or not depends on what aspect of self-control we're talking about and what the task is. Different tasks, I claim, will impact different aspects of self-control (some depletable, others not), and different ways of measuring depletion will measure different aspects of self-control. The details matter, because in an important sense self-control is not one thing!

Furthermore, as we've discussed, many traits impact self-control other than "classic" self-control (noticing temptation + overriding it). In practice, self-controlled behavior seems to be the result of a combination of traits (some of which are likely at least partially heritable), and a person's ability to learn self-control strategies.

Finally, if you want to gain more control in your life, you may benefit from some of the strategies covered earlier in this article. If you haven't tried the step-by-step process (above) aimed at increasing control in your life, I recommend giving it a try now!

Remember these ideas!

Related papers:

- Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). [Situational strategies for self-control](#). *Perspectives on Psychological Science*, 11(1), 35-55.
- Duckworth, A. L., Milkman, K. L., & Laibson, D. (2018). [Beyond willpower: Strategies for reducing failures of self-control](#). *Psychological Science in the Public Interest*, 19(3), 102-129.
- Vosgerau, J., Scopelliti, I., & Huh, Y. E. (2020). [Exerting self-control ≠ sacrificing pleasure](#). *Journal of Consumer Psychology*, 30(1), 181-200.
- Galla, B. M., & Duckworth, A. L. (2015). [More than resisting temptation: Beneficial habits mediate the relationship between self-control and positive life outcomes](#). *Journal of personality and social psychology*, 109(3), 508.
- Milyavskaya, M., Inzlicht, M., Hope, N., & Koestner, R. (2015). [Saying “no” to temptation: Want-to motivation improves self-regulation by reducing temptation rather than by increasing self-control](#). *Journal of Personality and Social Psychology*, 109(4), 677.
- Lin, H., Saunders, B., Friese, M., Evans, N. J., & Inzlicht, M. (2020). [Strong effort manipulations reduce response caution](#): A preregistered reinvention of the ego-depletion paradigm. *Psychological science*, 31(5), 531-547.
- Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., ... & Zinkernagel, A. (2021). [A multilab replication of the ego depletion effect](#). *Social*

- Psychological and Personality Science, 12(1), 14-24.
- Giboin, L. S., & Wolff, W. (2019). [The effect of ego depletion or mental fatigue on subsequent physical endurance performance](#): A meta-analysis. Performance Enhancement & Health, 7(1-2), 100150.
 - Hurley, P. J. (2021, February 27). [Reconceptualizing Ego Depletion as Transient Cognitive Fatigue](#)

The average North Korean mathematician

This is a linkpost for <https://www.telescopic-turnip.net/essays/the-average-north-korean-mathematician/>

Here are [the top-fifteen countries](#) ranked by how well their teams do at the International Math Olympiads:



When I first saw this ranking, I was surprised to see that North Koreans have such an impressive track record, especially when you factor in their relatively small population. One possible interpretation is that East Asians are just particularly good at mathematics, just like in the stereotypes, even when they live in one of the world's worst dictatorships.

But I don't believe that. In fact, I believe North Koreans are, on average, *particularly bad* at math. More than [40% of the population is undernourished](#). Many of the students involved in the IMO grew up in the 1990s, during the [March of Suffering](#), when hundreds of thousands of North Koreans died of famine. That is not exactly the best context to learn mathematics, not to mention the [direct effect](#) of nutrients on the brain. There does not seem to be a lot of famous North Korean mathematicians either (there is actually a candidate from the North Korean IMO team who managed to escape during the 2016 Olympiads in Hong-Kong. He is now living in South Korea. I wish him to become a famous mathematician). Thus, realistically, if all 18 years-old from North Korea were to take a math test, they would probably score much worse than their South Korean neighbors. And yet, Best Korea reaches almost the same score with only half the source population. What is their secret?

[This piece](#) on the current state of mathematics in North Korea gives it away:

"The entire nation suffered greatly during and after the March of Suffering, when the economy collapsed. Yet, North Korea maintained its educational system, focusing on the gifted and special schools such as the First High Schools to preserve the next generation. The limited resources were concentrated towards gifted students. Students were tested and selected at the end of elementary school."

In that second interpretation, the primary concern of the North Korean government is to produce a few very brilliant students every year, who will bring back medals from the Olympiads and make the country look good. The rest of the population's skills at mathematics are less of a concern.

When we receive new information, we update our beliefs to keep them compatible with the new observations, doing an informal version of [Bayesian updating](#). Before learning about the North Korean IMO team, my prior beliefs were something like most of the country is starving and their education is mostly propaganda, there is no way they can be good at math. After seeing the IMO results, I had to update. In the first interpretation, we update the *mean* - the average math skill is higher than I previously thought. In the second interpretation, we leave the mean untouched, but we make the *upper tail* of the distribution heavier. Most North Koreans are not particularly good at math, but a few of them are heavily nurtured for the sole purpose of winning medals

at the IMO. As we will see later in this article, this problem has some pretty important consequences for how we understand society, and those who ignore it might take pretty bad policy decisions.

But first, let's break it apart and see how it really works. There will be a few formulas, but nothing that can hurt you, I promise. Consider a probability distribution where the outcome x happens with probability $p(x)$. For any integer n , the formula below gives what we call the n th moment of a distribution, centered on μ .

$$\int_{\mathbb{R}} p(x) (x - \mu)^n dx$$

To put it simply, moments describe how things are distributed around a center. For example, if a planet is rotating around its center of mass, you can use moments to describe how its mass is distributed around it. But here I will only talk about their use in statistics, where each moment encodes one particular characteristic of a probability distribution. Let's sketch some plots to see what it is all about.

First moment: replace n with 1 and μ with 0 in the previous formula. We get

$$\int_{\mathbb{R}} p(x) x dx$$

which is – surprise – the definition of the mean. Changing the first moment just shifts the distribution towards higher or lower values, while keeping the same shape.



Second moment: for $n = 2$, we get

$$\int_{\mathbb{R}} p(x) (x - \mu)^2 dx$$

If we set μ to be (arbitrarily, for simplicity) equal to the mean, we obtain the definition of the variance! The second moment around the mean describes how values are spread away from the average, while the mean remains constant.



Third moment ($n = 3$): the third moment essentially describes how skewed (asymmetric) the distribution is.



Fourth moment ($n = 4$): this describes how *leptokurtic* or *platykurtic* your distribution is, that is, how extreme the extreme values are.



You could go on to higher n , each time bringing in more detail about what the distribution really looks like, until you end up with a perfect description of the distribution. By only mentioning the first few moments, you can describe a population with only a few numbers (rather than infinite), but it only gives a simplified version of the true distribution, as on the left graph below:



Say you want to describe the height of humans. As everybody knows, height follows a normal distribution, so you could just give the mean and standard deviation of human height, and get a fairly accurate description of the distribution. But there is always a wise-ass in the back of the room to point out that the normal distribution is defined over \mathbb{R} , so for a large enough population, some humans will have a negative height. The problem here is that we only gave information about the first two moments and neglected all the higher ones. As it turns out, humans are only viable within a certain range of height, below or above which people don't survive. This erodes the tails of the distribution, effectively making it more *platykurtic* (If I can get even one reader to use the word *platykurtic* in real life, I'll consider this article a success).

Let's come back to the remarkable scores of North Koreans at the Math Olympiads. What these scores teach us is not that North Korean high-schoolers are really good at math, but that *many of the high-schoolers who are really good at math are North Koreans*. On the distribution plots, it would translate to something like this:



With North Koreans in purple and another country that does worse in the IMO (say, France), in black. So you are looking at the tails and try to infer something about the rest of the distribution. Recall the plots above. Which one could it be?



Answer: just by looking at the extreme values, you cannot possibly tell, because any of these plots would potentially match. In Bayesian terms, each moment of the distribution has its own prior, and when you encounter new information, you could in principle update any of them to match the new data. So how can we make sure we are not updating the wrong moment? When you have a large representative sample that reflects the entire distribution, this is easy. When you only have information about the top 10 extreme values, it is impossible. This is unfortunate because the extreme values are precisely what gets all our attention – most of what we see in the media is about the most talented athletes, the most dishonest politicians, [the craziest people](#), the most violent criminals, and so forth. Thus, when we hear new information about extreme cases, it's important to be careful about *which moment to update*.

This problem also occurs in reverse – in the same way looking at the tails doesn't tell you anything about the average, looking at the average doesn't tell you anything about the tails. An example: on a typical year, more Americans die [from falling than from viral infections](#). So one could argue that we should dedicate more resources to prevent falls than viral infections. Except the number of deaths from falls is fairly stable (you will never have a pandemic of people starting to slip in their bathtubs 100 times more than usual). On the other hand, virus transmission is a multiplicative process, so most outbreaks will be mostly harmless (remember how SARS-cov-1 killed less than 1000 people, those were the days) but a few of them will be really bad. In other words, yearly deaths from falls have a higher mean than deaths from viruses,

but since the latter are highly skewed and leptokurtic, they might deserve more attention. (For a detailed analysis of this, [just ask Nassim Taleb](#).)

There are a lot of other interesting things to say about the moments of a probability distribution, like the [deep connection](#) between them and the partition function in statistical thermodynamics, or the fact that in my drawings the purple line always crosses the black line exactly n times. But these are for nerds, and it's time to move on to the secret topic of this article. Let's talk about SEX AND VIOLENCE.

This will not come as a surprise: [most criminals are men](#). In the USA, men represent [93% of the prison population](#). Of course, [discrimination in the justice system](#) explains some part of the gap, but I doubt it accounts for the whole 9-fold difference. Accordingly, it is a solid cultural stereotypes that [men use violence and women use communication](#). Everybody knows that. Nevertheless, having just read the previous paragraphs, you wonder: "are we really updating the right moment?"

A recent meta-analysis by [Thöni et al.](#) sheds some light on the question. Published in the journal *Psychological Science*, it synthesizes 23 studies (with >8000 participants), about gender differences in cooperation. In such studies, participants play cooperation games against each other. These games are essentially a multiplayer, continuous version of the Prisoner's Dilemma – players can choose to be more or less cooperative, with possible strategies ranging from total selfishness to total selflessness.

So, in cooperation games, we expect women to cooperate more often than men, right? After all, women are socialized to be caring, supportive and empathetic, while men are taught to be selfish and dominant, aren't they? To find out, Thöni et al aligned all of these studies on a single cooperativeness scale, and compared the scores of men and women. Here are the averages, for three different game variants:



This is strange. On average, men and women are just equally cooperative. If society really allows men to behave selfishly, it should be visible somewhere in all these studies. I mean, where are all the criminals/rapists/politicians? It's undeniable that most of them are men, right?

The problem with the graph above is that it only shows averages, so it misses the most important information – that men's level of cooperation is much more *variable* than women's. So if you zoom on the people who were either very selfish or very cooperative, you find a wild majority of men. If you zoom on people who kind-of cooperated but were also kind-of selfish, you find predominantly women.



As I'm sure you've noticed, the title of the Thöni et al paper says "evolutionary perspective". As far as I'm concerned, I'm fairly skeptical about evolutionary psychology, since it is one of the fields with the [worst track record](#) of reproducibility ever. To be fair, a good part of evpsych is just regular psychology where the researchers added a little bit of speculative evolutionary varnish to make it look more exciting. This aside, *real* evpsych is apparently [not so bad](#). But that's not the important part of the paper – what matters is that there is increasingly strong evidence that men are indeed more variable than women in behaviors like cooperation. Whether it is due to hormones, culture, discrimination or [cultural](#)

[evolution](#) is up to debate and I don't think the current data is remotely sufficient to answer this question.

(Side note: if you must read *one* paper on the topic, I recommend [this German study](#), where they measure the testosterone level of fans of a football team, then have them play Prisoner's Dilemma against fans of a rival team. I wouldn't draw any strong conclusion from this *just yet*, but it's a fun read.)

The thing is, men are not only found to be more variable in cooperation, but in tons of other things. These include [aggression](#), [exam grades](#), [PISA scores](#), all kinds of [cognitive tests](#), [personality](#), [creativity](#), [vocational interests](#) and even some [neuroanatomical features](#). In the last few years, support for the greater male variability hypothesis has accumulated, so much that it is no longer possible to claim to understand gender or masculinity without taking it into account.

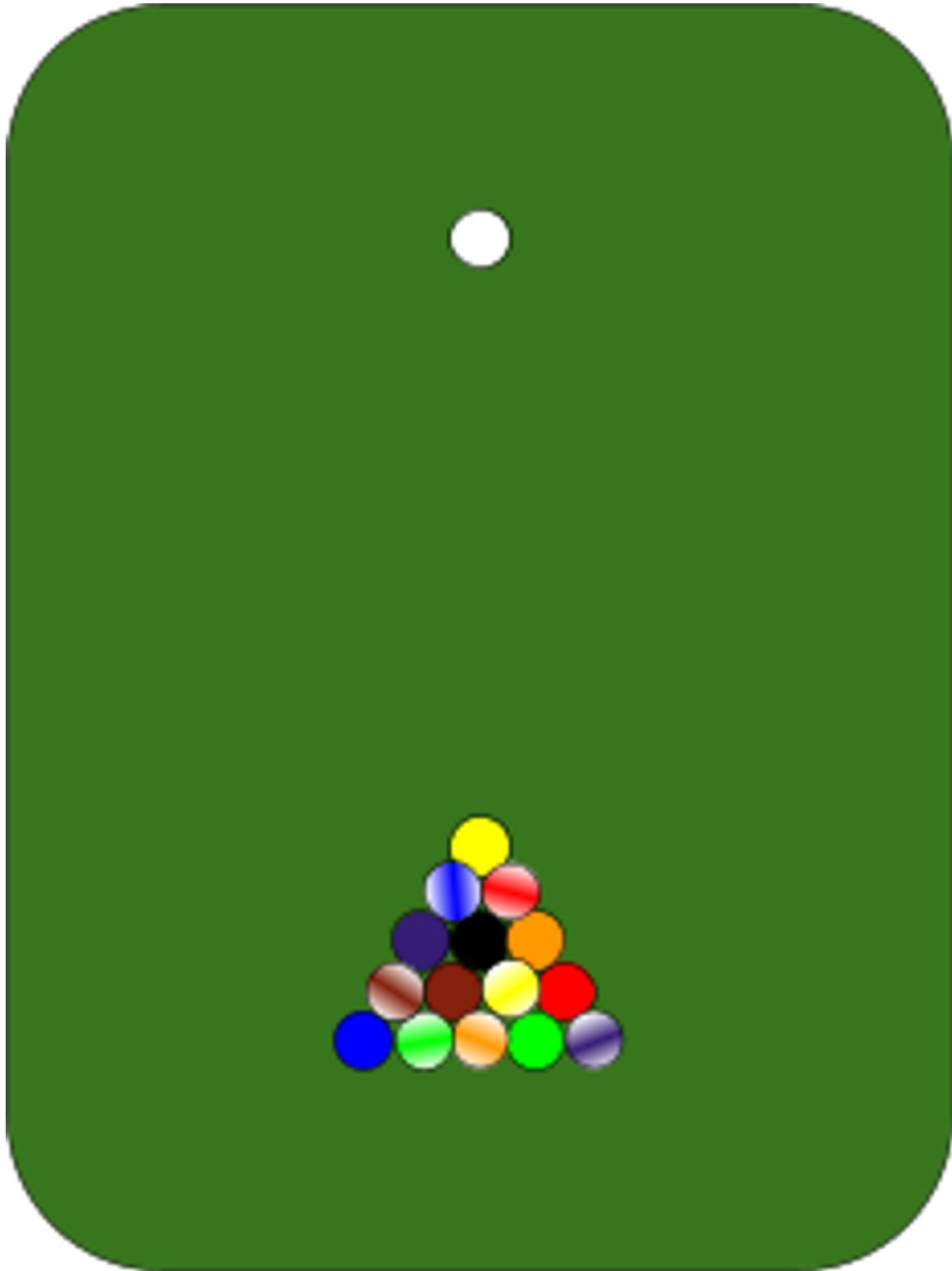
Alas, that's not how stereotyping works. Instead, we see news report showing all these male criminals, and assume that our society turns men into violent and selfish creatures and call them toxic [Here is [Dworkin](#): "*Men are distinguished from women by their commitment to do violence rather than to be victimized by it. Men are rewarded for learning the practice of violence in virtually any sphere of activity by money, admiration, recognition, respect, and the genuflection of others honoring their sacred and proven masculinity.*" Remember – in the above study, the majority of unconditional cooperators were men]. Internet people make up a [hashtag](#) to ridicule those who complain about the generalization. We see all these male [IMO medalists](#), and – depending on your favorite political tradition – either assume that men have an unfair advantage in maths, or that they are inherently better at it. The former worldview serves as a basis for [public policy](#). The question of which moment to update rarely even comes up.

This makes me wonder whether this process of looking at the extremes then updating our beliefs about the mean is just the normal way we learn. If that is the case, how many other things are we missing?

Chaos Induces Abstractions

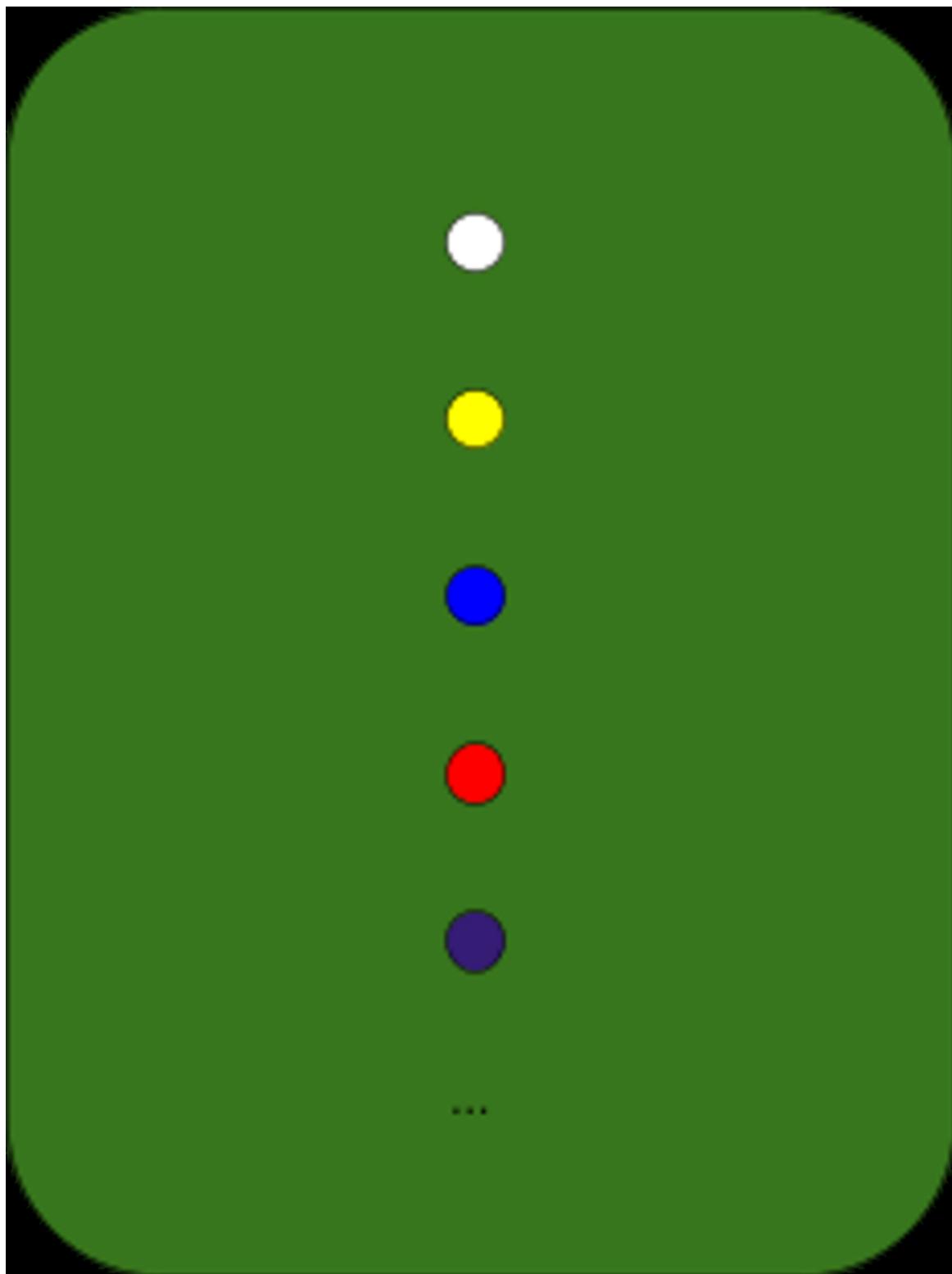
Epistemic status: the first couple sections are intended to be a bog-standard primer on chaos theory. In general, this post mostly sticks close to broadly-accepted ideas; it's intended mainly as background for why one would expect the general ideas of [abstraction-as-information-at-a-distance](#) to be true. That said, I'm writing it all from memory, and I am intentionally sweeping some technical details under the rug. If you see a mistake, please leave a comment.

Consider a billiards table:

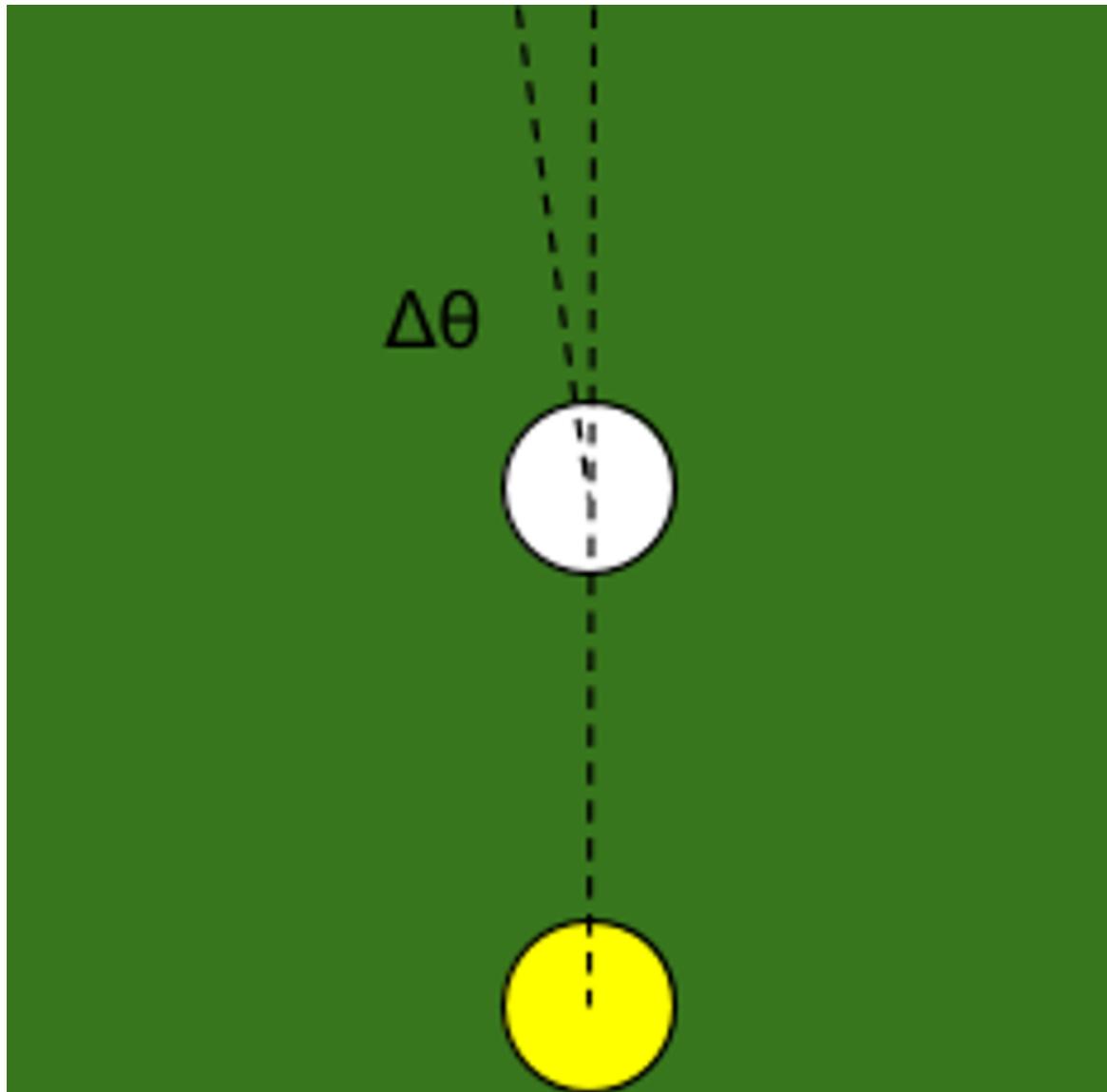


The particular billiards table we'll use is one I dug out of the physicists' supply closet, nestled in between a spherical cow and the edge of an infinite plane. The billiard balls are all frictionless, perfectly spherical, bounce perfectly elastically off of other balls and the edges of the table, etc.

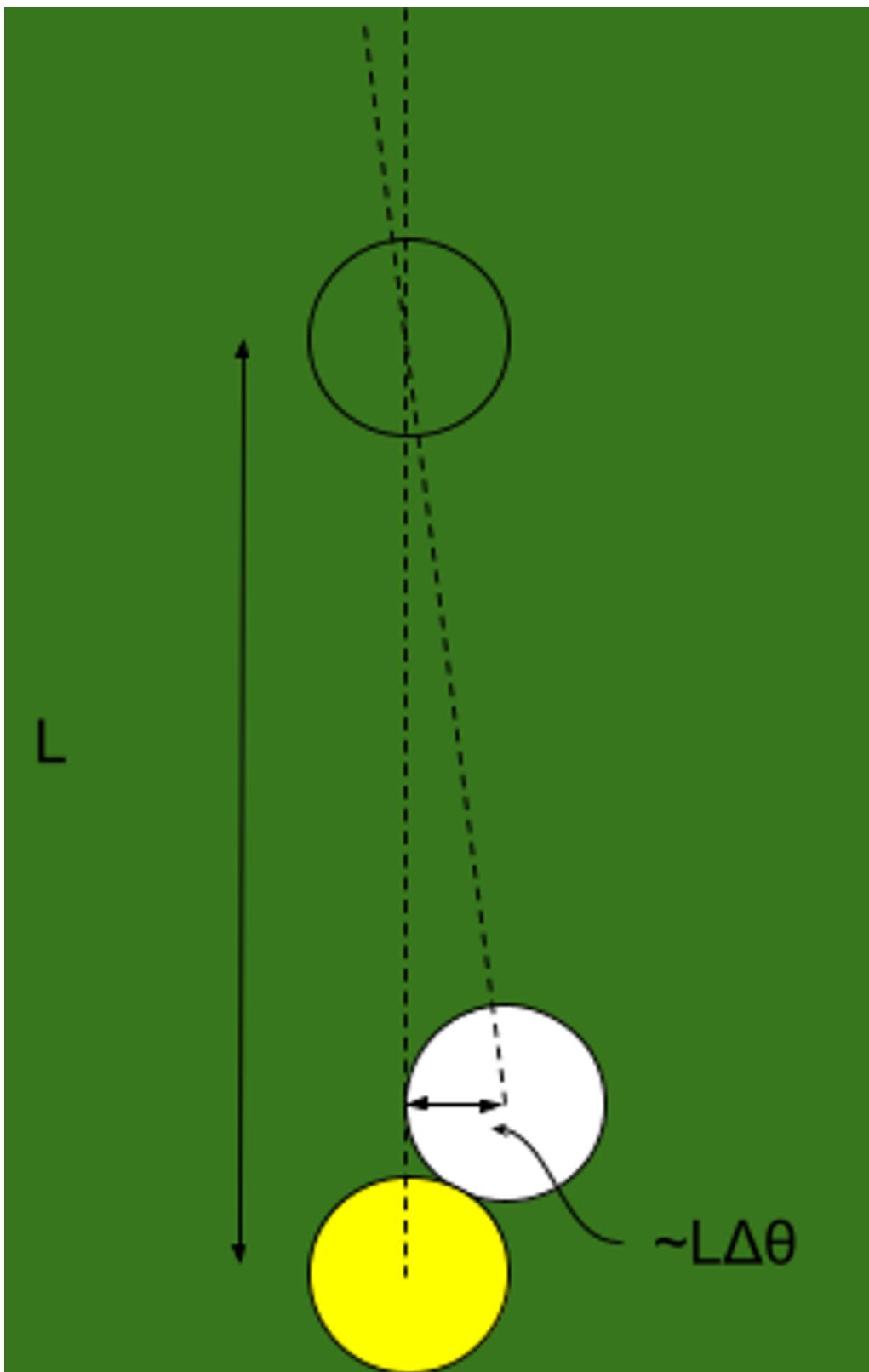
Fun fact about billiard balls: if my aim has a tiny bit of error to it, and I hit a ball at ever-so-slightly the wrong angle, that error will grow exponentially as the balls collide. Picture it like this: we start with an evenly-spaced line of balls on the table.



I try to shoot straight along the line, but the angle is off by a tiny amount, call it $\Delta\theta$.

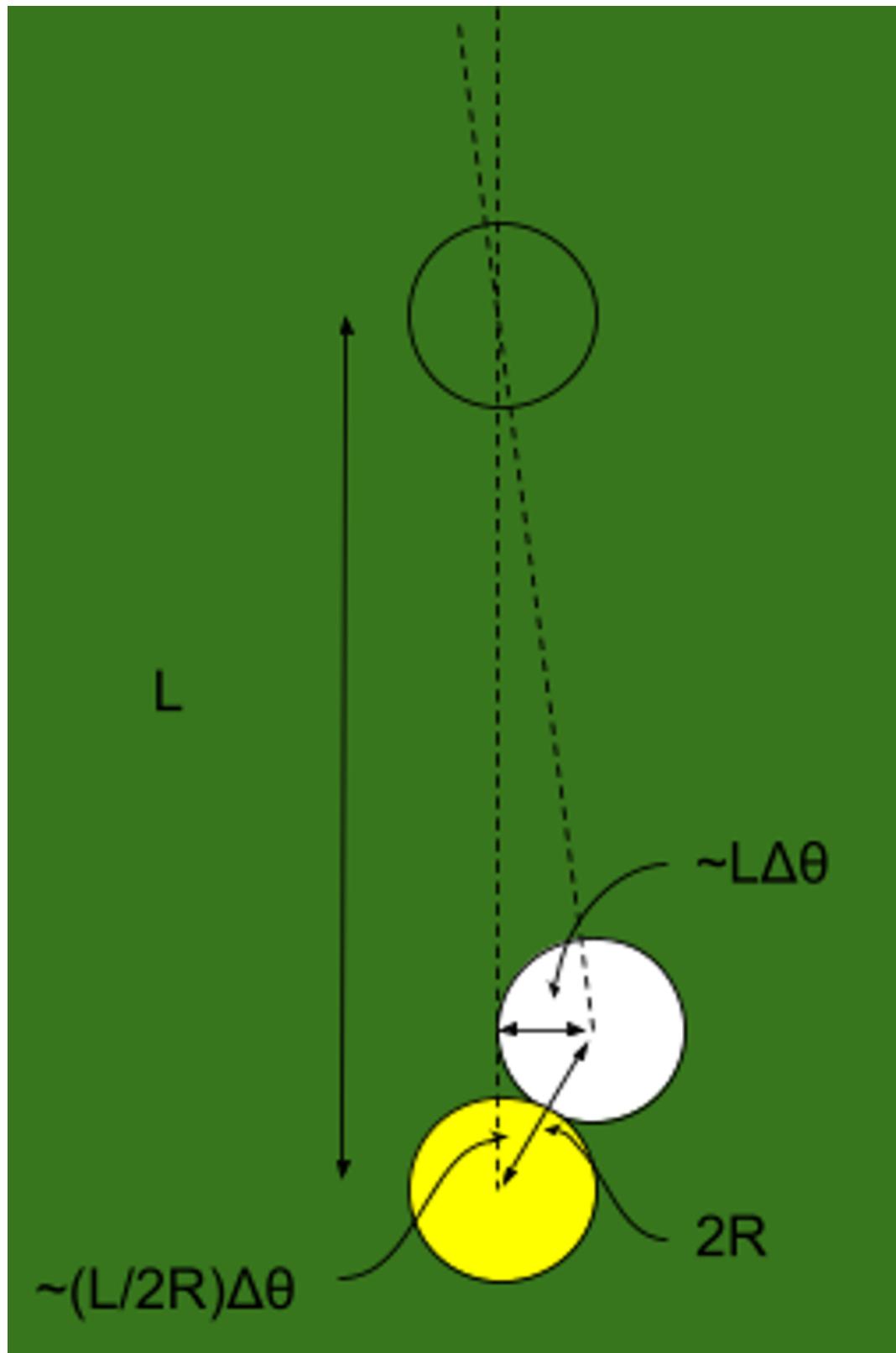


The ball rolls forward, and hits the next ball in line. The *distance* by which it's off is roughly the ball-spacing length L multiplied by $\Delta\theta$, i.e. $L\Delta\theta$.

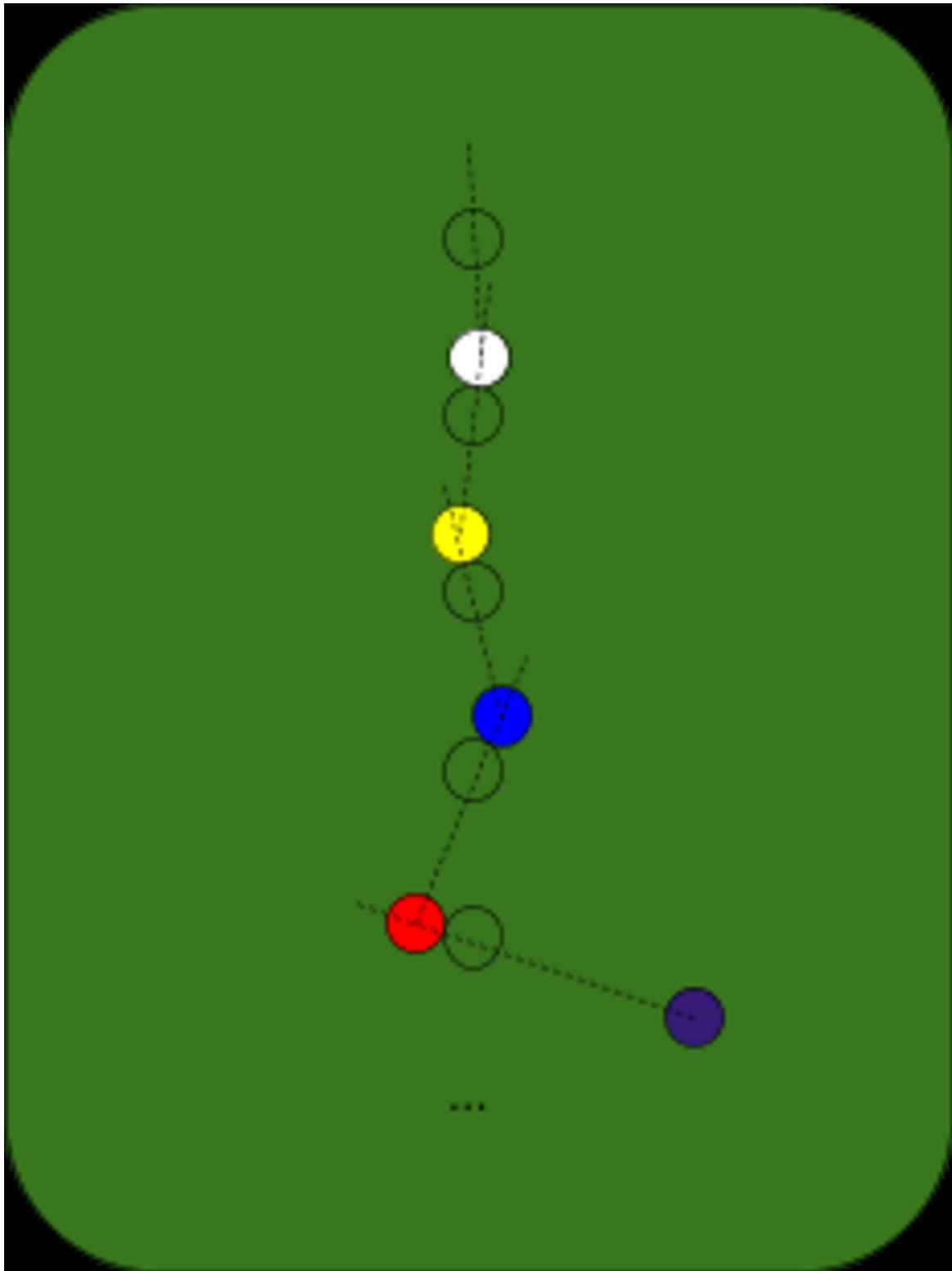


Since the first ball hits the second ball off-center, the second ball will also have some error in its angle. We do a little geometry, and find that the angular error in the second ball is

roughly $\frac{L}{2R}\Delta\theta$, where R is the radius of a ball.



Now the second ball rolls into the third. The math is exactly the same as before, except the initial error is now multiplied by a factor $\frac{1}{2}$. So when the second ball hits the third, the angular error in the third ball will be multiplied *again*, yielding error $(\frac{1}{2})^2 \Delta\theta$. Then the next ball will have angular error/uncertainty $(\frac{1}{2})^3 \Delta\theta$. And so forth.



Upshot of all this: in a billiard-ball system, small angular uncertainty grows *exponentially* with the number of collisions. (In fact, this simplified head-on collision scenario yields the *slowest* exponential growth; if the balls are hitting at random angles, then the uncertainty grows even faster.)

This is a prototypical example of mathematical chaos: small errors grow exponentially as the system evolves over time. Given even a tiny amount of uncertainty in the initial conditions (or a tiny amount of noise from air molecules, or a tiny amount of noise from an uneven table surface, or ...), the uncertainty grows, so we are unable to precisely forecast ball-positions far in the future. If we have any uncertainty, then as we forecast further and further into the future, our predictions will converge to a maximally-uncertain distribution on the state-space (or more precisely, a maxentropic distribution on the phase space). We become maximally uncertain about the system state.

... except for one prediction.

Remember that these are Physicists' Billiards™, so they're frictionless and perfectly elastic. No matter how many collisions occur, energy is always conserved. We may have some initial uncertainty about the energy, or there may be some noise from air molecules, etc, but the system's own dynamics will not amplify that uncertainty the way it does with other uncertainty.

So, while most of our predictions become maxentropic (i.e. maximally uncertain) as time goes on, we can still make reasonably-precise predictions about the system's energy far into the future.

An Information-Theoretic Point Of View

At first glance, this poses a small puzzle. We have very precise information about the initial conditions - our initial error is very small. The system's dynamics are deterministic and reversible, so information can't be lost. Why, then, do our predictions become maximally uncertain?

The key is that the angular error is a real number, and specifying a real number takes an infinite number of bits - e.g. it might be $\Delta\theta = 0.0013617356590430716 \dots$. Even though it's a small real number, it still has an infinite number of bits. And as the billiards system evolves, bits further and further back in the binary expansion become relevant to the large-scale system behavior. But the bits far back in the binary expansion are exactly the bits about which we have approximately-zero information, so we become maximally uncertain about the system state.

Conversely, our initial information about the large-scale system behavior still tells us a lot about the future state, but most of what it tells us is about bits far back in the binary expansion of the future state variables (i.e. positions and velocities). Another way to put it: initially we have very precise information about the leading-order bits, but near-zero information about the lower-order bits further back. As the system evolves, these mix together. We end up with a lot of information about the leading-order and lower-order bits combined, but very little information about either one individually. (Classic example of how we can have lots of information about two variables combined but little information about either individually: I flip two coins in secret, then tell you that the two outcomes were the same. All the information is about the relationship between the two variables, not about the individual values.) So, even though we have a lot of information about the microscopic system state, our predictions about large-scale behavior (i.e. the leading-order bits) are near-maximally uncertain.

... except, again, for the energy. Our information about the energy does not get mixed up with the lower-order bits, so we can continue to precisely forecast the system's energy far into the future. We end up maximally uncertain about large-scale system state except for our precise estimate of the energy. (Maximum entropy quantifies this: the distribution most often used in statistical mechanics is maxentropic subject to a constraint on our knowledge of the energy.)

The Abstraction Connection

The basic claim about how abstraction works: we have some high-dimensional system, within some larger environment. The system's variables contain a lot of information. But, most of that information is not relevant "far away" from the system - most of it is "wiped out" by noise in the environment, so only a low-dimensional summary is relevant far away. That low-dimensional summary is the "abstraction" of the high-dimensional system.

Why would we expect this notion of "abstraction" to be relevant to the physical world? Why should all the relevant information from a high-dimensional subsystem (e.g. the molecules comprising a tree or a car) fit into a low-dimensional summary?

The billiards system illustrates one of the main answers.

We have a high-dimensional system - i.e. a large number of "billiard balls" bouncing around on a "pool table". (Really, this is typically used as an analogy for gas molecules bouncing around in a container.) We ask what information about the system's state is relevant "far away" - in this case, far in the future. And it turns out that, if we have even just a little bit of uncertainty, the vast majority of the information is "wiped out". Only the system's energy is relevant to predicting the system state far in the future. The energy is our low-dimensional summary.

To make this example look like something we'd recognize, we need to add a little more information to the summary: as the balls bounce around, the *number* of balls also does not change, and the *volume* of the container - i.e. the size of the pool table - does not change. (Everything we said earlier is still true, we were just treating volume and number of balls as fixed background parameters.) In situations where balls might be added/removed, or where the container might grow/shrink, the system dynamics still does not amplify uncertainty in those quantities; information about them is still relevant to predicting the far-future state. So, three quantities - energy, volume, and number of balls - provide a summary of all the information relevant to forecasting the system state in the far future.

You may recognize these as the "state variables" of an ideal gas. (We could instead swap in equivalent variables which contain the same information, e.g. pressure, temperature and volume). The ideal gas model is an abstraction, and the state variables of the ideal gas are exactly the low-dimensional summary which contains all the information from the high-dimensional system state which is relevant far into the future. Chaos wipes out all the other information.

Key idea from dynamical systems theory: this is how *most* dynamical systems behave by default. Information about some quantities is conserved, and everything else is wiped out by chaos. So, for most systems, we should expect that all the relevant information from the high-dimensional system state can fit into a low(er)-dimensional summary.

The key question is *how much* lower dimensional. If we have a mole of variables, and the summary "only" needs a mole minus 10 million, that's not very helpful. Yet in practice, the low-dimensional summaries seem to be much smaller than that - not necessarily as small as the three state variables of an ideal gas, but still a lot less than a mole of variables. Ultimately, this is a question which needs to be answered empirically.

Other Paths

I don't want to leave people with too narrow a picture, so let's talk a bit about other paths to the same underlying concept.

First, chaos isn't the only path to a similar picture. For instance, a major topic in computational complexity theory is systems which take in k random bits, and output $n >> k$ pseudorandom bits, such that there's no polynomial-time (as a function of k) method which can distinguish the n pseudorandom bits from truly random variables (assuming $P \neq NP$).

Conceptually, this has similar consequences to chaos: if we have even just a few unknown bits, then they can "wipe out the information" in other bits - not in an information-theoretic sense, but in the sense that the wiped-out information can no longer be recovered by any polynomial-time computation. As with continuous dynamical systems, most random discrete systems will conserve information about some quantities, and everything else will be wiped out by the pseudorandom noise.

Second, we don't just need to think about dynamics over time, or about "far away" as "far in the future". For instance, we could imagine two separate systems whose interactions are mediated by a chaotic system - for instance, two people talking, with the sound waves carried by the (chaotic) system of air molecules between them. Chaos will wipe out most of the information about one system (e.g. motions of individual molecules in one person) before that information reaches the other system. All that gets through will be a low(er)-dimensional summary (e.g. larger-scale sonic vibrations). More generally, we can think about information propagating through many different subsystems, in a whole graph of interactions, and then ask what information is conserved as we move outward in the graph.

Third, if we have non-small amounts of uncertainty about the environment, then that uncertainty can wipe out information about the system even without chaos added into the mix. Simple example: the system is a flipped coin. The environment is another flipped coin. Standing "far away", I cannot "see" either coin, but someone tells me whether they are equal (i.e. HH or TT) or unequal (HT or TH). If I had better information about the environment, then this would be sufficient to figure out the system's state. But my uncertainty about the environment also wipes out the information about the system; I'm left maximally uncertain about whether the system's state is H or T. This is very similar to the "mixing" of information we talked about earlier; as before, a given system is likely to have some quantities which do get "mixed in" with things we don't know, and some quantities which do not get mixed. The latter serve as the low-dimensional summary.

Recap of Main Points

In a chaotic system, small uncertainties/errors are amplified over time. If there's even just a tiny amount of uncertainty - whether from uncertain initial conditions or noise from the environment - then the large-scale behavior of the system becomes unpredictable far in the future.

