

Best of LessWrong: May 2021

1. [There's no such thing as a tree \(phylogenetically\)](#)
2. [Your Dog is Even Smarter Than You Think](#)
3. [Curated conversations with brilliant rationalists](#)
4. [Small and Vulnerable](#)
5. [Finite Factored Sets](#)
6. [Saving Time](#)
7. [Less Realistic Tales of Doom](#)
8. [Agency in Conway's Game of Life](#)
9. [What will 2040 probably look like assuming no singularity?](#)
10. [Don't feel bad about not knowing basic things](#)
11. [April 15, 2040](#)
12. [Formal Inner Alignment, Prospectus](#)
13. [Knowledge Neurons in Pretrained Transformers](#)
14. [\[Prediction\] What war between the USA and China would look like in 2050](#)
15. [How counting neutrons explains nuclear waste](#)
16. [Covid 5/13: Moving On](#)
17. [Covid 5/6: Vaccine Patent Suspension](#)
18. [Book Review of 5 Applied Bayesian Statistics Books](#)
19. [Covid 5/27: The Final Countdown](#)
20. [Academia as Company Hierarchy](#)
21. [Teaching ML to answer questions honestly instead of predicting human answers](#)
22. [Did the Industrial Revolution decrease costs or increase quality?](#)
23. [Bayeswatch 2: Puppy Muffins](#)
24. [A Review and Summary of the Landmark Forum](#)
25. [SGD's Bias](#)
26. [Understanding the Lottery Ticket Hypothesis](#)
27. [Decoupling deliberation from competition](#)
28. [\(Trying To\) Study Textbooks Effectively: A Year of Experimentation](#)
29. [Bayeswatch 5: Hivemind](#)
30. [Concerning not getting lost](#)
31. [\[link\] If something seems unusually hard for you, see if you're missing a minor insight](#)
32. [Covid 5/20: The Great Unmasking](#)
33. [Bayeswatch 4: Mousetrap](#)
34. [Parsing Chris Mingard on Neural Networks](#)
35. [\[Weekly Event\] Alignment Researcher Coffee Time \(in Walled Garden\)](#)
36. [Sabien on "work-life" balance](#)
37. [Mundane solutions to exotic problems](#)
38. [Challenge: know everything that the best go bot knows about go](#)
39. [The Variational Characterization of KL-Divergence, Error Catastrophes, and Generalization](#)
40. [Abstraction Talk](#)
41. [Love on Cartesian Planes](#)
42. [Starting a Rationalist Meetup during Lockdown](#)
43. [AI Safety Research Project Ideas](#)
44. [Questions are tools to help answerers optimize utility](#)
45. [Two Definitions of Generalization](#)
46. [The case for hypocrisy](#)
47. [Death by Red Tape](#)
48. [The Homunculus Problem](#)
49. [Life and expanding steerable consequences](#)
50. [Open and Welcome Thread - May 2021](#)

Best of LessWrong: May 2021

1. [There's no such thing as a tree \(phylogenetically\)](#)
2. [Your Dog is Even Smarter Than You Think](#)
3. [Curated conversations with brilliant rationalists](#)
4. [Small and Vulnerable](#)
5. [Finite Factored Sets](#)
6. [Saving Time](#)
7. [Less Realistic Tales of Doom](#)
8. [Agency in Conway's Game of Life](#)
9. [What will 2040 probably look like assuming no singularity?](#)
10. [Don't feel bad about not knowing basic things](#)
11. [April 15, 2040](#)
12. [Formal Inner Alignment, Prospectus](#)
13. [Knowledge Neurons in Pretrained Transformers](#)
14. [\[Prediction\] What war between the USA and China would look like in 2050](#)
15. [How counting neutrons explains nuclear waste](#)
16. [Covid 5/13: Moving On](#)
17. [Covid 5/6: Vaccine Patent Suspension](#)
18. [Book Review of 5 Applied Bayesian Statistics Books](#)
19. [Covid 5/27: The Final Countdown](#)
20. [Academia as Company Hierarchy](#)
21. [Teaching ML to answer questions honestly instead of predicting human answers](#)
22. [Did the Industrial Revolution decrease costs or increase quality?](#)
23. [Bayeswatch 2: Puppy Muffins](#)
24. [A Review and Summary of the Landmark Forum](#)
25. [SGD's Bias](#)
26. [Understanding the Lottery Ticket Hypothesis](#)
27. [Decoupling deliberation from competition](#)
28. [\(Trying To\) Study Textbooks Effectively: A Year of Experimentation](#)
29. [Bayeswatch 5: Hivemind](#)
30. [Concerning not getting lost](#)
31. [\[link\] If something seems unusually hard for you, see if you're missing a minor insight](#)
32. [Covid 5/20: The Great Unmasking](#)
33. [Bayeswatch 4: Mousetrap](#)
34. [Parsing Chris Mingard on Neural Networks](#)
35. [\[Weekly Event\] Alignment Researcher Coffee Time \(in Walled Garden\)](#)
36. [Sabien on "work-life" balance](#)
37. [Mundane solutions to exotic problems](#)
38. [Challenge: know everything that the best go bot knows about go](#)
39. [The Variational Characterization of KL-Divergence, Error Catastrophes, and Generalization](#)
40. [Abstraction Talk](#)
41. [Love on Cartesian Planes](#)
42. [Starting a Rationalist Meetup during Lockdown](#)
43. [AI Safety Research Project Ideas](#)
44. [Questions are tools to help answerers optimize utility](#)
45. [Two Definitions of Generalization](#)
46. [The case for hypocrisy](#)
47. [Death by Red Tape](#)