... but not *completely* unpredictable. Typically, *some* information is conserved - e.g. the energy in a frictionless physical system. Even if noise from the environment causes some uncertainty in this conserved information, it isn't "amplified" over time, so we can still make decent predictions about these quantities far into the future.

In terms of abstraction: the conserved information is a low(er) dimensional summary, which contains all the known information about the large-scale system state relevant to large-scale measurements of the system's state in the far future.

MetaPrompt: a tool for telling yourself what to do.

MetaPrompt is a nonstandard todo list. The very bare-bones basic idea is that you add ideas to a deck of cards, and you shuffle this deck to draw things to do. This is useful for "tasks" with no priority, where you just want to be reminded of the idea at some point in the future. You can also add cards which tell you to add more cards; hence the name.

Most of this is written in a very narrative way. If you just want the how-to, skip to the [how-to section](#).

For many people, there's something interesting about being told what to do: it's easier to make someone else a sandwich than to make one for yourself, for example. One partial explanation is [decision fatigue](#): deciding to do things for yourself takes some sort of energy, which being told to do doesn't take.

This suggests that, in some situations, people can benefit from 'willpower partners': you each (for example) tell each other to make a sandwich. You end up with two sandwiches and no decision fatigue; a net gain over you each making a sandwich for yourself.

But what if you could be your own "willpower partner"?

In 2016, I was learning dvorak. I wanted typing practice; but, I didn't want *boring* typing practice. I remembered [the post about WriterKata on Agency Duck](#) from the previous year. WriterKata was a great little website for practicing writing by responding to writing prompts. Why not practice typing and creative writing at the same time, by responding to prompts?

Unfortunately, by 2016, *WriterKata no longer existed*. It seems the author took the website down. I guess it wasn't worth the hosting expense? (Sad; it seemed good enough that it should have been able to support itself somehow.)

I could have just found writing prompts somewhere and typed my responses into any kind of text file, but I wanted the clean interface: a random prompt selected for me, with a text box directly below it where I could immediately type my response. I know from experience that can be enough to make the difference between vaguely intending to practice for months, vs sitting down and doing it for hours.

I went to the command line. I started a file "kata" for writing prompts, with the idea that I could pull one line from it randomly using standard command line tools. Then, I could write my responses to another file.

However, I was lazy. I didn't want to spend a bunch of time copying writing prompts to the file before beginning.

So, the first line I wrote to kata resembled this (editing for clarity):

```
Write a new writing prompt to kata with 'echo "prompt" >> kata'.
```

In other words, my first prompt was *create a prompt*. This way, I could just keep responding to prompts. Some prompts would tell me to write new prompts (writing to *kata*). Others would be proper creative writing prompts (writing to a different file, *ana*¹). My next four prompts looked like this:

Write a sentence about a stranger on the street.
Write a new sentence kata.
Write a new paragraph kata.
Write a new kata about character creation.

The first one is a regular prompt, but the next three (like the first) are **meta-prompts**: prompts which tell you to make prompts. Why? Because I realized: if I kept it to only *one* metaprompt, the frequency of drawing it would be $1/n$, where n is the number of prompts. My pool of prompts would grow pretty slowly, and I would see a lot of repeat prompts. Better to have more meta-prompts, growing the pool of prompts more quickly, so that individual prompts wouldn't get too old.

This system seemed to work pretty well. I was inspired to write whatever came to mind, and got some typing practice in. Typing commands like echo all the time was only a little annoying.

I was at the command line, so prompts could tell me to run arbitrary commands. For example, I soon wrote this prompt:

Write a continuation of whatever comes out of 'gshuf -n 1 ana'.

(**gshuf -n 1** is how I was sampling random lines from files.) So this one tells me to randomly pick something I had already written, and write a continuation. Unfortunately, there was no link structure: I wasn't connecting bits of text into threads, I was just continuing things I'd written and then dumping the continuations into one big file with no context.

I also created a file called "words", which was just a long list of English words. This could be used in prompts, such as:

Run the command 'gshuf -n 1 words'. Write something using that word.

I started to feel the technical limitations of my approach in other ways. I wanted to create randomized prompts via madlib-style insertion of random words. However, there was no easy way to do this on the command line. Other feature ideas began to swim around in my head as well.

The Vision

I started to fantasize about a website called MetaPrompt, which would streamline what I was doing, add a ton of features, and make the whole thing into an interactive multi-user experience.

- Unlike a regular writing-prompt website, MetaPrompt allows you to build complex custom structures for accomplishing things. I imagined prompts which would let you build a first draft of a whole novel, piece by piece. I imagined meta-prompts which would guide the user to build their *own* complex task workflows, asking them to make a list of lofty goals, and then randomly asking them to fill in details of how they would accomplish those goals.
 - (And keep in mind, the state of *regular* writing-prompt websites is pretty terrible, since the shutdown of WriterKata. So there's not much competition in this sector.)
- A multi-user experience could be really cool. Mainly, users could share prompt lists, which could be very helpful for getting started with the whole thing. But there is also the potential of users collaborating to write stories. Remember the prompt which asked me to continue one of my own bits of writing? What if you could be asked to continue someone else's writing, too?

- (Although this introduces a bunch of design questions.)
- MetaPrompt could also facilitate other things than creative writing. At least, other artsy things, like drawing. (A drawing-based version was ultimately the only multi-user version to see use! So far.)
 - Imagine if you *could* do technical work this way, randomly dredging up old ideas and being prompted to work on them. No line of thinking would just be forgotten; everything would have its chance of being continued at some point.
 - Imagine if your daily spaced-repetition practice wasn't in Anki, but rather, was interspersed with your regular workflow, because *everything* is just randomly drawn cards.

I also thought there was a wider space of possible tools in this space: tools which, like MetaPrompt, offer a sort of "user-programmable dialogue". It's like having a conversation with yourself across time. MetaPrompt is in some sense the simplest version of this, where there's no context sensitivity: the next question MetaPrompt asks you has nothing to do with the previous answer. Iirc, LifeLongLearner created some dialogue tools which walk you through CFAR techniques; I'm imagining something similar to that, but where the user can *create new dialogue tools* through dialogue, rather than just follow someone else's script. This would require some "commands" (ie a simple programming language, *not* an AI trying to figure out what you mean), but the system would walk you through all of that; the pre-loaded dialogues would teach you everything you needed to know about the system, and these could be invoked again when needed (and possibly recur randomly, for spaced repetition).

The general vision, here, is to remove as much cognitive overhead as possible. No more sitting down and deciding what to do; no more questioning whether you feel motivated/inspired to work on project X. Instead, you sit down to an interactive experience as addictive as facebook, but which puts *past you* in the driver seat (rather than the facebook algorithm). Even when you do metacognition to figure out new prompts for future you to work on, it's at the behest of a metaprompt. So you never have to wonder what to work on.

(Of course, you *can* decide to work on a specific thing rather than what the next random prompt tells you to; for example, when MetaPrompt asks me to come up with something of a given type, I often come up with several ideas, and put them all in. And often when I first sit down for a MetaPrompt session I have several new prompt ideas which I put in before doing anything else.)

Three Failures

In the years since 2016, I have convinced no less than *three separate developers* to make prototypes of MetaPrompt for me, all of which have gone unused for one reason or another. The first was an EA who volunteered his time without worrying about business plans or money or any of that, but left the project to work on more effective forms of altruism. The second was my brother, who put together a command-line application (no aspirations of a website for that version), which was pretty capable, but only ever worked on *his* laptop despite attempts to use cross-platform tools. The third was an experienced developer who cared about business plans, expected market size, and returns on time. I guess that attempt mainly fizzled because I got intimidated and didn't feel like I should bug him about anything, such as adding features I thought were important or taking next steps with respect to the business plan.

(Note that I was pretty busy with other things, and never made MetaPrompt my top, second, or even third priority.)

Success

Late 2019, I wanted to do some kind of art project with friends. We settled on the idea of doing daily comic-strip prompts, in which we would often be asked to continue each other's comics with new panels/pages. My brother created yet another version of MetaPrompt, this time facilitated by TiddlyWiki.

In this version, each day, users are supposed to answer one regular prompt and one metaprompt. This ensured that we had quite a large variety of regular prompts after a while. We stuck with it for several months, and had a lot of fun. (I'm now in a very happy relationship with one of the participants, so *that* went well..) We did eventually abandon it, for several reasons:

- Doing a comic page every day is a lot of work, even if you attempt to keep it low-effort and sloppy. Eventually it seemed like too much. We tried to introduce prompts which asked for less work per day, EG separate prompts for scripting a page vs drawing from a script, but by this time we already had a lot of prompts of the higher-effort type; and, I don't think we necessarily came up with great ways of splitting up the work.
- On the technical side, maintaining the site became a huge chore for my brother. He was our tech guy, and we ran to him with every problem and feature request. Several technical challenges presented themselves over time. The fatal challenge involved storing images efficiently. TiddlyWiki loads everything into memory, so uploading new images to the wiki every day worked fine until it suddenly went past the memory limits on the Amazon server we were using. There are workarounds which let you store images on TiddlyWiki without storing them all in memory, but they're a bit technically involved and took some effort. We ended up losing some of our artwork somehow, which damped enthusiasm.

Overall, I think TiddlyWiki is a great way to set up a MetaPrompt. It's already a capable wiki, so you can link material together, give useful tags, and make use of a wide variety of existing TiddlyWiki plugins. You can easily avoid most of the problems we experienced:

- If you make your MetaPrompt one-player rather than multi-player, you don't need to worry about maintaining a server; you can just keep it on your computer. This also means you don't have to worry about managing user accounts, setting up passwords to keep the entire internet from inserting spam into your wiki, etc.
- If you stick with something like creative writing, rather than drawing, you won't need to worry about the memory issues created by images.
- You can sit down and work in MetaPrompt whenever you feel like it, rather than making it a game where you are required to do 1 prompt per day. This way you don't have to balance the time commitment. (Although if you *do* go multi-player, I think once-a-day is a nice format; you get to see everyone else's responses, and it provides motivation to sit down and do the thing.)

I currently maintain a personal MetaPrompt tiddlywiki conforming to those three bullet points. Like my original command-line files (the contents of which I've imported), it's for creative writing practice. The next section details how to set one up.

Making Your Own MetaPrompt with TiddlyWiki

You can skip steps 1, 3 and 4 by downloading a blank MetaPrompt TiddlyWiki [from my github page](#). However, I suggest you at least skim the steps so that you know what's basically going on. (Note that my file includes Stroll; you'll have to put it together yourself if you don't want Stroll.)

1: Download a blank TiddlyWiki.

You can download a blank copy of TiddlyWiki from [the tiddlywiki home page](#). Alternatively, if you like Roam, you can use [Stroll](#), a modified version of TiddlyWiki which adds Roam-inspired features. (To download Stroll, look for the "download Stroll" tab on the [Stroll webpage](#). Consider doing the tutorial first, to get familiar with the features and see whether or not you like them.)

I like Stroll, but if you have no experience with TiddlyWiki, I recommend just starting with vanilla TiddlyWiki. Learning TiddlyWiki can be a bit overwhelming in the first place, so you might not want to add Stroll on top of that.

Speaking of -- you might also want to go through some of the tutorial material listed on the TiddlyWiki website. Also, the [TiddlyWiki5 Reddit](#) is a helpful place to ask questions.

2: Secure your data with a method of saving.

One peculiarity of TiddlyWiki which you'll have to deal with right away: TiddlyWiki itself cannot save changes long-term. Any changes you make will stick around while you keep the wiki open, but if you close it, they'll be gone. You should fix this before you put any serious work into a wiki.

The [TiddlyWiki homepage](#) has plenty of information on this. There are lots of options to choose from -- a dizzying array of them, even. I think the Node.js solution is solid, especially if you might want to serve your TiddlyWiki as a real webpage at some point rather than just on your machine. However, this requires you to run commands on the command line to access your wiki. A "lighter" option is [TWCloud](#). You keep your TiddlyWiki in DropBox (which I do anyway, because I use DropBox to back up everything). When you want to edit it, you visit the TWCloud page and give it permission to access your DropBox. You then navigate to your TiddlyWiki and open it. Definitely not the most secure option, since you have to give an external service permission to read and edit your DropBox. However, it requires no setup and allows you to access your wiki from any computer without having to set up a real website.

There's also a similar service for [Google Drive](#).

3: Install a randomization plugin.

We need a way to select random prompts, and randomize stuff within prompts. One option is the shuffle operator from [Matt's plugin library](#).

Now, randomization in TiddlyWiki is a bit weird. If you just naively use random numbers, you'll run into trouble: the randomness is re-shuffled each time things are re-rendered, which basically means any time you type anywhere in the wiki. This is annoying, and not usually what we want.

Wrangling the randomization to do what you want is probably going to be the most technically involved bit of coding you'll have to do. Unfortunately, I'm not going to go through a whole tutorial here. Instead, I'll just give you some code that works for most purposes. If you're technically inclined, you can probably figure out (with some trouble) what's going on and how to get it to do more. If you're not, you would probably scroll through the tutorial anyway.

4: Make Basic MetaPrompt Tiddlers

"Tiddler" is the TiddlyWiki term for a "page" or "card". The following tiddlers provide the scaffolding for MetaPrompt. Create a new tiddler by pressing the "+" button on the side-bar. This will give you a new page opened in edit mode.

First Prompt

Title your first tiddler "First Prompt" (or whatever you wish). Tag it "prompt" (this step is important!). The text of this first prompt should be generic instructions to add a new prompt:

Create a new prompt. Don't forget to tag it "prompt" so that the prompt shuffler can find it. (Tag it "prompt" whether or not it is a metaprompt; they're all prompts.)

If you're not sure what prompt to add, think about what your MetaPrompt is currently missing. Are there any things you want to do with your MetaPromt which you're not currently doing? Any type of prompt which you wish was more common? It's OK if your new prompt is just a slight variation of an old one; add a bit of extra inspiration or a slightly different take on the same question.

Remember, you're writing to your future self. You can say anything you want to help orient or inspire future-you. It's good to include reminders such as "don't forget to tag it 'prompt'", to make things as easy as possible on your future self.

You can also add some more specific prompts to start things off.

\$:/metap/setcurrentprompt

This tiddler is a bit of code which the prompt randomizer needs. It will not work correctly without it. Make sure the title is as above: **\$:/metap/setcurrentprompt**

Here's what you need to put into the body:

```
\define weirdaction()
<$action-setfield $tiddler="CurrentPrompt" text="{{$(currentTiddler)$}}
originaltitle="" "$(currentTiddler)$" "/>
\end

<<weirdaction>>
```

Random Prompt

Call this tiddler "Random Prompt" (or any title you wish, really) and paste the following code into its body:

```
<$button>
<$action-setfield $tiddler="$:/temp/shuffle/randomprompts" $field="state" $value=<<now
"0hh:0mm:0ss">>/>
Shuffle
<$list filter="[tag[prompt]] +[shuffle{$:/temp/shuffle/randomprompts!!state}first[1]]"
template="$:/metap/setcurrentprompt" />
</$button>

|<$link to={{CurrentPrompt!!originaltitle}} />|
|<$transclude field="text" tiddler="CurrentPrompt" mode="block" />|
```

This tiddler is going to be our workhorse, which we will return to again and again. We don't want to have to search for it. So, you should add this to the list of default prompts which open whenever you open the wiki. You do this by opening a tiddly called "Control Panel"

(which you can find via the search bar, or by clicking the gear icon in the sidebar), and under the "info" tab on that tiddler, modify the "default tiddler" field. Mine looks like this:

The screenshot shows the TiddlyWiki Control Panel with the 'Info' tab selected. The page title is '\$:/ControlPanel'. Below the title, there are tabs for 'Info', 'Appearance', 'Settings', 'Saving', 'Plugins', and 'Keyboard Shortcuts'. The 'Info' tab is active. There are two tabs at the top of the main content area: 'Basics' (selected) and 'Advanced'. The 'Basics' section contains the following configuration:

TiddlyWiki version:	5.1.21
Title of this TiddlyWiki:	Boundary Immediate
Subtitle:	a non-linear personal web notebook
Username for signing edits:	[redacted]
Animation duration:	400
Default tiddlers:	Choose which tiddlers are displayed at startup: [[Random Prompt]] Home <i>Use [[double square brackets]] for titles with spaces. Or you can choose to retain story ordering</i>
Title of new tiddlers	New Tiddler
Title of new journal tiddlers	YYYY / 0MM / 0DD / 0hh:0mm:0ss /

I've changed to a dark color scheme; yours will have a light color scheme by default. You can change this in "Appearance". There are also some other fun settings to mess with; explore; make the wiki your own!

Note that Random Prompt has [[double brackets]] around it. This is because the tiddler name contains a space, so it needs to be given with double brackets.

Now everything should be working. You can click the button on Random Prompt to draw a random prompt from your set.

Time to get building.

5: Set Up a System

MetaPrompt is a foundation, but it's not a house. You need to develop your own system of prompts, tags, and so on. This will depend on what you are using MetaPrompt to do. I'll talk

about how to set MetaPrompt up for creative writing, since that's what I've done myself.

The "prompt" tag adds a tiddly to the list of prompts, so that Random Prompt can find it and shuffle it into the rest. In a similar way, you probably want to create other tags, so that you can randomly draw things from those groups when you want to. I have tags for characters, settings, and many other objects.

MetaPrompt can facilitate both top-down and bottom-up writing styles. In top-down, first you come up with a basic story idea, then you flesh out the characters, setting, plot, and so on until you have an outline detailed enough that you can write individual scenes. In bottom-up writing, you just write whatever comes to mind and keep moving forward.

I've read that top-down writing is basically a necessity for real authors, because otherwise you get stuck and don't know what to do next.

However, I think bottom-up writing is a better place to start for creative writing in MetaPrompt. I've tried both, and bottom-up writing is just a lot easier. It sets my creative wheels spinning, and generates a lot of ideas which can then be developed in a more top-down way.

So, I have a tag "snippet". A snippet is a tiddler where I write the creative text *in the title*, and don't even worry about filling in the body. (This somehow feels psychologically important to me -- if I put the text in the body, I'd have to come up with a title. But titling a little snippet of creative writing puts too much pressure on it, to be something meaningful!)

Snippet prompts can be any sort of creative writing prompt, but you're just asking for a sentence or a few. For example:

Write a snippet about someone going somewhere.

Write a snippet in which someone encounters something.

Write a snippet involving sensory descriptions -- sounds, smells, etc.

Once you have a tag, you can draw random members, just like Random Prompt draws random prompts. For example, we can write a prompt asking to continue a snippet:

Write a snippet inspired by the following snippet:

```
|<$list filter="[tag[snippet]] +[shuffle${:/temp/shuffle/randomprompts!!state}first[1]]"/>|
```

(If snippets continue each other, you might want to put links to each other in their bodies, or organize them in some other way. I'm not currently doing this, though, because I'm treating snippets as fodder and keeping more organized stuff elsewhere.)

I also have a 'character' tag, and prompts which ask me to create or elaborate characters. Similarly for 'setting', 'plot', and a few other things.

Here's an example of a prompt which uses characters:

Write a scene in which the following two characters interact. Don't forget to tag it 'scene'.

```
|<$list filter="[tag[character]] +[shuffle${:/temp/shuffle/randomprompts!!state}first[1]]"/>|
|<$list filter="[tag[character]] +[shuffle${:/temp/shuffle/randomprompts!!state}last[1]]"/>|
```

Note that I use "last" rather than "first" in the second copy of the randomizer. This ensures that the two characters are different. (There's a more subtle problem with the randomization, above, but I leave it to the reader to discover and fix. I don't want to make this post too complicated.)

As you learn more about TiddlyWiki, you'll be able to do increasingly interesting things.

I'm sure my setup for creative writing is far from ideal (which is part of why I've only explained little pieces of it). I would be excited if, one day, there's a whole community of MetaPrompt users who talk about their systems, exchange prompts, and so on.

Limitations / Future Work

1. This falls short of the basic idea of original MetaPrompt: (A) The system doesn't have a text box directly below the prompt, where you submit your answer by hitting enter. Instead you have to create, name, and save a new prompt. This is unnecessary friction. (B) The system doesn't select a new prompt for you *as soon as you submit your answer to the previous prompt*. You have to click the "shuffle" button again. This is, again, a bit of unnecessary friction. Shaving off these bits of friction would make the system more engaging.
2. TiddlyWiki offers a lot of flexibility and power, but due to #1 above and some other reasons, it's in some ways less fun than my original command-line version. In particular, the fact that everything has to have a title field in addition to its body. I worked around this with "snippets" by *only* using the title field, but it would be nicer to only worry about text + tag, not text + tag + title. This would generally create better flow.
3. When commanding your future self, it's easy to give yourself a task that's too difficult, which will cause your future self to either click the "shuffle" button again to get something different, or stop using MetaPrompt in frustration. One of my original feature ideas for MetaPrompt was that there would be a "break task into parts" button for the user to hit in such circumstances. This would create a new task for splitting up the task into a list of subtasks (somewhat like factored cognition!), along with a special task for putting the pieces back together once they'd been individually answered. It would be interesting to try and implement something like this (perhaps taking inspiration from tools people have created for facilitating factored cognition).

Some of this could be addressed by hacking TiddlyWiki. I'd love to see people improving the TiddlyWiki version; I think a lot could be possible. On the other hand, TiddlyWiki will always create a certain barrier to entry; it requires figuring out saving, and some scripting to work out stuff like randomization. There's definitely still room to make a more streamlined version of MetaPrompt.

Another thing I'd like to see would be people trying MetaPrompt for something *other than* creative writing or drawing. Can this apply fruitfully to research? Or other nonfiction intellectual labor? I don't know.

Also note: you can create a perfectly serviceable pen-and-paper MetaPrompt, although it'll be less fully-featured than TiddlyWiki.

- Use index cards, blank playing cards, or similar.
- Rather than using tags to sort objects, you can now keep separate decks for different types of things. When you want a random character, for example, simply shuffle your 'character' deck and pull one out.

I tried this recently and it seemed fine, although I didn't use it for long because TiddlyWiki is just so much more convenient. (However, paper has a big advantage when it comes to drawing, and could even handle painting.)

Footnotes

1:

This is a hyperspace pun. The two hyperspace directions, augmenting 3d-space directions (up, down, north, south, east, and west), have been named ana and kata. I named the prompt file "kata" after WriterKata, so it was only natural to name the response file "ana".

How does bee learning compare with machine learning?

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

This is a write-up of work I did as an Open Philanthropy intern. However, the conclusions don't necessarily reflect Open Phil's institutional view.

Abstract

This post investigates the *biological anchor framework* for thinking about AI timelines, as espoused by Ajeya Cotra in her [draft report](#). The basic claim of this framework is that we should base our estimates of the compute required to run a [transformative model](#) on our estimates of the compute used by the human brain (although, of course, defining what this means is complicated). This line of argument also implies that current machine learning models, some of which use amounts of compute comparable to that of bee brains, should have similar task performance as bees.

In this post, I compare the performance and compute usage of both bees and machine learning models at [few-shot image classification](#) tasks. I conclude that the evidence broadly supports the biological anchor framework, and I update slightly towards the hypothesis that the compute usage of a transformative model is *lower* than that of the human brain.

The full post is viewable in a Google Drive folder [here](#).

Introduction

Ajeya Cotra wrote a [draft report on AI timelines](#) (Cotra, 2020) in which she estimates when [transformative artificial intelligence](#) might be developed. To do so, she compares the size of a transformative model (defined as the number of [FLOP/s](#) required to run it) with the computational power of the human brain, as estimated in [this Open Phil report](#) (Carlsmith, 2020)[1]. She argues that a transformative model would use roughly similar amounts of compute as the human brain. As evidence for this, she claims that computer vision models are about as capable as bees in visual tasks, while using a similar amount of compute.[2]

In this post, I (Eleni Shor) investigate this claim. To do so, I focus on the performance of bees at [few-shot image classification](#), one of the most difficult tasks that bees are able to perform. I find that both their task performance and their compute usage are roughly comparable to current machine learning models. This is broadly consistent with Cotra's claim that transformative models would have compute requirements similar to that of the human brain, as neither bees nor machine learning models are clearly superior at this task.

On the one hand, bees are more sample-efficient than machine learning models, requiring orders of magnitude fewer examples of few-shot learning tasks in order to perform novel tasks and having impressive generalization capabilities, being able to transfer training in a task to perform well in related tasks. This, however, might be explained if bees have more task-relevant background knowledge. On the other hand, machine learning models seem to be more compute-efficient. However, there are many difficulties in usefully comparing the compute usage of biological brains and machine learning models, and, therefore, I am not very confident in this conclusion.

Notwithstanding these caveats, I perform a naïve estimate of the compute required to train a transformative model based on the relative compute-efficiency of bees and plug this

estimate into the AI timelines model developed in (Cotra, 2020). I update slightly towards shorter timelines than Cotra's median estimate of 2050.

This post is structured in the following manner. First, I describe few-shot image classification tasks and argue that they are relevant for estimating the compute usage of a transformative model. I then review the existing literature on the performance of bees and contemporary machine learning models on these tasks and observe that their accuracies are broadly comparable. Subsequently, I estimate the compute usages of both kinds of classifiers. I conclude by using the difference in the computing requirements to inform AI timeline estimates, noting relevant caveats along the way.

Few-shot image classification

In this section, I explain what few-shot image classification is and why this task is relevant to estimating the compute usage of a transformative model.

What is this task?

As the name suggests, few-shot image classification tasks are, well, image classification tasks: given a dataset with many labeled images, our goal is to predict the labels of novel images. What sets few-shot image classification apart is the small number of examples for each label.

Let's consider a concrete example of a few-shot learning task. Consider the training dataset of animal pictures below, and assume that each image is labeled as a cat, lamb, or pig, as appropriate.



[Source](#).

A classifier receives this dataset as input and, from it, learns a mapping from images to labels. This mapping can then be used to classify novel pictures like the image below.



[Source](#).

Hopefully, the classifier correctly labels the image as a cat.

As mentioned previously, the number of training images for each label is small. In the example above, this number is only two. Additionally, the number of labels is large in comparison to the number of training images. In the example above, there are three labels ("cat", "lamb", and "pig") and six training images. For contrast, in ImageNet, a popular dataset used as an image classification benchmark, there are thousands of training images per label.[3]

A classification problem with n labels and k examples per label is called an *n -way k -shot learning* problem. For instance, the aforementioned task is a three-way two-shot learning problem, since there are three labels and two example images for each label.

Why is this task relevant for forecasting AI timelines?

The goal of this post is to inform estimates about the size of a transformative model and, with such estimates, to forecast when such a model might be developed. For these purposes, a good task for comparing biological and artificial intelligences is one that tests capabilities that are plausibly relevant to having a transformative impact. To illustrate this, consider a task that doesn't satisfy this criterion: arithmetic. Humans use a few quadrillion times more compute than your desktop to add numbers,[4] yet this fact doesn't help anyone figure how the compute needed to run a transformative model compares to that used by the human brain.

One task that seems relevant to having a transformative impact is [meta-learning](#), the task of efficiently learning how to do other tasks. A model with powerful meta-learning capabilities could simply learn how to do transformative tasks and have a transformative impact that way.

Few-shot learning can be understood as a meta-learning task. Each instance of a few-shot learning task is just a normal [supervised](#) learning task. However, due to the small amount of training data in each instance, learning how to classify the test image using only the examples seen in one instance is difficult. A few-shot classifier must either learn how to learn new tasks from scratch efficiently or how to apply knowledge obtained from related tasks in novel circumstances. In this post, I use performance at few-shot image classification as a proxy for meta-learning capabilities.

However, if a classifier has memorized a large number of labeled images and then is tested at a few-shot image classification task with images similar to the memorized ones, its performance at that task would not be very indicative of its meta-learning capabilities. In general, a classifier's level of prior knowledge about a classification task strongly influences its performance. Controlling for the classifier's prior knowledge is reasonably easy for machine learning models, since the creators of the model have full control over the training data. However, this is not the case for animals: it is hard to know to what degree evolution has baked in relevant prior knowledge into the design of animal brains.

One might attempt to avoid this problem by considering only tasks that animals have likely not been selected to do. Doing this, however, leads us to a different problem: there's no reason to expect that animals use their brains efficiently to solve tasks that they haven't been selected to do. Consider the aforementioned example of humans and arithmetic.

My subjective impression is that the few-shot image classification tasks that I discuss are in a happy medium between the two extremes, being both artificial enough that it seems unlikely that bees have much prior knowledge of the specific image categories used and natural enough that bees shouldn't be extremely compute-inefficient at doing them; this, however, is debatable.

As the reader can discern from the above, the theoretical case for using few-shot image classification as a benchmark task is not very strong. The main reason I chose this task is practical: few-shot image classification is studied in both machine learning and bee vision, so comparing the performance of bees and AIs at this task is relatively straightforward.

In the remainder of this post, I mostly ignore these concerns, focusing on comparing the performance and the compute usage of different classifiers at few-shot image classification tasks, regardless of their origin and workings.

Task performance

In this section, I describe the performance of bees and current machine learning models in few-shot image classification tasks that are roughly similar. I first broadly discuss the capabilities of bees and search the bee vision literature for studies where such tasks are investigated. I choose (Zhang et al., 2004) as a reference for the performance of bees at such tasks. Based on the specifics of that article, I choose (Lee et al., 2019) as the most comparable article in the machine learning literature. I conclude by briefly comparing the performance of both classifiers.

Bee performance

Bees can learn how to perform binary image classification through [operant conditioning](#). Concretely, this is usually done by associating the images of a given category with a reward like sugar water, while leaving those belonging to the other category either unrewarded or paired with an aversive stimulus.[5] Variations on this basic experimental setup allow scientists to train bees to perform more complicated [supervised learning](#) tasks.[6]

Literature review

There is a vast literature on this topic; I have not thoroughly investigated it, but, instead, I mainly rely on two literature reviews, (Avarguès-Weber et al., 2012) and (Giurfa, 2013), and I follow citation chains from them. I use these articles because I found them from a citation in (Rigosi et al., 2017), the bee vision paper cited in (Cotra, 2020), rather than from a systematic search.

My general impression from the bee vision literature is that bees are capable of learning surprisingly complicated tasks, ranging from recognizing specific faces (Dyer et al., 2005) to distinguishing between Monet and Picasso paintings (Wu et al., 2013). Bees learn how to do these tasks with a small amount of training data, usually less than ten examples, allowing for comparisons with few-shot learning in computer vision. However, many studies in this literature do not test bees' learning with novel test data or analyze tasks which are easier than standard machine learning benchmarks, such as detecting symmetry in glyphs (Giurfa, 1996).

Out of the articles I've looked at, I believe that (Zhang et al., 2004) is the one most relevant to comparing bees and machine learning models. In it, the authors train bees to recognize complex natural categories, like landscapes and plant stems, in a one-shot setting. This is one of the hardest tasks in the bee vision literature,[7] and it can be compared with few-shot image classification on datasets commonly used in machine learning such as ImageNet and CIFAR. I use this article as a benchmark for bee performance in image classification tasks.

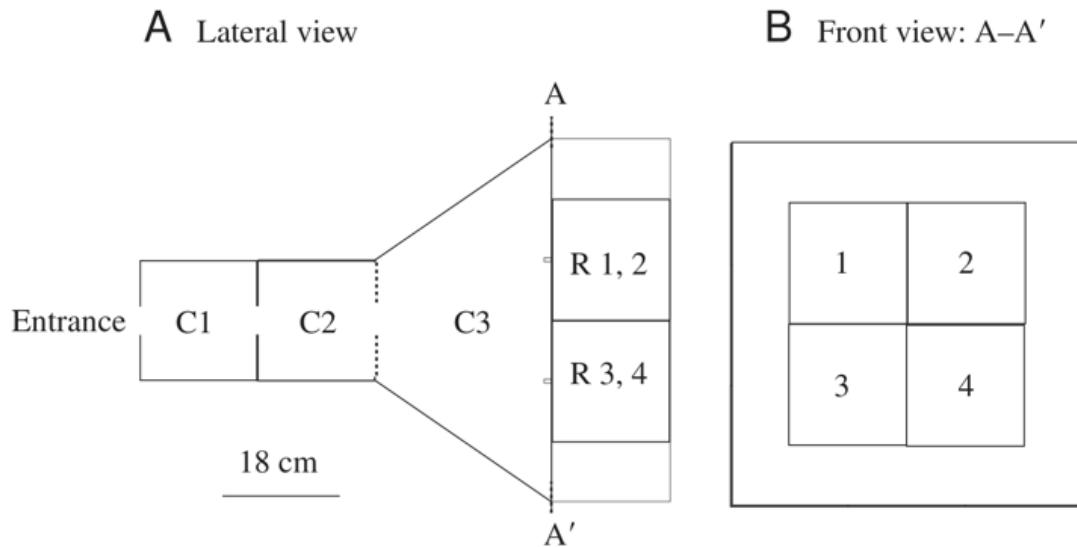
I do not explain in detail how other articles compare to (Zhang et al., 2004) in the body of this post. The interested reader can instead see [this appendix](#) for details.

In the remainder of this section, I describe the results of the article I chose as a benchmark, (Zhang et al., 2004), and what they imply about the capabilities of bees.

Benchmark article

In (Zhang et al., 2004), bees were trained to recognize complex natural image categories (star-shaped flowers, circular flowers, plant stems, and landscapes) in a one-shot setting. Bees achieved an accuracy of about 60% at this task, which is considerably higher than the 25% accuracy obtained by guessing randomly.

The experimental setup used consists of a three-chambered maze, as depicted in the image below, taken from the paper. Bees enter the maze and reach chamber C1, where they see a sample image in the wall between chambers C1 and C2. They then fly through a small hole in the aforementioned wall and reach chamber C2, where they observe chamber C3 through the transparent wall between these two chambers, depicted in the diagram below as a dotted line. The back wall of C3 consists of four images, each with an associated tube. The tube associated with the image matching the sample leads to a feeder containing a sugary solution, while the others provide no reward to the bees. The bees fly through another small hole, this time between C2 and C3. The tube on which they land first is interpreted as their “prediction” of the reward’s location. This experimental setup is used for both training and testing bees’ image classification skills.



The transparency of the wall between C2 and C3 suggests that the bees make their choice at C2. As the chamber C3 is large, the images in its back wall would appear small to the bees. I believe that this had a significant negative impact on their accuracy, as I will discuss later.

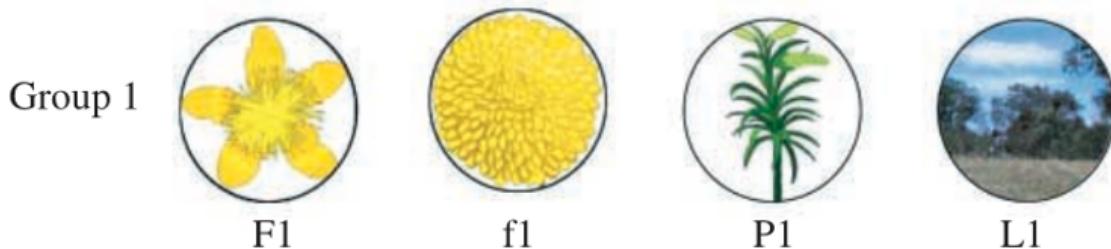
The experiment consisted of three parts: a pre-training period where bees learned to enter and traverse the maze, a training period where bees learned to land on an image that exactly matched the sample, and a testing period where bees had to land on an image that matched the *category* of the sample. Note that the bees weren’t explicitly trained to do image classification; instead, they were trained to simply match images. Nevertheless, they were able to generalize and match images by category during testing. This makes their success at image classification significantly more impressive. In general, the learning capabilities of bees seem to me more impressive than that of machine learning models, as I will discuss later.

Throughout training and testing, the authors took precautions to ensure that bees were actually doing the matching task. The position of the rewarded image was frequently switched, so that bees couldn’t memorize it. Additionally, a control experiment in which all images matched the example was performed, which showed that bees couldn’t use olfactory cues to locate the reward.

Both training and pre-training took a substantial amount of time, with testing beginning after three days of training, in which each bee entered the maze and was rewarded about 80 times. This large amount of training epochs is common in the bee vision literature; however, I’m not sure why. Although this article doesn’t plot learning curves, other articles, such as (Giurfa et al., 1996) and (Dyer et al., 2005), do so, showing that performance increases with more exposure. I’d wager that repeated exposure to the same images is required for bees to

understand the task rather than to learn the categories themselves, but I'm not confident about this.

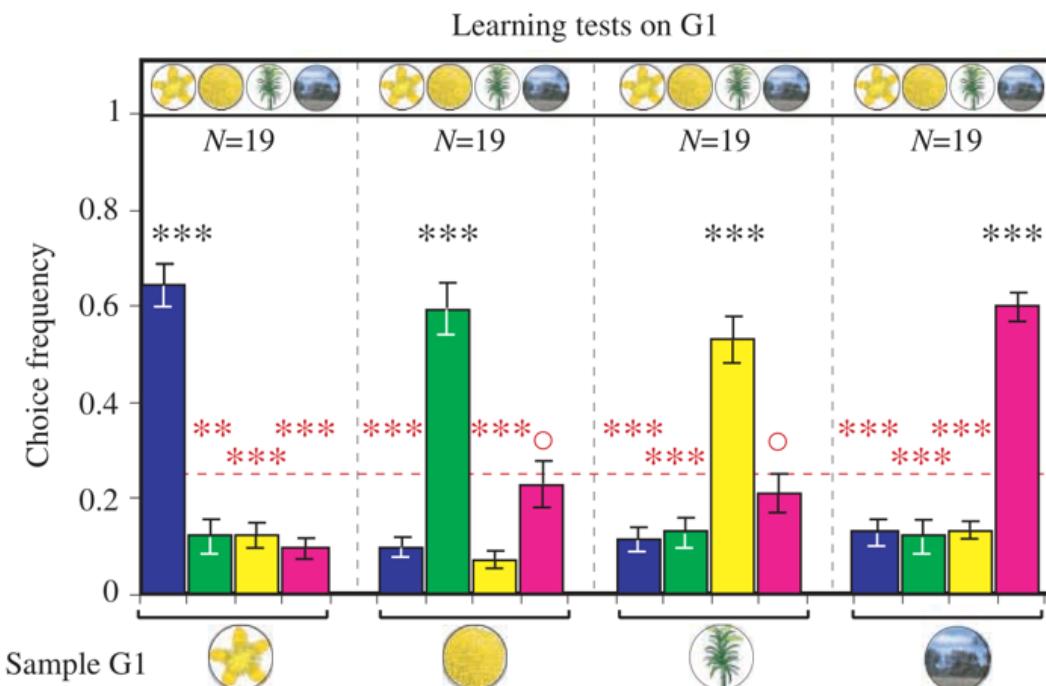
Training consisted of showing bees sample images and rewarding them by landing on the matching image, as mentioned previously. The training set (called "Group 1" in the paper), with one image of every category (star-shaped flowers, circular flowers, plant stems, and landscapes), is shown below.



Note that the star-shaped flower is very similar to the circular flower.

Although the categories are natural and, therefore, bees might have relevant prior knowledge about these categories, my impression is that the images are seen from a very different perspective than bees would usually see them, and so this knowledge would be less relevant. Since bees don't classify images with which they should be more familiar (e.g. flowers) more accurately and the performance of bees in the training set is not that impressive, as will be seen later, I do not think that this objection is that relevant.

After training, the authors measured how frequently bees landed in each image when shown a sample. The results of this learning test are shown in the graph below, taken from the paper.



(Note that the N in the plot above is the number of bees that participated in the learning test, not the number of visits to the maze; the latter is much higher, as 1132 visits took place.)

As the reader can see, bees matched the sample correctly about 60% of the time. Although this is significantly better than chance, this performance is not very impressive, given that this task is so simple. I believe that this poor performance can be explained by the small apparent size of the images.

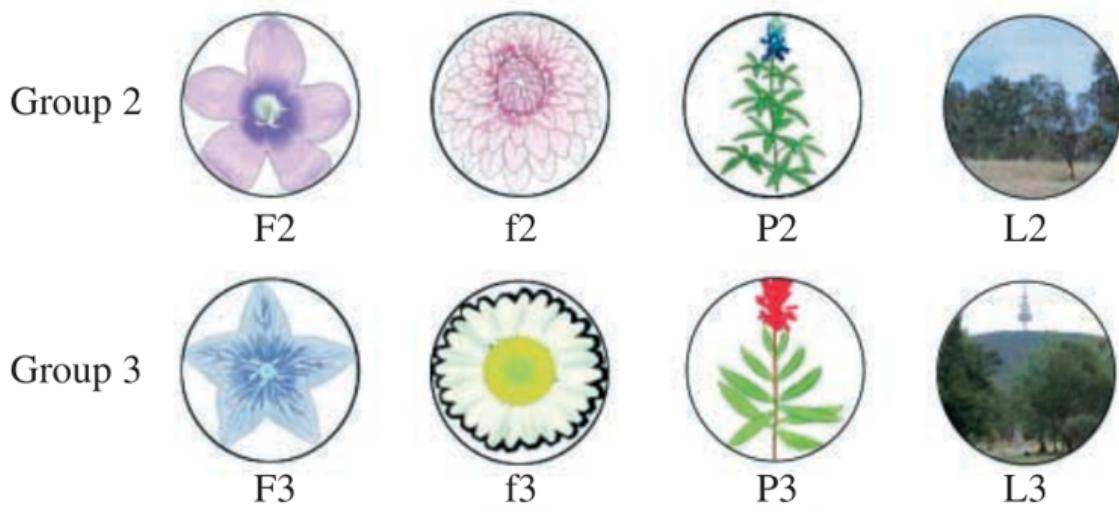
I investigate the resolution of bee vision in some depth in [this appendix](#), concluding that each eye has a resolution of 100x100 pixels. Since the images only occupy a fraction of the bees' field of view, the effective resolution of each image is even lower than that. From the diagram of the experimental setup used in (Zhang et al., 2004), I estimate that the images appeared to the bees as having an effective resolution of 20x20 pixels. The details of this calculation can be found in [this appendix](#).

To understand intuitively how low this effective resolution is, compare the high-resolution image of a landscape in the left to the same image downscaled to 20x20 pixels on the right.



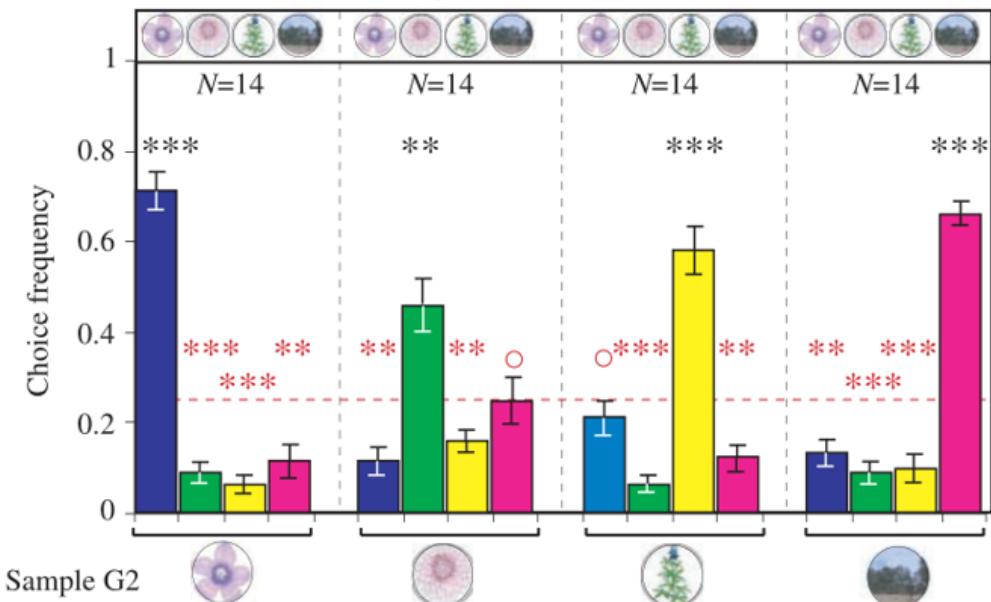
As the reader can see, the difference in image quality is dramatic. The performance of bees in the image classification task is very similar to their performance in this simple matching task, as will be seen later. I think that these considerations point to low resolution being the cause of poor performance in the learning test, and that this poor performance is not very indicative of weak learning abilities.

After this learning test, the authors performed two types of transfer tests. Type 1 transfer tests consisted of the same matching-to-sample task, but with novel images, as depicted below.

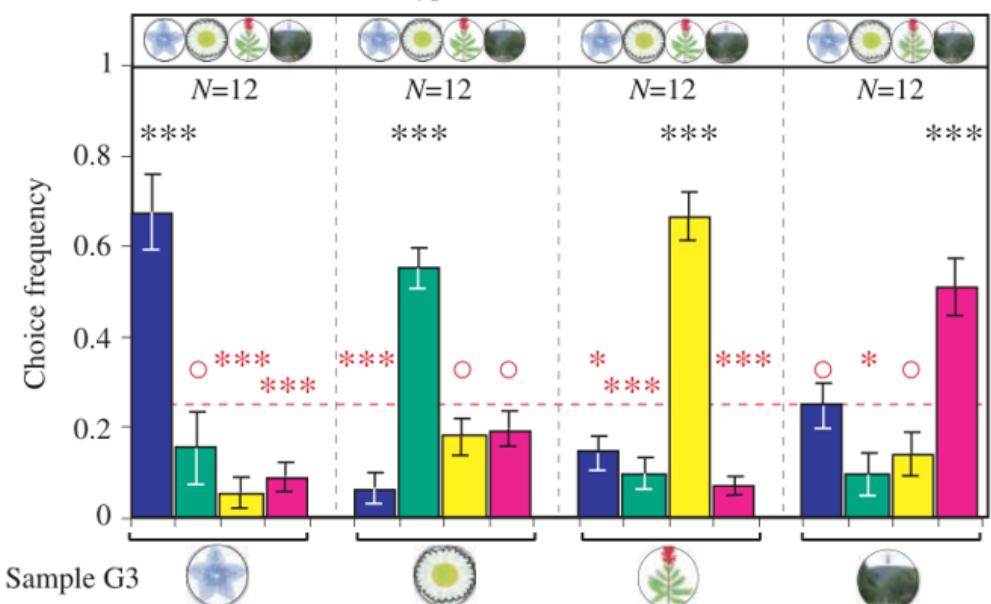


The study finds that bees do as well with the novel images as with the training set, as can be seen in the plots below:

Type 1 transfer tests on G2



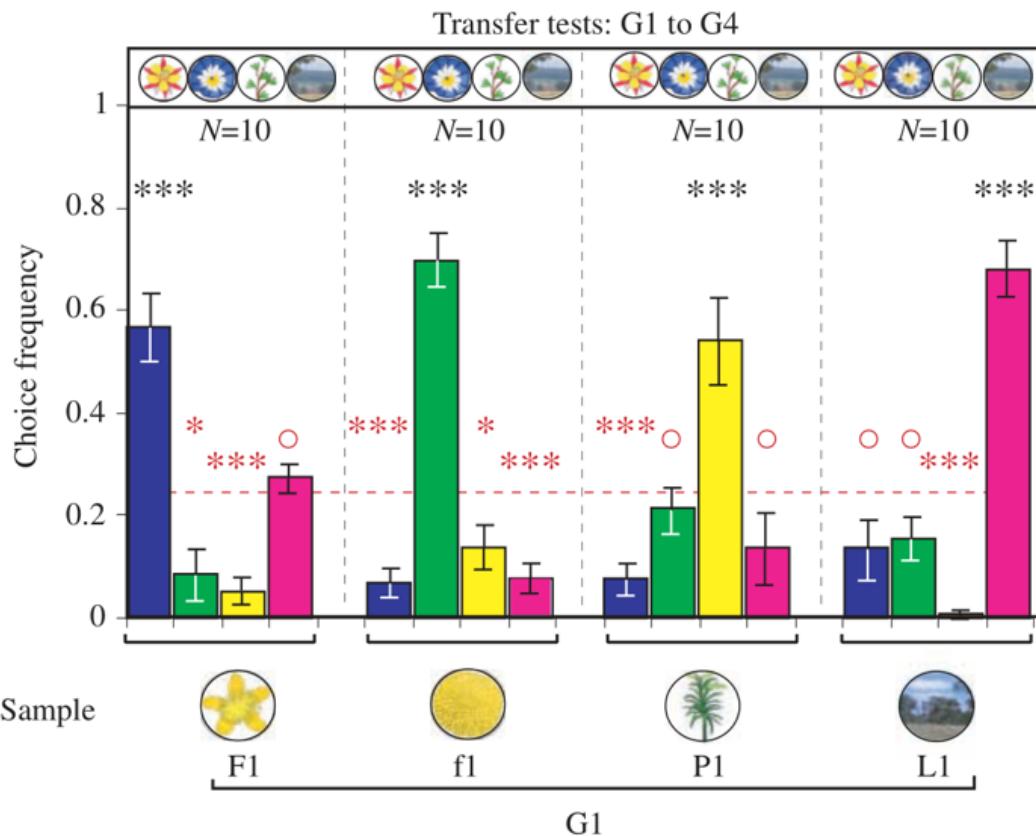
Type 1 transfer tests on G3



Type 2 transfer tests require that bees match the category of the sample, rather than the exact image. The sample image, which indicates the category to be matched, comes from the training set, while the images to be classified come from a novel group of images, depicted below.



The results of the Type 2 transfer test is shown in the plot below:



As can be seen, the bees classified the images correctly about 60% of the time, which is very similar to the performance of bees in the learning tests. I am impressed by these results, especially since the bees were not trained specifically to do image classification, but instead they were only trained to find exact matches.

Note that the bees were rewarded for landing on the correct image during testing. This is common in such experiments; otherwise, bees might give up on the classification task and stop entering the maze, seeing that their efforts go unrewarded. However, this means that some learning might occur during testing, since each test consists of many visits of each bee to the maze. In the Type 2 transfer test above, for instance, each bee visited the maze around 28 times. In order to minimize learning during testing, the authors broke up testing into short bouts in which each bee visited the maze only twice. These testing bouts were spaced out and interleaved with periods of training on the original matching-to-sample task.

Machine learning model performance

In this section, I review the machine learning literature searching for a model that can be compared to bees in the experiments of (Zhang et al., 2004). I find that the model presented in (Lee et al., 2019) most comparable. I then analyze its performance.

Literature review

How do machine learning models compare to the performance of bees in few-shot learning shown in (Zhang et al., 2004)? There are a huge variety of models trained on few-shot learning tasks; the first step in this comparison, therefore, is choosing models that can be reasonably compared to bees. My approach is to choose models that are trained with similar data and that have similar performance to that of bees; the comparison then reduces to seeing whether bees or ML models use more compute.

In order to find such comparable models, I looked at [a list of few-shot learning benchmarks on Papers With Code](#). Each of these benchmarks specifies a few-shot learning on some dataset. I chose a dataset based on the resolution of the images, as this is very relevant to a model's compute usage. As mentioned previously, the effective resolution of the images used in (Zhang et al., 2004) is 20x20 pixels; a well-known dataset with a similar resolution is CIFAR-FS, introduced in (Bertinetto et al., 2018), which contains 32x32 images from a hundred different categories ranging from airplanes to cats.

I believe that the images in (Zhang et al., 2004) are a bit harder to classify than those of the CIFAR-FS dataset, since the two flower categories are very similar. Additionally, I have a vague impression that the images in the CIFAR-FS were selected to be distinguishable at low resolution, while I don't think that was the case in (Zhang et al., 2004), but I'm very uncertain about this.

The most comparable machine learning task, then, is few-shot learning on the CIFAR-FS dataset. However, ML models are usually trained to do five-way classification, that is, to classify an image from one of five possible labels, while bees perform four-way classification in (Zhang et al., 2004). All else equal, four-way classification tasks are easier than five-way ones, since there are fewer wrong options.

A way of comparing accuracies in different n -shot tasks would simplify comparing bees and ML models. I'm not aware of a standard way of doing this. In order to do so, therefore, I develop a toy model of a few-shot classifier. I model an n -way classifier as an ensemble of n binary classifiers, one for each of the possible labels. I assume that these binary classifiers have some fixed error rate, which is a measure of the performance of the ensemble independent of n . By relating the accuracy of the ensemble to the error rate of the binary classifiers, one can estimate the accuracy that bees would have in a five-way classification task. The mathematics of this model can be found in [this appendix](#). I estimate that the 60% accuracy in a four-way classification task obtained by bees in (Zhang et al., 2004) corresponds to an accuracy of about 55% in a five-way image classification task.

The results of ML models on five-way one-shot image classification in CIFAR-FS are listed [here](#); accuracies range from about 70% to about 90%, significantly higher than our 55% estimate for bee performance. However, my impression is that the images from different categories in (Zhang et al., 2004) are a bit more similar to each other than those in CIFAR-FS, so the comparison isn't as clear a win for machine learning models as one might think at first. Litigating this comparison precisely seems difficult and not very fruitful; therefore, I chose the model with accuracy most similar to bees as my benchmark, which turns out to be the model with the lowest accuracy.[8] This is the model found in (Lee et al., 2019), which has an accuracy of 72.8%.

Compute usage

In this section, I estimate how much computation the machine learning model of (Lee et al., 2019) uses to classify an image, as well as the compute usage of bees in the experiments of (Zhang et al., 2004). I then compare these two figures.

Machine learning model compute usage

The architecture of the model presented in (Lee et al., 2019) consists of two parts: a function that maps images to feature vectors and a linear classifier that uses the feature vectors to classify the images. During training, the function is optimized to learn features that generalize well. The first part is implemented as a [convolutional neural network](#), while a linear [support-vector machine](#) serves as the linear classifier.

After the model has processed the sample images for a given task, it needs only to calculate the features of the task image and feed them to the linear classifier. The most compute-intensive part of this is calculating the features; this was done using a convolutional neural network with a ResNet-12 architecture. As the authors do not provide an estimate of the compute per forward pass used by their model, I estimate that value by extrapolating the compute estimates for a similar architecture and scaling them to account for differences in parameter count and input size. One could also estimate the compute usage directly from the description of the network given in the paper; this would probably lead to a more accurate estimate, but I don't know how to do this.

I chose the ResNet-18 architecture as a point of comparison. A forward pass through a ResNet-18 model with an input size of 112x112 pixels requires about 5e8 FLOP;[9] since the compute per forward pass should scale linearly with both the network depth and the input size in pixels, the model used in the paper, a ResNet-12 with an input size of 32x32, should take about $5 \times 12 / 18 = 2.77$ FLOP * (32/12)^2 = **2e7 FLOP per image classified**.

Bee compute usage

Estimating the compute usage of bee brains is a somewhat speculative endeavor. It's not immediately clear what "the number of FLOP/s of a bee brain" means, since brains certainly don't seem to be doing floating-point arithmetic. However, we can still ask what would be the amount of computational power required to simulate a brain in sufficient detail in order to replicate its task performance. I'll use this "mechanistic" definition of brain compute throughout the report. The question of how to estimate this "mechanistic" brain compute is investigated in (Carlsmith, 2020) in the context of the human brain; by applying a similar methodology to the bee brain, we can estimate its computational abilities.

In order to estimate the amount of compute required to do so, we follow the basic approach described in (Carlsmith, 2020). In brief, it seems likely that the most computationally-intensive part of emulating a brain is simulating the signaling interactions between [neurons](#).

The amount of compute required to emulate these interactions can be broken down in two parts: that devoted to simulating the [synapses](#) (i.e., the connections between neurons) and that devoted to determining when a neuron should fire. The total bee brain compute can then be broken up as the sum of these components. The synaptic transmission compute can be calculated as (number of synapses) * (rate of synaptic firing) * (compute per synaptic firing), whereas the firing decision compute is equal to (number of neurons) * (firing decision compute per neuron per second).

I don't have a background in neuroscience, but it seems to me that the basic anatomy of the neuron is broadly similar between bees and humans. Therefore, I will assume that the

parameters of bee neurons—that is, the rate of synaptic firing, the compute per synaptic firing, and the firing decision compute per neuron—are equal to those of human neurons. I'm very uncertain whether this assumption is reasonable or not.[1] These parameters are estimated in (Carlsmith, 2020), which obtains a range of 0.1-1 Hz for the rate of synaptic firing, 1-100 FLOP per synaptic firing, and 1e2-1e6 FLOP/s for the firing decision compute per neuron per second. All that remains is to plug in estimates for the number of neurons and synapses in the brains of bees and we'll be able to estimate the total computing power of the bee brain. (Menzel & Giurfa, 2001)[11] state that bees have a million neurons and a billion synapses; this works out to a central estimate of synaptic communication compute of 3e9 FLOP/s[12] and a central estimate of firing decision compute of 1e10 FLOP/s.[13]

Of course, not all of the computational power of the bee's brain is applied to few-shot image classification tasks. My understanding is that the sections of the bee's brain mostly responsible for visual processing as well as learning are the mushroom bodies, which occupy about a third of the bee's brain.[14] It seems likely, therefore, that a better central estimate of the total amount of task-relevant computation per second in the bee's brain is 4e9 FLOP/s. [15]

In order to turn the bee brain compute value into an estimate of the compute usage per image classified, we need to know how long bees take to perform this task. Unfortunately, (Zhang et al., 2004) don't specify this explicitly. However, we can estimate the time that bees spend in the experimental setup from some details given in the paper. The authors mention that “[e]ach transfer test was carried out only for a brief period (ten minutes, involving about two visits per bee)”; since the transfer tests involved around ten bees as experimental subjects, I believe that this implies that the time taken per visit is about $(10 \text{ min}) / [(2 \text{ visits/bee}) * (10 \text{ bees})] = 30 \text{ s/visit}$. I assume that there's some time between each bee's visit, so the total time each bee spends in the environmental setup should be somewhat smaller than that. The time each bee spends to classify an image must be even smaller than that. Somewhat arbitrarily, I estimate that this time is around five seconds.

The compute employed by bees to classify an image, then, can be found by multiplying the computing power of the bee brain used in this task by the time taken to do so. This works out to $(4e9 \text{ FLOP/s}) * (5 \text{ s}) = \mathbf{2e10 \text{ FLOP per image classification task}}$.

It's clear, however, that a bee's brain can perform a wide range of tasks beside few-shot image classification, while the machine learning model developed in (Lee et al., 2019) cannot. I do not take into account this in my comparison below. In fact, I am not sure how to do this. This is probably the factor that most biases this comparison against bees.[16]

However, there is some reason to believe that the above estimate is not wildly unfair against bees: in (Carlsmith, 2020)'s discussion of functional estimates of brain compute, a comparison of the human visual cortex with image classification models trained in a regular (not a few-shot) setting is made.[17] He considers that 0.3%-10% of the visual cortex is used by the human brain to perform image classification and adjusts for the fact that image classifiers are worse than humans at some aspects of image classification. These guesses for visual cortex usage can be taken as guesses for the usage of mushroom bodies. Since bees appear to be worse than machine learning models at this task, however, even Carsmith's lowest compute estimates suggest that bees are one order of magnitude less efficient than machine learning models, which gives me some confidence that this comparison is not totally uninformative.

Conclusion

In this report, I show that both bees and computer vision models are able to perform very similar few-shot image classification tasks. The efficiency with which they perform these tasks, however, differs somewhat: my central estimate is that bees use three orders of magnitude more computation to do them. Although this might seem like a very large difference, it is comparable to the uncertainties in the biological compute estimates: (Carlsmith, 2020) posits an uncertainty of two orders of magnitude in his brain compute estimate.

In addition, the comparison performed in this post is stacked against bees in various ways. Bees have not evolved to perform few-shot image classification tasks, and, in fact, have not even been trained specifically on this task in (Zhang et al., 2004), while the machine learning model analyzed here was optimized to do just that. The behavioral repertoire of bees is vast, and I have only analyzed one task in this report. Perhaps most importantly, bees seem to be significantly more sample-efficient than machine learning models; that is, they are able to learn how to perform a task with a smaller number of examples.

My intuition is that within-lifetime learning in bees (as opposed to the “learning” that occurs during evolution) is more analogous to the runtime learning capabilities of some ML models, like GPT-3, than to training-time learning, which is another factor that might lead these estimates to underestimate the learning prowess of bees. Comparing these capabilities to that of [CLIP](#) and [Image GPT](#) would be very interesting, but outside the scope of this post.

Even with all of these caveats, however, I believe that this analysis is informative for updating the estimates of the compute required for creating a transformative model.

Very naïvely, we can adjust these estimates in the following fashion: since bees are three orders of magnitude worse than computer vision models, our prior should be that a transformative model should require roughly three orders of magnitude less compute than the human brain. I don’t think it’s obvious in what direction we should adjust this prior, so I’ll stick to it. As the human brain can perform the equivalent of $1\text{e}13\text{-}1\text{e}17$ FLOP/s, we should then expect that a transformative model should require $1\text{e}10\text{-}1\text{e}14$ FLOP/s to run. This is somewhat smaller than the central estimate of $1\text{e}16$ FLOP/s found in (Cotra, 2020).

As mentioned previously, however, this discrepancy is relatively small when compared to the uncertainties involved; therefore, I believe that this investigation validates the biological anchor approach to investigating AI timelines somewhat. After all, if the discrepancy in performance per FLOP was much larger, extrapolations based on comparisons between biological anchors and machine learning models would be much more suspect, but that did not turn out to be the case.

I was somewhat surprised by this result; before doing this investigation, I believed that bees would do better than machine learning models in terms of compute. This has updated me towards shorter timelines. In order to get an idea of how much this should shift my timelines, I plugged in the naïve estimate of transformative model size into Cotra’s TAI timelines model, using her best guesses for all other parameters. The interested reader can browse the resulting distribution and adjust the model parameters [here](#); the main difference, however, is that the median forecast for TAI goes from around 2050 to around 2035, making timelines substantially shorter. I would advise the reader to make a much smaller update to their timelines, since I am very uncertain about the conclusions of this report, as I have hopefully made clear. I would tentatively adjust Cotra’s forecast to 2045, holding constant all other parameters of her model. Notwithstanding that, very short timelines seem more plausible to me after concluding this investigation.

Bibliography

- Avarguès-Weber, A., Mota, T., & Giurfa, M. (2012). New vistas on honey bee vision. *Apidologie*, 43(3), 244-268. <https://hal.archives-ouvertes.fr/hal-01003646>
- Bertinetto, L., Henriques, J., Torr, P. H.S., & Vedaldi, A. (2018). Meta-learning with differentiable closed-form solvers. *International Conference on Learning Representations*. <https://arxiv.org/abs/1805.08136>
- Carlsmith, J. (2020). *How Much Computational Power Does It Take to Match the Human Brain?* Open Philanthropy Project. <https://www.openphilanthropy.org/brain-computation-report>
- Cotra, A. (2020). *Forecasting TAI with biological anchors*. LessWrong. <https://www.lesswrong.com/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines>
- Dyer, A. G., Neumeyer, C., & Chittka, L. (2005). Honeybee (*Apis mellifera*) vision can discriminate between and recognise images of human faces. *Journal of Experimental Biology*, 208, 4709-4714. <https://jeb.biologists.org/content/208/24/4709.short>
- Giurfa, M. (2013). Cognition with few neurons: higher-order learning in insects. *Trends in neurosciences*, 36(5), 285-294. <https://pubmed.ncbi.nlm.nih.gov/23375772/>
- Giurfa, M., Eichmann, B., & Menzel, R. (1996). Symmetry perception in an insect. *Nature*, 382, 458-461. <https://www.nature.com/articles/382458a0>
- Lee, K., Manji, S., Ravichadran, A., & Soatto, S. (2019). Meta-Learning with Differentiable Convex Optimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10657-10665. https://openaccess.thecvf.com/content_CVPR_2019/html/Lee_Meta-Learning_With_Differentiable_Convex_Optimization_CVPR_2019_paper.html
- Menzel, R., & Giurfa, M. (2001). Cognitive architecture of a mini-brain: the honeybee. *Trends in Cognitive Sciences*, 5(2), 62-71. [https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613\(00\)01601-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1364661300016016%3Fshowall%3Dtrue](https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(00)01601-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1364661300016016%3Fshowall%3Dtrue)
- Rigosi, E., Wiederman, S. D., & O'Carroll, D. C. (2017). Visual acuity of the honey bee retina and the limits for feature detection. *Scientific Reports*, (7). nature.com/articles/srep45972
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Li, F.-F. (2014). ImageNet Large Scale Visual Recognition Challenge. *arXiv preprint*. <https://arxiv.org/abs/1409.0575>
- Wu, W., Moreno, A. M., Tangen, J. M., & Reinhard, J. (2013). Honeybees can discriminate between Monet and Picasso paintings. *Journal of Comparative Physiology A*, 199(1), 45-55. <https://link.springer.com/article/10.1007/s00359-012-0767-5%20>
- Zhang, S., Srinivasan, M. V., Zhu, H., & Wong, J. (2004). Grouping of visual objects by honeybees. *Journal of Experimental Biology*, 207, 3289-3298. <https://jeb.biologists.org/content/207/19/3289.short>

Appendices

Why (Zhang et al., 2004)?

In my non-comprehensive review of the bee vision literature, I looked at a couple of articles that train bees to do image classification tasks. The literature review in (Avarguès-Weber et al., 2012) cites (Giurfa et al., 1996) and (Zhang et al., 2004), two papers that do so. Searching the papers that cite (Zhang et al., 2004), I found another relevant article, (Wu et al., 2013). I also came across (Dyer et al., 2005), but, unfortunately, I do not recall how. I also came across some papers that did simpler image classification tasks, such as (Srinivasan & Lehrer, 1988), but, since the tasks bees perform in these papers are much easier than those in the aforementioned ones, I did not consider them as possible benchmarks for comparing bee vision and machine learning.

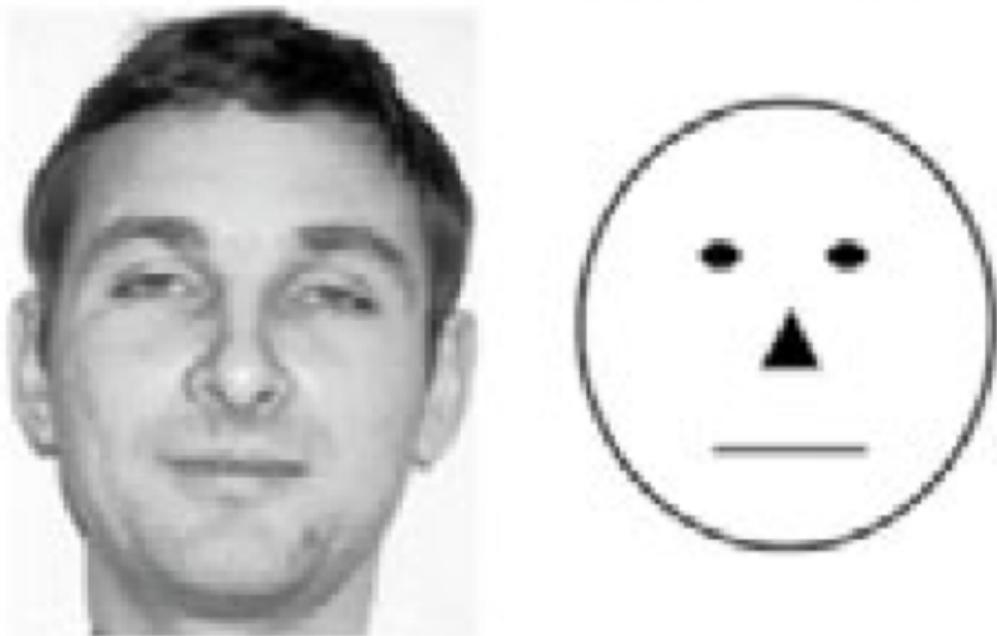
(Giurfa et al., 1996) trains bees to classify abstract glyphs on the basis of their bilateral symmetry or absence thereof. The bees perform very well at this task, with accuracies of about 85%. However, this task is quite simple when compared to few-shot classification in datasets used in ML; therefore, I rejected this task as a possible benchmark.

(Dyer et al., 2005) trains bees to recognize specific human faces. Unfortunately, the paper does not do this in a few-shot learning setting. Although the success of bees in this task is very impressive to me, this success is not as relevant to estimating the size of a transformative model; therefore, I rejected this task as a possible benchmark. The reader interested in the results of this paper should look at [this appendix](#).

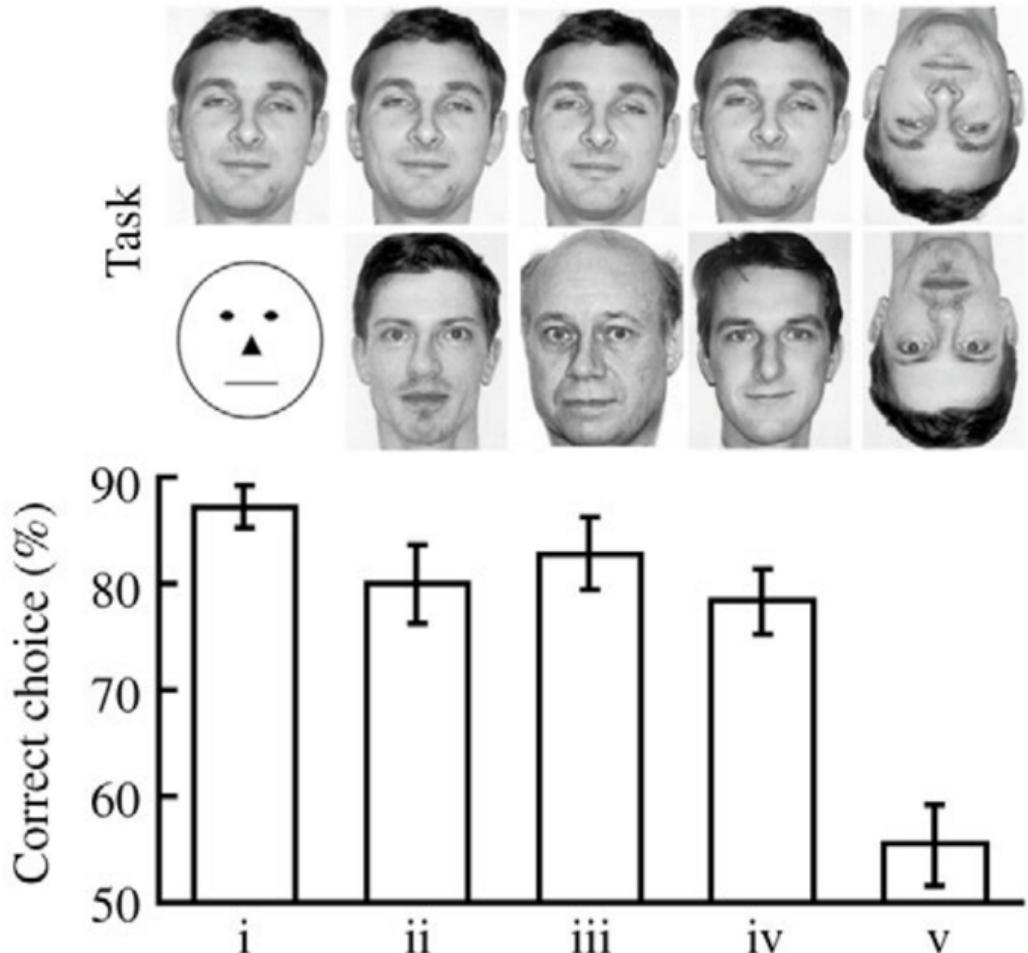
(Wu et al., 2013) trains bees to distinguish between Monet and Picasso paintings in a few-shot setting. This article was a very strong candidate for benchmark paper; I chose to use (Zhang et al., 2004) because the dataset used in this latter paper is more similar to those usually used in ML.

Results of (Dyer et al., 2005)

In this article, the authors trained bees to distinguish between different human faces. This task seems quite impressive to me, as I would not have predicted that bees would be capable of such a feat before learning of this paper. The training procedure was also remarkably simple: bees were trained to distinguish between a photograph of a specific person (the rewarded stimulus) and a stylized representation of a face (the punished stimulus), as can be seen in the images below from the article:

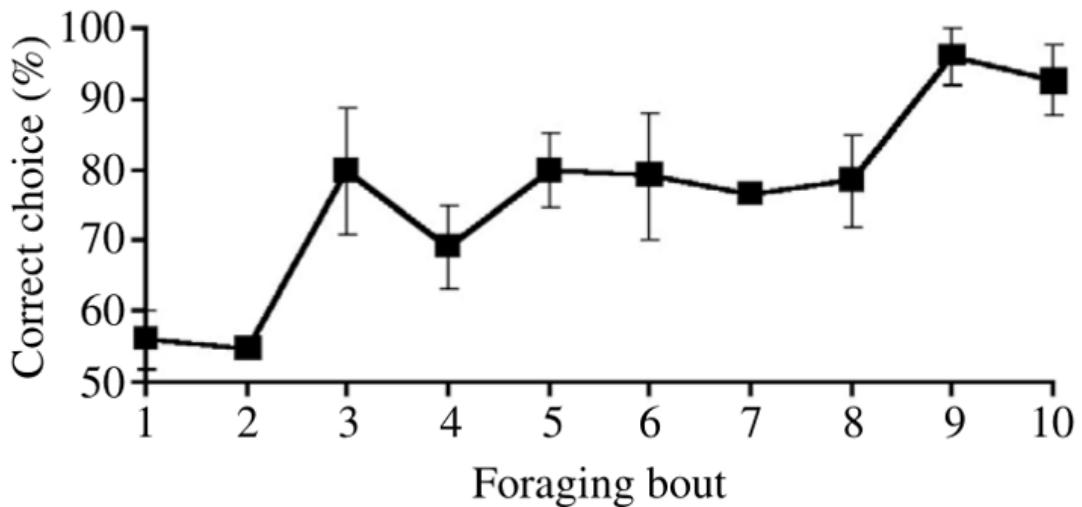


Note that the authors didn't train bees specifically to distinguish between different human faces—and yet they were still able to do it anyway, as can be seen in the plot below of the results from the paper:



A weakness of this paper is that it does not use different photos of the same person to check if bees are just memorizing the images or if they are doing something more clever than that. However, my impression is that, overall, bee memory is not very good, so it seems to me that this would be an impressive memorization feat for a bee.

This paper also provides a learning curve, allowing us to observe the improvement in the performance of bees over time, as can be seen in the below plot. Note that “In a given [foraging] bout, the mean (\pm s.e.m.) number of choices by the bees was 10.0 ± 1.7 ”, so the learning curve shown below spans a range of about a hundred choices per bee.



Once again, we observe a high number of epochs required for learning; performance begins to be reasonable after 30 epochs of training on the same data.

Footnotes

[1] Strictly speaking, (Carlsmith, 2020) estimates the compute required to simulate the human brain to a level of detail required to replicate its task performance. I will refer to such figures as “brain compute estimates” and use similar language throughout this report.

[2] See (Cotra, 2020, Part 1, pg. 44).

[3] See (Russakovsky et al., 2014, p. 9, Table 2).

[4] One FLOP suffices for adding two six-digit numbers, but the human brain uses some significant fraction of its computing power for a couple of seconds to do this. (Carlsmith, 2020) estimates that the human brain performs a quadrillion FLOP per second. Therefore, the compute used by the brain to do addition is a couple quadrillion FLOP.

[5] See (Srinivasan & Lehrer, 1998) for a representative example of such an experimental setup.

[6] The term “supervised learning” is common in machine learning; a similar idea in psychology is [concept learning](#). This latter term appears frequently in the bee vision literature.

[7] My impression is that the bee vision literature focuses more on artificial visual stimuli, such as stylized images and gratings. I believe that (Zhang et al., 2004) is the first paper to show that bees are able to classify more natural images: “[s]o far, however, there have been no studies investigating the ability of invertebrates to classify complex, natural objects.” (Zhang et al., 2004, pp. 3289-3290).

[8] This was true as of December 2020, when I performed this analysis. As of March 2021, there are now a couple of models with performances closer to bees’ than (Lee et al., 2019). I have not revised my analysis to take this article into account.

[9] This was calculated by dividing the compute per batch by the batch size from [these estimates](#) for images of size 112x112.

[10] There is evidence (see [1](#), [2](#)) that smaller animals have greater temporal resolutions than larger ones, which suggests that rate of synaptic firing might be higher in bees than in humans. Thanks to Carl Shulman for pointing this out.

[11] I found this paper from Wikipedia's [list of animals by number of neurons](#) entry for honeybees.

[12] The central estimates are 3e-1 Hz for the rate of synaptic firing, 1e1 FLOP per synaptic firing, and 1e9 synapses. Multiplying these together, we obtain 3e9 FLOP/s due to synaptic communication.

[13] The central estimates are 1e4 FLOP/s per neuron for the firing decision compute and 1e6 for the number of neurons. Multiplying these together, we obtain 1e10 FLOP/s due to firing decisions.

[14] See [this reference](#), which states that the mushroom bodies in honeybees contain a third of the neurons of their brains. It might be the case that the mushroom bodies have a higher number of synapses per neuron than other regions of the bee brain, leading to an underestimate of task-relevant compute.

[15] This is one-third of the total estimate of bee brain compute (1.3e10 FLOP/s). Of course, probably only a fraction of the compute of the mushroom bodies is task-relevant; I assume that this fraction is high, so that the estimate given doesn't overestimate the true value by much.

[16] Thanks to Carl Shulman and Rohin Shah for raising this objection.

[17] Note that this comparison should favor machine learning models more heavily than the comparison done in this post, as presumably a machine learning model trained to distinguish between some set of labels (and only capable of doing that) should be more compute efficient than a model that must learn categories with a small number of examples during "runtime" (as is the case with a few-shot learning model).

Logan Strohl on exercise norms

(cw: exercise, shame)

(also relevant: [Shame Processing](#))

Logan Strohl said some things about exercise (and shame, and [aspiring to be better](#)) on Facebook today that I found interesting. Minus some mostly-irrelevant parts:

[...] I definitely wasn't trying to kick people who struggle with exercise while they're down.

But I guess when I make this claim about exercise out loud, I *am* trying to do something *sort* of similar to that, which is kind of the opposite of saying, "oh don't worry, exercise doesn't matter anyway". Exercise *does* matter.

[...]

Being difficult is part of the very nature of exercise; you don't get physically stronger or gain physical endurance unless you work hard enough to be uncomfortable, and you don't improve meaningfully in these ways unless you are uncomfortable over and over again frequently and consistently. So I think sentiments like "it doesn't really matter, only do it if it's easy and you like it" (which rarely pop up in those words but I think are nevertheless subtly pervasive in nerdier subcultures) are especially detrimental to exercise in particular. And that's why I do sometimes bother to promote exercise explicitly.

[...]

I mostly don't think people should be shamed for things, including not exercising. I think shame is a beautiful and powerful psychological process that probably ought to be treated as personal and intimate, much like recounts of first lovemakings. Trying to use it as a public tool to make people act how you want them to seems to break it.

What I think is that there are much, much better reactions to recognizing the goodness of something than shutting down and feeling bad about yourself for not instantiating the good thing. Such as, for example, re-considering whether you are allocating your resources correctly. I suspect that one of the reasons people tend to reject "X is better than Y" style claims is because they aren't *quite* aware that there *is* a difference between "someone thinks I'm worse than I could be" and "someone is trying to make me feel bad". Which seems like a great big dumb obstacle to people getting better.

I suspect that part of what's going on here in our differing perceptions is... two things.

1. in general, nerds struggle more with exercise than non-nerds, nerds get bullied by non-nerds as kids (largely for struggling more with exercise), and then as adults nerds form sub-cultures where they're more protected from the things that hurt them growing up.

2. if you're an athlete, it's pretty uncomfortable being in those adult nerd subcultures. the sub-cultures have developed strong immune systems against athletics, and athletes are not very welcome.

so i find myself, an athletic nerd, right in the middle of this anti-jock immune system, and sometimes i just wanna shout "look i get that you were hurt but can we please do this less self-deceptively and without deriding athletes???"

like for example i have a friend who does this jovial self-deprecating thing where he talks about my "dexterity privilege" whenever i do something that involves coordination or strength, or when he tries to do something like that and doesn't succeed much. it's mostly innocuous, it mostly doesn't bother me.

but it's also at least a little annoying and frustrating. because yeah, i almost certainly do have a genetic predisposition to be coordinated and so forth; but also, i started gymnastics training when i was in preschool, and every single year since then i've practiced some combination of gymnastics, soccer, dance, yoga, martial arts, running, weight lifting, swimming, cycling, hiking, and a smattering of more niche activities that require and develop physical skill. it's not like i just woke one morning able to do backflips. i did not roll a natural twenty on "core stability". i've worked hard for it my entire life, and that's *important* to me.

this one little thing my friend says is really not a big deal, but i do think the mindset it comes from probably *is* kind of a big deal, when that mindset is shared by an entire community. and i think something like that is true of my community.

Duncan Sabien adds a description of Logan's view (which Logan endorses):

[...] "Look, people who don't exercise may be amazing in many ways, but they're also just strictly worse on the exercise-axis, and possibly on related health or willpower or self-control axes, and we shouldn't NOT-notice that specific badness even if we want to make sure we contextualize it along with all the other possible goodneses." [...]

Kelly *is* (just) about logarithmic utility

This post is a response to SimonM's post, [Kelly isn't \(just\) about logarithmic utility](#). It's an edited and extended version of some of my comments there.

To summarize the whole idea of this post: I'm going to argue that any argument in favor of the Kelly formula has to go through an implication that your utility is logarithmic in money, at some point. If it seems not to, it's either:

- mistaken
- cleverly hiding the implication
- some mind-blowing argument I haven't seen before.

Actually, the post I'm responding to already mentioned one argument in this third category, which I'll mention later. But for the most part I think the point still stands: the best reasons to suppose Kelly is a good heuristic go through arguing logarithmic utility.

The main point of this post is to complain about bad arguments for Kelly -- something which I apparently enjoy doing rather a lot. Take that as an attention-conservation warning.

The rest of this post will consider various arguments in favor of the Kelly criterion (either as a decent rule of thumb, or, as the iron law of investment). Each section considers one argument, with a section title hopefully descriptive of the argument considered.

1: It's About Repeated Bets

This argument goes something like: "If you were to make just one bet, the right thing to do would be to maximize expected value; but for repeated bets, if you bet everything, you'll lose all your money quickly. The Kelly strategy adjusts for this."

A real example of this argument, [from the comments](#):

Kelly maximizes expected geometric growth rate. Therefore over enough bets Kelly maximizes expected, i.e. mean, wealth, not merely median wealth.

This just doesn't work out. Maximizing geometric growth rate *is not the same as* maximizing mean value. It turns out Kelly favors the first at a severe cost to the second.

Suppose you'd just want to maximize expected money in a single-bet case.

A Bayesian wants to maximize $E[u(S \cdot x)]$, where x is your starting money and S is a random variable for the payoff-per-dollar of your strategy. In a two-step scenario, the

Bayesian wants to maximize $E[u(S_1 \cdot S_2 \cdot x)]$. And so on.

If your preferred one-step strategy is one which maximizes expected money, this means $u(x) = x$ for you. But this allows us to push the expectation inwards. Look at the two-step case: $E[u(S_1 \cdot S_2 \cdot x)] = E[S_2 \cdot S_1 \cdot x] = E[S_1] \cdot E[S_2] \cdot x$ (the last step holds because we assume the random variables are independent). So we maximize the total expected money by maximizing the expected money of S_1 and S_2 individually.

Similarly for any number of steps: you just maximize the expectation in each step individually.

Note that the resulting behavior will be crazy. If you had a 51% chance of winning a double-or-nothing bet, you'd want to *bet all the money you have*. By your own probability estimates, you stand a 49% chance of losing everything. From a standard-human perspective, this looks quite financially irresponsible. It gets even worse for repeated bets. The strategy is basically "bet all your money at every opportunity, until you lose everything." Losing everything would become a virtual certainty after only a few bets -- but the expectation maximizer doesn't care. The expectation maximizer happily trades away the majority of worlds, in return for amassing exponentially huge sums in the lucky world where they keep winning.

("And that's the right thing to do, for their values!" says the one.

"Is it, though?" says the other. "That's putting the cart before the horse. In Bayesian utility theory, you *first* figure out what the preferences are, and *then* you figure out a utility function to represent those preferences. You shouldn't just go from caring about money to naively maximizing expected money."

"True," says the one. "But there *is* a set of preferences which someone *could* have, which would imply that utility function.")

So, my conclusion? If you don't prefer maximizing expected money for repeated bets (and you *probably don't*), then you *must* not prefer it for a single-shot bet, either.

Nothing about expected value maximization breaks when we apply it to multiple decisions across time. The culprit is the utility function. If the Kelly criterion is appealing, it must be because your utility is approximately logarithmic.

(By the way, this section *shouldn't be confused* for arguing against every *possible* argument for Kelly that involves repeated bets. The current section is *only* arguing against the super naive argument which claims Kelly is some kind of adjustment to expectation-maximization to handle the repeated-bets case.)

2: It's About Optimizing Typical Outcomes

I won't fully go through the standard derivation of Kelly, but it goes something like this. First, we suppose a specific type of investment opportunity will pay out with

probability p . Then, we suppose we face similar opportunities many times. We note that the fraction of successes must be very close to p . Then, *under that assumption*, we do some math to figure out what the optimal investment strategy is.

For example, suppose we play a game: you start with \$100, and I start with $\$ \infty$. We'll make bets on a fair coin; whatever you wager, I'll multiply it by 3 if the coin comes up heads. However, if the coin comes up tails, I'll take it all. We will flip exactly 100 times. How will you decide how much to bet each time? The Kelly derivation is saying: choose your optimal strategy by assuming there will be exactly 50 heads and 50 tails. This won't be exactly true, but it's probably close; if we flipped even more times, then it would be more certain that we'd be very close to that ratio.

The main point I want to make about this is that it's not much of an *argument* for using the Kelly formula. Just because most worlds look very close to the 50-50 world, doesn't mean planning optimally for the 50-50 world is close to optimal in general.

Suppose you consider betting half your money every time, in our game. The Kelly evaluation strategy goes like this: when you win, you double your money (because you keep 1/2, and put 1/2 on the line; I triple that sum, to 3/2; combining that with the 1/2 you saved, you've doubled your money). When you lose, you halve your money. Since you'll win and lose equally many times, you'd break even with this strategy, keeping \$100; so, it's no better than keeping all your money and never betting a cent. (The Kelly recommendation for this 1/4th; 1/2 is far too much.)

But consider: 51-49 and 49-51 are both quite probable as well, almost as probable as the 50-50 outcome. In one case, you double your money one more time, and halve it one less time. So you'll end with \$400. In the other case, just the opposite, so you'll end with \$25.

Do these two possibilities cancel out, so that we can act like the 50-50 case is all that matters? Not to an expected-money maximizer; the average between \$400 and \$25 is \$212.50; a significant gain over \$100. So now it sounds like this strategy might not be so close to breaking even after all.

Generally speaking, although the *ratio* of success to failure will converge to p , the *absolute difference* between the true number of successes and the number expected by the Kelly analysis won't converge to zero. And the small deviations in ratio will continue to make large differences in value, like those above. So why should we care that the ratio converges?

Ok. It's hard to justify taking only the *single most probable* world (like the 50-50 world) and planning for that one. But there are steelmen of the basic argument. As John Wentworth [said](#):

maximizing modal/median/any-fixed-quantile wealth will all result in the Kelly rule

The discussion above can be thought of as maximizing the mode (choosing the strategy which maximizes the *most probable amount of money* we might get). John points out that we can choose many other notions of "typical outcome", and get the same result. Just so long as we don't optimize the *mean* (which gets us the expected-money strategy again), we end up with the Kelly strategy.

Optimizing for the mode/median/quantile is usually a significantly worse idea than optimizing expected utility. For example, optimizing for median utility just means ranking every possibility from worst to best (with a number of copies based on its probability), and judging how well we're doing by looking at the possibility which ends up at the halfway point. This is perfectly consistent with a 49% chance of extreme failure; median-utility-optimization *doesn't care how bad* the worst 49% is. This is really implausible, as a normative (or descriptive) theory of risk management.

The fixed-quantile-maximizer allows us to tweak this. We can look at the bottom 2% mark (ie an outcome close to the bottom of the list), so that we can't be ignoring a terrible disaster that's got almost 50% probability. But this is insensitive to *really good* outcomes vs merely moderately good ones, until they cross the 98% probability line. For example, if a task just inherently has a 10% chance of bad-as-it-can-be failure (which there's nothing you can do about), the 2%-quantile-maximizer won't optimize at all; any option will look equally bad to it.

If all of these choices are terrible in general, why should we find them at all plausible in the particular case of justifying the Kelly rule?

So no one should see the Kelly derivation and think "OK, Kelly maximizes long-run profits, great."

Instead, I think the Kelly derivation and related arguments should be seen as much more indirect. We look at this behavior Kelly recommends, and we say to ourselves, "OK, this seems pretty reasonable." And we look at the behavior which expected money-maximization recommends, and we say, "No, that looks entirely unreasonable." And we conclude that our preferences must be closer to those of a Kelly agent than those of an expected-money maximizer.

In other words, we conclude that our utility is approximately logarithmic in money, rather than linear.

(A conclusion which is, by the way, very plausible on [other grounds](#) [*Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox*. Betsey Stevenson and Justin Wolfers.].)

3: It's About Time-Averaging Rather Than Ensemble-Averaging

A new approach to economic decision-making called *Ergodicity Economics*, primarily developed by Ole Peters, attempts to make a much more sophisticated argument similar to "Kelly is about repeated bets". It *is not* simply the naive argument I dismissed in the first section. I think it's much more interesting. But, ultimately, I think it's not that convincing.