48. [The Homunculus Problem](#)
49. [Life and expanding steerable consequences](#)
50. [Open and Welcome Thread - May 2021](#)

There's no such thing as a tree (phylogenetically)

This is a linkpost for <https://eukaryotewritesblog.com/2021/05/02/theres-no-such-thing-as-a-tree/>

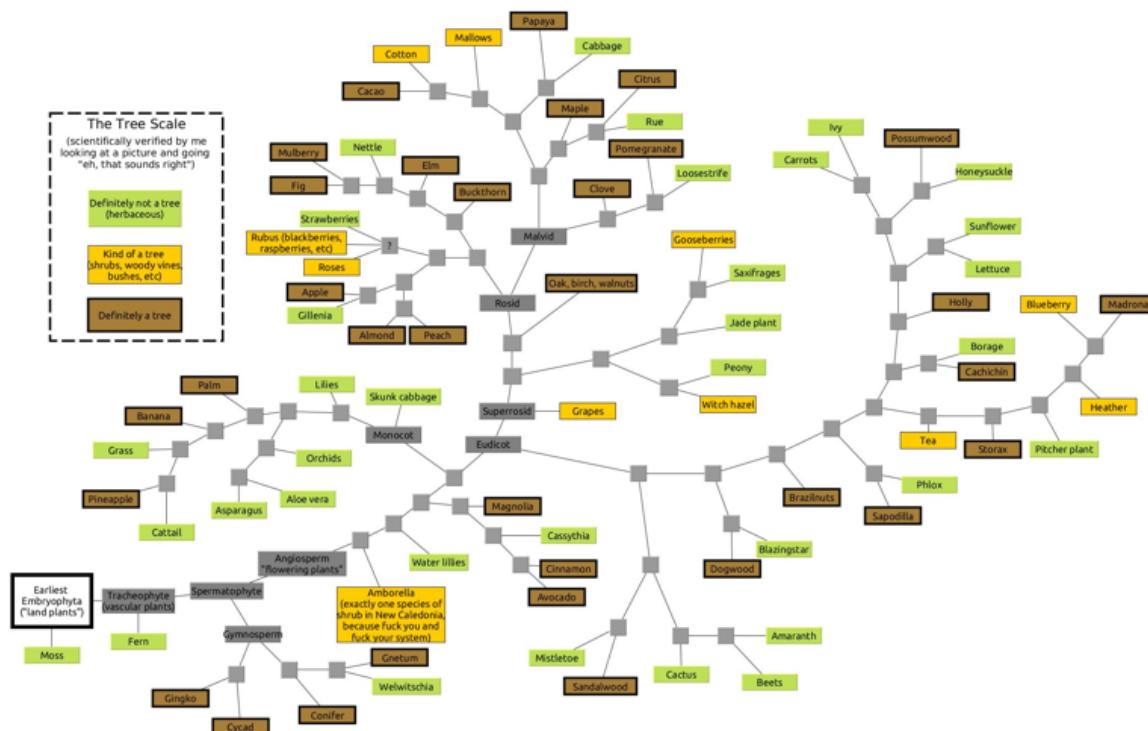
[Crossposted from Eukaryote Writes Blog.]

So you've heard about how fish aren't a monophyletic group? You've heard about carcinization, the process by which ocean arthropods convergently evolve into crabs? You say you get it now? Sit down. *Sit down.* Shut up. Listen. You don't know nothing yet.