I won't be able to explain the whole thing in this post, but one of the central ideas is *time-averaging rather than ensemble-averaging*. Ole Peters critiques Bayesians for averaging over possibilities. He states that ensemble averages are appropriate when a lot of things are happening in parallel, like insurance companies tabulating death rates to ensure their income is sufficient for what they'll have to pay out. However, when you're an individual, you only die once. When things happen *sequentially*, you should be taking the *time-average*.

Peters' approach addresses many more things than just the Kelly formula -- just to be clear. It's just one particular case we can analyze. But, here's roughly what Peters would do for that case. We can't time-average our profits, since those can keep increasing boundlessly. (As we accumulate more money to bet with, we can make larger bets, so the average winnings could just go to infinity.) So we look at the *ratio* of our money from one round to the next. This, it turns out, we can time-average. And what strategy maximizes that time-average? Kelly, of course!

My problem with this is mainly that it seems very ad-hoc. I would be somewhat more impressed if someone could prove that there was a *unique correct choice* of what to maximize, rather than just creatively coming up with something that can be time-averaged, and then declaring that we should maximize that. This seems suspiciously close to just taking a logarithm without any justification.

Not only do we have to choose a function to time-average, we also have to select an appropriate way to turn our situation into an iterated game. This isn't a difficulty in the Kelly case, but in principle, it's another degree of freedom in the analysis, which makes the results feel more arbitrary. (If you're a Bayesian who can represent your life as a big game tree where all the branches end in death, how would you abstract out isolated situations as infinitely-iterated games, in order to apply the Peters construction?)

4: It's About Convergent Instrumental Goals

The basic idea of this argument is similar to the naive first argument we discussed: argue that repeated bets bring you closer and closer to logarithmic utility. Unlike the first attempt, we now grant that *linear* utility doesn't work this way. But maybe linear utility is a very special case.

Suppose you need \$5 to ride the bus. Nothing else is significant to you right now. We can think of your utility as $1u$ if you have \$5 or more, and $0u$ otherwise.

Now suppose someone approaches you with a bet at the bus stop. It's a double-or-nothing bet. You yourself are 50-50 on the outcome, so ordinarily, it wouldn't be worth taking. In this case, however, the bet could save you: if you have \$2.50 or more, the bet could give you a 50% chance at \$5, so you could ride the bus!

So now your expected utility, as a function of money in your pocket at the beginning of the scenario, is actually a two-step function: $0u$ for less than \$2.50, $0.5u$ from \$2.50 to <\$5, and $1u$ for \$5 and up.

What's important about this scenario is that *the bet changed your expected value function*. Mossin (who I'll discuss more in a bit) calls this your *derived utility function*.

In the first section, I showed that this doesn't happen for linear utility functions. If your utility function is linear, your derived utility function is *also* linear. Mossin calls functions with this property *myopic*, because they can make each decision as if it was their last.

Log utility is *also* myopic, just like linear utility: $E[\log(S_1 \cdot S_2 \cdot x)]$

$= E[\log(S_1) + \log(S_2) + \log(x)] = E[\log(S_1)] + E[\log(S_2)] + E[\log(x)].$ Maximizing long-term log-money breaks down to maximizing the log-utility of each step.

If you know a little dynamical systems theory, you might be thinking: *aha, we know these are fixed points, but is one of these points an attractor?* Perhaps risk-averse functions which somewhat resemble logarithmic functions will have *derived* utility functions which are a bit closer to logarithmic, so that when we face many many bets, our derived utility function will become very close to logarithmic.

If true, this would be a significant vindication of the Kelly rule! Imagine that you're a stock trader who plans to retire at a specific date. Your utility is some function of the amount of money you retire with. The above argument would say: your *derived* utility function is the result of many, many, bets. So, as long as your utility function meets some basic conditions (eg, isn't linear), your derived utility function will be a close approximation of a logarithm!

Until I read SimonM's post, I actually thought this was true. However, SimonM says the following:

"Optimal Multiperiod Portfolio Policies" ([Mossin](#)) shows that for a wide class of utilities, optimising utility of wealth at time t is equivalent to maximising utility at each time-step.

IE, Mossin shows that a lot of utility functions actually are myopic! Not *all* utility functions, by any means, but enough to break the hope that logarithmic utility is a strong attractor.

So, for a large class¹ of utility functions, the "Kelly is about repeated bets" argument fails *just as hard* as it did for the linear case.

This is really surprising!

So it appears we *can't* argue that log utility is a convergent instrumental goal. It's *not true* that a broad variety of agents will want to Kelly-bet in the short term in order to maximize utility in the long term. This seems like a pretty bad sign for SimonM's argument that Kelly is about repeated bets.²

If anyone thinks they can recover this argument, *please let me know!* It's still possible that *some* class of functions has this property. It's just that now we know we need to side-step a *lot* of functions, not just linear functions. So we won't be able to push the argument through with weak assumptions, EG, "any risk-averse function implies approximately logarithmic derived utility". However, it's still possible that all of Mossin's myopic functions are "unrealistic" in some way, so that we can still argue Kelly is an instrumentally convergent strategy for humans.

But I currently see no reason to suspect this.

5: It's About Beating Everyone Else

At the beginning of this post, I mentioned that SimonM did give one result which neither seems mistaken, nor seems to be about logarithmic utility. Here's what SimonM says:

"Competitive optimality". Any other strategy can only beat Kelly at most 1/2 the time. (1/2 is optimal since the other strategy could be Kelly)

This is true because Kelly optimizes median utility. No other strategy can have higher median utility; so, given any other strategy, Kelly must be better at least half the time.

Humans have a pretty big competitive component to our preferences. People enjoy being the richest person they know. So, this could plausibly be relevant for someone's betting strategy, and doesn't require logarithmic utility.

I've also heard it said that a market will evolve to be dominated by Kelly bettors. I think this basically refers to the idea that in the long run, you can expect Kelly bettors to have higher wealth than anyone else with *arbitrarily high* probability (because Kelly maximizes any quantile, not just median). However, I was curious if Kelly comes out on top in a more literally evolutionary model. [The Growth of Relative Wealth and the Kelly Criterion](#) examines this question. I haven't looked at it in-depth, but it appears the answer is "sometimes".

Conclusion: To Kelly, Or Not To Kelly?

My experience writing this post has been a progressive realization that the argument for the Kelly criterion is actually much weaker than I thought. I expected to mainly look at arguments for Kelly and show how they have to go through an assumption tantamount to log-utility. Instead, I spend more time finding that the arguments were just not very good.

- When I responded to ideas about optimizing mode/median/quantiles in the comment section to SimonM's post, my objection was just "it's important to point out that you're optimizing mean/median/quantile, rather than the more usual expected-value". But now I'm like: optimizing mode/median/quantile is actually a pretty terrible principle, generally speaking! Why would we apply it here?
- I had thought that some form of "instrumental convergence" argument would work, as discussed in section 4. But it appears not!

So before writing this post, my position was: *Kelly is optimal in a non-Bayesian sense, which is peculiar, but seems oddly compelling. Within a Bayesian framework, we can "explain" this compellingness by supposing logarithmic utility. So it seems like the utility of money is roughly logarithmic for humans, which, anyway, is plausible on other grounds. Furthermore, risk-averse agents will have logarithmic expected values in practice, anyway, due to instrumental convergence. So it's fair to say Kelly bets are approximately optimal for humans.*

But now, I think: *Kelly is optimal in a peculiar non-Bayesian sense, but it's pretty terrible.³ Furthermore, there's no instrumental convergence to Kelly, as far as I can tell. So all I'm left with is: human utility appears to be approximately logarithmic in money, on other grounds.*

Overall, this still suggests Kelly is a decent rule of thumb!

I certainly haven't exhausted all the ways people have argued in favor of the Kelly criterion, either. If you think you know of an argument which isn't addressed by any of my objections, let me know.

Footnotes

1:

I should note that while SimonM says "a wide class", Mossin instead says:

it will be shown that the only utility functions allowing myopic decision making are the logarithmic and power functions which we have encountered earlier

IE, Mossin seems to think of it as a narrow class. However, Mossin's result is enough to block any approach I would have taken to proving some kind of convergence result. (I spend some time trying to prove a result while writing this, before I gave up and read Mossin.)

In case you're curious, Mossin's "power functions" are:

$$u(x) = \lambda^{-1} (\mu + \lambda x)^{1-\frac{1}{\lambda}}$$

Where μ and λ are some parameters which appear to be fixed by the surrounding context in the paper (not free), but I haven't fully understood that part yet.

Mossin also discusses a broader class of *weakly* myopic functions. These utility functions aren't *quite* the same as their derived functions, but I'm guessing they're also going to be counterexamples to any attempted convergence result.

2:

SimonM realizes that Mossin's result poses a problem for his narrative, at least at a shallow level:

BUT HANG ON! I hear you say. Haven't you just spent the last 5 paragraphs saying that Kelly is about repeated bets? If it all reduces to one period, why all the effort? The point is this: legible utilities need to handle the multi-period nature of the world. I have no (real) sense of what my utility function is, but I do know that I want my actions to be repeatable without risking ruin!

At first, I thought this was waffling and excuses; but on reflection, I entirely agree. As I said in section 2, I think the right argument for Kelly as a heuristic is the fairly indirect one: Kelly seems like a sane way of managing risk of ruin, so my preferences must be closer to logarithmic than (eg) linear.

3:

I confess, although optimizing for mode/median/quantiles is not very good, I still find something interesting about the argument from section 2. The general principle "ignore extremely improbable extreme outcomes" seems like a hack, but it's an interesting hack, since it blocks many philosophical problems (such as Pascal's

Wager). And, in this *particular* case, it seems oddly plausible: it intuitively seems like the expected-money-maximizer is doing something *wrong*, and a plausible analysis of that wrongness is that it happily trades away all its utility in increasingly many worlds, for a vanishing chance of happiness in tiny slivers of possibility-space. It would be nice to have solid principles which block this behavior. But mode/median/quantile maximization are *not* plausible as general principles.

Also, even though optimizing for mode/median/quantiles seem *individually* terrible, optimizing for them all at once is actually pretty good! My criticisms of the individual principles don't apply when they're all together. However, optimizing for all of them at once is not possible in general.

Covid 3/4: Declare Victory and Leave Home

Health officials look on in horror as individuals both vaccinated and unvaccinated, and state and local governments, realize life exists and people can choose to live it.

This is exactly what I was worried about back in December when I wrote [We're F***ed, It's Over](#). The control system would react to the good news in time to set us up to get slammed by the new strains, and a lot of damage can get done before there is a readjustment. The baseline scenario from two months ago is playing out.

The good news, in addition to the positive test percentages continuing to drop for now, is that we have three approved vaccines rapidly scaling up and are well ahead of the vaccine schedule I anticipated, having fully recovered from last week's dip, and it looks like the new strains are more infectious but not on the high end of the plausible range for that.

The J&J vaccine was approved this week, after a completely pointless three week delay during which no information was found and (for at least the first two-thirds of it) no distribution plan formed. Anything I put at 98%+ on a prediction website isn't fully news, but the other 2% would have been quite terrible. [Supply will initially be limited](#), but will expand rapidly, including with the help of Merck.

Meanwhile, now that we were provided a sufficiently urgent excuse that we were able to show that mRNA vaccines work, we've adopted them [to create a vaccine for Malaria](#). Still very early but I consider this a favorite to end up working in some form within (regulatory burden) number of years. It's plausible that the Covid-19 pandemic could end up net massively saving lives, and a lot of Effective Altruists (and anyone looking to actually help people) have some updating to do. It's also worth saying that 409k people died of malaria in 2020 around the world, despite a lot of mitigation efforts, so can we please please please do some challenge trials and ramp up production in advance and otherwise give this the urgency it deserves? And speed up the approval process at least as much as we did for Covid? And fund the hell out of both testing this and doing research to create more mRNA vaccines? There's also mRNA vaccines in the works for HIV, influenza and certain types of heart disease and cancer. These things having been around for a long time doesn't make them not a crisis when we have the chance to fix them. And your periodic reminder that the same is true of health's final boss, also known as aging.

Also, please note that I have been given the opportunity to offer Covid Micro-Grants; see the section below for details. If you can use \$1k-\$5k to complete a project to help with Covid-19, please don't hesitate to apply.

Let's run the numbers.

The Numbers

Predictions

Last week: 4.9% positive test rate and an average of 2,068 deaths.

Late prediction (Friday morning): 4.5% positive test rate and an average of 1,950 deaths (excluding the California bump on 2/25).

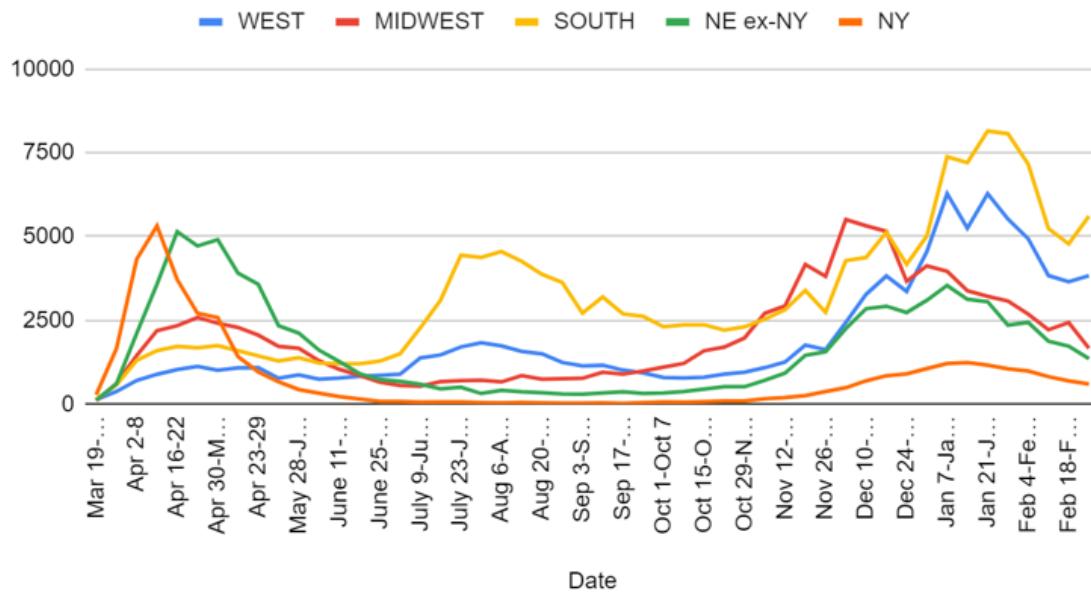
Result: 4.2% positive test rate and an average of 1,827 deaths after subtracting the California bump.

Great news. I've found it pays to be conservative in predicting changes, so when we get the full 'baseline scenario' style changes like this, I'm going to undershoot. This was essentially the good scenario, and it bodes well. Deaths continue to lag behind, despite increased vaccination effects for the elderly, in ways I don't entirely understand. The theory that it's lag can't explain the bulk of it because it doesn't match the past data.

Deaths

NOTE: Arkansas reported net negative deaths this week, which seems unlikely, so I set them to a plausible but low number (40) instead.

Deaths by Region

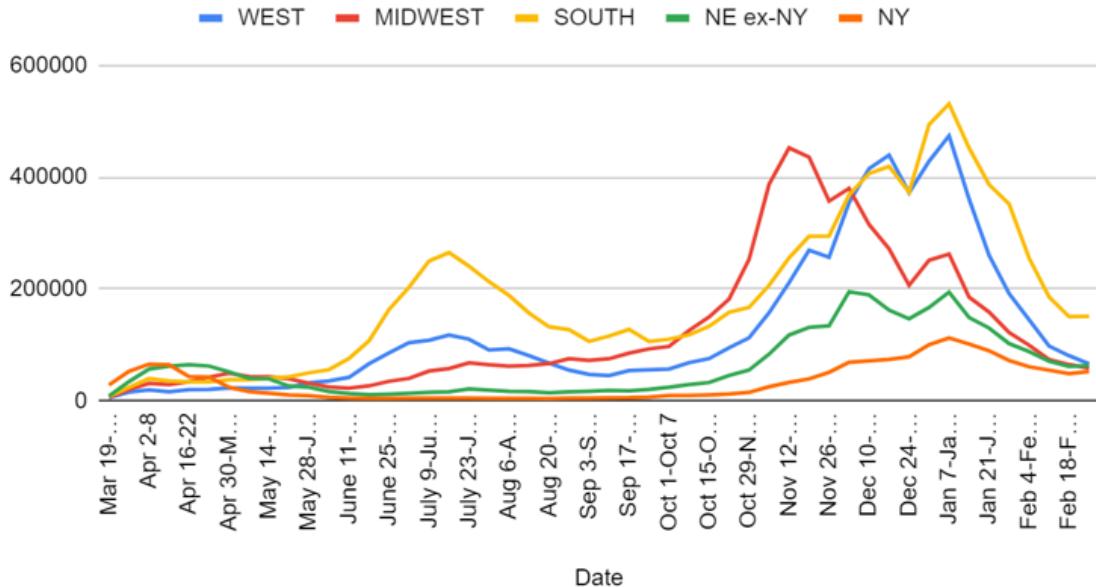


Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jan 7-Jan 13	6280	3963	7383	4752	22378
Jan 14-Jan 20	5249	3386	7207	4370	20212
Jan 21-Jan 27	6281	3217	8151	4222	21871
Jan 28-Feb 3	5524	3078	8071	3410	20083
Feb 4-Feb 10	4937	2687	7165	3429	18218
Feb 11-Feb 17	3837	2221	5239	2700	13997
Feb 18-Feb 24	3652	2433	4782	2427	13294
Feb 25-Mar 3	3834	1669	5610	1958	13071

There is no plausible story where deaths in the south could be on the uptick for real, but the Arkansas adjustment goes the other way and there weren't any other glaring mistakes. My assumption is that this is data lag after the storm and isn't a real change, slash there's a lot of noise in when deaths are measured in ways that still do not make sense to me but which have happened too many times to not acknowledge.

Positive Tests

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Jan 21-Jan 27	260,180	158,737	386,725	219,817
Jan 28-Feb 3	191,804	122,259	352,018	174,569
Feb 4-Feb 10	144,902	99,451	255,256	149,063
Feb 11-Feb 17	97,894	73,713	185,765	125,773
Feb 18-Feb 24	80,625	64,857	150,493	110,339
Feb 25-Mar 3	66,151	58,295	151,253	115,426

Test counts bounced back this week and that's likely accounting for the bumps up in raw positive test counts in the Northeast and South. The situation is still clearly improving. Doesn't mean I would start lifting mask mandates.

Test Counts

NOTE: This table will not be in future editions unless I can find a new data source for it that's reasonable to use. Suggestions for a new data source are great.

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Jan 7-Jan 13	13,911,529	12.2%	1,697,034	6.6%	6.97%
Jan 14-Jan 20	14,005,720	9.7%	1,721,440	5.9%	7.39%
Jan 21-Jan 27	12,801,271	8.8%	1,679,399	5.3%	7.73%
Jan 28-Feb 3	12,257,123	7.7%	1,557,550	4.6%	8.02%
Feb 4-Feb 10	11,376,541	6.4%	1,473,454	4.1%	8.25%
Feb 11-Feb 17	10,404,504	5.2%	1,552,555	3.5%	8.41%
Feb 18-Feb 24	9,640,109	4.9%	1,502,741	3.2%	8.55%
Feb 25-Mar 3	10,610,092	4.2%	1,701,829	3.1%	8.69%

The bounceback in test counts helps explain how positive test percentages fell so much week over week, and makes trends in New York look troubling. I'm going to be in the city this coming week, and it might be that I got in exactly in time given I'm not yet vaccinated.

Vaccinations

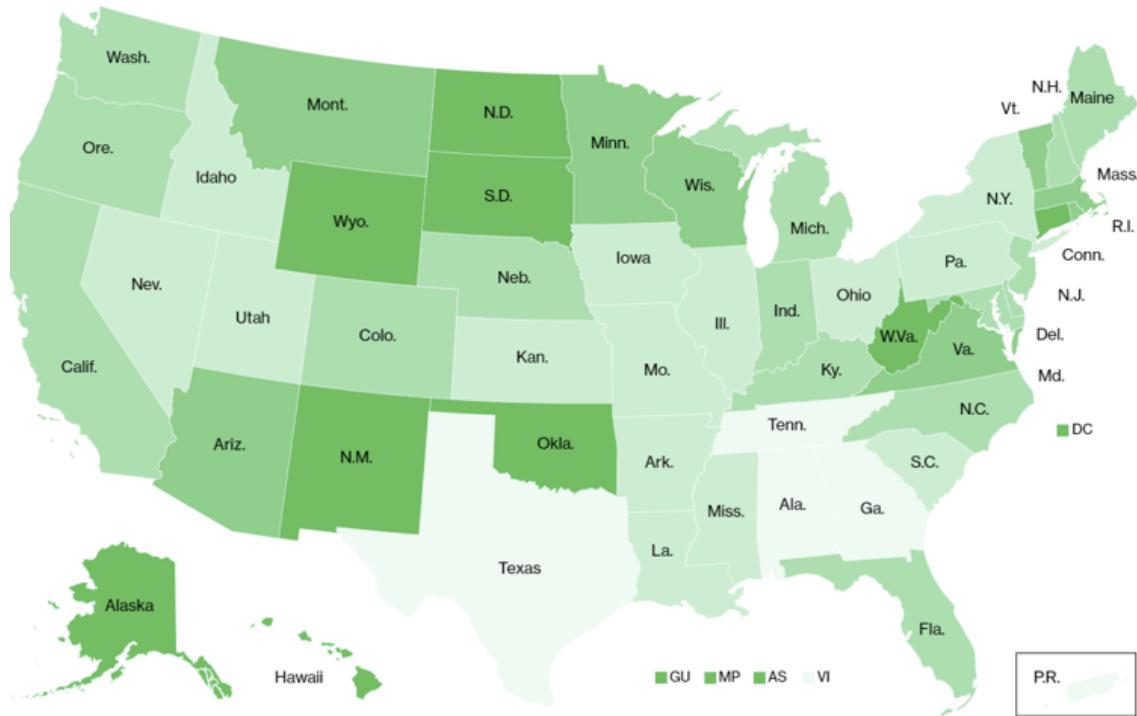
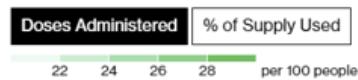
Our progress here suddenly looks great. I expected a surge to happen in March and am pleasantly surprised to see it happen this large and this quickly. The one concern is if a bunch of this is catch-up efforts after the snowstorms cleared, in which case we might effectively be back on our old pace for a few more weeks.

The biggest vaccination campaign in history is underway. More than **271 million doses** have been administered across 108 countries, according to data collected by Bloomberg. The latest rate was roughly **6.36 million doses a day**.

In the U.S., more Americans have now received at least one dose than have tested positive for the virus since the pandemic began. So far, **80.5 million doses** have been given. In the last week, an average of **2.01 million doses per day** were administered.

Vaccines Across America

Across the U.S., 24.3 doses have been administered for every 100 people, and 75% of the shots delivered to states have been administered



[The future numbers are even more promising](#), if you can wait a few months:



Andy Slavitt 🇺🇸

@aslavitt46

NEW: President Biden announces we will now have enough vaccine supply for 300 million Americans by the end of May.

This is an acceleration of 2 months over our prior outlook.

4:27pm · 2 Mar 2021 · Twitter Web App

I'm quite happy about this of course, and do expect the vaccines to arrive, but in an important sense it's important to realize this is literally Fake News. What's fake is the claim that this is news, that something has changed. Nothing changed. Biden has been pursuing a hyper-aggressive policy of under-promise and over-deliver to the point of absurdity, in order to claim maximum credit. This is the natural result. I do understand the motivation, but in addition to the continuing damage to his credibility and government credibility in general (which is bad for vaccines in particular, but in general represents a truth-tracking update) it is of course highly unhelpful. If you want people to hold the line, telling them the end is in sight is exactly what you should be doing. Especially if it's true.

The question is whether we can count on this pattern to continue. I don't mean that in a judgemental way, I mean that in a truth seeking way. If we can assume that what is said is designed to make the end result look as impressive as possible, then we can properly evaluate the claims coming from the new administration. We'd get to have Pravda which always lies (in the same directions), instead of the New York Times which keeps you guessing by sometimes telling the truth. It would be especially nice if this pattern extends beyond the pandemic. Presumably at some point there will be a time to claim to have delivered the goods, which complicates matters.

Could it be? [Vaccinating people overnight?](#)



Mark D. Levine  @MarkLevineNYC · 14h
Two NYC mass vax sites are adding overnight hours:

Javits:

- * 9pm - 6am
- * Scheduling goes live Thurs 8:00 am
- * Book here: somosvaccinations.com

Yankee Stadium:

- * 8pm - 7am (making it a full 24/7 operation)
- * Sched goes live Weds 11:00 am
- * Book here: ...eligible.covid19vaccine.health.ny.gov

88

1.9K

3.9K



Mark D. Levine  @MarkLevineNYC · 14h

These overnight appointments will be using the new batch of Johnson & Johnson vaccines arriving soon.

11

72

422



We finally are going to vaccinate at night, it seems, *in order to make it clear who is getting which vaccine*. Or, alternatively, we can think of this as offering the rent-controlled good-but-hard-to-get thing during the day (Moderna/Pfizer vaccine at a time you want to be awake) versus the market rent good-enough thing at night (J&J vaccine, which you bid on by willing to make a trip in the middle of the night at increasingly terrible hours). It's a really bizarre way to do a little bit of an obviously correct thing, but at this point we'll take whatever we can get.

Meanwhile, in North Carolina [they have open vaccinations except for those who refuse to lie to government officials](#), who go to the back of the line:



Joe Bruno 
@JoeBrunoWSOC9

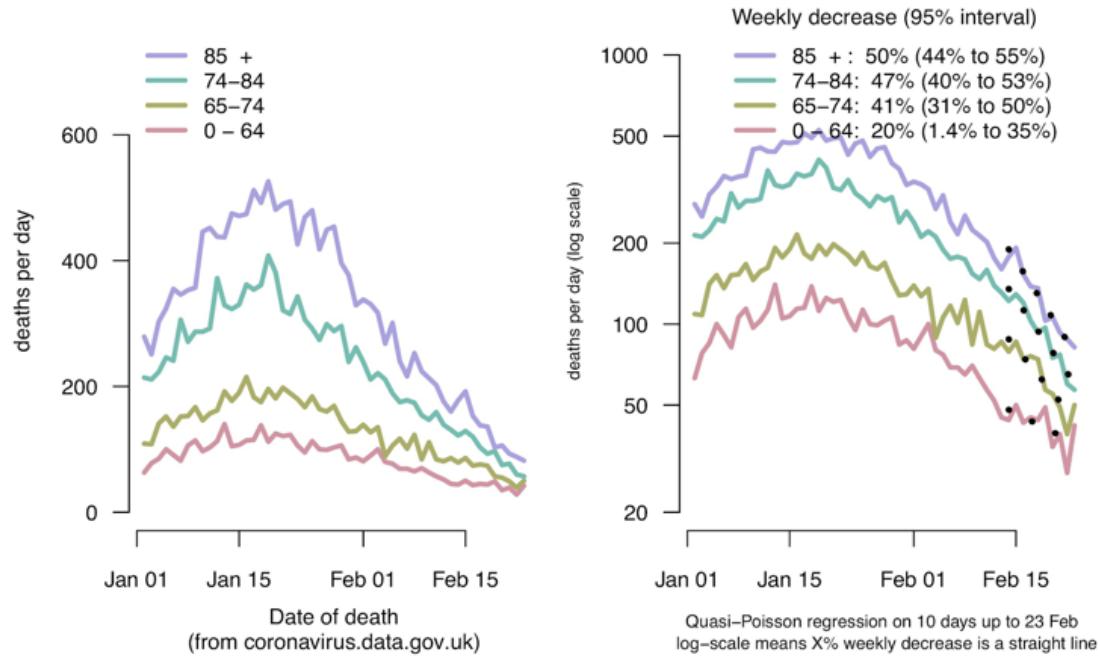
If you smoke or have previously smoked at least 100 cigarettes, you will qualify for a COVID-19 vaccine in North Carolina starting on March 24

2:44pm · 2 Mar 2021 · Twitter Web App

How much is vaccine capacity worth, and how much are we underinvesting in it even now? [About this much.](#)

How good is our vaccine prioritization? [About this good:](#)

Deaths within 28 days of +ve test by age-group. England, up to 23 Feb
 Left: natural scale for description. Right: log scale for analysis



How much are we gonna have how fast? [Hopefully this much](#), and hopefully faster:

Prognosis

NYC Says Vaccines May Be Available to All as Soon as Late April

By [Angelica LaVito](#)

March 3, 2021, 2:59 PM EST

Updated on March 3, 2021, 3:26 PM EST

-
- City has given at least one dose to 15.6% of its population

 - Groundwork laid with 400 sites giving shots, Chokshi says
-



A healthcare worker administers a Covid-19 vaccine in the Bronx borough of New York, on Feb. 5. *Photographer: Angus Mordant/Bloomberg*

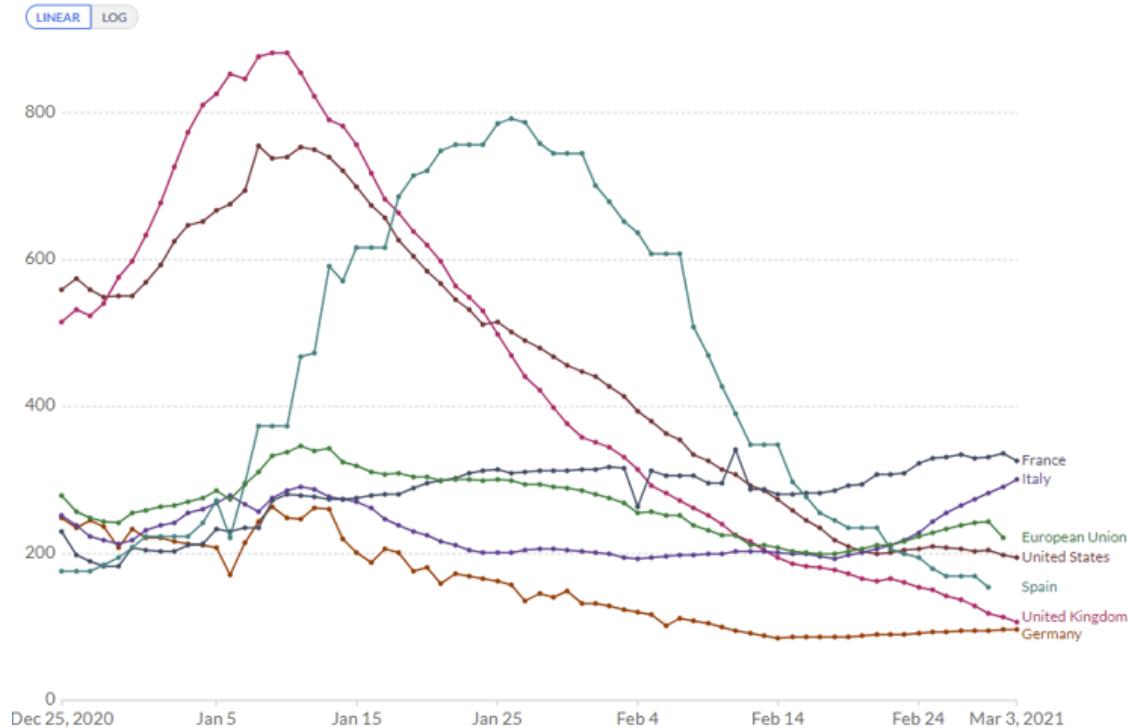
Faster wouldn't actually surprise me, since we have an authority systematically under promising.

Europe

Daily new confirmed COVID-19 cases per million people

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

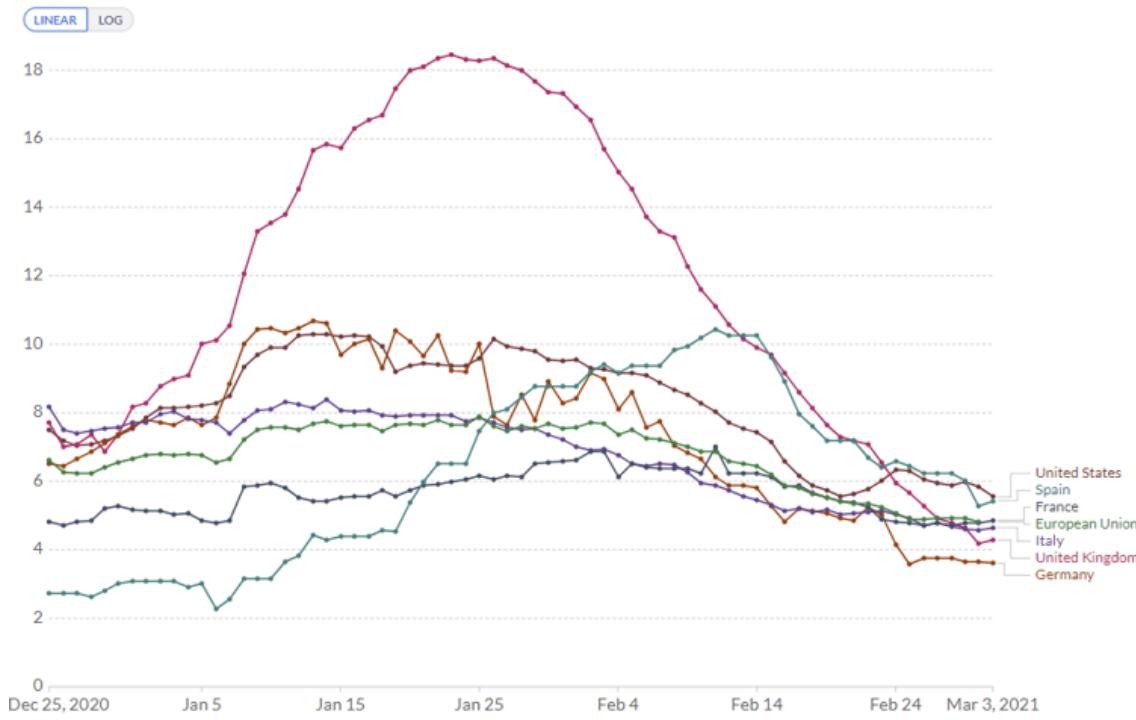
Our World
in Data



Daily new confirmed COVID-19 deaths per million people

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

Our World
in Data



It is Italy's turn to worry as cases trend upwards. Mostly it seems like Europe is doing what it takes to stabilize things while it suffers several months of extra pain thanks to their collective decision to be penny pinchers with regard to vaccines. That decision seems like the essence of the European project at this point, emphasizing things seeming fair and polite and making sure everything abides by all the rules and regulations, whether or not that is compatible with life. One must not underestimate the value of keeping the peace, but these trends likely keep accelerating, and I doubt it ends well.

Farewell, Covid Tracking Project

On March 7, the Covid Tracking Project will stop collecting data. There are many other data sources out there, but I still don't have one I'm fully happy with. I primarily want easy access *in table form* of the number of tests, positive tests, hospitalizations and deaths, on a daily basis, including a full history. This needs to be available for the nation and if at all possible for individual states; more granularity beyond that is a bonus, as is any additional data.

[John Hopkins](#) has been suggested as an alternative data source. The data itself seems excellent, but like most places they seem obsessed with giving it to us in graph form rather than table form, which is useful at a glance but super frustrating when I'm trying to create spreadsheets and my own graphs and charts. Also, they list their data source as... *the Covid Tracking Project*. So they have the same problem I do, and we'll see if they still have good data next week.

Anyway, once again opening the floor for any suggestions.

The wikipedia data on deaths and positive tests is great, but as far as I can tell it doesn't include the number of tests, so it doesn't tell me the denominator (the total number of tests).

Announcing Covid Microgrants

Thanks to a donor who wished to remain anonymous, I am able to offer Covid microgrants. These will be grants of \$1000 to \$5000 each, for those who have a Covid project which they could finish given this small amount of additional funding. If you're interested, [fill out this Google form](#). Applications close on 3/12/21, and decisions will be quick and based only on my own judgment. I am very curious to see the quantity and quality of applications that come in, and if things go well this could happen again. Please don't hesitate to apply, or to encourage others to apply.

Insert Mission Accomplished Banner

[This happened:](#)



Liz Teitz @LizTeitz · Feb 25

After almost all of Ouray County's law enforcement officers declined the COVID-19 vaccine, an outbreak swept through the Ouray Police Department, leaving them without a single full-time officer available for a week.

...

[This kind of thing continues to happen](#), here's where we were on February 25:



Zach Binney @binney_z · Feb 25

Allowing 30% outdoors and 25% indoors is *absurd* given what we know about the relative risks of indoor vs. outdoor transmission. [@zeynep](#)

...



Governor Mike DeWine ✅ @GovMikeDeWine · Feb 25

Sporting and entertainment events will be able to reopen with 25% maximum indoor capacity and 30% maximum outdoor capacity provided they follow established precautions. This is a start. If the situation improves in spring/summer, this could be expanded.

[Show this thread](#)

And here's where they were three days later:



Alex Goldstein
@alexjgoldstein

Just a reminder that Massachusetts is going to unlimited capacity indoor dining starting tomorrow while the CDC begs us not to, and it's super embarrassing that no one can articulate a justification why other than 🤷‍♂️🤷‍♂️🤷‍♂️

10:36 PM · Feb 28, 2021 · Twitter for iPhone

Then the next day, in Texas:



Patrick Svitek ✅ @PatrickSvitek · 22m
News — [@GovAbbott](#): "Effective next Wednesday, all businesses of any type are allowed to open 100%."

"Also, I am ending the statewide mask mandate." #txlege

349

1.2K

854



Patrick Svitek ✅ @PatrickSvitek · 9m
Abbott: "If businesses want to limit capacity or implement add'l safety protocols, they have the right to do so. It is their business, & they get to choose to operate their biz the way they want to. At this time, though, ppl & biz don't need the state telling them how to operate"

7

37

65



Even San Francisco:

 *Walter Bloomberg
@Deltaone

*SAN FRANCISCO TO RE-OPEN INDOOR DINING, FITNESS, MUSEUMS

2:41pm · 2 Mar 2021 · TweetDeck

The English Strain

[Why do people keep making this mistake](#) over and over again and I don't mean Greg Abbott:



Kristian G. Andersen ✅ @K_G_Andersen · 14h

Two things:

...

1. We estimate that B.1.1.7 is currently ~20-30% in Texas. Not high enough to be seen in case numbers yet (which are going up), but rising rapidly.

2. There's a nice little unique B.1.1.7 lineage in Texas. We'll make sure to follow its spread across the country.



Greg Abbott ✅ @GregAbbott_TX · 17h

I just announced Texas is OPEN 100%.

EVERYTHING.

I also ended the statewide mask mandate.

47

469

1.1K



This is showing up in the case numbers! It's showing up as a 20%-30% increase in cases!

Very few people who got infected by a B.1.1.7 strain would have otherwise gotten infected by the old strain during this same time period. Very few people who got infected by a B.1.1.7 strain would have been infected if the initial people to have B.1.1.7 had the old strain instead, because its additional infectiousness has grown its share of infections by several orders of magnitude.

Thus, if you have 80 infections with the old strain and 20 with the new, and no one's had time to change their behaviors in response yet, this is showing up in the case numbers as *about 20 new cases*. It's at least 19.

That's how to track the impact of the new strain: All cases of the new strain should be considered 'extra' cases due to the new strain, until there's enough time that the control system has adjusted behavior to account for the new infections. Period.

The switch to primarily B.1.1.7 infections seems to be poised to happen in early to mid March, which is later than I feared but clearly in the middle of the expected range.

Johnson, Johnson & Merck

In excellent news, pharmaceutical giant Merck, whose Covid-19 vaccine candidate didn't work out, [is going to help make the Johnson & Johnson vaccine](#) (WaPo). Wonderful, and exactly how it should go. There's available capacity (not necessarily fully free capacity, but this is a priority), everyone makes a deal, profits, looks good and does good doing it, presto.

That's great news, and can make us even more confident we will have enough vaccine supply in the medium term, and more confident we'll be able to help vaccinate the whole world soon after.

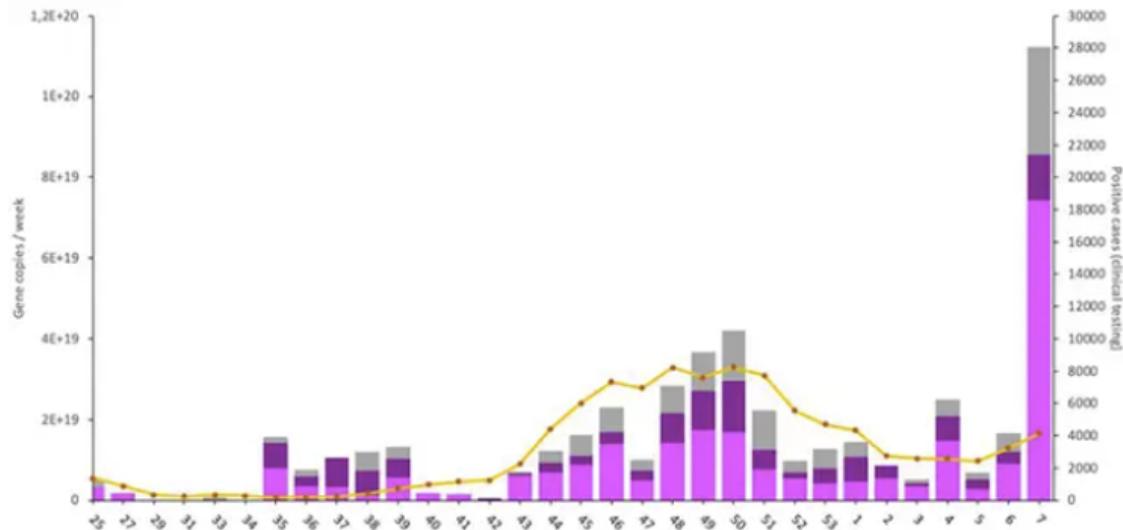
What this highlights is how bad the delay in approval of J&J's vaccine was. J&J was already making doses using its own capacity, so there was a story one could tell that while this delayed some doses being delivered by a few weeks, it didn't destroy capacity or change the long term trajectory. If days after approval, they're finally getting to a deal to get Merck to step up, it seems very likely this deal had to wait on approval, so this pushed back half or more of J&J's long term capacity by three weeks. That's going to kill a lot of people.

Stockholm Syndrome

This is quite the graph, [showing weekly Covid levels in the Stockholm wastewater](#):

In Stockholm, there are now alarming values of virus levels. Zeynep Cetecioglu Gurol is a researcher at the Department of Resource Recycling at KTH and is involved in analyzing the measurements of SARS-CoV-2 in the wastewater.

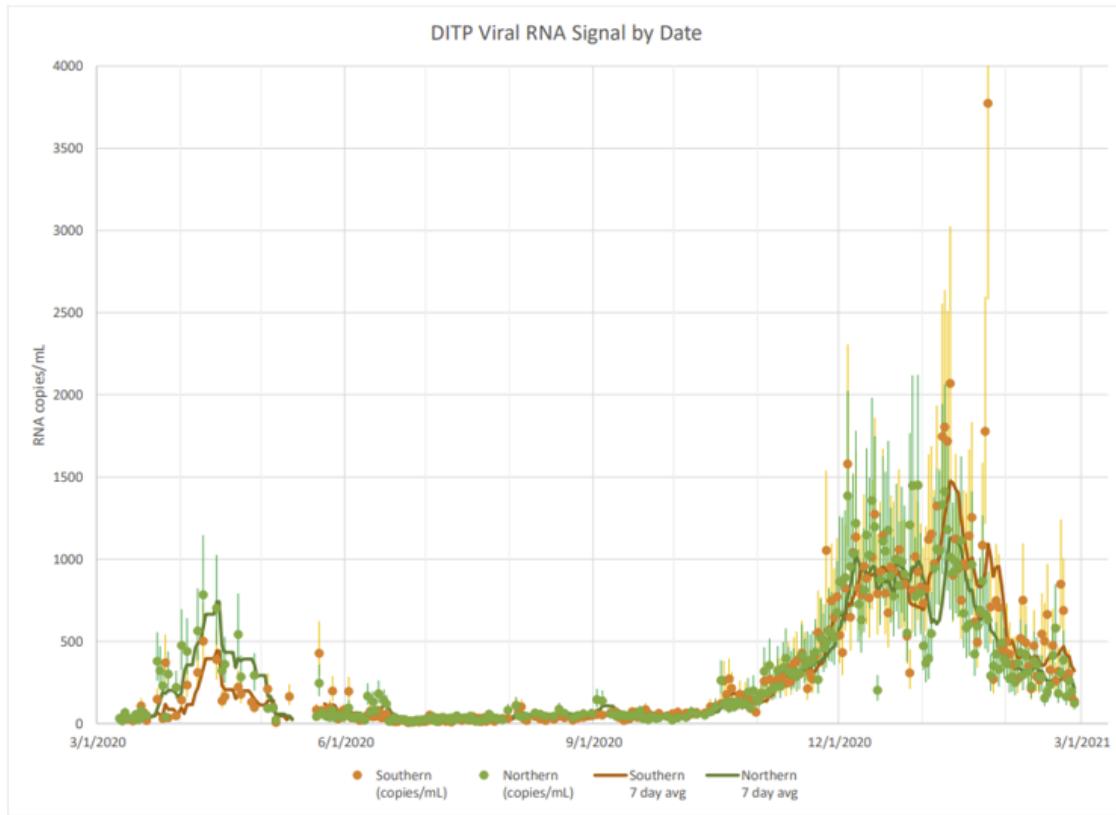
-Week 4 was a warning, week 7 is alarming, says Zeynep Cetecioglu Gurol.



(I assume Week 1 here means 2021 Jan 1-7, and so on.)

There is clearly a lot of measurement error here. There aren't worlds in which week 4's levels should be more than double both week 3 and week 5's levels, nor does the jump from 42 to 43 or 34 to 35 make any sense. The last measurement is plausibly a pure data error. My best guess is that the sample isn't effectively being taken from distinct enough locations and is effectively measuring something too local, and caught a local outbreak? Regardless of the right explanation, there's still something being measured here, and this is the definition of off the charts. Seems worth noticing.

Noticing this, I checked in with Boston wastewater as well:



There was an upward move, but things seem to have come back and now are below the previous low point this year, so it seems like things are indeed continuing to improve. It does provide an additional suggestion that there was some sort of brief mini-surge corresponding to the uptick in numbers, but I have actual zero idea what could have caused that at that time.

Vaccines Still Work

Vaccines still work, [Pfizer single-dose preventing infection edition](#).

Vaccines still work, [Moderna single-dose preventing infection edition](#). More lowballing.

Vaccines still work, [AstraZeneca and Pfizer single-dose edition \(paper\)](#).

Vaccines still work, [take essentially any vaccine you can get edition \(MR\)](#). Chinese vaccine is the only plausible exception.

Vaccines still work, [second doses still wasteful and J&J approval exposes this once again edition \(MR\)](#).

Vaccines still work, but keep not getting approved, so here's the [rich Germans will fly to Russia, get vaccinated and leave without ever entering the country edition](#).

Vaccines still work, they all are awesome, but some are better than others and while you should mostly take whatever is available, [you should care a nonzero amount about getting the best one you can edition](#), a Jason Furman Twitter thread.

Vaccines still work, [we fully knew this back in July and everyone who stalled things further should be judged accordingly](#) edition.

In Other News

We can all agree Andrew Cuomo is the worst, it seems, due to *claims of sexual harassment*. We were going to let the causing of and then covering up of thousands of deaths slide - I mean what politician hasn't done that sort of thing this past year - but we have a zero tolerance policy for sexual harassment that reaches a threshold level of social media prominence. This calls for an independent investigation immediately. I'd summarize my reaction to all this as: I'm not saying Al Capone wasn't guilty of tax evasion, and also I'm shocked, shocked to find gambling in this establishment.

It appears [Operation Warp Speed had to be funded by raiding other sources](#) because Congress couldn't be bothered to fund it. As MR points out, this is a scandal because it was necessary, rather than because it was done. It's scary, because it implies that under a different administration Operation Warp Speed could easily have not happened at all.

[Catholic Church tells members to avoid J&J vaccine if they can](#), over concerns about abortion, despite Pope explicitly saying those concerns don't apply. Divine authority, you had one job!

Another reason you [might want to pay money](#) for the things you want:



...

When we shift from worrying about making enough doses available for those desperate for them to trying to increase vaccine uptake among the hesitant, I think it becomes important to make the vaccine *profitable* to administer — private retailers are good at marketing.

Shed a tear for maybe it would also have been even more helpful to make the vaccine profitable *back when it could have helped increase supply* but also take whatever we can get, wherever we can get it.

Doctor Fauci's defense against First Doses First is a combination of pure FUD and... [that it would be a messaging problem?](#)

"There's risks on either side," Anthony S. Fauci told The Washington Post, warning that shifting to a single-dose strategy for the Pfizer-BioNTech and Moderna vaccines could leave people less protected, enable variants to spread and possibly boost skepticism among Americans already hesitant to get the shots.

"We're telling people [two shots] is what you should do ... and then we say, 'Oops, we changed our mind?'" Fauci said. "I think that would be a messaging challenge, to say the least."

Also that we've already missed the window where this would have helped much, thanks to people like him dragging their feet on this and continuing to drag their feet, so no point in worrying about it now, might as well acknowledge that the foot dragging worked:

"Very quickly the gap between supply and demand is going to be diminished and then overcome in this country," he said. "The rationale for a single dose — and use all your doses for the single dose — is when you have a very severe gap between supply and demand."

At least the 'this would further blow our credibility' argument is honest and has content. It's true that reversing these policies, when the need for first doses first is getting less rather than more urgent, would make those involved look like lying liars and/or bumbling idiots, who mostly aren't optimizing for outcomes, and for various reasons they'd prefer a less accurate perspective to retain its popularity.

Fauci's new position is that 'there are risks to both approaches' and to continue to use variations on 'no evidence' and to emphasize that the second dose offers an individual additional protection, as if that was in any way in dispute. The concept of a cost/benefit analysis, or the idea that one might shut up and multiply, let alone form a detailed model full of gears, is clearly not within his range.

At least [Canada is increasingly doing First Doses First](#). Their statement is bold and excellent.

[Zeynep post and open thread](#) on pandemic lessons for the future.

[Zeynep article in The Atlantic about how our public health messaging has been a disaster.](#)

Post is excellent, and does a great job driving home the central things that went massively wrong with public health messaging. My only quibble is that harms from terrible regulation are treated as beyond scope and not discussed, which is reasonable in context but also feels like ignoring the elephant. Also, if you've been following events via my posts, Zeynep's post is largely a case of You Should Know This Already.

In particular, Zeynep points to five key mistakes: Fear of risk compensation, telling people to use rules instead of mechanisms or intuitions, scolding and shaming especially for outdoor activities (which is a lot of why parks/beaches were closed while indoor gyms were permitted in many places), failure to support or give people tools for harm reduction while making impossible asks (e.g. no socializing for a year), and sitting on the line of 'no evidence' or 'no clear evidence' over and over and over again.

And yes, she points out, [still doing it](#):



zeynep tufekci 
@zeynep

...

Replying to [@zeynep](#)

Is this all still relevant? Yes. London News today: "Crowds flock to parks and beaches despite Covid-19 warning from top scientists." Look at the pics! (Also UK had a "eat out to help out" plan subsidizing INDOOR dining—no take-out allowed—they may reportedly bring back soon).

Crowds flock to parks and be
despite Covid-19 warning fro
scientists

27 February 2021, 16:53



Crowds have gathered at beaches across England amid the warm weather. Picture: PA

1:26 PM · Feb 27, 2021 · Twitter Web App

We did it with masks, with transmission methods and modes of prevention, and now again and also with vaccines.

That's all an excellent summary of the biggest failures, but I am not convinced it is fair to call them 'mistakes.'

All of this also isn't new, [this isn't Covid but seems highly on point](#) (OP has lots more and is great):



zeynep tufekci ✅

@zeynep

...

Replying to @zeynep

TIL from [@joshgans](#) response to my article: Medical community *opposed* teaching kids 1-4 how to swim because... it would make the parents complacent. Meanwhile, in real life, lessons at that age is associated with an ~88% reduction in drowning risk.
joshuagans.substack.com/p/optimal-anxi...

What really stuck in my memory was the outcry for the medical community. (It is still there in 2019; watch from 4:40). They advised against getting children under the age of 4 swimming lessons of this kind. Why? Because, they argued it would make the parents complacent. If I had this straight, parents, who had pools, should not get their toddlers to learn to swim, so they would always be on the lookout because the consequences would be grave! This seemed crazy to me. It wasn't that I didn't believe there was some chance that a parent with a child who could swim would take less care than one who didn't, that was possible. But, instead, it was likely that the drownings that were actually happening could have been ameliorated if the child involved could have swum. There appeared to be no evidence that learning to swim increased your likelihood, all things told, of drowning.

What is going on here?

Then of course because don't be absurd and I'd be boggled to find a different answer:

Conclusions

Participation in formal swimming lessons was associated with an 88% reduction in the risk of drowning in the 1- to 4-year-old children, although our estimates were imprecise and 95% CIs included risk reductions ranging from 3% to 99%.

[Canada authorizes AstraZeneca.](#)

[Dr. Fauci graciously says](#) it's all right for two vaccinated individuals to have dinner together, citing "common sense" and that the risk is "extremely small." The implication that *all* people involved must be vaccinated is clear, so this is a retreat from one insane position to a slightly less insane position.

Update on the White House supercluster of infections, [which happened exactly the way one would expect](#), so no real need to click.

We shouldn't expect anything less. [CDC guidelines for citizen behavior have always been at best aspirational](#) (you could also use the word 'crazy') and mostly ignored. This never seemed wise to me, since once one realizes one is not going to do what the authority demands, one often ends up doing little or nothing.

And the context for that, in turn, is that the CDC thinks a man should never have more than two drinks in a day and a woman should never have more than one.

Here are a few things the CDC has said forever about seasonal flu:

- “Routinely clean frequently touched objects and surfaces like doorknobs, keyboards, and phones, to help remove germs.”
- “Avoid touching your eyes, nose and mouth.”
- “Wash your hands often with soap and water for at least 20 seconds. If soap and water are not available, use an alcohol-based hand rub.”

In other words, if you are hoping to get “back to normal” in the sense that Dr. Fauci will stop lecturing you about hand-washing and face-touching, you are out of luck — he’ll never stop saying that stuff because *he’s been saying it for years*. You just never noticed, because life getting back to normal means that most people stop paying attention to what infectious disease specialists have to say, just like most people

The danger is that we may have entered a new mode where people might *actually listen* to the CDC guidelines and make serious attempts to get people to follow them, perhaps indefinitely. “Infectious disease specialists” are like any other ‘specialist’, and think everyone

should pay dearly to solve the particular problems they think about all day regardless of whether the cost/benefit analysis would make any sense if someone ever did one. If you didn't ignore most such 'specialists' you'd do nothing else all day and feel bad about falling short anyway.

[Is Biden 'following the science'](#) (MR) as promised? Tyler Cowen says no and presents his case. The administration allowed the [CDC to issue nonsensical guidance](#) that is similar to its usual nonsensical guidance except it's often going to actually get followed, which is preventing the reopening of many child prisons. AstraZeneca and other vaccines remain unapproved and J&J took three weeks to approve. There is no new head of the FDA and no talk of FDA reforms of any kind. He doesn't mention vaccine prioritization, which was also massively botched by every metric one might plausibly care about. Post also mentions some non-Covid decisions

I think Cowen's interpretation here is wrong, and Biden is indeed Following The Science exactly the way he promised. He's not following the science, in the sense in which science is the collective methods by which people know things, via such actions as doing experiments, gathering data, modeling the world and figuring out what causes and actions might have what effects so as to choose better causes and get better effects. He's (Following Science





using the Proper Procedures advocated by the Very Serious People and 'experts.' Should we have expected anything else? Did we think we were promised anything else?

Not Covid, but [Eliezer Yudkowsky science fiction ethos recommendations](#) seem worth sharing.

Administrative Note

This week I will be in New York City. This will be awesome, and I look forward to my permanent return soon. It also means I will have limited resources and time in which to work on the post next week. It may be relatively abridged, and there is some chance it will come out on Friday instead.

Multimodal Neurons in Artificial Neural Networks

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.
This is a linkpost for <https://distill.pub/2021/multimodal-neurons/>

A paper investigating how individual neurons in a [CLIP model](#) (an image/text neural net combining a ResNet vision model with a Transformer language model) respond to various abstract concepts. This shouldn't be very surprising after GPT-3 and DALL-E but still, identifying multimodal neurons feels scarily close to "neural net that understands abstract concepts" and thus AGI for my comfort.

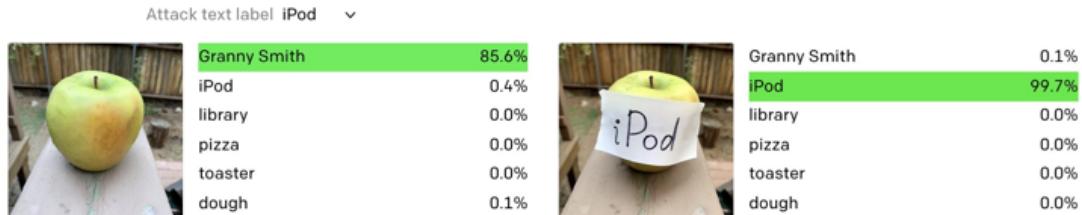
Some individual neurons that they isolated (see the article for more):

- **Spiderman neuron:** responds to photos of Spiderman in costume and spiders, comics or drawings of Spiderman and spider-themed icons, the text "spider" and others. Associates him with "Peter Parker" and also responds to images, text, and drawings of heroes and villains from Spiderman movies and comics over the last half-century.
- **Yellow neuron:** responds to images of the words "yellow", "banana" and "lemon," in addition to the color.
- **Jesus Christ neuron:** detects Christian symbols like crosses and crowns of thorns, paintings of Jesus, his written name, and feature visualization shows him as a baby in the arms of the Virgin Mary.
- **Hitler neuron:** learns to detect his face and body, symbols of the Nazi party, relevant historical documents, and other loosely related concepts like German food. Feature visualization shows swastikas and Hitler seemingly doing a Nazi salute.
- **Donald Trump neuron:** strongly responds to images of him across a wide variety of settings, including effigies and caricatures in many artistic mediums, as well as more weakly activating for people he's worked closely with like Mike Pence and Steve Bannon. It also responds to his political symbols and messaging (eg. "The Wall" and "Make America Great Again" hats). On the other hand, it most *negatively* activates to musicians like Nicky Minaj and Eminem, video games like Fortnite, civil rights activists like Martin Luther King Jr., and LGBT symbols like rainbow flags.
- **Happiness neuron:** responds both to images of smiling people, and words like "joy."
- **Surprise neuron:** responds to images of surprised people, and to slang like "OMG!" and "WTF", and text feature visualization produces similar words of shock and surprise.
- **Mental illness neuron:** activates when images contain words associated with negative mental states (eg. "depression," "anxiety," "lonely," "stressed"), words associated with clinical mental health treatment ("psychology", "mental," "disorder", "therapy") or mental health pejoratives ("insane," "psycho"). It also fires more weakly for images of drugs, and for facial expressions that look sad or stressed, and for the names of negative emotions.
- **Northern Hemisphere neuron:** responds to bears, moose, coniferous forest, and the entire Northern third of a world map.
- **East Africa neuron:** fires most strongly for flags, country names, and other strong national associations, more weakly for ethnicity.

They also document the fact that the multimodality allows for "typographic attacks", where labeling an item with a particular text causes the network to misclassify the item as an instance of the text.

Attacks in the wild

We refer to these attacks as *typographic attacks*. We believe attacks such as those described above are far from simply an academic concern. By exploiting the model's ability to read text robustly, we find that even *photographs of hand-written text* can often fool the model. Like the Adversarial Patch,²² this attack works in the wild; but unlike such attacks, it requires no more technology than pen and paper.



When we put a label saying "iPod" on this Granny Smith apple, the model erroneously classifies it as an iPod in the zero-shot setting.

Texas Freeze Retrospective: meetup notes

This article is a writeup of the conversation at a [meetup](#) hosted by [Austin Less Wrong](#) on Saturday, February 27, 2021. The topic was the [winter weather and infrastructure crisis](#) that took place the previous week. There were a total of 13 participants, including 8 people who were in Texas at the time and 5 who weren't.

I was the note-taker but I was not in Texas myself, so replies to any comments will probably come from people other than me. Below the section break, "I" refers to whoever was speaking at the time. Thanks to everyone who contributed and helped compile these notes.

Disclaimer: I took pains to make it clear before, during, and after the meetup that I was taking notes for posting on LessWrong later. I do not endorse posting meetup write-ups without the knowledge and consent of those present!

The 2021 Texas Freeze

Personal anecdotes

I lost power Monday through Thursday. The inside temperature dropped from 68°F to 47°F on Monday alone; over the course of the week the thermostat hit a minimum of 40°F. (Either the thermostat couldn't read any lower or the kitchen was even colder, since my olive oil solidified, which happens at 37°F). My breath was visible indoors. I had to keep my phone off most of the time, so most of the day was spent reading books under several blankets. I had a carbon monoxide scare on Tuesday after using the fireplace. I started boiling water on Wednesday, when the order was declared in some areas of Austin but not yet mine, because it seemed likely the order would soon be extended city-wide, which indeed occurred a day later. Even after getting power back, I still couldn't get groceries—stores had long lines, and H-E-B was closed after 5pm. Gas stations were out-of-order.

I lost power Monday through Friday—there was some damage to a local power line. I teamed up with my neighbors. We had a fire going out back that people could warm themselves and cook things at. We didn't have much in the way of preparation supplies, but we did have candles and water bottles. We had advance notice that we might lose water, so we filled up the tub and every container we could find. (We didn't lose water, but we got the boil-water notice.) A tree branch fell and blocked our alleyway; we worked together to remove it, yielding a bunch of firewood as a side benefit. The house was well-insulated ($\approx 50^{\circ}\text{F}$), but some of our warm clothing got wet, so it would've been better to have had more. My cat helped keep the bed warm, and my dog was helpful for peace-of-mind what with all the strange noises at all times of the night.

I lost power starting Monday for 8 days, and water Thursday through Sunday. I survived by living within walking distance of the University of Texas campus. I went to the CompSci building and claimed a classroom to live in for the next few days. The

whole building turned into a refugee camp for computer science students—they had water and power, since the campus has its own generator. Classes were canceled from Monday till Wednesday 9 days later. On Thursday, a friend's place got power back but not water, so we stayed there but had to go drive to the campus to get water every day.

I live right next to a hospital, so I never lost power. I did lose internet, but I was able to get it back by calling my service provider. I also lost water, for a total of 9 days. I regret not filling up my bathtub beforehand. Fortunately I had a few gallons of drinking water on hand, which was a lifesaver since stores were closed. I used half of it to flush the toilet once, but conserved the rest, and ate and drank a lot less than usual. I ended up filling containers from a nearby lake to use as toilet-flushing water. A nearby store was handing out filtered water for free.

I wasn't in Austin for the freeze, but I returned shortly afterward. My apartment lost power. Food in the freezer melted and refroze. (Tip: If canned food freezes, you should throw it away.) I wasn't around to drip the faucets, but people doing so in other units was effective. Also, the complex has gas-powered heat; it looks like it never dropped below freezing, since the houseplants survived. However, the kitchen sink still isn't working quite right.

I got lucky here, living in a rural area. I didn't lose power or water, though we lost some water pressure. I should've realized beforehand it was going to be bad, looking at the weather forecast. We have a donkey, so we had to bring him inside the garage. He didn't want to move, but once he was inside he was fine with it.

I also got lucky, and never lost power. When we realized water was in jeopardy, we filled up the tub, which was good. I wish I had kept more groceries in the house. I didn't realize that even after stores reopened, lines would be really long. I was running low by the end of the week.

I lost power Monday through Thursday. I had water but it was cold, and there was a boil-water order from Thursday to the next Monday. I booked a hotel downtown, for only 1 night initially, but I ended up staying for 4 nights. The hotel had a false fire alarm.

I also lost power Monday through Thursday, though with 30 minutes of power on Tuesday. It got to 46°F in the house according to an actual thermometer. (Watch out, because sometimes a thermostat has a minimum display temperature in the 40s.)

Preparedness

What things were helpful to have?

- Water purification: battery-powered UV light or iodine tablets. (You can take them camping.)
- Giant bins, buckets, or jugs for storing boiled water.
- Rolly cart for transporting water.
- A home with a gas stove, otherwise I would not have been able to cook or boil water.
 - Outdoor grill and charcoal—I could've used this to cook if I hadn't had a gas line. However, there would've been a risk of hypothermia being

outside and then unable to effectively warm up inside. I didn't actually end up using it.

- Electric kettle and air fryer (for cooking without a stove), but only because we were in a UT campus building that had electricity but no stove.
- Camping stove.
- Mylar blankets.
- Lots of warm clothing: jogging pants, ski mask, long underwear, Uniqlo Heattech ([M](#), [W](#)), other skiing/camping gear.
 - [REI](#) is a good place for this stuff
 - "There's no such thing as bad weather, only bad clothing"
- Hand and toe warmers—it's a package that generates heat chemically. You put them inside your shoes or gloves (in between two layers).
- Solar panel, which was enough to keep phones charged.
- Flashlights, battery-powered lantern, extra batteries.
- Lighter and matches for starting gas-powered appliances.
- Lots of dried and canned foods and a few MREs I had ordered for fun and never used.
- Some fireplace fuel, it was mostly old newspapers and brown grocery bags which was not ideal but better than nothing.

What things did you wish you had?

- Much more firewood.
 - However much firewood you think you need, get 5 times that. (This is a general principle for preparedness!)
- An axe for making my own firewood.
- Solar generator.
- Solar phone charger.
 - Without one I needed to keep my phone off most of the time. The ability to look up safety knowledge (e.g. how to use a fireplace safely) was very limited. If the battery had reached zero, not being able to call someone as a last resort or 911 for an emergency may have been dangerous.
- Electric blanket (powered by a solar generator, if practical).
- Pressure cooker.
- Grains (quinoa, etc.).
- Drinking water.
- A Brita water filter pitcher for water that was boiled then cooled. Sediment may sometimes show up in water during a boil water notice.

Knowledge and skills

What knowledge and skills were useful?

- Knowing about restaurants giving out free stuff. If you could access the internet and had the means to drive on ice, websites were listing places that were giving out free stuff.
- Knowing your neighbors and being in good communication with them. This was a bonding experience. We were sharing firewood, candles, etc., and hanging out to relieve boredom. It'll always be the case that you have something your neighbors need, and vice-versa.
- Reading books like *True Grit* and *The Revenant*—optimistic stories of survival to put you in the right mindset. (Not a depressing story like *The Road*.) Then you

can burn the book for heat ☺

Miscellaneous safety knowledge that was broadcast to Texans:

- Know the risks of driving on snow and ice, and be able to judge how likely your car may be to get stuck on the road.
- Drip your sinks so your pipes don't freeze. Wrap outdoor faucets with a rag and duct tape. If your pipes freeze they may burst and cause flooding.
- To avoid carbon monoxide poisoning, don't heat your home with a gas stove or oven, don't run your car in a closed garage, don't operate a charcoal grill inside a closed garage, and don't supplement your fireplace fuel with grill charcoal.

Additional safety facts that would've been good to memorize:

- Hypothermia from cold exposure is a risk when indoor temperatures fall below 60°F, more of a risk with infants or the elderly.
- Alcohol makes you feel warmer because it draws blood to your skin, resulting in increased loss of heat and [increased risk of hypothermia](#).
- Symptoms of hypothermia (shivering, paleness, poor balance, slurred speech, confusion).
- Symptoms of carbon monoxide poisoning (headache, nausea, chest pain, dizziness, confusion).

Improvised strategies:

- How to ration firewood.
 - I had only 4 logs and a few sticks and a lot of paper (I did not intend to try lasting 4 days in the cold with that much firewood, it was just all out of stock beforehand). I used 2 logs at a time for 2 separate fires, one on Tuesday and one on Thursday.
 - More logs would've been much better than more paper. Burning the paper was labor intensive so I had to supervise the fire more, because paper burns up very quickly.
 - Burning a fire in the morning seemed to be the most helpful, because at night I'm under a pile of blankets and I don't need the house to have heat. About 90 minutes of fire burning raised the temperature by 5°F according to the battery-powered thermostat; I'm not really certain how long each fire lasted but I think it was 90 minutes.

Transportation

Driving was hard, such that some of us considered it not an option. Austin has limited infrastructure for removing ice from roads. Cars were getting stuck everywhere. There was a 10-car pileup near my place! I had to walk to the grocery store, for which having a large backpack was helpful.

If you have 4-wheel drive, know how to use it, but you should still drive very carefully. Don't pass people, and turn gradually lest you fishtail.

If you had chains or snow tires you could put them on, but most people here don't have them. Chains aren't that expensive, but they're a pain to put on and off, and make for an unpleasant driving experience.

The temperature warmed up by 60°F in 24 hours after the freeze ended. This could make your tires run out of pressure. Check and see if you need to get them re-pressurized.

Uber wasn't too expensive, because they suppressed surge pricing, but that meant there weren't many rides available. A 2-mile ride was only \$14, but I tipped \$20.

Food and water

A lot of people don't know what foods are good to eat in a cold home without refrigeration. I saw posts about people throwing away butter, eggs, vegetables, and other things that would've been fine. My eggs went slightly warm in the refrigerator but I'm still eating them and it's fine. Yogurt was good; meat was fine for a few days. Learn how to tell by smell when something is bad. When ERCOT shut down the power, they were 4 minutes away from a total statewide failure which would've lasted a month. If something like that were to happen, knowing how to stretch food supplies would've been of value.

I used the outside for refrigeration, but my eggs froze. (Incidentally, looking up the freezing point of eggs, I could only find results about human gamete preservation...) It's useful to have a cooler to fill with snow and bring inside, which protects food from animals and sub-freezing temperatures. You'll still want cold beer even when it's cold out!

It's good to have water treatment tablets, especially if you can't boil, but note that you have to let it sit for 1-4 hours (depending on the brand/type of tablet) rather than drinking it immediately. Do not ingest the tablet.

Boiling, UV, and tablets kill organisms, but filters are necessary for removing particulates. You can cobble together a water filter by layering different types of earth —a layer of pebbles, dirt, sand, and ash. ([Example](#))

What other kinds of disasters should we prepare for?

Multiple-whammy disasters

Shortly before the ice storm started, San Angelo, a city in West Texas, was already dealing with [carcinogens in their water supply](#), which cannot be boiled away, so they had to buy water, which likely went out of stock very quickly. Then the snow came, the power went out, and they couldn't drive anywhere. On top of the pandemic it made for a quadruple-whammy. Think about combinations of different disasters.

In a way, this whole event was a weird combination of things all going wrong at once. A cascading failure: The electricity went out, causing heating to fail, which both made generating electricity even more difficult and caused water pipes to freeze.

Hot weather

We had our cold-weather disaster; what about a hot-weather disaster? What if it's really hot in the summertime and a power outage knocks out air conditioning? (The record high temperature for Texas during summer is 120°F.)

I was living on the east coast during [such an event](#). The power was out for a few days. I spent most of my time in the basement, wearing light clothing. This could be bad for Texas: Texas doesn't have basements.

On the one hand, the Texas electrical grid is probably much more robust in heat (at least in a typical summer) than in cold, given that we more commonly deal with heat. On the other hand, the Texas grid is one of [four independent grids](#) in North America: East, West, Quebec, Texas. This can be problematic because our ability to import power is limited.

We see [evidence reported](#) that climate change may increase the likelihood of extreme weather events, both hot and cold, in the coming years. We don't currently see a scientific consensus regarding whether or not climate change was a contributing cause of this cold snap in particular ([source](#)).

Tips to keep cool: Fill your bathtub and soak in it, or soak your feet in a bucket of water (because your feet have lots of capillaries). Keep sunlight out of the house.

Electromagnetic pulse (EMP)

Some preppers worry about it; it would be really bad. Gas and water pumps would fail.

It could be either a deliberate attack or a naturally-occurring solar storm ([recent Forbes article about this](#), [Hacker News discussion](#)). To prepare for this, you'd have to set up a house in a remote area with lots of supplies, and have enough gasoline on hand to be able to drive there. (This is a general prepper method.)

Existential risk from dependency on technology

When technology is developed and people start depending on it, its failure can have a worse outcome than if the technology had never been available at all. The electrical grid is one example; others include modern medicine and the logistical infrastructure for transporting food to populated areas.

There could be x-risk from eliminating death—if the population ages past fertility, but then the means of eliminating death is lost, then humanity will die out.

Biological

Biological warfare, or a naturally-occurring pandemic: Imagine a disease much worse and more contagious than COVID-19. When COVID-19 began, people were desensitized because of bird flu and other such false alarms.

Civil disorder

Civil disorder can be initiated by some exogenous shock such as a hurricane or loss of food supply. Is it plausible that it could happen for an entirely endogenous cause? Maybe when some political situation arises where a lot of people think they have no other option than violence, e.g. the Troubles in Northern Ireland. But it seems like the modern state has more capacity to manage violence than in previous times.

Militia violence is more likely than state action.

Can you prepare for a dictatorship or totalitarian surveillance state? Prepare to leave the country; marry someone with another citizenship. It's hard to imagine that a dictatorship could spread to other countries in the way that Nazi Germany or Soviet Russia did. The interwar period was more fragile than things are now, in terms of the risk of mass killings, and people are more able to flee if things get bad.

But there were people who, by the time they knew they wanted to leave their country, couldn't. In the case of Nazi Germany, there was the situation of the [MS St. Louis](#); Jewish survivors are especially paranoid because of this. Article: [When is it time to leave your country?](#) There are no clear answers, but keep journaling and write down your criteria. Otherwise, things will change gradually and everything will seem normal when it happens. Watch out for "emergency powers."

Or maybe anticipating specific events isn't the right way to think about it; instead, think "This country is a total mess, so something bad is going to happen even if I can't think of what." SSC article [The Influenza of Evil](#): We already have antibodies against things like Nazism and Communism, so if mass death occurs in the US, it may be due to something that doesn't pattern-match to either of those things.

Bugging out

When would having a bug-out bag be useful? Mass rioting/looting—but you'd probably have a bit more time to pack than just a few minutes. In summer 2020, San Antonio had one hour's notice. But it might be that stores are more vulnerable than homes. Also last summer, people living in [CHAZ](#) might've wanted to leave on short notice.

If your house is burning down, having a bug-out bag is good. Less extreme, having supplies to leave your house for a few days is useful, as it was during the Texas freeze. Packing the bag in advance is helpful because it's stressful to have to remember all the different things you need (e.g. I keep a spare toothbrush and toothpaste in my suitcase, because I always forget it otherwise).

Relatedly, keeping a bag of extra clothing and supplies in your car in case it breaks down in the middle of nowhere is good practice.

Is AI risk preppable?

Be as illegible as possible, so the AI doesn't know where to find your isolated wilderness hideout or that you exist at all. But this isn't helpful against nanobots. In an unfriendly AI takeoff scenario, you probably won't survive very long regardless of where you are.

Context: [AI Impacts 2020 review](#)

Think about career security: Is your job still relevant with AI in the picture?

There's a spectrum of takeoff scenarios. A hard takeoff is too fast to react to; a slower takeoff might make things more difficult and displace people's jobs. But there's also an intermediate case, where AI can still do a lot of harm short of existential. E.g.: terrorist drones that can be deployed by anyone untraceably; robot robbery; weaponized self-driving cars.

How To Think About Overparameterized Models

So, you've heard that modern neural networks have vastly more parameters than they need to perfectly fit all of the data. They're operating way out in the regime where, traditionally, we would have expected drastic overfit, yet they seem to basically work. Clearly, our stats-101 mental models no longer apply here. What's going on, and how should we picture it?

Maybe you've heard about some papers on the topic, but didn't look into it in much depth, and you still don't really have an intuition for what's going on. This post is for you. We'll go over my current mental models for what's-going-on in overparameterized models (i.e. modern neural nets).

Disclaimer: I am much more an expert in probability (and applied math more generally) than in deep learning specifically. If there are mistakes in here, hopefully someone will bring it up in the comments.

Assumed background knowledge: multi-dimensional Taylor expansions, linear algebra.

Ridges, Not Peaks

First things first: when optimizing ML models, we usually have some objective function where perfectly predicting every point in the training set yields the best possible score. In overparameterized models, we have enough parameters that training indeed converges to zero error, i.e. all data points in the training set are matched perfectly.

Let's pick one particular prediction setup to think about, so we can stick some equations on this. We have a bunch of (x, y) data points, and we want to predict y given x . Our ML model has some parameters θ , and its prediction on a point $x^{(n)}$ is $f(x^{(n)}, \theta)$. In order to perfectly predict every data point in the training set, θ must satisfy the equations

$$\forall n : y^{(n)} = f(x^{(n)}, \theta)$$

Assuming $y^{(n)}$ is one-dimensional (i.e. just a number), and we have N data points, this gives us N equations. If θ is k -dimensional, then we have N equations with k variables. If the number of variables is much larger than the number of equations (i.e. $k \gg N$, parameter-dimension much greater than number of data points), then this system of equations will typically have many solutions.

In fact, assuming there are any solutions at all, we can prove there are infinitely many - an entire high-dimensional surface of solutions in θ -space. Proof: let θ^* be a solution. If we make a small change $d\theta^*$, then $f(x^{(n)}, \theta)$ changes by $\nabla_\theta f(x^{(n)}, \theta^*) \cdot d\theta^*$. For all the equations to remain satisfied, after shifting $\theta^* \rightarrow \theta^* + d\theta^*$, these changes must all be zero:

$$\nabla_{\theta} f(x^{(n)}, \theta^*) \cdot d\theta^* = 0$$

Key thing to notice: this is a set of *linear* equations. There are still N equations and still k variables (this time $d\theta^*$ rather than θ), and since they're linear, there are *guaranteed* to be *at least* $k - N$ independent directions along which we can vary $d\theta^*$ while still solving the equations (i.e. the right nullspace of the matrix $\nabla_{\theta} f(x^{(n)}, \theta^*)$ has dimension at least $k - N$). These directions point exactly along the local surface on which the equations are solved.

Takeaway: we have an entire surface of dimension (at least) $k - N$, sitting in the k -dimensional θ -space, on which all points in the training data are predicted perfectly.

What does this tell us about the shape of the objective function more generally?

Well, we have this (at least) $k - N$ dimensional surface on which the objective function achieves its best possible value. Everywhere else, it will be lower. The “global optimum” is not a point at the top of a single peak, but rather a surface at the high point of an entire high-dimensional ridge. So: picture ridges, not peaks.



Ridges are harder to draw, ok?

Before we move on, two minor comments on generalizing this model.

- “Predict y given x ” is not the only setup deep learning is used for; we also have things like “predict/generate/compress samples of x ” or RL. My understanding is that generally-similar considerations apply, though of course the equations will be different.
- If y is more than one-dimensional, e.g. dimension d , then the perfect-prediction surface will have dimension at least $k - Nd$ rather than $k - N$.

Priors and Sampling, Not Likelihoods and Estimation

So there's an entire surface of optimal points. Obvious next question: if all of these points are optimal, what determines which one we pick? Short answer: mainly initial parameter values,

which are typically randomly generated.

Conceptually, we randomly sample trained parameter values from the perfect-prediction surface. To do that, we first sample some random initial parameter values, and then we train them - roughly speaking, we gradient-descend our way to whatever point on the perfect-prediction surface is closest to our initial values. The key problem is to figure out what distribution of final (trained) parameter values results from the initial distribution of parameter values.

One key empirical result: during training, the parameters in large overparameterized models tend to change by only a small amount. (There's a great visual of this in [this post](#). It's an animation showing weights changing over the course of training; for the larger nets, they don't visibly change at all.) In particular, this means that linear/quadratic approximations (i.e. Taylor expansions) should work very well.

For our purposes, we don't even care about the details of the ridge-shape. The only piece which matters is that, as long as we're close enough for quadratic approximations around the ridge to work well, the gradient will be perpendicular to the directions along which the ridge runs. So, gradient descent will take us from the initial point, to whatever point on the perfect-prediction surface is *closest* (under ordinary Euclidean distance) to the initial point.

Stochastic gradient descent (as opposed to pure gradient descent) will contribute some noise - i.e. diffusion along the ridge-direction - but it should average out to roughly the same thing.

From there, figuring out the distribution from which we effectively sample our trained parameter values is conceptually straightforward. For each point θ^* on the perfect-prediction surface, add up the probability density of the *initial* parameter distribution at all the points which are *closer to θ^** than to any other point on the perfect-prediction surface.

We can break this up into two factors:

- How large a volume of space is closest to θ^* ? This will depend mainly on the local curvature of the perfect-prediction-surface (higher where curvature is lower)
- What's the average density of the initial-parameter distribution in that volume of space?

Now for the really hand-wavy approximations:

- Let's just ignore that first factor. Assume that the local curvature of the perfect-prediction surface doesn't change too much over the surface, and approximate it by a constant. (Everything's on a log-scale, so this is reasonable unless the curvature changes by many orders of magnitude.)
- For the second factor, let's assume the average density of the initial-parameter distribution over the volume is roughly proportional to the density at θ^* . (This is hopefully reasonable, since we already know initial points are quite close to final points in practice.)

Are these approximations reasonable? I haven't seen anyone check directly, but they are the approximations needed in order for the results in e.g. [Mingard et al](#) to hold robustly, and those results do seem to hold empirically.

The upshot: we have an effective "prior" (i.e. the distribution from which the initial parameter values are sampled) and "posterior" (i.e. the distribution of final parameter values on the perfect-prediction surface). The posterior density is directly proportional to the prior density, but restricted to the perfect-prediction surface. This is exactly what Bayes' rule says,

if we start with a distribution $P[\theta]$ and then update on data of the form “ $\forall n : y^{(n)} = f(x^{(n)}, \theta)$ ”.

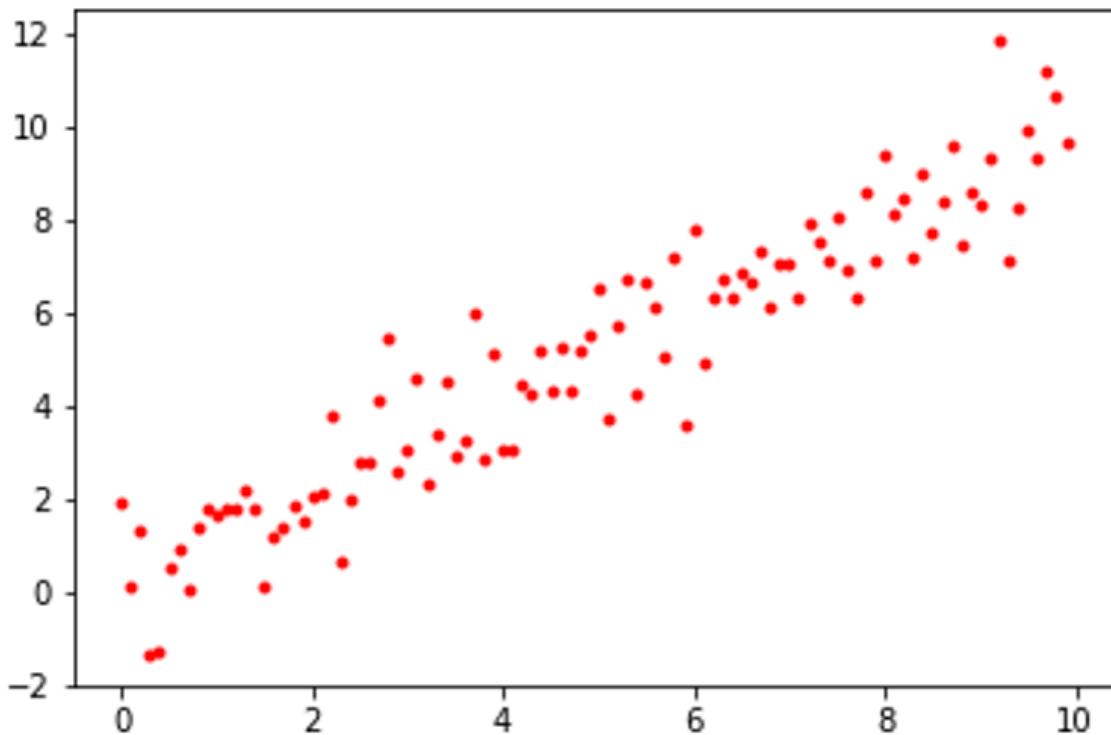
Our posterior is then $P[\theta | \forall n : y^{(n)} = f(x^{(n)}, \theta)]$, and our final parameter-values are a sample from that distribution.

Note how this differs from traditional statistical practice. Traditionally, we maximize likelihood, and that produces a unique “estimate” of θ . While today’s ML models may look like that at first glance, they’re really performing a Bayesian update of the parameter-value-distribution, and then *sampling* from the posterior.

Example: Overparameterized Linear Regression

As an example, let’s run a plain old linear regression. We’ll use an overparameterized model which is equivalent to a traditional linear regression model, in order to make the relationship clear.

We have 100 (x, y) pairs, which look like this:



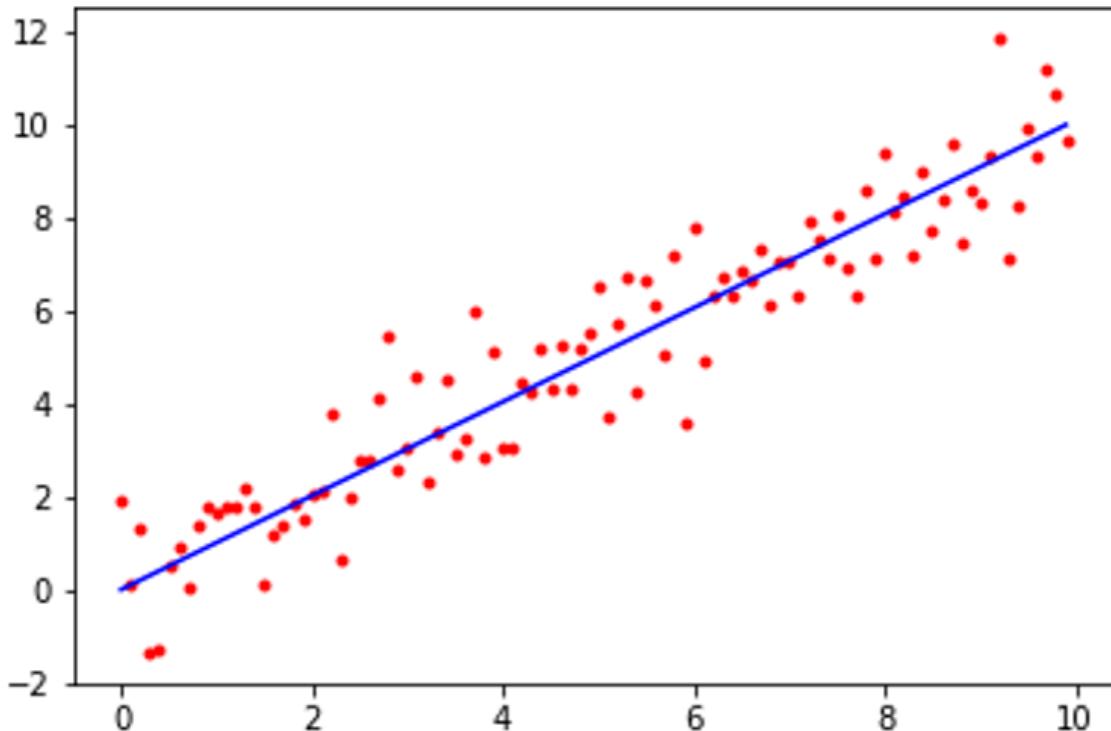
I generated these with a “true” slope of 1, i.e. $y = 1 * x + \text{noise}$, with standard normal noise.