"Trees" are not a coherent phylogenetic category. On the evolutionary tree of plants, trees are regularly interspersed with things that are *absolutely, 100%* not trees. This means that, for instance, either:

- The common ancestor of a maple and a mulberry tree was not a tree.
- The common ancestor of a stinging nettle and a strawberry plant was a tree.
- And this is true for most trees or non-trees that you can think of.

I thought I had a pretty good guess at this, but the situation is far worse than I could have imagined.



[CLICK TO EXPAND](#). Partial phylogenetic tree of various plants. TL;DR: Tan is definitely, 100% trees. Yellow is tree-like. Green is 100% not a tree. Sourced mostly from Wikipedia.

I learned after making this chart that [tree ferns](#) exist (h/t seebz), which I think just emphasizes my point further. Also, h/t kithpendragon for suggestions on improving accessibility of the graph.

Why do trees keep happening?

First, what is a tree? It's a big long-lived self-supporting plant with leaves and wood.

Also of interest to us are the non-tree "woody plants", like lianas (thick woody vines) and shrubs. They're not trees, but at least to me, it's relatively apparent how a tree could evolve into a shrub, or vice-versa. The confusing part is a tree evolving into a dandelion. (Or vice-versa.)

Wood, as you may have guessed by now, is also not a clear phyletic category. But it's a reasonable category - a lignin-dense structure, usually that grows from the exterior and that forms a pretty readily identifiable material when separated from the tree. (...Okay, not the most explainable, but you know wood? You know when you hold something in your hand, and it's *made of wood*, and *you can tell that*? Yeah, that thing.)

[All plants](#) have lignin and cellulose as structural elements - wood is plant matter that is dense with both of these.

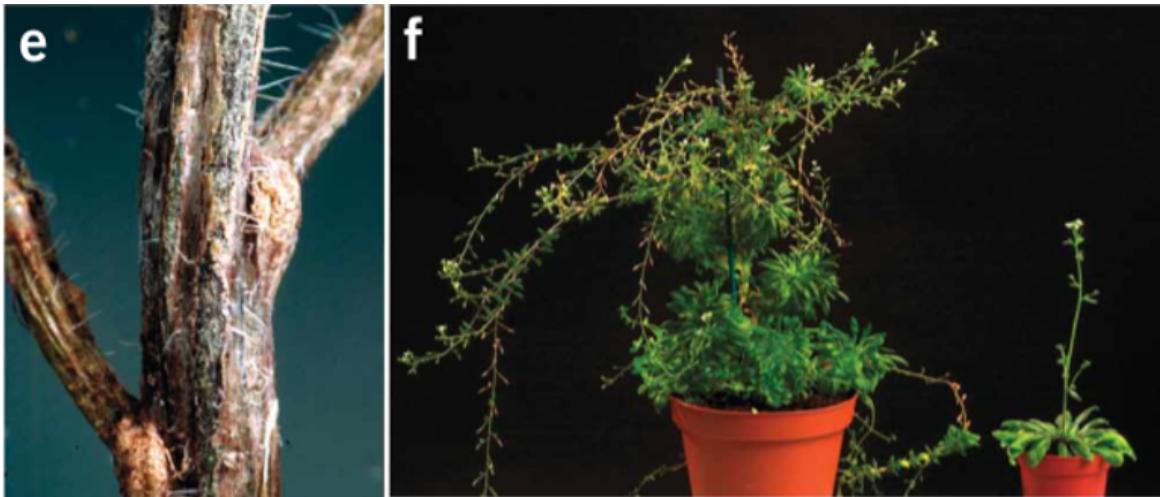
Botanists don't seem to think it only could have gone one way - for instance, the common ancestor of flowering plants is theorized to have been woody. But we also have pretty clear evidence of recent evolution of woodiness - say, a new plant arrives on a relatively barren island, and some of the offspring of that plant becomes treelike. Of plants native to the Canary Islands, wood independently evolved [at least 38 times!](#)

One relevant factor is that all woody plants do, in a sense, begin life as herbaceous plants - by and large, a tree sprout shares a lot of properties with any herbaceous plant. Indeed, botanists call this kind of fleshy, soft growth from the center that elongates a plant "primary growth", and the later growth from towards the outside which causes a plant to thicken is "secondary growth." In a woody plant, secondary growth also means growing wood and bark - but other plants sometimes do secondary growth as well, like potatoes (in roots)

[This paper](#) addresses the question. I don't understand a lot of the closely genetic details, but my impression of its thesis is that: Analysis of convergently-evolved woody plants show that the genes for secondary woody growth are similar to primary growth in plants that don't do any secondary growth - even in unrelated plants. And woody growth is an adaption of secondary growth. To abstract a little more, there is a common and useful structure in herbaceous plants that, when slightly tweaked, "dendronizes" them into woody plants.