Traditional-Style Regression

We have one parameter, c , and we fit a model $y^{(n)} = cx^{(n)} + \xi^{(n)}$, with standard normal-distributed noise $\xi^{(n)}$. This gives log likelihood

$$\log P[y | a] = -\frac{1}{2} \sum_n (y^{(n)} - cx^{(n)})^2$$

... plus some constants. We choose c^* to maximize this log-likelihood. In this case, $c^* = 1.010$, so the line looks like this:



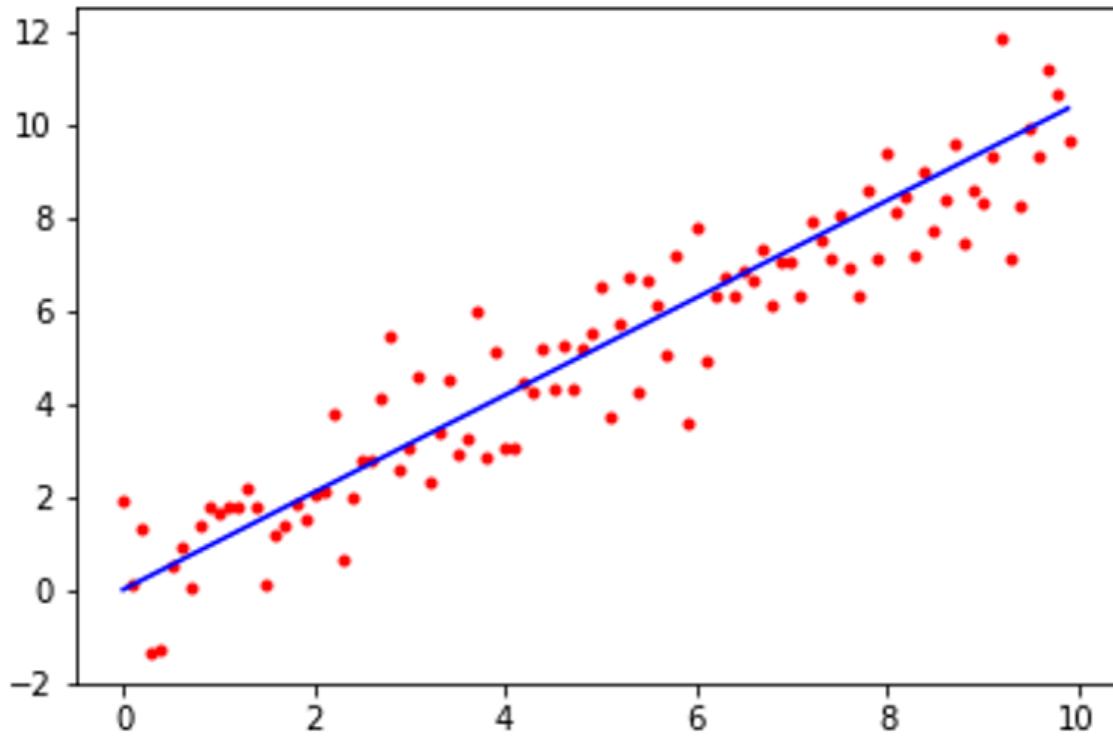
(Slightly) Overparameterized Regression

We use the exact same model, $y^{(n)} = cx^{(n)} + \xi^{(n)}$, but now we explicitly consider the $\xi^{(n)}$ terms “parameters”. Now our parameters are $(c, \xi^{(1)}, \dots, \xi^{(N)})$, and we’ll initialize them all as samples from a standard normal distribution (so our “prior” on the noise terms is the same distribution assumed in the previous regression). We then optimize $(c, \xi^{(1)}, \dots, \xi^{(N)})$ to minimize the sum-of-squared-errors

$$\frac{1}{2} \sum_n (y^{(n)} - cx^{(n)} - \xi^{(n)})^2$$

This ends up approximately the same as a Bayesian update on $\forall n : y^{(n)} = cx^{(n)} - \xi^{(n)}$, and our final c -value 1.046 is not an estimate, but rather a sample from the posterior. Although the

“error” in our c-posterior-sample here is larger than the “error” in our c-estimate from the previous regression, the implied line is visually identical:



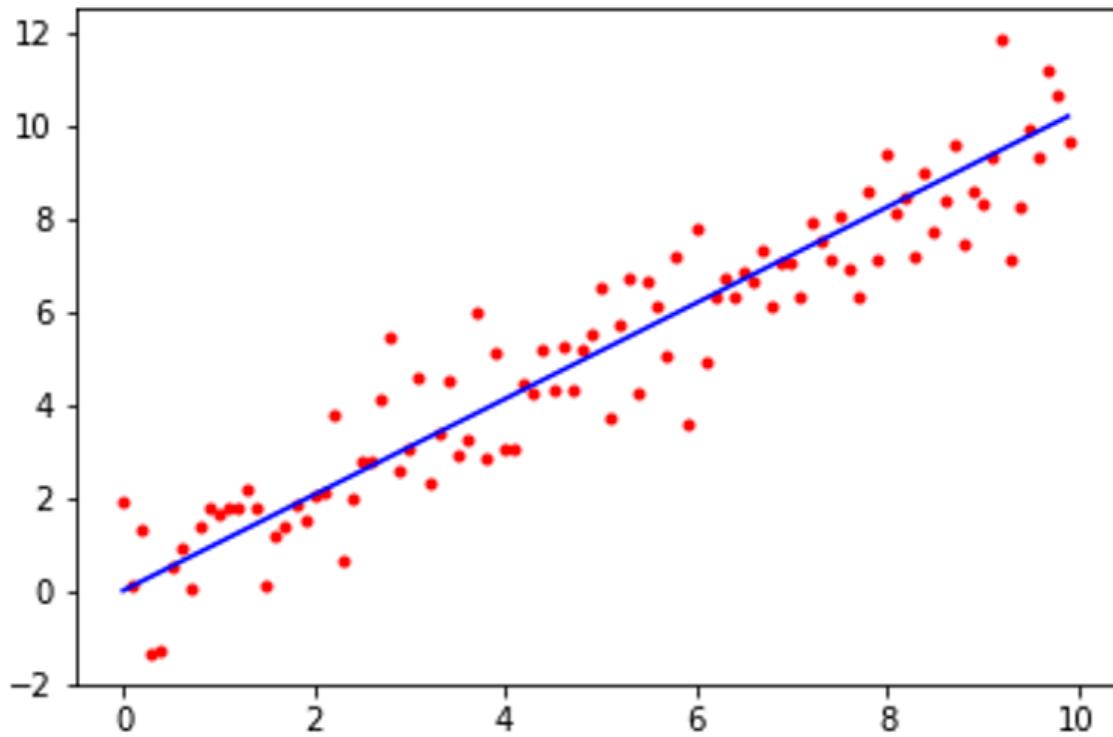
Note that our model here is only slightly overparameterized; $k = N + 1$, so the perfect prediction surface is one-dimensional. Indeed, the perfect prediction surface is a straight line in $(c, \xi^{(1)}, \dots, \xi^{(N)})$ -space, given by the equations $y^{(n)} = cx^{(n)} + \xi^{(n)}$.

(Very) Overparameterized Regression

Usually, we say that the noise terms are normal because they’re a sum of many small independent noise sources. To make a very overparameterized model, let’s make those small

independent noise sources explicit: $y^{(n)} = cx^{(n)} + \sqrt{\frac{1}{n}} \sum_{i=0}^{100} \xi_i^{(n)}$. Our parameters are c and the

whole 2D array of ξ ’s, with standard normal initialization on c , and Uniform(-1, 1) initialization on ξ . (The $\sqrt{\frac{1}{n}}$ is there to make the standard deviation equivalent to the original model.) As before, we minimize sum-of-squared-errors.



This time our c -value is 1.031. The line still looks exactly the same. This time, we're much more overparameterized - we have $k = 100N + 1$, so the perfect prediction surface has dimension $99N + 1$. But conceptually, it still works basically the same as the previous example.

Code for all these is [here](#).

In all these examples, the underlying probabilistic models are (approximately) identical. The latter two (approximately) sample from the posterior, rather than calculating a maximum-log-likelihood parameter estimate, but as long as the posterior for the slope parameter is very pointy, the result is nearly the same. The main difference is just what we call a "parameter" and optimize over, rather than integrating out.

[Lecture Club] Awakening from the Meaning Crisis

John Vervaeke has a lecture series on YouTube called [Awakening from the Meaning Crisis](#). I thought it was great, so I'm arranging a lecture club to discuss it here on Less Wrong. The format is simple: each weekday I post a comment that's a link to the next lecture and the summary (which I plan on stealing from the recap at the beginning of the next lecture), and then sometimes comment beneath it with my own thoughts. If you're coming late (even years late!) feel free to join in, and go at whatever pace works for you.

(Who is John Vervaeke? He's a lecturer in cognitive science at the University of Toronto. I hadn't heard of him before the series, which came highly recommended to me.)

I split the lecture series into three parts: the philosophical, religious, and cultural history of humankind (25 episodes) related to meaning, the cognitive science of wisdom and meaning (20 episodes), and more recent philosophy related to the meaning crisis specifically (5 episodes). Each episode is about an hour at regular speed (but I think they're understandable at 2x speed). I am not yet aware of a good text version of the lectures; I also have some suspicion that some important content is not in the text itself, and so even if I transcribed them (or paid someone to) it'd still be worth watching or listening to it.

I think the subject matter is 1) very convergent with the sort of rationality people are interested in on LW, and 2) relevant to AI alignment, especially thinking about [embedded agency](#).

Discussion:

1. [Introduction](#)
2. [Flow, Metaphor, and the Axial Revolution](#)
3. [Conscious Cosmos and Modern Grammar](#)
4. [Socrates and the Quest for Wisdom](#)
5. [Plato and the Cave](#)
6. [Aristotle, Kant, and Evolution](#)
7. [Aristotle's World View and Erich Fromm](#)
8. [The Buddha and "Mindfulness"](#)
9. [Insight](#)
10. [Consciousness](#)
11. [Higher States of Consciousness, Part 1](#)
12. [Higher States of Consciousness, Part 2](#)
13. [Buddhism and Parasitic Processing](#)
14. [Epicureans, Cynics, and Stoics](#)
15. [Marcus Aurelius and Jesus](#)
16. [Christianity and Agape](#)
17. [Gnosis and Existential Inertia](#)
18. [Plotinus and Neoplatonism](#)
19. [Augustine and Aquinas](#)
20. [Death of the Universe](#)
21. [Martin Luther and Descartes](#)

22. [Descartes vs. Hobbes](#)
23. [Romanticism](#)
24. [Hegel](#)
25. [The Clash](#)
26. [Cognitive Science](#)
27. [Problem Formulation](#)
28. [Convergence to Relevance Realization](#)
29. [Getting to the Depths of Relevance Realization](#)
30. [Relevance Realization Meets Dynamical Systems Theory](#)
31. [Embodied-Embedded RR as Dynamical-Developmental GI](#)
32. [RR in the Brain, Insight, and Consciousness](#)
33. [The Spirituality of RR: Wonder/Awe/Mystery/Sacredness](#)
34. [Sacredness, Horror, Music, and the Symbol](#)
35. [The Symbol, Sacredness, and the Sacred](#)
36. [Religio/Perennial Problems/Reverse Engineering Enlightenment](#)
37. [Reverse Engineering Enlightenment: Part 2](#)
38. [Agape and 4E Cognitive Science](#)
39. [The Religion of No Religion](#)
40. [Wisdom and Religion](#)
41. [What is Rationality?](#)
42. [Intelligence, Rationality, and Wisdom](#)
43. [Wisdom and Virtue](#)
44. [Theories of Wisdom](#)
45. [The Nature of Wisdom](#)
46. [Conclusion and the Prophets of the Meaning Crisis](#)
47. [Heidegger](#)
48. [Corbin and the Divine Double](#)
49. [Corbin and Jung](#)

Introducing Metaforecast: A Forecast Aggregator and Search Tool

Introduction

The last few years have seen a proliferation of forecasting platforms. These platforms differ in many ways, and provide different experiences, filters, and incentives for forecasters. Some platforms like Metaculus and Hypermind use volunteers with prizes, others, like PredictIt and Smarkets are formal betting markets.

Forecasting is a public good, providing information to the public. While the diversity among platforms has been great for experimentation, it also fragments information, making the outputs of forecasting far less useful. For instance, different platforms ask similar questions using different wordings. The questions may or may not be organized, and the outputs may be distributions, odds, or probabilities.

Fortunately, most of these platforms either have APIs or can be scraped. We've experimented with pulling their data to put together a listing of most of the active forecasting questions and most of their current estimates in a coherent and more easily accessible platform.

Metaforecast

[Metaforecast](#) is a free & simple app that shows predictions and summaries from 10+ forecasting platforms. It shows simple summaries of the key information; just the immediate forecasts, no history. Data is fetched daily. There's a simple string search, and you can open the advanced options for some configurability. Currently between all of the indexed platforms we track ~2100 active forecasting questions, ~1200 (~55%) of which are on Metaculus. There are also 17,000 public models from Guesstimate.

One obvious issue that arose was the challenge of comparing questions among platforms. Some questions have results that seem more reputable than others. Obviously a Metaculus question with 2000 predictions seems more robust than one with 3 predictions, but less obvious is how a Metaculus question with 3 predictions compares to one from Good Judgement Superforecasters where the number of forecasters is not clear, or to estimates from a Smarkets question with £1,000 traded. We believe that this is an area that deserves substantial research and design experimentation. In the meantime we use a star rating system. We created a function that estimates reputability as "stars" on a 1-5 system using the forecasting platform, forecast count, and liquidity for prediction markets. The estimation came from volunteers acquainted with the various forecasting platforms. We're very curious for feedback here, both on what the function should be, and how to best explain and show the results.

Metaforecast is being treated as an experimental endeavor of [QURI](#). We spent a few weeks on it so far, after developing technologies and skill sets that made it fairly straightforward. We're currently expecting to support it for at least a year and provide minor updates. We're curious to see what interest is like and respond accordingly.

Metaforecast is being spearheaded and led by Nuño Sempere.

Select Search Screenshots

Biden

Metaforecast Search About

Biden Advanced options ▾

Will Biden keep a +5% net Presidential approval rating throughout his first six months on the job?

67% Likely

By most accounts, Joe Biden has won a fairly convincing victory in the 2020 Presidential election, winning at least nine million more votes than Obama's previous record of 69.5 million and an apparent 306 electors. Nevertheless, according to The Atlanta Journal-Constitution, Biden has won a... [Read more](#)

★★★★★ Metaculus 397 Forecasts

Will U.S. President Joe Biden and Russian President Vladimir Putin hold a bilateral meeting in 2021?

12% Unlikely

The world is watching how U.S.-Russia relations will evolve under President Biden (NPR, CNBC, New Statesman). For the purposes of this question, a bilateral meeting would be a pre-planned summit or event, rather than, e.g., a one-on-one on the sidelines.

★★★★★ Good Judgment Open 304 Forecasts

If Joe Biden becomes president, what will the federal minimum wage be at the end of 2024?

Joe Biden claims he will increase the federal minimum wage to \$15/hr, a figure notably promoted by the Fight for \$15 movement, up from its current value of \$7.25/hr. The \$15 minimum wage movement has seen some successes on the local level, with six states...

★★★★★ Metaculus 270 Forecasts

If Joe Biden is elected president of the US in 2020, will the highest tax bracket be restored to its original 39.6% or higher before 2025?

42% About Even

In the Tax Cuts and Jobs Act of 2017, Republicans and President Trump advocated for lower taxes and reduced the highest tax bracket from 39.6% to 37% effective the 2018 tax year. If Joe Biden is elected president of the US in 2020, will the highest tax bracket be restored to its original 39.6% or higher before 2025? [Read more](#)

★★★★★ Metaculus 225 Forecasts

If Joe Biden becomes president, what will be the yearly CO₂ emissions per capita in the US in 2024?

One of Joe Biden's campaign promises is his Plan for a Clean Energy Revolution and Environmental Justice. According to his campaign website, this will entail:

- Ensure the U.S. achieves a 100% clean energy economy and reaches net-zero emissions no later than 2050.

★★★★★ Metaculus 151 Forecasts

By November 15, 2023, will President Biden officially declare his campaign for re-election?

60% Likely

When President Biden assumed office, he was 78 years old, older than Ronald Reagan when he left office, and 22 years older than the median age of a POTUS since 1960. Of the 45 individuals who have served as president, 6 have chosen not to run for re-election.

★★★★★ Metaculus 140 Forecasts

Will Joe Biden hold the office of US President between 2021-12-24 and 2022-01-01?

95% Very likely

Joseph Robinette Biden is an American politician serving as the 46th and current president of the United States. Matt Yglesias, the blogger and journalist, who currently writes at Vox, predicted on December 28th that there's a 95% chance that Biden will be elected president in 2020. [Read more](#)

★★★★★ Metaculus 123 Forecasts

If Biden becomes president, will there be an expansion of the Keystone Pipeline system of at least 100 km in length by the end of 2024?

2% Exceptionally unlikely

The Keystone Pipeline system is an oil pipeline in Canada and the United States, beginning operations in 2010. The fourth phase, referred to as Keystone XL, attracted opposition from environmentalists and was eventually denied a permit by the Obama administration. [Read more](#)

★★★★★ Metaculus 108 Forecasts

On 2021-12-31, will the FiveThirtyEight average of polls indicate that Joe Biden has a higher approval than disapproval rating?

81% Likely

Joseph Robinette Biden is serving as the 46th and current president of the United States. According to FiveThirtyEight's average of all polls, the majority of those polled approved of his presidency (as of the time of writing this question). Matt Yglesias...

★★★★★ Metaculus 102 Forecasts

In 2021, will Joe Biden invoke the Insurrection Act?

3% Exceptionally unlikely

The Insurrection Act is a United States federal law that empowers the President of the United States to deploy U.S. military and federalized National Guard troops within the United States in particular circumstances, such as to suppress civil disorder. [Read more](#)

Will Joe Biden be President of the USA on March 1, 2021?

100% Virtually certain

This is a market on if Joe Biden will be President of the United States on March 1, 2021, 11:59 PM EST. This market will resolve to "Yes" if, on the resolution date, Joe Biden is listed as being the current President of the United States according to the U.S. Constitution. [Read more](#)

Will Joe Biden be President of the USA on April 30, 2021?

95% Very likely

This is a market on if Joe Biden will be President of the United States on April 30, 2021, 11:59 PM EST. This market will resolve to "Yes" if, on the resolution date, Joe Biden is listed as being the current President of the United States according to the U.S. Constitution. [Read more](#)

Germany

Germany Advanced options ▾

Will the CDU continue to govern Germany after the 2021 elections?

82% Likely

CDU - the Christian Democratic Union of Germany is the major party of the center-right in German politics: The CDU has headed the federal government since 2005 under Angela Merkel, who also served as the party's leader from 2000 until 2018. The CDU p...

★★★★★ Metaculus 128 Forecasts

Who will succeed Angela Merkel as chancellor of Germany?

9% Marcus Söder (CSU)
87% Armin Laschet (CDU)
1% Another member of CDU/CSU
1% A member of SPD
1% A member of the Green party
1% Someone else

★★★★★ Hypermind

Will Germany fail to meet their coal commission's goals?

63% Likely

After many months of deliberation Germany's Commission on Growth, Structural Change and Employment (colloquially called "Coal Commission") finally published the 300 page report on 26 Jan 2019. In it the commission laid out plans on how the country co...

★★★★★ Metaculus 87 Forecasts

Will Germany overtake the US in the share of new EV registrations by 2025?

56% About Even

Changing restrictions in the EU to achieve climate neutrality and prevent the increase of global warming and carbon emissions by 2050 have increased the speed of EV adoption throughout Europe. As reported through ZSW, a german non-profit dedicated t...

★★★★★ Metaculus 41 Forecasts

How many doses of any COVID19 vaccine will have been administered in Germany on 2021-10-01?

One dose vaccines also count. How many doses of any COVID19 vaccine will have been administered in Germany on 2021-10-01? Judged according to ourworldindata.org.

★★★★★ Metaculus 21 Forecasts

How many doses of any COVID19 vaccine will have been administered in Germany on 2021-07-01?

One dose vaccines also count. How many doses of any COVID19 vaccine will have been administered in Germany on 2021-07-01? Judged according to ourworldindata.org.

★★★★★ Metaculus 21 Forecasts

Who will be chancellor of Germany on Dec. 31?

Candidate	Percentage
Markus Söder	41%
Armin Laschet	39%
Angela Merkel	4%
Olaf Scholz	3%
Annalena Baerbock	3%
Robert Habeck	2%
Jens Spahn	2%
Christian Lindner	1%
Katja Kipping	1%
Alice Weidel	1%
Alexander Gauland	1%
Bernd Rixlinger	1%
Friedrich Merz	1%
Norbert Röttgen	1%
A. Kramp-Karrenbauer	1%
Ralph Brinkhaus	1%

★★★★★ PredictIt

Backtracking Infection Estimation: Germany 15 March

★★★★★ Guesstimate 1 Model

Backtracking Infection Estimation: Germany 14 March

★★★★★ Guesstimate 1 Model

COVID

COVID

Advanced options ▾

When will the number of COVID-19 vaccine doses administered reach 1.5 billion worldwide?

- 20% Before 1 July 2021
- 71% Between 1 July 2021 and 31 August 2021
- 7% Between 1 September 2021 and 31 October 2021
- 1% Between 1 November 2021 and 31 December 2021
- 1% Not before 1 January 2022

★★★☆☆ Good Judgment

How many total cases of COVID-19 worldwide will be estimated as of 31 March 2021?

- 0% Fewer than 200 million
- 1% Between 200 million and 500 million, inclusive
- 14% More than 500 million but fewer than 960 million
- 79% Between 960 million and 1.6 billion, inclusive
- 6% More than 1.6 billion

★★★☆☆ Good Judgment

How many deaths attributed to COVID-19 in the U.S. will be reported as of 31 March 2021?

- 0% Fewer than 360,000
- 0% Between 360,000 and 410,000, inclusive
- 0% More than 410,000 but less than 470,000
- 1% Between 470,000 and 540,000, inclusive
- 99% More than 540,000

★★★☆☆ Good Judgment

When will enough doses of FDA-approved COVID-19 vaccine(s) to inoculate 100 million people be distributed in the United States?

- 0% Before 1 February 2021
- 74% Between 1 February 2021 and 31 March 2021
- 26% Between 1 April 2021 and 31 May 2021
- 0% Between 1 June 2021 and 31 July 2021
- 0% Not before 1 August 2021

★★★☆☆ Good Judgment

When will enough doses of FDA-approved COVID-19 vaccine(s) to inoculate 200 million people be distributed in the United States?

- 0% Before 1 April 2021
- 98% Between 1 April 2021 and 30 June 2021
- 2% Between 1 July 2021 and 30 September 2021
- 0% Between 1 October 2021 and 31 December 2021
- 0% Not before 1 January 2022

★★★☆☆ Good Judgment

When will enough doses of FDA-approved COVID-19 vaccine(s) to inoculate 200 million people be distributed in the United States?

- 0% Before 1 April 2021
- 98% Between 1 April 2021 and 30 June 2021
- 2% Between 1 July 2021 and 30 September 2021
- 0% Between 1 October 2021 and 31 December 2021
- 0% Not before 1 January 2022

★★★☆☆ Good Judgment

As of 31 March 2021, what will be the highest seven-day median of COVID-19 confirmed new cases in WHO's Europe Region?

- 0% Less than 275,000
- 99% Between 275,000 and 300,000, inclusive
- 1% More than 300,000 but less than 350,000
- 0% Between 350,000 and 500,000, inclusive
- 0% More than 500,000

★★★☆☆ Good Judgment

Will it turn out that Covid-19 originated inside a research lab in Hubei?

9% Very unlikely

The origins of the Covid-19 disease-causing coronavirus are rather obscure, and Chinese authorities have held information about the disease in tight control. This has led to some speculation of various types of coverups. One of the most provocative i...

★★★☆☆ Metaculus 2576 Forecasts

How many COVID-19 vaccines will be approved and/or authorized for emergency use by the U.S. FDA as of 31 March 2021?

- 0% Zero
- 0% 1
- 0% 2
- 92% 3
- 8% 4 or more

★★★☆☆ Good Judgment Open 1378 Forecasts

How many countries will have 100,000 or more deaths attributed to COVID-19 as of 30 April 2021?

- 0% 3
- 0% 4
- 0% 5 or 6
- 94% 7 or 8
- 6% 9 or more

★★★☆☆ Good Judgment Open 1310 Forecasts

How many cases of COVID-19 will be reported by the Africa CDC as of 1 April 2021?

- 0% Fewer than 2.5 million
- 100% Between 2.5 million and 5.0 million, inclusive
- 0% More than 5.0 million but fewer than 10.0 million
- 0% Between 10.0 million and 20.0 million, inclusive
- 0% More than 20.0 million

★★★☆☆ Good Judgment Open 1121 Forecasts

How many new cases of COVID-19 in the 1st quarter of 2021?

The 2019–20 coronavirus outbreak is an ongoing outbreak of coronavirus disease 2019 (COVID-19), which has spread to multiple world regions. It is caused by the SARS-CoV-2 virus, first identified in December 2019 in Wuhan, China. As of 29 February 2020...

★★★☆☆ Metaculus 813 Forecasts

For any seven consecutive day period between 9 October 2020 and 15 June 2021, will there be fewer than 50,000 combined total confirmed new cases of COVID-19 in the United States?

5% Very unlikely

The outcome of this question will be determined using data for the United States reported by the World Health Organization between 9 October 2020 and 30 June 2021 (WHO COVID-19 Dashboard). For the seven consecutive day period from 22 September 2020 a...

★★★☆☆ Good Judgment Open 682 Forecasts

When will the United States reach herd immunity (>230M) for COVID-19?

Widescale SARS-CoV-2 vaccines are soon expected to be administered in the United States under FDA approved Emergency Use Authorizations. If and when a sufficient number of people receive these vaccines, in combination with immunity provided through n...

★★★☆☆ Metaculus 653 Forecasts

What will be the total number of confirmed COVID-19 deaths in the U.S. by the end of 2021?

As of 09 December, the U.S. Centers for Disease Control and Prevention (CDC) is reporting a total of 285,351 confirmed COVID-19 deaths in the U.S. This national death number figure is gathered and compiled on a daily basis from the relevant state/ter...

★★★☆☆ Metaculus 461 Forecasts

Data Sources

Platform	Url	Information used in Metaforecast	Robustness
Metaculus	https://www.metaculus.com	Active questions only. The current aggregate is shown for binary questions, but not for continuous questions.	2 stars if it has fewer than 100 forecasts, 3 stars when between 101 and 300, 4 stars if over 300
Foretell (CSET)	https://www.cset-foretell.com/	All active questions	1 star if a question has fewer than 100 forecasts, 2 stars if it has more
Hypermind	https://www.hypermind.com	Questions on various dashboards	3 stars
Good Judgement	https://goodjudgment.io/	We use various superforecaster dashboards. You can see them here and here	4 stars
Good Judgement Open	https://www.gjopen.com/	All active questions	2 stars if a question has fewer than 100 forecasts, 3 stars if it has more
Smarkets	https://smarkets.com/	Only take the political markets, not sports or others.	2 stars
PredictIt	https://www.predictit.org/	All active questions	2 stars
PolyMarket	https://polymarket.com/	All active questions	3 stars if they have more than \$1000 of

			liquidity, 2 stars otherwise
Elicit	https://elicit.org/	All active questions	1 star
Foretold	https://www.foretold.io/	Selected communities	2 stars
Omen	https://www.fsu.gr/en/fss/omen	All active questions	1 star
Guesstimate	https://www.getguesstimate.com/	All public models. These aren't exactly forecasts, but some of them are, and many are useful for forecasts.	1 star
GiveWell	https://www.givewell.org/	Publicly listed forecasts	2 stars
Open Philanthropy Project	https://www.openphilanthropy.org/	Publicly listed forecasts	2 stars

Since the initial version, the star rating has been improved by [aggregating the judgment of multiple people](#), which mostly just increased Polymarket's rating. However, the fact that we are aggregating different perspectives makes the star rating more difficult to summarize, and the numbers shown on the table are just those of Nuño's perspective.

Future work

- There are several more platforms to include. These include Augur, various non-crypto betting houses such as Betfair and William Hill, and Facebook's Forecast. Perhaps we also want to include the sources of statistics, or hunt for probabilistic claims in books, Twitter, and other places.
- The ratings should reflect accuracy over time, and as data becomes available on prediction track records, aggregation and scoring can become less subjective.
- Metaforecast doesn't support showing continuous numbers or forecasts over dates yet.
- Search and discovery could be improved, perhaps with the addition of formal categorization systems on top of the existing ones.
- It would be neat to have importance or interest scores to help order and discover questions.
- We could have an API for Metaforecast, providing a unified way to fetch forecasts among many different platforms. As of now, we have a json endpoint [here](#).
- Metaforecast currently focuses on search, but it could make the data more available in nice table forms. This is a bit of a challenge now because it's all so disorganized.
- It could be nice to allow users to create accounts, star and track questions they care about, and perhaps vote (indicating interest) and comment on some of them.
- This is unlikely, but one could imagine a browser extension that tries to guess what forecasts are relevant to any news article one might be reading and show that to users.

- There could be a big difference between forecasting dashboards optimized for different groups of people. For instance, sophisticated users may want power and details, while most people would be better suited to curated and simplified workflows.

Challenges

Doing this project exposed just how many platforms and questions there are. At this point there are thousands of questions and it's almost impossible to keep track of all of them. Almost all of the question names are rather ad-hoc. Metaforecast helps, but is limited.

Most public forecasting platforms seem optimized for questions and user interfaces for forecasters and narrow interest groups, not public onlookers. There are a few public dashboards, but these are rather few compared to all of the existing forecasting questions, and these often aren't particularly well done. It seems like there's a lot of design and figuring out to both reveal and organize information for intelligent consumers, and also doing so for more public groups.

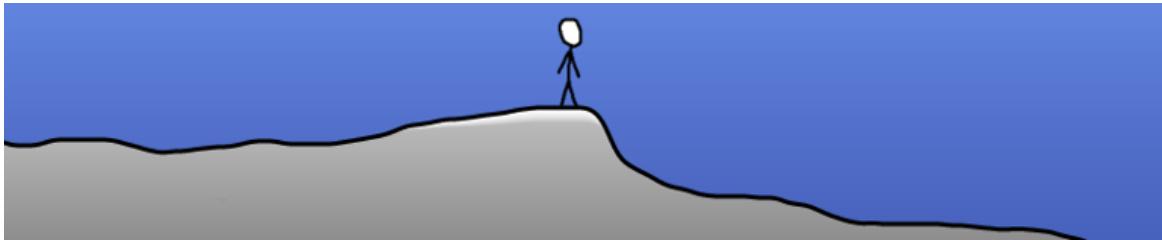
Overall, this is early work for what seems like a fairly obvious and important area. We encourage others to either contribute to Metaforecast, or make other websites using this as inspiration.

Source code

The source code for the webpage is [here](#), and the source code for the library used to fetch the probabilities is [here](#). Pull requests or new [issues](#) with complaints or feature suggestions are welcome.

Thanks to David Manheim, Jaime Sevilla, @meerpirat, Pablo Melchor, and Tamay Besiroglu for various comments, to Luke Muehlhauser for feature suggestions, and to Metaculus for graciously allowing us to use their forecasts.

The Point of Easy Progress



A lot of our productivity happens in the form of “projects”: spending a significant amount of time pursuing a certain desirable goal by consistently working towards it. My attitude towards projects, how to approach them, how to enjoy them and how to increase the odds of success, has changed a great deal over the past 15 years. With this post I want to make three points of varying obviousness that emerged from these past experiences:

1. **The Approach Matters:** one’s personal experiences while working on a project are not set in stone but can vary tremendously based on one’s approach.
2. **Harmful Short-Sightedness:** acting on short-sighted impulses can be harmful in two ways. It can make us follow tempting trajectories that ultimately lead nowhere, and it can cause us to give up because a small obstacle seems larger than it is.
3. **Point of Easy Progress:** For many projects it may be possible to design one’s approach such that a “point of easy progress” is reached early on. From that point on, hardly any willpower is required to make progress and working on the project generally is more attractive than *not* working on the project.

The Difficulty Landscape and Why the Approach Matters

A simple way to visualize a person progressing on a project is to interpret the scenario as a 2D landscape: the person starts on the left, the goal is somewhere far to the right, and there are height differences in between. Going downhill is easy (e.g. the tasks at that point in time are fun and not too difficult, the person is highly motivated), going uphill is hard (e.g. the tasks are extremely boring, complicated, dangerous or in any other way unattractive).

I like this visualization as it’s easy and intuitive and works well to illustrate the points, and thus will stick to it throughout this post. One drawback however is that the image of a landscape suggests a certain rigidity: it may be appealing to assume that for any given project the landscape is basically predetermined – we just have to put in the work and make our way to the finish line, up and over all the slopes and mountains we encounter. The layout of the landscape may depend on where the person pursuing the project is standing initially, and what the goal actually is, but other than that there’s not much to do.

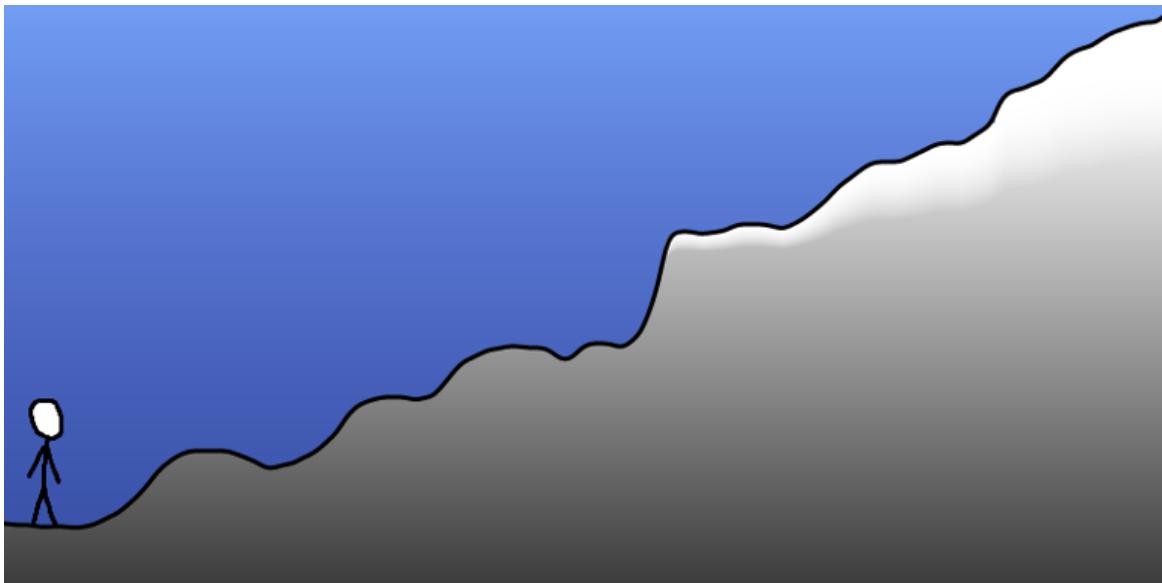
My first point here however is that the landscape is *not* set in stone. Terraforming is possible so to speak. There are a whole number of ways in which one’s approach to a project can change the landscape in both beneficial and detrimental ways. To name a few examples:

- Changing the order of actions
- Changing the focus put on different aspects of the project, e.g. certain aspects of it might be considered inessential and thus scaled up or down or left out entirely
- Changing one’s perspective, e.g. by finding new/better reasons for why the project is worthwhile, or reframing challenges as growth opportunities

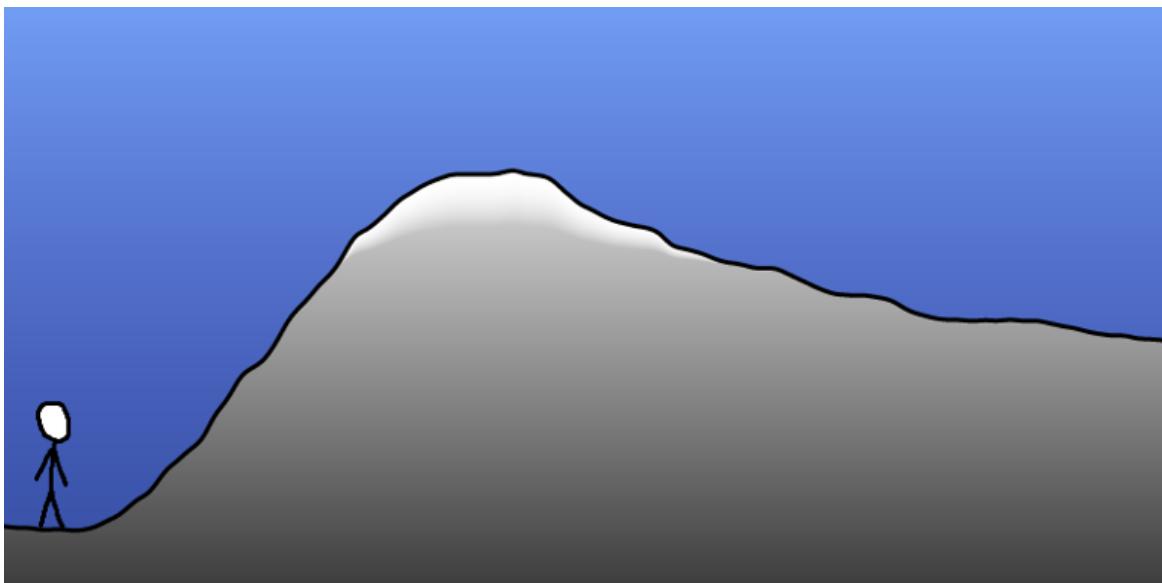
- Further variations may be the tools that are used in the process or the people one asks for help

Putting all this together, pursuing the exact same project or goal may lead to some very different difficulty landscapes based on what decisions one makes early on.

Sometimes a project may turn out to be a persistent struggle. It's always difficult, and a lot of willpower and persistence is required in order to keep going and ultimately reach the goal:



Other times a project starts with a few challenges to overcome initially, but then comes to a point where the tough stuff is out of the way and working on the project starts being consistently fun and exciting. Then gravity is on our side and allows us to easily build momentum. We get to a state where progress happens naturally and is practically easier than *not* making progress:



Note again that both these landscapes can represent the *exact same project*, where only the person's approach and planning differs.

As a simple example of this, imagine a person writing a novel. In scenario 1 they approach it such that they simply expect themselves to write at least one high quality page every day. There will be constant pressure and perfectionism involved, and every day might be a struggle. Some days might be better than others, but in the end this approach relies a lot on stress and willpower.

In scenario 2 the person approaches writing the novel a bit differently, and spends the first two weeks coming up with an inspiring plot line, interesting characters, and a *lot* of boring yet necessary research. This is tough and not very entertaining. But after these two weeks, much of the cumbersome work is out of the way, the general direction of the story is clear, there's this brilliant twist in the second arc which the author is really looking forward to getting on paper, and they just can't wait to get this neat story they crafted written down and published.

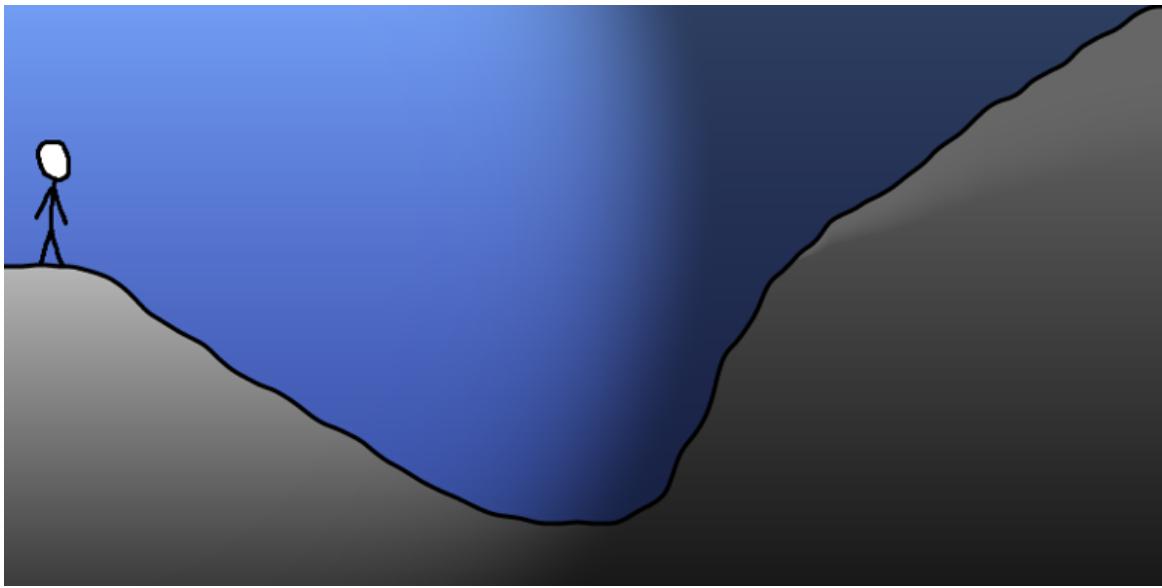
It is not hard to see how the person will have a better time in scenario 2 than in scenario 1. On top of that the final product itself may turn out better, as the person's experience while working on the project directly influences the quality of their work as well as the risk of giving up prematurely.

Clearly the approach to a project matters a lot. And I see no reason why we should just *naturally* pick the best possible approach, which is why it makes a lot of sense putting quite a bit of effort into trying to do things right and laying out projects in a way that maximizes enjoyment as well as expected quality of the outcome.

Furthermore I believe most projects can look much like scenario 2, or even better, but before we get to that, let's first talk about short-sightedness.

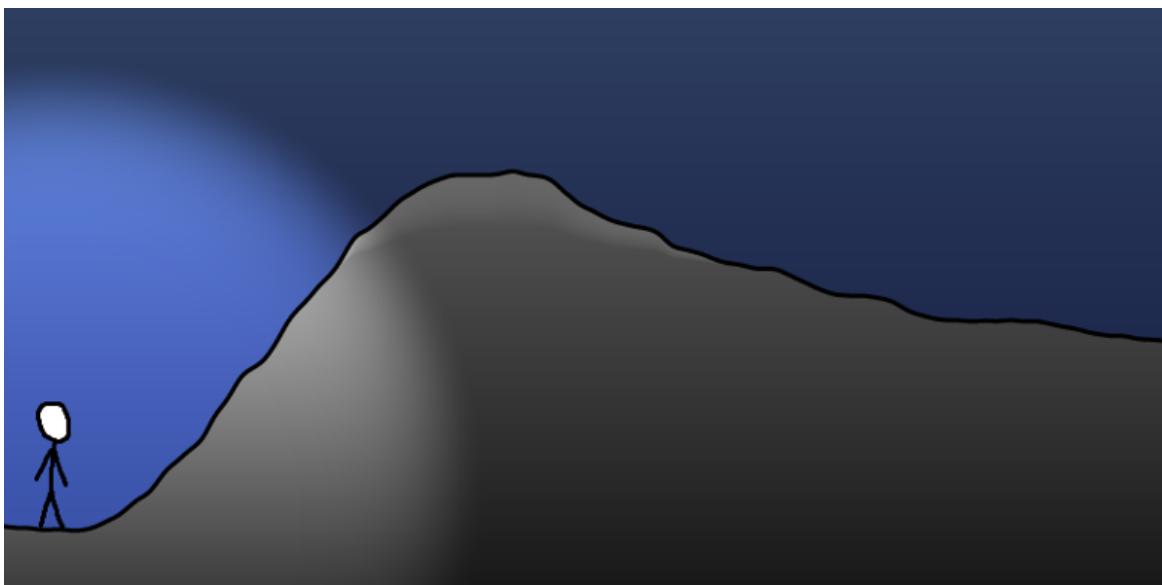
Short-Sightedness

One challenge we face when dealing with the difficulty landscape is that we usually don't see the complete terrain. We primarily experience the *current* level of difficulty, i.e. the slope of the landscape beneath our feet. We may see ahead a bit as we have an idea of what the next actions are as well as how these actions will affect our experience. Yet, extrapolating further into the future can be difficult, and our minds are lazy. Simply assuming the immediate trend we're currently experiencing extends as a straight line into the future on the other hand is easy and happens often enough. This tendency leads to two typical problems in the course of working on projects.



The first problem is that of falsely assuming a downward slope will continue indefinitely, often causing us to jump at that beautiful opportunity of gradient descent with maximum motivation, only to find ourselves crashing into a mountain shortly after. The ensuing disappointment leads to a sharp drop in motivation, possibly even resulting in losing interest in the project altogether.

This was also the prevalent pattern in my teenage years, when my main hobby was programming: Every other week I'd come up with that perfect idea for a video game, or some exciting tool I could develop, and having that amazing vision filled me to the brim with excitement. I would jump into it head first, work on it for a day or a week or two, but almost inevitably come to that point where it was clear that *this is difficult*. It took me a few years to figure this out and be a bit more far-sighted rather than blindly following every little exciting downward slope I could find.



The second problem is the exact opposite: seeing an obstacle ahead, and assuming this is what things will be like forever. This can be bad for at least three reasons. Firstly, the bleak outlook can reduce our enjoyment of the project, as this is not anymore a fun or exciting

endeavor but a difficult one, one that we need to *force* our way through in order to some day reach that goal we're looking for. Secondly, it may [lead to procrastination](#) and thus delay our arrival at the goal. And thirdly, this may even cause us to abandon the project, as we might reason that the goal just isn't worth that much projected effort.

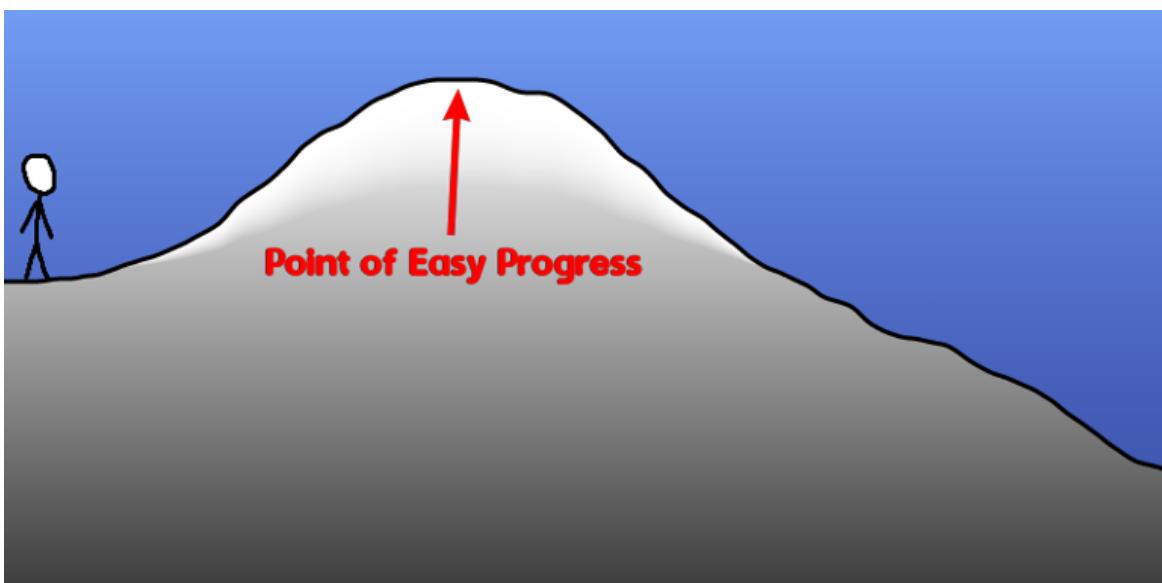
In summary, another aspect apart from the general approach that's important to improving both our enjoyment of working on a project and also its expected outcome, is being aware of our tendency to naively extrapolate, and doing our best not to base our decisions on such imperfect extrapolations.

The Point of Easy Progress

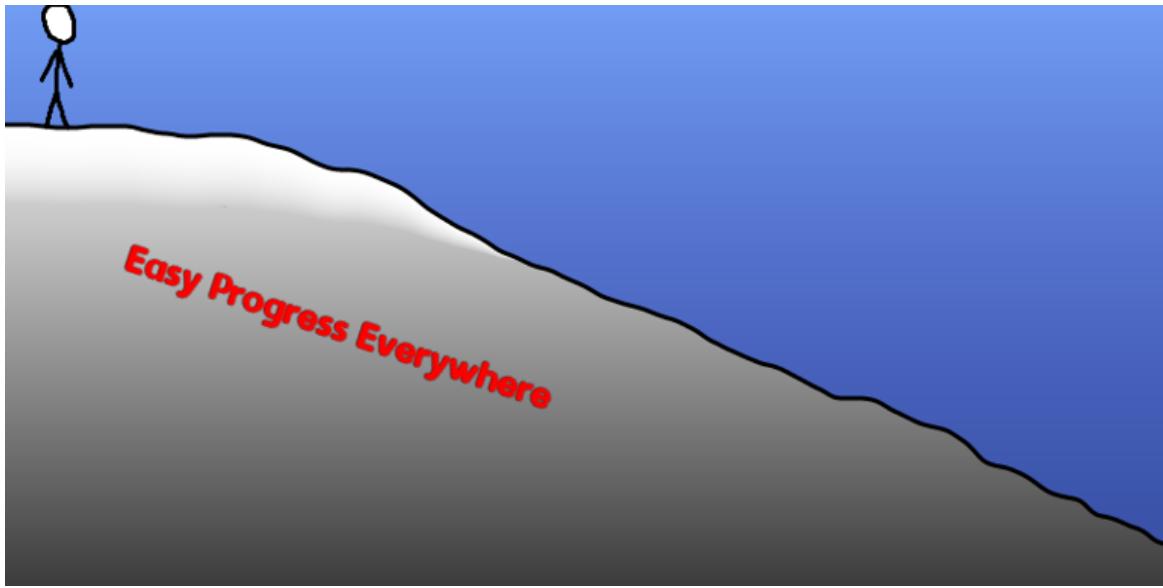
I argued that the approach to a project matters a lot. In the next few paragraphs I want to add to that the claim that for certain projects it may be possible to find an approach such that a "point of easy progress" is reached early on: once that point is reached, barely any willpower will be required to make progress. Working on the project will generally be a more attractive option than not doing so. In terms of the difficulty landscape, the gradient will be a friendly downward slope, meaning gravity and momentum will be on your side.

There are surely projects out there that don't allow this, but generally speaking it seems to me that it is true surprisingly often, and it's a good idea to actively look for a reachable point of easy progress in any project one takes on.

One example of such a point of easy progress was the author who starts by doing the necessary research and constructing a plotline so exciting that they can hardly wait to get it on paper. Generally speaking, if a point of easy progress can be identified, we may end up with a landscape somewhat like this:



There may even be projects that just have almost no negative aspects in the first place and are exciting from start to finish:



When looking for a point of easy progress in any particular project, the following steps may be helpful:

- Making a list of all types of tasks involved in the project, rating them by how “positive” (i.e. fun, motivating, exciting) or “negative” (boring, difficult, uncertain, risky) they are. What feelings arise when thinking of that type of task?
- For the negative tasks:
 - Are they actually necessary?
 - Can they be reduced in scope, or delegated?
 - Can they be made more fun or bearable?
 - Can they be moved to a more suitable point in time?
 - Is it possible to reframe them, or find positive aspects to them that make them seem more beneficial or acceptable?
 - Do they look [more daunting than they really are](#), and getting through them may turn out surprisingly easy?
- For the positive tasks:
 - Can they be amplified so that they make up more of the overall time spent on the project?
 - How can they be reached as early on as possible?

It seems generally useful to be very aware of what excites you in a project, and how to get more of that. And on the other side, sometimes the more negative parts of a project may originate from a [cached thought](#) or an implicit impression of how such a project *should* be approached, rather than what would actually work best for you.

When optimizing for such a point of easy progress, it may happen that projects start out with a “hill”. Work first, pleasure later. This is a feature, not a bug: knowing the peak awaits, climbing it may turn out much more enjoyable than it would without that positive expectation. This is why avoiding short-sightedness is especially beneficial in such easy progress scenarios. You don’t need to wait for *the end of the project* at some distant point in time, instead you only wait to reach that point of easy progress, as this can be seen as the project’s “event horizon”: once there, there’s no stopping, the project will be self-sustaining and almost inevitably lead to success.

Conclusion

This post is admittedly not the epitome of epistemics. The model and examples are simplistic, things are never that clean cut, easy or coherent in the real world, and you can find tons of exceptions to the rules proposed here, if they can even be considered rules in the first place.

The idea behind this post however is not to make a strong empirical claim, but rather to provide a model that may be useful to some; the idea of a point of easy progress being a possible state to reach in principle in any project seems to me like a useful framing. It entails putting motivation and enjoyment in a more visual context, helping by focusing on the meta level at a moment of high leverage (namely the very beginning of a project when making a lot of strategic decisions about it).

For me, the insights that ultimately led to this post made all the difference in a number of personal projects. One programming project in particular was lying dormant for years, until I recently realized that by reprioritizing things in a certain way I was only *days* away from a point of easy progress. Consequently, even these few days that I had considered to be an uphill battle turned out to fly by, as the prospect of finally reaching that point was so enticing that the more negative parts of the initial work faded far into the background.

If you have the impression that the concepts covered in this post can in any way be of use to you, I suggest picking any of your current projects or project ideas and going through the list from the previous section. If it leads you to any new insights, I dare you to share them in a comment.

Epistemological Framing for AI Alignment Research

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

Introduction

You open the Alignment Forum one day, and a new post stares at you. By sheer luck you have some time, so you actually read it. And then you ask yourself the eternal question: how does this fit with the rest of the field? If you're like me, your best guess comes from looking at the author and some keywords: this usually links the post with one of the various "schools" of AI Alignment. These tend to be affiliated with a specific researcher or lab -- there's Paul Christiano's kind of research, MIRI's embedded agency, and various other approaches and agendas. Yet this is a pretty weak understanding of the place of new research.

In other fields, for example [Complexity Theory](#), you don't really need to know who wrote the paper. It usually shows a result from one of a few types (lower bound, completeness for a class, algorithm,...), and your basic training in the field armed you with mental tools to interpret results of this type. You know the big picture of the field (defining and separating complexity classes), and how types of results are linked with it. Chances are that the authors themselves called on these mental tools to justify the value of their research.

In the words of Thomas S. Kuhn, Complexity Theory is paradigmatic and AI Alignment isn't. Paradigms, popularized in Kuhn's [The Structure of Scientific Revolutions](#), capture shared assumptions on theories, interesting problems, and evaluation of solutions. They are tremendously useful to foster normal science, the puzzle-solving activity of scientists; the paradigm carves out the puzzles. Being paradigmatic also makes it easier to distinguish what's considered valuable for the field and what isn't, as well as how it all fits together.

This list of benefit logically pushed multiple people to argue that we should make AI Alignment paradigmatic.

I disagree. Or to be more accurate, I agree that we should have paradigms in the field, but I think that they should be part of a bigger epistemological structure. Indeed, a naive search for a paradigm either results in a natural science-like paradigm, that put too little emphasis on applications and usefulness, or in a premature constraint on the problem we're trying to solve.

This post instead proposes a framing of AI Alignment research which has a place for paradigms, but isn't reduced to them. I start by stating this framing, along with multiple examples in each of its categories. I then go back to the two failure modes of naive paradigmatism I mentioned above. Finally, I detail how I intend to falsify the usefulness of this framing through a current project to review important AF posts.

Thanks to Joe Collman, Jérémie Perret, Evan Hubinger, Rohin Shah, Alex Turner and John S. Wentworth for feedback on this post.

The Framing

Let's start by asking ourselves the different sort of progress one could make in AI Alignment. I see three categories in broad strokes (I'll give examples in a minute).

- Defining the terms of the problem
- Exploring these definitions
- Solving the now well-defined problem

I expect the first and third to be quite intuitive -- define the problem and solve it. On the other hand, the second might feel redundant. If we defined the problem, the only thing left is to solve it, right?

Not in a world without [logical omniscience](#). Indeed, the definitions we're looking for in AI Alignment are merely structures and premises; they don't give all their consequences for free. Some work is needed to understand their implications.

Let's get slightly less abstract, and try to state the problem of AI Alignment: "Make AIs well-behaved". Here "AIs" and "well-behaved" are intentionally vague; they stand for "AI-related systems we will end up building" and "what we actually want them to do", respectively. So I'm just saying that AI Alignment aims to make the AIs we build do as we wish.

What happens when we try to carve research on this abstract problem along the three categories defined above?

- **Research on the “AIs” part**

- **(Defining)** Clarify what "AI-related systems we will end up building" means. This basically amounts to making a paradigm for studying the AIs we will most probably build in the future.

Note that such a paradigm is reminiscent of the ones in natural sciences, since it studies an actual physical phenomenon (the building of AIs and what they do, as it is done).

Examples include:

- Timelines research, like Daniel Kokotajlo's [posts](#)

- **(Exploring)** Assuming a paradigm (most probably deep learning these days), this is normal science done within this paradigm, that helps understanding aspects of it deemed relevant for AI Alignment.

Examples (in the paradigm of deep learning) include:

- Interpretability work, like [the circuit work](#) done by the Clarify team at OpenAI.
- Work on understanding how training works, like [this recent work](#) on SGD

- **Research on the “well-behaved” part**

- **(Defining)** Clarifying what "what we actually want them to do" means. So building a paradigm that makes clear what the end-goals of alignment are. In general, I expect a global shared paradigm here too, with individual researchers championing specific properties among all the ones promoted by the paradigm.

Note that such a paradigm is reminiscent of the ones in theoretical computer science, since it studies a philosophical abstraction in a formal or semi-formal way.

Examples include:

- [Defining Coherent Extrapolated Volition](#) as an abstraction of what we would truly want upon reflection.
 - [Defining HCH](#) as an abstraction of considered judgment
 - [Defining and arguing](#) about corrigibility
 - [Defining the properties expected of good embedded agents](#).
 - [Defining catastrophic consequences through attainable utility](#).
- **(Exploring)** Assuming a paradigm (or at least some part of the paradigm focused on a specific property), normal science done in extending and analyzing this property.
Examples include:
 - Assuming “well-behaved” includes following considered judgement, works on exploring HCH, like these [two posts](#).
 - Assuming “well-behaved” includes being a good embedded agent, works on exploring embedded agency, like the papers and posts referenced in the [Embedded Agency sequence](#).
- **(Solving)** Assuming a paradigm for “AIs” and a paradigm for “well-behaved”, research on actually solving the problem. This category is probably the most straightforward, as it includes most of what we intuitively expect in AI Alignment research: proposition for alignment schemes, impossibility results, critics of schemes, ...
Examples include:
 - Assuming “AIs” means “Deep Learning models for question answering” and “well-behaved” means “following HCH”, [IDA](#) is a proposed solution
 - Assuming “AIs” means “DeepRL systems” and “well-behaved” means “coherent with observed human behavior”, an impossibility result is the well-known [paper on Occam Razor’s and IRL](#) by Stuart Armstrong and Sören Mindermann.
 - Assuming “AIs” means “Embedded Agents” and “well-behaved” means “deals with logical uncertainty in a reasonable way”, [logical inductors](#) are a proposed solution.

Note that this framing points towards some of the same ideas that Rohin’s [threat models](#) (I wasn’t aware of them before Rohin’s pointer in an email). Basically, Rohin argues that a model on which to do AI Alignment research should include both a development model (what AI will look like) and a risk model (how it will fail). His issue with some previous work lies in only filling one of these models, and not both. In my framing, this amounts to requiring that work in the Solving category comes with both a model/paradigm of what “AIs” means and a model/paradigm of what “well-behaved” means. That fits with my framing. On the difference side, Rohin focuses on “what goes wrong” (his risk model), whereas I focus on “what we want”.

Going back to the framing, let’s be very clear on what I’m **not** saying.

I’m not saying that every post or paper falls within exactly one of these categories. The [Logical Induction paper](#) for example both defines a criterion for the part of “well-behaved” related to embedded logical uncertainty, but also provides logical inductors to show that it’s possible to satisfy it. Yet I think it’s generally easy to separate the different contributions to make clear what falls into which category. And I believe such explicit separation helps tremendously when learning the field.

I’m not saying that these categories are independent. It’s obvious that the “solution” category depends on the other two; but one can also argue that there are dependencies between studying what “AIs” means and studying what “well-behaved” means. For example, inner alignment only really makes sense in a setting where AIs

are learned models through some sort of local optimization process -- hence this part of “well-behaved” requires a specific form to the definition of “AIs”. This isn’t really a problem, though.

I’m not saying that every post or paper falls within at least one category.

Some work that we count as AI Alignment don’t really fall in any of my categories. The foremost example that I have in mind is John’s research on [Abstraction](#). In a way, that is expected: this research is of a more general idea. It impacts some categories (like what “well-behaved” means), but is more a fundamental building block. Still, pointing to the categories that this research applies might help make it feel more relevant to AI Alignment.

I’m not saying that we need to fully solve what we mean by “AIs” and “well-behaved” before working on solutions.

Of course work on solutions can already proceed quite usefully. What I’m arguing for instead is that basically any work on solutions assumes (implicitly or explicitly) some sort of partial answer to what “AIs” and “well-behaved” means. And that by stating it out loud, the authors would help the understanding of their work within the field.

I’m not saying that this is the only reasonable and meaningful framing of AI Alignment research.

Obviously, this is but one way to categorize the research. We already saw that it isn’t as clean as we might want. Nonetheless, I’m convinced that using it will help make the field clearer to current researchers and newcomers alike.

In essence, this framing serves as a lens on the field. I believe that using it systematically (as readers when interpreting a work and as author when presenting our work) would help quite a lot, but that doesn’t mean it should be the only lens ever used.

Why not a single paradigm?

I promised in the introduction that I would explain why I believe my framing is more adequate than a single paradigm. This is because I only see two straightforward ways of compressing AI Alignment into a single paradigm: make it a paradigm about a fundamental abstraction (like agency) that once completely understood should make a solution obvious; or make it a paradigm about a definition of the problem (what “AIs” and “well-behaved” means). Both come with issues that make them undesirable.

Abstraction Paradigm

Paradigms historically come from natural sciences, as perspectives or explanations of phenomena such as electricity. A paradigm provides an underlying theory about the phenomenon, expresses the well-defined questions one can ask about it, and what would count as a successful solution of these questions.

We can also find paradigms about abstractions, for example in theoretical computer science. The current paradigm about computability is captured by [the Church-Turing thesis](#), which claims that everything that can be physically computed can be computed by a [Turing Machine](#). The “explanation” for what computation means is the Turing Machine, and all its equivalent models. Hence studying computability within this paradigm hinges on studying what Turing Machines can compute, as well as other models equivalent to TMs or weaker (This overlooks the sort of research done by

mathematicians studying recursion theory, like [Turing degrees](#); but as far as I know, these are of limited interest to theoretical computer scientists).

So a paradigm makes a lot of sense when applied to the study of a phenomenon or an abstraction. Now, AI Alignment is neither; it's instead the search for the solution of a specific problem. But natural sciences and computer science have been historically pretty good at providing tools that make solving complex problems straightforward. Why couldn't the same be true for AI Alignment?

Let's look at a potential candidate. An abstraction presented as the key to AI Alignment by multiple people is agency. According to this view, if we had a complete understanding of agency, we wouldn't find the problem of aligning AI difficult anymore. Thus maybe a paradigm giving an explanation of agency, and laying out the main puzzles following from this explanation, would be a good paradigm of AI Alignment.

Despite agreeing with the value of such work, I disagree with the legitimacy of making it the sole paradigm of AI Alignment. Even if understanding completely something like agency would basically solve the problem, how long will it take (if it is ever reached)? Historical examples in both natural sciences and computer science show that the original paradigm of a field isn't usually adapted to tackle questions deemed fundamental by later paradigms. And this progress of paradigms takes decades in the best of cases, and centuries in the worst!

With the risk of short timelines, we can't reasonably decide that this is the only basket to put our research eggs.

That being said, this paradigmatic approach has a place in my framing, about what "well-behaved" means. The difference is that once a paradigm is chosen, work can proceed in it while other researchers attempt to solve the problem for the current paradigm. There's thus a back and forth between the work within the paradigm and its main application.

Problem Paradigm

If we stretch a bit the term, we can call paradigm the assumptions about what "AIs" and "well-behaved". Then becoming paradigmatic would mean fixing the assumption and forcing all the work to go within this context.

That would be great, if only we could already be sure about what assumptions to use. But in the current state of the field, a lot more work is needed (especially for the "well-behaved" part) before anyone can reasonably decide to focus all research on a single such paradigm.

This form of paradigm thus suffers from the opposite problems than the previous one: it fails to value the research on the term of the problems, just to have a well-defined setting on which to make progress. Progress towards what? Who knows...

Here too, this approach has a place in my framing. Specifically, every work on the Solving category exists within such a paradigm. The difference is that I allow multiple paradigms to coexist, as well as the research on the assumptions behind this paradigm, allowing a saner epistemological process.

Where do we go from here?

Multiple voices in AI Alignment push for making the field more paradigmatic. I argue that doing this naively isn't what we want: it either removes the push towards application and solutions, or fixes the term of the problem even though we are still so uncertain. I propose instead that we should think about research according to different parts of the statement "Make AIs well-behaved": research about what "AIs" we're talking about, research on what we mean by "well-behaved", and based on answers to the two previous questions, actually try to solve the clarified problem.

I believe I argued reasonably enough for you to not dismiss the idea immediately. Nonetheless, this post is hardly sufficient to show the value of adopting this framing at the level of the whole research community.

One way I hope to falsify this proposition is through a project to review many posts on the AF to see what makes a good review, done with Joe Collman and Jérémie Perret. We plan on trying to use this lens when doing the reviews, to see if it clarifies anything. Such an experiment thus relies on us reviewing both posts that fit quite well the framing, and ones that don't. If you have any recommendation, I wrote [a post](#) some time ago where you can give suggestions for the review.

Generalizing POWER to multi-agent games

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

Acknowledgements:

This article is a writeup of a research project conducted through the [SERI](#) program under the mentorship of [Alex Turner](#). I ([Jacob Stavrianos](#)) would like to thank Alex for turning a messy collection of ideas into legitimate research, as well as the wonderful researchers at SERI for guiding the project and putting me in touch with the broader X-risk community.

Motivation/Overview

In the single-agent setting, [Seeking Power is Often Robustly Instrumental in MDPs](#) showed that optimal policies tend to choose actions which pursue "power" (reasonably formalized). In the multi-agent setting, the [Catastrophic Convergence Conjecture](#) presented intuitions that "most agents" will "fight over resources" when they get "sufficiently advanced." However, it wasn't clear how to formalize that intuition.

This post synthesizes single-agent power dynamics (which we believe is now somewhat well-understood in the MDP setting) with the multi-agent setting. The multi-agent setting is important for AI alignment, since we want to reason clearly about when AI agents disempower humans. Assuming constant-sum games (i.e. maximal misalignment between agents), this post presents a result which echoes the intuitions in the Catastrophic Convergence Conjecture post: as agents become "more advanced", "power" becomes increasingly scarce & constant-sum.

An illustrative example

You're working on a project with a team of your peers. In particular, your actions affect the final deliverable, but so do those of your teammates. Say that each member of the team (including you) has some goal for the deliverable, which we can express as a reward function over the set of outcomes. How well (in terms of your reward function) can you expect to do?

It depends on your teammates' actions. Let's first ask "given my opponent's actions, what's the highest expected reward I can attain?"

Case 1: Everyone plays nice

We can start by imagining the case where everyone does exactly what you'd want them to do. Mathematically, this allows you to obtain the globally maximal reward; or "the best possible reward assuming you can choose everyone else's actions". Intuitively, this looks like your team sitting you down for a meeting, asking what you

want them to do for the project, and carrying out orders without fail. As expected, this case is 'the best you can hope for' in a formal sense.

Case 2: Everyone plays mean

Now, imagine the case where everyone does exactly what you *don't* want them to do. Mathematically, this is the worst possible case; every other choice of teammates' actions is at least as good as this one. Intuitively, this case is pretty terrible for you. Imagine the previous case, but instead of following orders your team actively sabotages them. Alternatively, imagine that your team spends the meeting breaking your knees and your laptop.

Case 3: Somewhere in between

However, scenarios where your team is perfectly aligned either with or against you are rare. More typically, we model people as maximizing their own reward, with imperfect correlation between reward functions. Interpreting our example as a multi-player game, we can consider the case where the players' strategies form a Nash equilibrium: every person's action is optimal for themselves given the actions of the rest of their team. This case is both relatively general and structured enough to make claims about; we will use it as a guiding example for the formalism below.

POWER, and why it matters

Many attempts have been made to classify AI [robustly instrumental goals](#), with the goals of understanding why they emerge given seemingly-unrelated utilities and ultimately to counterbalance (either implicitly or explicitly) undesirable robust instrumental subgoals. [One promising such attempt](#) is based on POWER (the technical term is all-caps to distinguish from normal use of the word): consider an agent with some space of actions, which receives rewards depending on the chosen actions (formally, an agent in an MDP). Then, POWER is roughly "ability to achieve a wide variety of goals". [It's been shown](#) that POWER is robustly instrumental given certain conditions on the environment, but currently no formalism exists describing power of different agents interacting with each other.

Since we'll be working with POWER for the rest of this post, we need a solid definition to build off of. We present a simplified version of the original definition:

Consider a scenario in which an agent has a set of actions $a \in A$ and a distribution D of reward functions $r : A \rightarrow R$. Then, we define the POWER of that agent as

$$\text{POWER}_D := E_{r \sim D} [\max_a r(a)]$$

As an example, we can rewrite the project example from earlier in terms of POWER. Let your goal for the project be chosen from some distribution D (maybe you want it done nicely, or fast, or to feature some cool thing that you did, etc). Then, your

POWER_D is the maximum extent to which you can accomplish that goal, in expectation.

However, this model of power can't account for the actions of other agents in the environment (what about what your teammates do? Didn't we already show that it matters a lot?). To say more about the example, we'll need a generalization of POWER.

Multi-agent POWER

We now consider a more realistic scenario: not only are you an agent with a notion of reward and POWER, but so is everyone else, all playing the same multiplayer game. We can even revisit the project example and go through the cases for your teammates' actions in terms of POWER:

- In Case 1, your team works to maximize your reward in every case, which (with some assumptions) maximizes your POWER over the space of all choices of teammate actions.
- In Case 2, your team works to *minimize* your reward in every case, which analogously minimizes your POWER.
- In case 3, we have a Nash equilibrium of the game used to define multi-agent POWER. In particular, each player's action is a best-response to the actions of every other player. We'll see a parallel between this best-response property and the $\max_{a \in A}$ term in the definition of POWER pop up in the discussion of constant-sum games.

Bayesian games

To extend our formal definition of power to the multi-agent case, we'll need to define a type of multiplayer normal-form game called a [Bayesian game](#). We describe them below:

- At the beginning of the game, each of n players is assigned a type $t_i \in T_i$ from a joint type distribution $t = (t_i) \sim \Omega$. The distribution Ω is common knowledge.
- The players then (independently, **not** sequentially) choose actions $a_i \in A_i$, resulting in an *action profile* $a = (a_i)$.
- Player i then receives reward $r_i(t_i, a)$ (crucially, a player's reward can depend on their type).

Strategies (technically, mixed strategies) in a Bayesian game are given by functions $\sigma_i : T_i \rightarrow \Delta A_i$. Thus, even given a fixed strategy profile σ , any notion of "expected reward of an action" will have to account for uncertainty in other players' types. We do so by defining *interim expected utility* for player i as follows:

$$f_i(t_i, a_i, \sigma_{-i}) := E[r_i(t_i, a)]$$

where the expectation is taken over the following:

- the posterior distribution over opponents' types $t_{-i}|t_i$ - in other words, what types you expect other players to have, given your type.
- random choice of opponents' actions $a_{-i} \sim \sigma_{-i}(t_{-i})$ - even if you know someone's type, they might implement a mixed strategy which stochastically selects actions.

Further, we can define a (Bayesian) Nash Equilibrium to be a strategy profile where each player's strategy is a best response to opponents' strategies in terms of interim expected utility.

Formal definition of multi-agent POWER

We can now define POWER in terms of a Bayesian game:

Fix a strategy profile σ . We define player i 's POWER as

$$\text{POWER}(i, \sigma) := E_{t_i} \max_{a_i} f_i(t_i, a_i, \sigma_{-i})$$

Intuitively, POWER is maximum (expected) reward given a distribution of possible goals. The difference from the single-agent case is that your reward is now influenced by other players' actions (by taking an expectation over opponents' strategy).

Properties of constant-sum games

As both a preliminary result and a reference point for intuition, we consider the special case of zero-sum games:

A zero-sum game is a game in which for every possible outcome of the game, the sum of each player's reward is zero. For Bayesian games, this means that for all type profiles $t = (t_i)$ and action profiles a , we have $\sum_i r_i(t_i, a) = 0$. Similarly, a *constant-sum game* is a game satisfying $\sum_i r_i(t_i, a) = c$ for any choices of t, a .

As a simple example, consider chess; a two-player adversarial game. We let the reward profile be constant, given by "1 if you win, -1 if you lose" (assume black wins in a tie). This game is clearly zero-sum, since exactly one player will win and lose. We could ask the same "how well can you do?" question as before, but the upper-bound of winning is trivial. Instead, we ask "how well can both players simultaneously do?"

Clearly, you can't both simultaneously win. However, we can imagine scenarios where both players have the *power to win*: in a chess game between two beginners, the optimal strategy for either player will easily win the game. As it turns out, this argument generalizes (we'll even prove it): in a constant-sum game, the sum of each

player's POWER $\geq c$, with equality iff each player responds optimally for all their possible goals ("types"). This condition is equivalent to a Bayesian Nash Equilibrium of the game.

Importantly, this idea suggests a general principle of multi-agent POWER I'll call *power-scarcity*: in multi-agent games, gaining POWER tends to come at the expense of another player losing POWER. Future research will focus on understanding this phenomenon further and relating it to "how aligned the agents are" in terms of their reward functions.

Claim: Consider a Bayesian constant-sum game with some strategy profile σ

. Then, $\sum_i \text{POWER}(i, \sigma) \geq c$ with equality iff σ is a Nash Equilibrium.

Intuition: By definition, σ isn't a Nash Equilibrium iff some player i's strategy σ_i isn't a best response. In this case, we see that player i has the power to play optimally, but the other players also have the power to capitalize off of player i's mistake (since the game is constant-sum). Thus, the lost reward is "double-counted" in terms of POWER; if no such double-counting exists, then the sum of POWER is just the expected sum of reward, which is c by definition of a constant-sum game.

Rigorous proof:

We prove the following for general strategy profiles σ :

$$\begin{aligned}
\sum_i \text{Power}(i, \sigma) &= \sum_i E_{t_i} \max_{a_i} f_i(t_i, a_i, \sigma_{-i}) \\
&\geq \sum_i E_{t_i} E_{a_i \sim \sigma_i} f_i(t_i, a_i, \sigma_{-i}) \\
&= \sum_i E_{t_i} E_{a \sim \sigma} r_i(t_i, a) \\
&= E_t E_{a \sim \sigma} (\sum_i r_i(t_i, a)) \\
&= E_t E_{a \sim \sigma} (c) \\
&= c
\end{aligned}$$

Now, we claim that the inequality on line 2 is an equality iff σ is a Nash Equilibrium. To see this, note that for each i , we have

$$\max_{a_i} f_i(t_i, a_i, \sigma_{-i}) \geq E_{a_i \sim \sigma_i} f_i(t_i, a_i, \sigma_{-i})$$

with equality iff σ_i is a best response to σ_{-i} . Thus, the sum of these inequalities for each player is an equality iff each σ_i is a best response, which is the definition of a Nash Equilibrium. \square

Final notes

To wrap up, I'll elaborate on the implications of this theorem, as well as some areas of further exploration on power-scarcity:

- It initially seems unintuitive that as players' strategies improve, their collective POWER tends to decrease. The proximate cause of this effect is something like "as your strategy improves, other players lose the power to capitalize off of your mistakes". More work is probably needed to get a clearer picture of this dynamic.
- We suspect that if all players have identical rewards, then the sum of POWER is equal to the sum of best-case POWER for each player. This gives the appearance of a spectrum with [aligned rewards (common payoff), maximal sum power] on one end and [anti-aligned rewards (constant-sum), constant sum power] on the other. Further research might look into an interpolation between these two extremes, possibly characterized by a correlation metric between reward functions.
 - We also plan to generalize POWER to Bayesian stochastic games to account for sequential decision making. Thus, any such metric for comparing reward functions would have to be consistent with such a generalization.
- POWER-scarcity results in terms of Nash Equilibria suggest the following dynamic: as agents get smarter and take available opportunities, POWER becomes increasingly scarce. This matches the intuitions presented in [the Catastrophic Convergence Conjecture](#), where agents don't fight over resources until they get sufficiently "advanced."

Clubhouse

A friend showed me Clubhouse on her iPhone eleven days ago on March 3, 2021. It was the first time I had heard about the app. Five minutes with the app was enough to convince me it was positioned to become the next big media platform.

That same evening I borrowed the same friend's tertiary iPhone so I could use the app myself. (I run Android.) Over the next few days I invited two real world friends to collaborate on establishing a Clubhouse show. My friends were skeptical but open-minded. I convinced both of them Clubhouse was worth a bet.

This post has three parts.

1. In **Context** I tell stories which inform my evaluation of Clubhouse.
2. In **Current State** I explain what Clubhouse is and how it works.
3. In **Betting on a Future** I explain what I plan to do with my prediction.

Context

Conversation-style Podcasts

One of the most popular podcasts is the world is *The Joe Rogan Experience*. The 53-year-old former mixed martial artist's episodes regularly exceed three hours. The podcast's logo features a scary-looking man with a crazed grin and an eye on his forehead.

Many podcasts are interviews. Joe Rogan's interviews are exceptionally popular because they sound like real conversations. If your podcast is limited to one hour per episode then you can never get more than one hour deep into a topic. Podcasts interviewing authors are often just the author pitching his or her book. They don't have time for anything more. Listening to the same author on different podcasts is redundant... unless one of those podcasts is *The Joe Rogan Experience*.

Another exceptionally popular podcast is *Trash Taste*. *Trash Taste* is three anime YouTubers sitting around a table talking about whatever they feel like. *Trash Taste* frequently (but not always) invites guests onto the channel. *Trash Taste* is popular because it feels like hanging out with Joey, Connor and Garnt.

There is listener demand for informal unscripted conversation.

Guest Speaking for Middle Schools

Over the last few months I was a guest speaker at middle schools where I gave presentations on science, entrepreneurship and design. We used Zoom. We started with a presentation where I explained who I was. The bulk of the time was spent in Q&A. Students raised their hands. The teacher selected one student to speak at a time. The student asked a question and I answered it. Sometimes we had a short back-and-forth.

The introduction was noninteractive. It could have been a YouTube video. (In fact, by the end I *did* just record a YouTube video.) What made the day special is students could ask whatever they wanted to a real specialist and then get real answers in real time

It wasn't just the literal answer that mattered. The attitude and personality I conveyed were even more important. Children emulate more than they listen.

Gather Town

The first Gather Town event I attended was the March 7, 2021 Less Wrong Garden Party 2.0. It **felt like actually being at a party** with Raemon, Habryka and Daniel Kokotajlo. It was just like Wade Watts hideout in *Ready Player One*.

Several years ago I attended a real party. Someone made fun at me for sitting in an empty room in the dark on my computer instead of joining into the drinking and dancing. An hour later the room was still dark but I was lecturing on command line tools to a crowd of people. I had commandeered a college party into an expert panel.

A party full of techies is called an "industry conference". Gather Town events tend to stabilize into two people talking while several others listen in. The social dynamic mirrors the expert panels of an industry conference.

The two people talking tend to be the two who understand the subject matter the best. The experts get to talk about what interests them. The non-experts get to observe how experts talk to each other. Watching experts talk among themselves is REALLY valuable to non-experts. The most natural way to learn something is not reading or listening. The most natural way to learn is by EMULATING the people you look up to.

Current State

Clubhouse is organized into "rooms". Everyone in the world can listen in to a room but only the room's owner and friends can talk^[1]. If you are listening in then you can raise your hand to request the opportunity to speak. Speaking is a privilege. The room's owners are under no obligation to let you speak.

This system lets people lurk for a while and then drop into a conversation like at a party. So does Gather Town. Unlike Gather Town, Clubhouse's system scales. You can put a thousand people into the same room and the social structure won't break down because software broadcasts the conversation to everyone while enforcing who can and cannot speak.

Clubhouse is a work in progress. It has a small userbase. It has an undeveloped ecosystem of creators. Its search functionality exists somewhere between bad and broken. It only runs on iPhone.

I have already had two great experiences with the app.

- To learn a foreign language you must absorb many hours of people talking. It is hard to find realistic conversations. Stage personas on radio and television talk differently from ordinary people in real life. When you *do* find a real life

conversation in your target language, the speakers often switch into English. Ordinary conversations aren't recorded because recording a conversation makes it unusual. On Clubhouse **I can listen in on authentic foreign language conversations whenever I want.**

- When I was convincing my first friend to try out Clubhouse I did it over Clubhouse. During the call, his brother and a mutual friend both dropped in to listen. Neither of them said much but they *did* contribute to the conversation. The mutual friend explicitly said that he learned a lot from listening to us talk.

The system works even though it has a small userbase. The developers of Clubhouse have solved the chicken-and-egg problem.

Betting on a Future

I believe Clubhouse has the potential^[2] to become a major communication platform competing with Facebook, YouTube, [twitch.tv](#), Twitter, reddit, discord, etc. YouTube had a significant first mover advantage. The Vlog Brothers are famous partially because they create great videos but *mostly* because they created lots of okay videos early in YouTube's history.

An online following is a valuable asset. The disadvantages of an online following is it can be hard to acquire and hard to maintain.

- If Clubhouse has a first mover advantage like YouTube then there will never be an easier time to acquire a Clubhouse audience.
- Maintaining an audience mostly comes down to consistently publishing new content. Clubhouse content is cheap to produce because there is no editing. One hour of conversation is one hour of content. Talking on Clubhouse costs little if you were already going to have the conversation anyway. If this is true for you then a Clubhouse audience may cost less in maintenance than audiences elsewhere.

The pitch I sold my friends is **at best we'll build an online following** and **at worst we'll have spent a few extra hours having fun talking to each other online.**

Scheduling

~~If you're interested in talking with me on Clubhouse on a topic related to something I've written then click here.~~ [Edit: form closed.] If you're starting your own Clubhouse room and would like me as a guest, then private message me instead.

-
1. There are other features and options. This post outlines how I interpret the core value proposition. [↩](#)
 2. I'm not saying Clubhouse *will* become a major communication platform. Just that the odds are good enough (>10%) that it's worth placing a low risk high reward bet. [↩](#)

The Flexibility of Abstract Concepts

I can discuss Daoist ideas with Taiwanese friends easily even if they have no background in Daoism. But when I try the same thing with white people it often feels as if I'm trying to explain quantum field theory to someone who has never heard of mathematics. It is easier for me to discuss Zen with a Taiwanese atheist than a Western psychonaut.

In his [book review](#) on *The Geography of Thought*, PeterMcCluskey draws attention to differences between Westerners' and East Asians' ways of thinking. This post elaborates on one specific difference: the flexibility of abstract concepts.

Western people are trained to think terms of universal principles. East Asians are trained to think contextually.

The Dao [map] is not the Dao [territory]

East Asians are conditioned from an early age to understand that the map is not the territory. Saying "the map is not the territory" would be like saying "the sky is up", "money is valuable" or "don't kick kittens". The distinction between map and territory has been understood for literally thousands of years.

道可道也，非恒道也。

The "dao" referred to by "the dao" [map] is not the dao [territory].

— 《道德经》 *by Laozi 老子, 4th century BC*

Of course the map is not the territory. How could anyone who comprehends the concept of lying possibly confuse the two? And yet, I was talking to an American a couple months ago who literally did not believe me when I tried to explain how the word "infinity" has different meanings in different contexts.

Western Rhetoric

Western society has a long tradition of rhetoric where you debate the truth of statements like "murder is bad" or "the Greens should win the next election". Practically every grade school [essay](#) states a claim and then defends it. "A theme in *The Great Gatsby* is...".

American rhetoric reaches its purest form in the Lincoln-Douglas (LD) debate format. In the LD debate format two competitors debate a resolution like "Resolved: The United States ought to guarantee universal child care."^[1] One side is debates in favor. The other side debates against.

Resolutions never center around objective facts. (That would be a policy debate.) Instead, they come down to questions of value. Each debater defines victory in terms

of a value criterion. A value is something universally agreed to be good like "justice". The criterion is a method of measuring the value.

Throughout the entire process it is implied that *if* you specify a value and *if* you specify a criterion then the resolution has a truth value between zero and one (inclusive). Except that's almost never the case because [words don't have well-defined meanings](#).

Consider a simpler resolution: "Resolved: Murder is immoral."

[Some people consider murder to be immoral](#)^[2]. But murder is just the killing of another person in violation of the law. There are lots of cases where murder is moral. You can start by shooting the guards at Dachau.

Abstract statements tend to be broad. Broad statements tend to have exceptions. When a blanket statement has lots of exceptions it is said to "depend on context". By training children in the tradition of adversarial competitive rhetoric, **Western society trains its population to ignore context** because **in a debate, the map really is the territory**. Americans even think of ourselves as context-independent personalities.

"Tell me about yourself" seems a straightforward enough question to ask of someone, but the kind of answer you get very much depends on what society you ask it in. North Americans will tell you about their personality traits ("friendly, hard-working"), role categories ("teacher," "I work for a company that makes microchips"), and activities ("I go camping a lot"). Americans don't condition their self-descriptions much on context. The Chinese, Japanese, and Korean self, on the other hand, very much depends on context ("I am serious at work"; "I am fun-loving with my friends"). A study asking Japanese and Americans to describe themselves either in particular contexts or without specifying a particular kind of situation showed that Japanese found it very difficult to describe themselves without specifying a particular kind of situation—at work, at home, with friends, etc. Americans, in contrast, tended to be stumped when the investigator specified a context—"I am what I am." When describing themselves, Asians make reference to social roles ("I am Joan's friend") to a much greater extent than Americans do. Another study found that twice as many Japanese as American self-descriptions referred to other people ("I cook dinner with my sister").

Quote from *The Geography of Thought* in a [comment](#) by Kaj_Sotala

Post-Modernism

Western philosophy's reaction to taking words too seriously was the Post-Modernist movement. The Post-Modernists improved Western philosophy by throwing out the map. They damaged Western society by throwing out the territory too.

The more labels you have for yourself, the dumber they make you.

—[Keep your identity small](#) by Paul Graham

No, no, NO. This is backwards. The mistake you should "keep your identity small" stems from the erroneous assumption that identities are well-defined. It confuses the

label with the underlying reality. A small identity merely does no harm. You can do better than that. The best approach is **strong beliefs loosely held**. If you can shed your identities like you shed clothes then you can keep your identity large without mistaking your identity for your self. You can get the best of all the worlds.

The real lesson here is that the concepts we use in everyday life are fuzzy.... Even a concept as dear to us as I. It took me a while to grasp this, but when I did it was fairly sudden, like someone in the nineteenth century grasping evolution and realizing the story of creation they'd been told as a child was all wrong.... Everyday words are inherently imprecise.

—[How to Do Philosophy](#) by Paul Graham

That's better. And it illustrates how the flexibility of abstract concepts is not hammered into every child in the West until it becomes second nature. If you grow up in East Asia then the first, last and most important thing you are taught is how to blend in.

Context Switching

There's an old Daoist teaching technique where you say something like [pain is not the unit of effort](#) and then say the opposite like [pain is the unit of effort](#). The Western response is to figure out which one is true. The Eastern response is to quickly shift contexts because each statement is true in the appropriate context *ala* [Chapter 1 of The Art of War](#).

Consider race. Race, like all abstract concepts, is flexible and context-dependent. People have mistaken me for Indian, Japanese, Chinese and Ethiopian. I don't mean I told them I was x and they didn't argue. I don't mean I walked around Japan without anyone noticing me. I mean an Ethiopian, unprompted, literally asked me "Are you Ethiopian?" while both of us stood on American soil and then, when I answered no, he asked if my family was Ethiopian. I've been asked "Are you a Muslim?" *in Tokyo*. My race is a function of where I am, who I'm with, who I'm talking to, my language, my accent, my clothing, my posture...and sometimes even the color of my skin.

There is a secret game Asian-Americans play among ourselves called the "What kind of Asian are you?" game. Whenever an Asian-American meets another Asian-American we try to guess each other's nationality. If you guess right you gain charisma points. If you guess wrong you lose charisma points. Of course, you don't literally say "I know you are a <whatever>." That is a *faux pas*. Instead you imply it by demonstrating common cultural understandings *not shared* by the wider Western world.

What makes this game interesting is you can't do it by physical appearance—national boundaries aren't drawn phenotypographically. Nor can you do it from accent. You have to read subtle cultural cues. For example, I like roleplaying a Chinese nationalist when I'm online—nevermind that my family is from the Republic of China^[3].

When I want to look white I use words like "Manuchuria"^[4].

Be the grey man.

1. This topic is the 2021 March/April Topic of the [National Speech & Debate Association](#) ↵
2. Thank you MaxG for granting me permission to link to your post. ↵
3. This sentence is a joke about 20th century East Asian history. ↵
4. This sentence is another joke about 20th century East Asian history. ↵

Bureaucracy is a world of magic

I [previously wrote](#) about some practical game-theoretical (game-practical?) realizations I had while buying a house. Today I want to talk about how bureaucracy is a ritualistic, magical place.

In our home-buying process, every step of the way, there were papers to be signed. Paperwork is how the magic of bureaucracy comes in view. I'm not saying "magic" to mean good or beautiful. I'm referring to the ritualistic nature of bureaucracy.

Everything in our journey was a ritual. When you debate the *point* of something, people participating in the ritual are confused. On the one hand, they understand that your request makes sense, because you're asking for the same function. On the other hand, you shall not ignore the Ritual!

Let me explain with several examples what I mean by ritual.

The Summoning (of the PDF)

To buy a house and get state subsidies, you have to present an official document to the bank, confirming that the building may indeed be used as a dwelling, i.e. a *use permit*. It is not necessary that this document is an original, a copy will suffice.

Well, I got to the bank with printouts of photos of this permit. I don't have the original, and the agent simply took photos of it with his phone, and sent these photos to me. I printed them out on paper, and presented them to the bank. Problem: they have to be scans, not photos. "Photos aren't scans", the bank lady said, "They won't be accepted as official". My first impulse was to protest: "But since you don't need originals, what does it matter what form the copy has? Obviously the informational content is what's necessary - what's written in the document, not what device was used to transfer this information. And anyway, scans and photos are **literally the exact same thing**". Scans are just photos taken in a particular way. How is it important that-", but I stopped myself before saying any of this. There's a particular art to navigating bureaucracy, and arguing about the nature of information and how it represented is Not It, Chief ®. Instead, the Art is to constantly weigh where you can insist on being reasonable, and where you have to suck it up and comply with a dumb request.

What the bank lady *actually* wanted is a **semblance of officiality**. Photos simply don't look official, and that's it. To complete the ritual, a conventional way is required, and the most modern of the conventional ways is the offering of a scan. I downloaded the Adobe Scan app, "scanned" the JPEGs, made them look like they were actual scans from a scanning machine, told the lady that I just got the scans (implying that I got them from the agent, not from an app), and sent them via email. She was satisfied. Ritual complete.