Dendronization - Evolving into a tree-like morphology. (In the style of "[carcinization](#)".) From 'dendro', the ancient Greek root for tree.

Can this be tested? Yep - knock out a couple of genes that control flower development and change the light levels to mimic summer, and [researchers found that](#) *Arabidopsis* - rock cress, a distinctly herbaceous plant used as a model organism - grows a woody stem never otherwise seen in the species.



The tree-like woody stem (e) and morphology (f, left) of the gene-altered *Arabidopsis*, compared to its distinctly non-tree-like normal form (f, right.) Images from Melzer, Siegbert, et al. ["Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*."](#) *Nature genetics* 40.12 (2008): 1489-1492.

So not only can wood develop relatively easily in an herbaceous plant, it can come from messing with some of the genes that regulate annual behavior – an herby plant’s usual lifecycle of reproducing in warm weather, dying off in cool weather. So that gets us two properties of trees at once: woodiness, and being long-lived. It’s still a far cry from turning a plant into a tree, but also, it’s really not that far.

To look at it another way, as [Andrew T. Groover](#) put it:

“Obviously, in the search for which genes make a tree versus a herbaceous plant, it would be folly to look for genes present in poplar and absent in *Arabidopsis*. More likely, tree forms reflect differences in expression of a similar suite of genes to those found in herbaceous relatives.”

So: There are no unique “tree” genes. It’s just a different expression of genes that plants already use. Analogously, you can make a cake with flour, sugar, eggs, sugar, butter, and vanilla. You can also make frosting with sugar, butter, and vanilla – a subset of the ingredients you already have, but in different ratios and use

But again, the reverse also happens – a tree needs to do both primary and secondary growth, so it’s relatively easy for a tree lineage to drop the “secondary” growth stage and remain an herb for its whole lifespan, thus “poaizing.” As stated above, it’s hypothesized that the earliest angiosperms were woody, some of which would have lost that in become the most familiar herbaceous plants today. There are also some plants like [cassytha](#) and [mistletoe](#), herbaceous plants from tree-heavy lineages, who are both parasitic plants that grow on a host tree. Knowing absolutely nothing about the evolution of these lineages, I think it’s reasonable to speculate that they each came from a tree-like ancestor but poaized to become parasites. (Evolution is very fond of parasites.)

Poaization: Evolving into an herbaceous morphology. From ‘*poai*’, ancient Greek term from Theophrastus defining herbaceous plants (“Theophrastus on Herbals and Herbal Remedies”).

(I apologize to anyone I’ve ever complained to about jargon proliferation in rationalist-diaspora blog posts.)

The trend of staying in an earlier stage of development is also called neotenizing. Axolotls are an example in animals – they resemble the juvenile stages of the closely-related tiger salamander. Did you know very rarely, or when exposed to hormone-affecting substances, axolotls “grow up” into something that looks a lot like a tiger salamander? Not unlike the gene-altered *Arabidopsis*.



A normal axolotl (left) vs. a spontaneously-metamorphosed “adult” axolotl (right). [Photo of normal axolotl from By th1098 – Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=30918973>. Photo of metamorphosed axolotl from deleted reddit user, via this thread: https://www.reddit.com/r/Eyebleach/comments/etg7i6/this_is_itzi_he_is_a_morphe_d_axolotl_no_thats_not/]

Does this mean anything?

A friend asked why I was so interested in this finding about trees evolving convergently. To me, it's that a tree is such a familiar, everyday thing. You know birds? Imagine if actually there were amphibian birds and mammal birds and insect birds flying all around, and they all looked pretty much the same – feathers, beaks, little claw feet, the lot. You had to be a real bird expert to be able to tell an insect bird from a mammal bird. Also, most people don't know that there isn't just one kind of "bird". That's what's going on with trees.

I was also interested in culinary applications of this knowledge. You know people who get all excited about “don’t you know a tomato is a fruit?” or “a blueberry isn’t *really* a berry?” I was one once, it’s okay. Listen, forget all of that.

There is a kind of botanical definition of a fruit and a berry, talking about which parts of common plant anatomy and reproduction the structure in question is derived from, but they’re definitely not related to the culinary or common understandings. (An apple, arguably the most central fruit of all to many people, is not truly a botanical fruit either).

Let me be very clear here – mostly, *this is not what biologists like to say*. When we say a bird is a dinosaur, we mean that a bird and a *T. rex* share a common ancestor that had recognizably dinosaur-ish properties, and that we can generally point to some of those properties in the bird as well – feathers, bone structure, whatever. You can analogize this to similar statements you may have heard – “a whale is a mammal”, “a spider is not an insect”, “a hyena is a feline”...