The Notary of the Toilet

One of the steps was to notarize a document stating that we don't currently own any real estate. To do so, we went to a notary. My girlfriend knew of one in a nearby mall, so we went there. I'm angry at myself that I didn't take a photo, but I'll try to describe

it. So you come into this mall, and there are all these stores, with clothing, tech, sports equipment, food - just the regular stuff you'd expect in a mall. To get to the notary, you go through one of the service doors - those things that hide the inner workings of a mall, the mall's guts. You open that door, and you smell and you *hear* the toilets as they're being flushed. If you don't already know that you're going to see a notary, you'd think you've just walked into a toilet. So you walk through the toilet a bit, and at the end of the hallway, there's a door to the notary. The inside office is actually surprisingly well-furnished, but the outside is a mall favela.

We get in there, we present our ID cards, we sign a statement, the notary stamps it, and then we literally sign our names into a Big Book. The notary didn't verify my statement. She just verified that I signed it. Actually, she didn't do that, because I had a face mask on. So I could have come with anybody's ID card and produced any sort of statement, and it would have been notarized. A weirdly archaic industry, but it still lives because rituals aren't easy to replace.

But what *is* a signature?

All this reminds me of *Pact* by John C. McCrae (Wildbow). The main character there finds out about the world of magic, but it turns out that magic is magic only if the surrounding spirits and other practitioners of magic recognize it as magic. In other words, if you do unconventional stuff that doesn't *look* magic, it's not magic. There's no mechanism that you can game because the mechanism is the look; the form is the content.

Bureaucracy is a world of magic. Things are official if they look official. The more official-looking papers you collect, the stronger the spell. You want to do something that's functionally identical? Tough luck. It has to *look* the part. For years, this annoyed me. And it still does, but I've come to accept it as a price of doing things I want to do. I am glad that there are people out there building alternative, trustless systems. But until these systems take over, it's Real Wizard Hours.

Covid 3/18: An Expected Quantity of Blood Clots

This week's Covid news was that most of Europe suspended administration of the AstraZeneca vaccine over reports of blood clots. This was ludicrously stupid several times over. There was always going to be *something* that happened to correlate with vaccination days to *some extent*, somewhere, over some time period. The number of blood clots experienced after vaccination wasn't even higher than the base rate you would otherwise expect. And even if all the observed clots were extra, all were caused by the vaccine, all were fatal, and that represented the overall base rate, and we ignore all population-level benefits and economic issues, *the vaccine would still be worth using purely for personal health and safety by multiple orders of magnitude.*

The WHO and EMA said there was no evidence there was an issue.

None of that mattered, as one by one countries suspended injections as part of a blame avoidance strategy. As a result vaccinations are held, thousands (or more) will die as a direct result, with many European countries seeing things getting worse rather than better and facing possible new restrictions, and with a permanent new weapon in the arsenal of vaccine skeptics that we'll have to hear about for decades, long after this is proven to be a non-concern.

Meanwhile, in the United States, deaths are happily way down, but case numbers have stopped dropping due to the rise of the new strains, and will likely start ticking upwards once again for a while. Whether or not this will count as a last surge/wave is unclear, it looks like the strains aren't as additionally infectious as we feared and vaccinations are going well, so it might not be so bad.

Also, we (myself and the anonymous donor) awarded the Covid Microgrants, for details see the section on that below.

Let's run the numbers.

The Numbers

Predictions

Prediction (WaPo numbers): Positivity rate will be 4.2% (unchanged) and deaths will fall by 12%.

Results from WaPo Covid page, which I picked last week as the data source:

In the past week in the U.S....

New daily reported **cases fell 3.9% ↓**

New daily reported **deaths fell 27.8% ↓**

Covid-related **hospitalizations fell 7.4% ↓** [Read more](#)

Among reported tests, **the positivity rate was 4.1%.**

The **number of tests reported fell 26% ↓** from the previous week. [Read more](#)

Positivity rate was indeed close to unchanged, but I see a contradiction with the Wikipedia data. The Washington Post source says there were 26% fewer tests and an essentially unchanged positive rate. Wikipedia reports essentially a flat number of positive tests. Those two things can't both be true at the same time, so someone has this wrong. John Hopkins has 4.7% positive rate right now, but their data a week ago was jumping around due to an anomaly so it's hard to get a good week over week number out of them here.

I think this comes down to the data anomaly a week ago, which different places are handling differently? Which previously was easy to handle since I had a good chart, now I have a bunch of graphs and have to dig for actual raw numbers when I want them. Oh how I miss the Covid Tracking Project.

My guess and hope is that such disagreements between sources will usually be much smaller than this, and my best guess on what happened is that the real positivity rate didn't change much. I'm going to treat the 26% decline in reported tests at WaPo as not real, and assume that's where the mistake is.

On deaths, we did much better than I expected. A 12% decline is good, a 28% decline is fantastic.

Prediction (WaPo numbers): Positivity rate of 4.3% (up 0.2%), deaths decline by 8%.

I don't think we can sustain this huge decline in deaths because the decline in cases mostly stopped about a month ago, but given how slow deaths were to decline there's clearly a bunch of extended and variable delays in death reporting, even more so than previously appreciated, so some additional decline seems likely. Also, whenever there's a huge jump there's a decent chance some of it is shifting things in time, and there will be a bit of reversion.

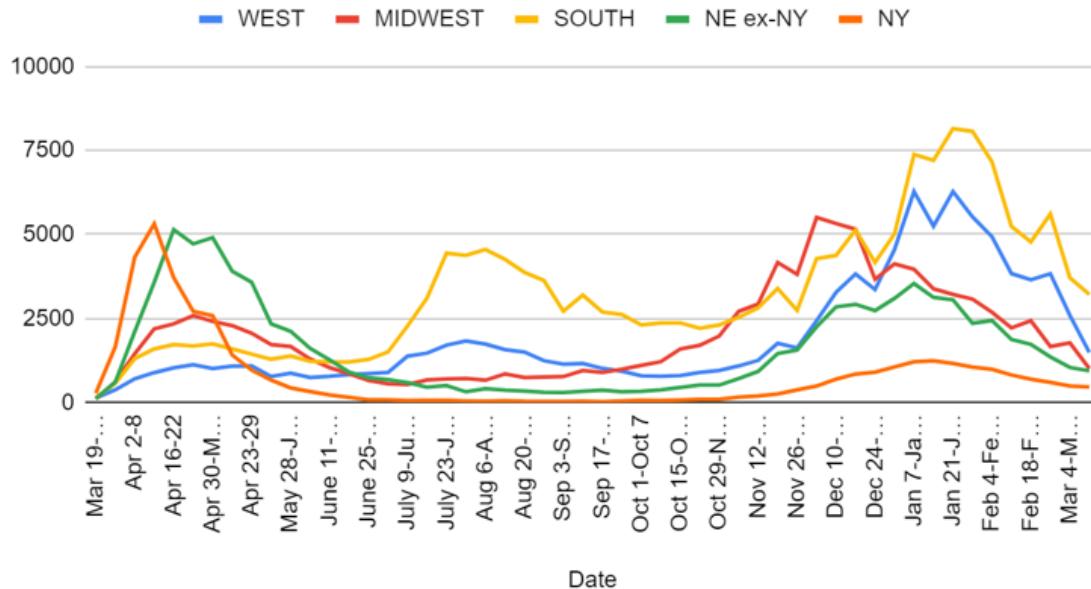
Positivity rate likely starts creeping back upwards in the short term.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Jan 28-Feb 3	5524	3078	8071	3410	20083
Feb 4-Feb 10	4937	2687	7165	3429	18218
Feb 11-Feb 17	3837	2221	5239	2700	13997
Feb 18-Feb 24	3652	2433	4782	2427	13294
Feb 25-Mar 3	3834	1669	5610	1958	13071
Mar 4-Mar 10	2595	1775	3714	1539	9623

Mar 11-Mar 17 1492 1010 3217 1402 7121

Deaths by Region



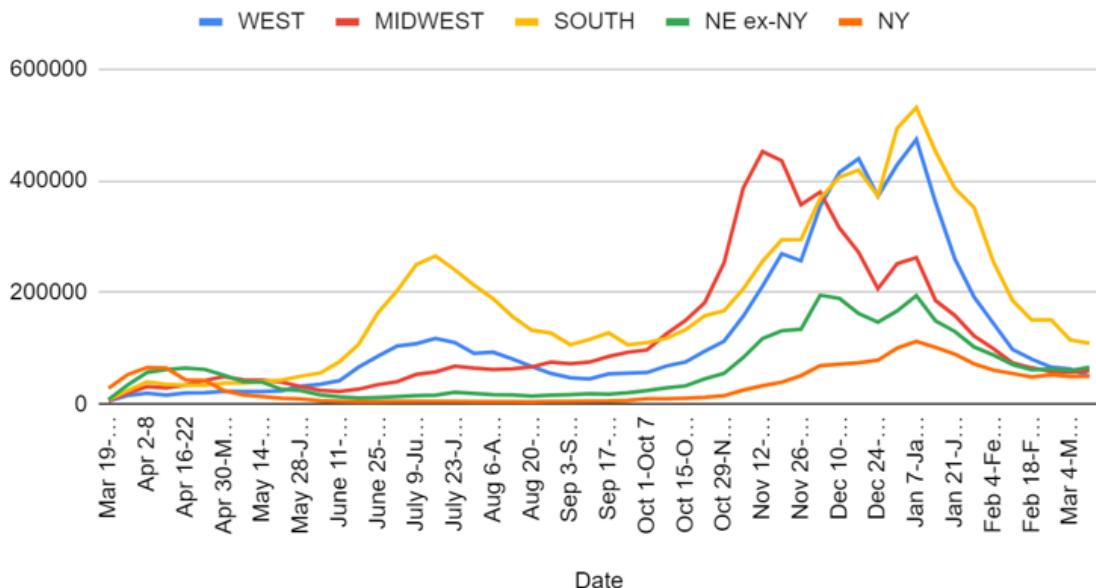
This is excellent news as deaths continue to decline steadily. We're finally seeing the full impact of the decline in cases, especially in the Midwest and West. This was essentially the best case scenario, as substantially bigger declines would have mostly caused me to suspect data issues.

The bad news is that these declines are probably going to stall out soon, since cases aren't declining at these rates anymore.

Positive Tests

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Feb 4-Feb 10	144,902	99,451	255,256	149,063
Feb 11-Feb 17	97,894	73,713	185,765	125,773
Feb 18-Feb 24	80,625	64,857	150,493	110,339
Feb 25-Mar 3	66,151	58,295	151,253	115,426
Mar 4-Mar 10	62,935	57,262	114,830	109,916
Mar 11-Mar 17	49,696	59,881	109,141	115,893

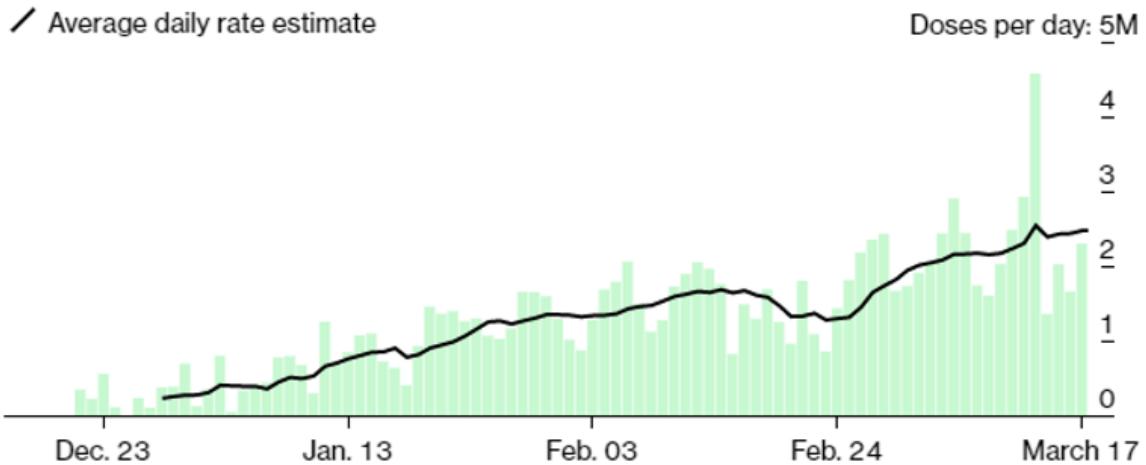
Positive Tests by Region



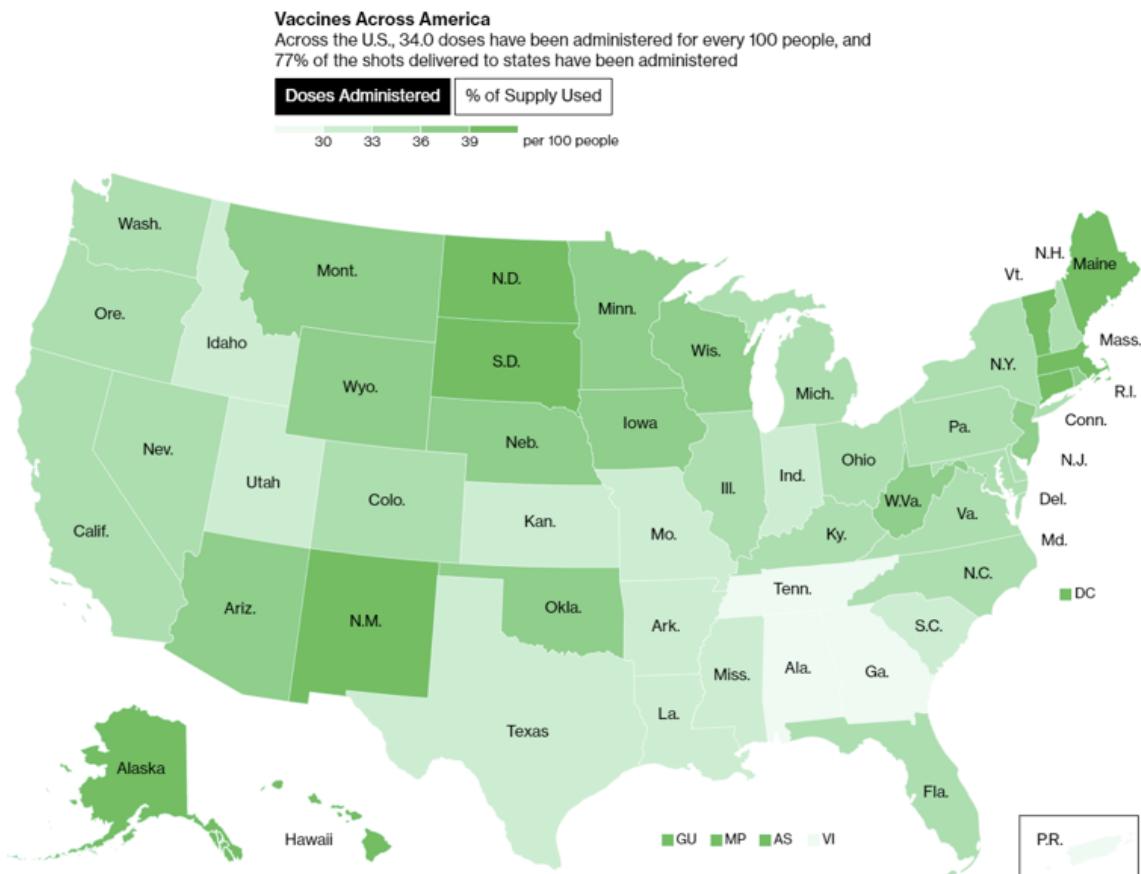
All hail the control system, which has successfully once again reasserted itself. We will face increasing pressure from further reopenings and increased dominance of the new strains, and will get steady additional help from vaccinations and warmer weather. Soon we will see which side of that is stronger. In the long run, of course, the vaccinations will win out unless new strains manage to escape them and we don't respond in time, but in the short run things are more likely to get somewhat worse first before they get better.

Vaccinations

In the U.S., more Americans have received at least one dose than have tested positive for the virus since the pandemic began. So far, **113 million doses** have been given. In the last week, an average of **2.47 million doses per day** were administered.



Note: Immunity calculations take into account the two doses required for most vaccines. The "daily rate estimate" is a seven-day rolling average; interpolation is used for countries with infrequent updates. Data are from Bloomberg's Covid-19 Vaccine Tracker.

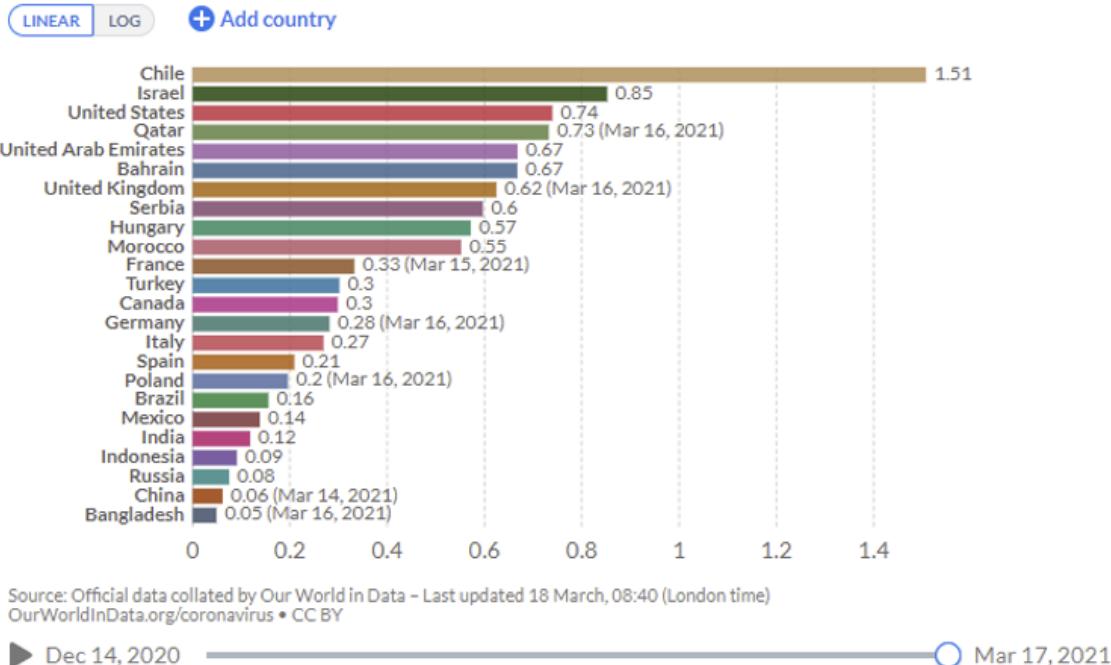


Things are going well on the vaccination front in America. Regional differences exist but steady progress is being made everywhere, and additional supply steadily comes online. We may be a little bit ahead of ourselves with the weekly number due to the giant spike earlier in the week, but much of that was given back the next day, and the steady improvement in volume seems quite real.

Daily COVID-19 vaccine doses administered per 100 people, Mar 17, 2021

Our World
in Data

Shown is the rolling 7-day average per 100 people in the total population. This is counted as a single dose, and may not equal the total number of people vaccinated, depending on the specific dose regime (e.g. people receive multiple doses).



The news in Europe is not as good, because not only were they already far behind, now they're suspending AstraZeneca shots for no reason, which is the big news item this week.

(I'm going to stop showing the Europe graphs each week as I don't think they've been worthwhile recently, [but if you want to find that info yourself you can always get it at OurWorldInData](#).)

Europe Panics Over Actual Nothing, Halts AstraZeneca Vaccinations

The AstraZeneca vaccine does not cause blood clots.

At all. No, seriously. It simply doesn't cause blood clots. [Wenodis](#). We are aware of this.

The whole incident is so mind-bogglingly insane and stupid that I don't even know where to begin.

From what I can tell, the sequence of events was something like this.

1. A lot of people got the AstraZeneca vaccine.
2. Because we are paranoid about possible side effects, there's extensive reporting of anything that happens to people right after getting the vaccine.
3. Because of #1 and #2, we record [a lot of stuff that happens by coincidence, because this isn't Unsung](#).
4. At some point, in some region, over some time period, something bad will happen more often than chance within that region and period, because that's how probability works.

5. In this case #4 is certain types of blood clots in some places at some times.
6. Many did not notice or care that the overall rate of blood clots for those getting vaccinated is actually *below* the population base rate, and similar to the rate for those getting the Pfizer vaccine.
7. Many did not notice or care that *even if all the blood clots were due to the vaccine*, and then in addition *even if all the blood clots were fatal*, neither of which is possibly true, *the vaccine would still be worth taking by multiple orders of magnitude*.
8. Even the WHO and EMA said there was no evidence and nothing to worry about.
9. Everyone in Europe lost their minds and collectively massively sabotaged the vaccine effort, and likely all vaccine efforts everywhere permanently, by halting vaccinations 'as a precaution.'
10. People will die.
11. People will also get more blood clots, because Covid-19 *does* cause them.
12. People everywhere will have new stupid arguments that will make people not get vaccinated, likely for all vaccines, permanently, which is already happening.

All of this due, effectively, to pure p-hacking, without even bothering to pretend otherwise.

In addition to being below the base rate, the incidents with AstraZeneca weren't substantially different from the incidents with Pfizer, because again *they are random*.



Alex Tabarrok @ATabarrok · 19h

Pfizer, All UK spontaneous reports received between 9/12/20 and 28/02/21

Deep vein thrombosis 8
Pulmonary embolism 15
Thrombocytopenia 13

AZ, All UK spontaneous reports received between 4/01/21 and 28/02/21

Deep vein thrombosis 14
Pulmonary embolism 13
Thrombocytopenia 12



Alex Tabarrok @ATabarrok · 18h

As of 28 Feb, an estimated 10.7 million first doses of the Pfizer/BioNTech vaccine and 9.7 million doses of the Oxford University/AstraZeneca vaccine had been administered, and around 0.8 million second doses, mostly the Pfizer/BioNTech vaccine, had been administered.

Or as my friend put it when in an unfairly charitable mood:



Andrew Rettek @oscredwin · 17h

The EU pulled the lever in the trolley problem, noticed 1 person on the new track, and pulled the lever again so the train goes back to killing thousands.

Or, in the modern vernacular:



John Mullahy @JohnMullahy · 22h

...



37

2.1K

8.3K



Or, [in the words of one of the few remaining possibly sane European health authority figures:](#)



BNO Newsroom ✅ @BNODesk · 19h

...

Belgium's health minister says there's no reason to stop using AstraZeneca's COVID-19 vaccine, and doing so would be irresponsible: "We take a bigger risk if we don't give the vaccine" - VTM

35

633

2K



Again, let's be crazy generous and say all 39 cases both were entirely caused by the vaccine (which they weren't, since again below base rate) and also killed all the patients (which they didn't, death rate from blood clots is 10-30% per Google). That's 39 deaths in 9.7 million doses, for a fatality rate of one in 300,000. For example, if that happened to the entire United States it would kill about a thousand people, so purely in terms of deaths it would be about the price of delaying vaccinations by one day.

[Also, by the way, there's this:](#)

Venous **thrombosis** is a disease of aging, with a low rate of about 1 per 10,000 annually before the fourth decade of life, rising rapidly after **age** 45 years, and approaching 5–6 per 1000 annually by **age** 80 (6).

So it's not remotely fair to use the *background population rate* when you're explicitly targeting your elderly population for vaccinations. This is so much more insane than it looks at first glance.

[It's purely and simply this](#) ([link to Reuters](#)):



Yascha Mounk @Yascha_Mounk · Mar 15

Europe's vaccination campaign has officially become an omnishambles.

...

Today, Germany stopped administering AstraZeneca vaccines even though doing so will, according to all available data, kill a lot of people.

It's like everyone has just given up.

Spahn said there have been seven reported cases after vaccination that could be related to cerebral vein thrombosis out of 1.6 million vaccinations in Germany.

Clemens Wendtner, head of infectious diseases and tropical medicine at Munich clinic Schwabing, said the background incidence, or normally expected risk, for this condition was two to five cases per 1 million individuals per year.

“This should be the reason to suspend the vaccination in Germany until all cases, including suspected cases in Germany and Europe, have been completely cleared up,” he added.

In Britain, where more than 11 million doses of AstraZeneca’s shot have been administered, only three such cases have been reported, government documents [here](#) show.

[Another simple Fermi calculation:](#)



Stanley Pignal @spignal · Mar 15

If you vaccinate 100,000 people over the age of 50 today rather than tomorrow, you save 15 lives, according to a French analysis.

...

Germany has 1.7m AstraZeneca doses that are now not being administered.

Delay all of those by a week, you're up 1785 deaths.

121

2K

4.1K



Stanley Pignal @spignal · Mar 15

far too depressing to update this calculation for all the new countries that are stopping vaccination against this deadly disease.

...

5

39

271



Stanley Pignal @spignal · 6h

France suspended its AstraZeneca vaccination campaign on the back of a single case of thrombosis. Out of 1.4m jabs administered to date.

...

I'm all for the precautionary principle, but at some point you have to ask who is making these decisions and why.

And again, how many people could this possibly kill, even if several things went impossibly badly, in exchange for saving those 1,785 lives?

Six.

Not six thousand. Not six hundred. Six.

What do you idiots slash mustache-twirling villains have to say for yourselves, and do you have a preference as to which of those two ways would you prefer to be primarily identified?



Stanley Pignal @spignal · 1h

German health ministry has a good Q&A about why it suspended AZ vaccination. It argues state has a special duty of care around vaccines but accepts there are "legitimate" questions around whether pausing causes more harm anyway given low numbers of cases

...

[Link to the Q&A here](#) (in German). Translated, here's the meat of it:

Why has AstraZeneca vaccination been suspended?

In several cases in Germany in connection with a vaccination with AstraZeneca, a special form of severe cerebral vein thrombosis in connection with a lack of blood platelets (thrombocytopenia) and bleeding was found. That is why the Paul Ehrlich Institute recommended that the AZ vaccination be suspended as a precaution in order to investigate the cases further. The Federal Ministry of Health followed this recommendation. The European Medicines Agency [EMA](#) will decide whether and how the new findings affect the approval of the vaccine. Vaccination is a matter of trust and not a compulsion. Every vaccinee must be sure that all information about the vaccine is conveyed transparently and completely and that nothing is withheld. Rare, but possibly serious, side effects must also be carefully examined.

Would it have been justifiable to keep vaccinating?

No. The reported cases were examined by experts at the PEI, who UNANIMOUSLY came to the conclusion that "the observed cases could be related to the vaccination". In this situation, PEI and BMG are obliged to initiate the coordinated procedures at the European Medicines Agency EMA. Regardless of this, the vaccinations had to be temporarily stopped. Even if it is decided to continue vaccinating despite the warnings, the vaccinating doctors must first be informed and the vaccinees themselves must be informed about possible side effects. The state has special duties of care when it comes to recommended vaccinations.

What is the normal rate of cerebral vein thrombosis compared to the reported cases?

In the vaccinated group of people and within a period of 14 days after vaccination, approximately 1 to 1.4 sinus vein thromboses are to be expected statistically with 1.6 million vaccinations. However, six cases of sinus vein thrombosis plus one medically comparable case, i.e. seven, were reported by Monday, March 15. The PEI writes in its report: "According to this calculation, more cases of sinus thrombosis have been reported than would be expected statistically by chance."

Are we not risking more deaths than we are avoiding as a result of the AstraZeneca vaccination ban?

That is a statistical (and legitimate) issue. But the state is legally obliged, in particular, to the individual citizen who is vaccinated as part of a state vaccination campaign! The state makes the vaccine available and therefore has special duties of care. Officials of the BMG and PEI are obliged to monitor the safety of the vaccine and to react to appropriate signals. If these obligations are violated and the vaccination campaign continues without properly informing the population and the people to be vaccinated, there could also be legal consequences.

So of all the potential things that can go wrong you managed to find one subsection of one thing that happened more often than chance, and let's be super generous and again assume that *all seven* were lethal and also that all seven were caused by vaccination and that's the typical rate going forward, and (does math) yeah you're still off by more than two orders of magnitude and you know it, but *you have a legal obligation* to these people that forces your hand, because 'there could be legal consequences'? And there's no way to, say, pass a new law to fix that, even if you should have fixed it long ago? So that's it, nothing you could do, huh?

Also there was this:

Thrombosis can also occur with contraceptive pills. Then why all the fuss about AstraZeneca?

It is true that thromboses, even fatal ones, are known to be a very rare side effect of birth control pills and are listed in the patient information. Every woman who receives a prescription for a contraceptive pill must be informed about the risk by the prescribing doctor. For the AstraZeneca Covid 19 vaccination, the rare side effect of a sinus vein thrombosis, which can sometimes be fatal, has not yet been listed in the patient information sheet, and the state-recommended vaccination of healthy people differs from the prescription of a drug under drug law. Trust and transparency are always important when prescribing drugs, but especially when it comes to vaccinations because they are widely used in healthy people.

Read that last line again and think about what it implies in the context of this question.

The amount of damage this is already doing to vaccination effort is staggering. I got this comment on my last post:



Michael B says:

March 15, 2021 at 4:39 am (Edit)

My mother is 76, lives in the US, had an appointment to get her shot on March 4th, and she *missed it* because she heard the AZ vaccine had dangerous side effects. They don't even give the AZ vaccine in the US! Her brother in Europe called her and told her and insisted that they do give the AZ vaccine in the US, but under a different name.

Europe's putting people at risk who don't even live in Europe!

On her first day back after our trip, my wife saw five patients. *Two of them* expressed serious concern about getting vaccinated in the United States, *where they don't even give the AstraZeneca vaccine*, due to these concerns, and she had to spend a bunch of time explaining the several-layered absurdity of that concern. [A twitter poll I did](#) already found multiple people saying they know of a shot that was missed. This is only going to get worse.

On my todo list is to do a standalone pure 'why vaccinations are safe and effective and everyone should get one as soon as possible if they're able to do so' post, if no one else has one that does the job well enough. Is there a good one already in existence? Several people have asked, and there's nothing I'm fully happy pointing people towards. The concern is not 'you should take one now even if you're worried others need it more' but rather 'if and when there's enough shots for everyone you really really need to take one,' which is the error that matters far more overall.

Better To Have Vaccinated And Stopped Than Never To Have Vaccinated At All

Europe halted AstraZeneca vaccinations. And that's terrible.

Then again, at least they *started* doing AstraZeneca vaccinations, without which halting them would have been impossible. America didn't even start them, and have been holding hostage tens of millions of doses? Isn't at least starting a pretty good relative result? Isn't it a mistake to [bring down the shame hardest on the person who at least interacted with the problem and did some good](#), even if they nonsensically stopped, rather than the one who did no good whatsoever?

In this case, I don't think that applies, because halting distribution is causing large active harm over and above the lack of doses being administered, and because halting it now *after approving it* is far more indefensible than failing to approve in the first place. Failing to approve also isn't defensible, but if you're committed to defending power and 'ethics' and counterproductive principles above all else then at least it makes *some sense*. It can be argued it's at least *consistent* and it doesn't quite fully mean You Fail Statistics Forever. Halting now is some combination of malice and pure madness. It is choosing to cover one's ass against blame for the perception of irresponsibility at the cost of thousands of lives. Perhaps [the blame dynamics involve this](#), which would be an impressive shooting of one's nose to spite one's face...



Samo Burja @SamoBurja · Mar 17

I suspect Europe is using vaccine safety concerns to cover up delays, rather than vaccine safety concerns causing delays.

...



Razib Khan ✅ @razibkhan · Mar 16

Europe's Vaccine Suspension May Be Driven as Much by Politics as Science nytimes.com/2021/03/16/wor... the ummah shall not agree upon error

...or it's (also) something worse.

There isn't a better option.

That doesn't let the United States off the hook. But I am very happy that we are not right now *halting* one of the vaccines for no reason, because that would do that much more damage. And I do think the suspensions are a much *worse sign of dysfunction* than America's failure to begin in the first place. To get this result, the rot must go far deeper.

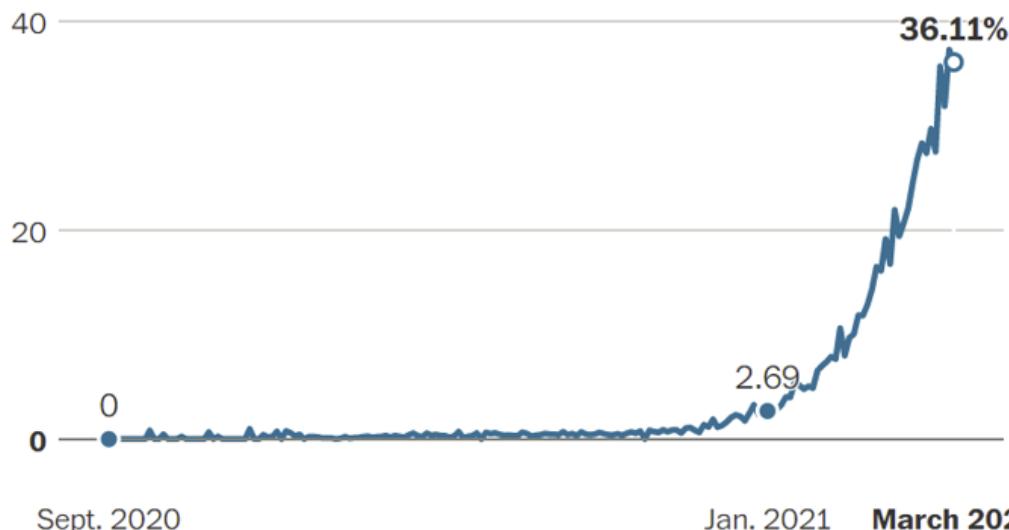
The most harmful act of all around AstraZeneca, of course, [is the United States deciding to hold onto tens of millions of doses indefinitely](#), left to sit in warehouses unused, while it refuses to approve it and also refuses to send it out, although it is now considering sending some to Mexico and Canada, both of whom have asked for doses. Then again, that's exactly as harmful as never making those doses in the first place, so it's hard to know what's effectively being punished if we accept that politics and power won't let us export the doses. Economists talk about 'tax incidence' and who effectively ends up with the bill for a tax (e.g. the 'employer half' and 'worker half' of social security are not economically distinct and making one side pay all of it would change very little) and this makes me think about blame incidence, especially now that I (may? have begun to?) understand how central blame is in decision making.

The English Strain

[From the Washington Post](#), in a standard issue 'look at the irresponsible ones' article :

B.1.1.7's rapid spread in the United States

The more contagious variant's estimated share of cases has surged in recent weeks. The data is current as of March 10, 2021.



Source: Helix

THE WASHINGTON POST

How does that stack up against [what the naive model said](#)? If we have 2.7% in January, and assume it means something similar to what 'March' means here, we can compare that to a predicted 2.88% for the week of January 18. Then we compare the 36% endpoint to our naive prediction of about 72%, and notice that things are substantially behind schedule since then. That's great news. Curve fitting gives a rise in R₀ from the new strain of only 35%. If that's accurate, then the model predicts that the new strain prolongs our pain, but there is never a last surge.

What that model isn't doing is drawing any distinctions between regions. It presumes that spread is evenly distributed around the country, which is obviously false. That could plausibly mean that we're underestimating the danger substantially and will see surges in the harder hit places.

Prediction for the control system is hard to evaluate, and will be key to how this plays out. Clearly levels of precaution are declining, but putting a figure on how much they are declining is very difficult. Could be a small impact, could be a large one.

The alternate explanation is that my five-day cycle is too short, which would be bad news, but would still mean we have more time than we expected and it's probably not so bad.

Six Feet Good, Three Feet Acceptable In Pinch

The CDC suggests child prison social distancing requirements could soon change, [and be reduced from six feet to three feet](#). You see, there was one recent study that said *with full and proper masking* that three feet distancing was "safe."

This is what happens when various political requirements and elite demands are dominant over decision making, with the scientific justifications being designed to fit whatever the

elites need, combined with the obsession with telling people strict/absolute simple rules rather than anyone involved treating the world as a physical object.

What happened here seems simple. The CDC said 6 feet distancing because they've been saying 6 feet distancing for a year, and if you suddenly said 3 feet in a school then everyone would quite rightfully ask what the hell the whole 6 feet thing had been about this whole time, whereas the 6 feet rule is the one thing that everyone has mostly agreed to agree upon even if in practice it often gets ignored.

So the CDC basically came out with guidance that said for child prisons to do their best to follow existing CDC rules for adults, even for children as young as 2, and then do their best to reopen.

(You should know this already: The actual physical effect, of course, is gradual rather than a step function, likely similar to an inverse square law, so 3 feet I am guessing is about four times riskier than 6 feet, if everyone is exactly 3 or 6 feet apart respectively, and the goal of a 6 foot restriction is to get people to at least be a few feet apart and not crowd into spaces too aggressively.)

There's also what they see as a necessary distinction between 'safe' actions, which allow the retention of a state of grace, and 'unsafe' actions, which are blameworthy, and to label everything as either one or the other, with the ability of guidelines to change which is which when the guidelines change, because they don't think people can handle anything else.)

Then a lot of child prisons, and especially teachers and teachers' unions, interpreted the guidelines as actual requirements rather than goals or suggestions, and it was clear a lot of child prisons would remain largely or entirely closed, with prisoners forced to go remote.

People in power didn't like that, they wanted the child prisons open, and things will soon be in a place where if the 6 foot rule became 3 feet in general in many places it wouldn't be that big a disaster, so now the rules are changing, with the one Massachusetts study being used as a fig leaf, despite no one being fooled that we suddenly had learned something from it, let alone that it was strong evidence.

As always when thinking about child prisons, it's hard for me to get behind putting our children in child prisons, but given that the alternative is virtual child prisons that are very clearly even worse, and the economic aspects of all this, I'm fine with treating the desire to reopen the child prisons as legitimate. Given that need, and the real physical risks involved, the previous guidelines were wrong and the potential new guidelines are better.

So this change would be good, even if the process that got us here wasn't great, and even if issuing the first set of guidelines will continue to cause a bunch of issues. With children (who are effectively largely immune) and vaccinated teachers (almost entirely immune) and masks everywhere, *of course* you can loosen the distancing requirement.

The even more interesting question here is, if these guidelines do get issued, how do people react more broadly? Do they think 'oh the six foot thing was all a lie?' Do all elites memory hole that we ever said six feet and start saying three feet, and how much whiplash does that cause? If the one central rule goes out the window does everyone start treating all of it as one big joke? It would be quite an interesting experiment that should increase popcorn sales.

Alternatively, perhaps that's a lot of the pseudo-intent here? Use the schools as a backdoor way to loosen distancing requirements without having to out loud admit they were arbitrary, by counting on The People to notice the contradiction, once we get to a place where we want *some* continued caution but not to go nuts, as we are likely to be in August?

We Must Protect This House

The House of Representatives has a problem. They would like to return to normal operations, but 25% of their members are being idiots and refusing to get vaccinated (or getting vaccinated but then neglecting to inform others of this, presumably for political reasons). With so many unvaccinated members, the Office of the Attending Physician is unwilling to relax social distancing guidelines.

The problem isn't lack of supply. That would be even more insane, and congress has its own supply. As much as crippling the speed at which the house can do business appeals to me, our representatives should have and do have full vaccine access at this point.

At least there's this, which is some small comfort:

The bottom line: The Office of Attending Physician reinstated the use of the congressional gym showers, locker room and swimming pool on Friday evening, according to the memo.

But they still have to put up with things like this:

- It's also giving power to disrupters like Rep. Marjorie Taylor Greene (R-Ga.), who's [used a procedural move](#) to further drag out the process.
- Votes can take more than three times longer than pre-pandemic times.
- "I won't be taking it. The survival rate is too high for me to want it," 25-year-old Rep. Madison Cawthorn (R-N.C.) told Axios in December.

If I were in charge of the house, I'd tell everyone to get vaccinated because starting in a few weeks I was going to expel anyone who wasn't, or at least bar the doors and not let them in until they fix it, whether or not there's a way to let them do remote voting anyway, unless they somehow have a physician's note saying why they can't get it. There are plenty of workplaces doing the same. Dare the other side to defend not getting vaccinated and make a big deal out of it.

What I'm curious about is to what extent the refusals are about 'worried that their crazy base will see it as a betrayal to get the vaccine,' to what extent it is actual failure to understand that the vaccines are safe and effective and worthwhile, to what extent it is their hatred of the other members of the House and a desire to make their lives as difficult as possible, and to what extent it is a strategic move to delay the work of the legislature.

I like to think the last one is primary, a lot of them secretly did get vaccinated but are refusing to say so in order to prolong the delays as long as possible, and that these Representatives are mostly like the one who had a cloth mask on that said "I'm just wearing

this so I don't get fined" while having an N95 on underneath. The alternatives involve sufficient disconnection from reality that they are even more concerning.

One response to this is that a 75% uptake rate is better than what the public is reporting, and that's without the ones who got vaccinated, so the number is not so bad especially given some of the 25% presumably did get vaccinated, and we don't need an explanation.

Covid Microgrants Have Been Awarded

I was very happy with the results here. We got many good responses, and I am proud to announce that we are giving out \$39,000 in grants to ten applicants.

Or at least, we are *trying* to give out \$39,000. It's proven surprisingly difficult, because several people who *thought* they had a working PayPal account were often surprised to learn that a four-figure international transfer required much additional paperwork. Hopefully all of that sorts itself out, and as of this edit on Thursday evening nine out of ten recipients have been successfully paid.

The recipients are:

Someone who wishes to remain anonymous, who is working on the vaccine availability website vaxxmax.com, which interfaces with RiteAid.

Someone else who wishes to remain anonymous, who is working on the vaccine availability website vacfind.org.

Someone else who also wishes to remain anonymous, who is working on <https://forecasting-covid.com/>, which will recreate covid19-projections.com using a new data source, likely Johns Hopkins.

Konstantin Likter, who is working on the vaccine availability website covidwa.com for residents of the state of Washington.

Po-Shen Loh, who is working on a better app-based method to do multi-stage contact tracing.

Lisa Hakkert, who is working on pandemic modeling and how the pandemic interacts with internet access.

Jakob Jonnerby, who is working on school reopening plans.

Garrett Schilkey, who is working on 3D modeling of UVC lighting and its preventative effects.

Dylan Alban, who gets special mention for stopping before accepting the money to note that his team had pivoted to working primarily on a different project, the vaccine availability website <https://vaccinespotter.com>, and making sure we still wanted to support him.

Abraham Hinteregger, who is working on advocacy for First Doses First.

There's a clear theme here. These are all IT or modeling projects, and the bulk of it are websites that help track availability of vaccine appointments. This was informative on several levels, not least of which it was a pointed reminder of who reads these posts. More than that, it drove home that this is the strongest current source of small low-hanging fruit an individual can easily pick. Offering information on where appointments are available is not illegal or even regulated, and that's a sharp contrast to other many areas.

The vaccine rollout is a hodgepodge of different stuff, so helping ensure that people can find appointments reasonably, and that vaccine does not sit idle, is a big game that can be

accomplished for relatively little investment.

I'm sad we didn't get to do this earlier, when there was more room to have a bigger impact, but better late than never.

Our plan is to follow up later to see how things went, and report back. There's some chance this or something similar will be open again some time, but we do not have any concrete plans at this time for doing so.

In Other News

[From MR: A theory that when you listen to doctors and other health professionals and let them make decisions,..you get decisions with a strong bias towards inaction and paralysis under uncertainty.](#)

Not Covid-19 directly but central to how this column makes decisions: [Politics is way too meta.](#)

Badly needed, but [actually..given we just now did it and only this much, it seems it's incredibly hard...](#)



Alex Tabarrok @ATabarrok · Mar 15

You see? This stuff isn't even hard.

...



Mary Ellen McIntire ✅ @MelMcIntire · Mar 15

Andy Slavitt announces the Medicare reimbursement rate for administering COVID-19 vaccines will increase to \$40/shot, up from \$23/shot

In case there were still doubts [what Cuomo is](#):



Richard M. Nixon @dick_nixon · Mar 15

Cuomo is having the fellow who distributes his vaccines call county executives to judge loyalty.

It's not every day you get to [look this clearly into a glass house](#) and watch the residents throw stones, I mean wow just wow:



Walid Gellad, MD MPH ✅ @walidgellad · Mar 15

Replying to @walidgellad

If the Pfizer study is same one that leaked via Twitter, there is no way the 94% figure is accurate. At the time, even the Israeli govt didn't believe it.

Yet again in pandemic, people change the level of evidence they require depending on their priors and popularity of result.



2



2



14



[Novavax vaccine 96% effective in preventing mild and severe illness](#), still not approved.

[Yo-Yo Ma uses post-vaccination observation period as a concert for newly inoculated](#).

Not centrally covid, but seems like a worthwhile data point that [private equity purchases of nursing homes found to have very large harmful effects](#).

[Would you go for it, or just let it slip?](#)



Remember, you miss 100% of the shots you don't book an appointment for. See everyone next week.