But this is *not* what's happening with fruit. Most "fruits" or "berries" are not descended from a common "fruit" or "berry" ancestor. Citrus fruits are all derived from a common fruit, and so are apples and pears, and plums and apricots – but an apple and an orange, or a fig and a peach, do not share a fruit ancestor.

Instead of trying to get uppity about this, may I recommend the following:

- Acknowledge that all of our categories are weird and a little arbitrary
- Look wistfully at pictures of *Welwitschia*
- Send a fruit basket to your local botanist/plant evolutionary biologist for putting up with this, or become one yourself



While natural selection is commonly thought to simply be an ongoing process with no "goals" or "end points", most scientists believe that life peaked at *Welwitschia*. [Photo from By Sara&Joachim on Flickr - Flickr, CC BY-SA 2.0, <https://commons.wikimedia.org/w/index.php?curid=6342924%5D>]

Some more interesting findings:

- A mulberry (left) is not related to a blackberry (right). They just... both did that.



[Mulberry photo by Cwambier – Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=63402150>. Blackberry photo by By Ragesoss – Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=4496657>]

- Avocado and cinnamon are from fairly closely-related tree species.
- It's possible that the last common ancestor between an apple and a peach was not even a tree.
- Of special interest to my Pacific Northwest readers, the Seattle neighborhood of [Magnolia](#) is misnamed after the local madrona tree, which Europeans confused with the (similar-looking) magnolia. In reality, these two species are only very distantly related. (You can find them both on the chart to see exactly how far apart they are.)
- None of [cactuses, aloe vera, jade plants, snake plants, and the succulent I grew up knowing as "hens and chicks"] are related to each other.
- [Rubus](#) is the genus that contains raspberries, blackberries, dewberries, salmonberries... that kind of thing. (Remember, a genus is the category just above a species – which is kind of a made-up distinction, but suffice to say, this is a closely-related groups of plants.) Some of its members have 14 chromosomes. Some of its members have 98 chromosomes.
- Seriously, I'm going to hand \$20 in cash to the next plant taxonomy expert I meet in person. God knows bacteriologists and zoologists don't have to deal with this.

And I have one more unanswered question. There doesn't seem to be a strong trend of plants evolving into grasses, despite the fact that grasses are quite successful and seem kind of like the most anatomically simple plant there could be – root, big leaf, little flower, you're good to go. [But most grass-like plants are in the same group](#). Why don't more plants evolve towards the "grass" strategy?

Let's get personal for a moment. One of my philosophical takeaways from this project is, of course, "convergent evolution is a hell of a drug." A second is something like "taxonomy is not automatically a great category for regular usage." Phylogenetics are absolutely fascinating, and I do wish people understood them better, and probably "[there's no such thing as a fish](#)" is a good meme to have around because most people do not realize that they're genetically closer to a tuna than a tuna is to a shark – and "no such thing as a fish" invites that inquiry.

(You can, at least, say that a tree is a strategy. Wood is a strategy. Fruit is a strategy. A fish is also a strategy.)

At the same time, I have this vision in my mind of a clever person who takes this meandering essay of mine and goes around saying “did you know there’s no such thing as wood?” And they’d be *kind of right*.

But at the same time, insisting that “wood” is not a useful or comprehensible category would be the most fascinatingly obnoxious rhetorical move. Just the pinnacle of choosing the interestingly abstract over the practical whole. A perfect instance of missing the forest for – uh, the forest for ...

...

... Forget it.

Related:

Timeless Slate Star Codex / Astral Codex Ten piece: [The categories were made for man, not man for the categories.](#)

Towards the end of writing this piece, I found that actual botanist Dan Ridley-Ellis made [a tweet thread](#) about this topic in 2019. See that for more like this from someone who knows what they’re talking about.

Your Dog is Even Smarter Than You Think

Epistemic status: highly suggestive.

[EDIT: Added more info on research methods. Addressed some common criticism. Added titles for video links and a few new vids. Prevented revolution with a military coup d'état]

A combination of surveys and bayesian estimates^[1] leads me to believe this community is interested in autism, cats, cognition, philosophy, and moral valence of animals. What I'm going to show you checks every box, so it boggles my mind that I don't see anyone talk about it. It has been bothering me so much that I decided to create an account and write this article.

I have two theories.

1. The community will ignore fascinating insight just because its normie coded. Cute tiktok-famous poodle doesn't pattern match to "this contains real insight into animal cognition".
2. Nobody tried to sell this well enough.

I personally believe in the second one^[2] and I'll try to sell it to you.

Stella

There's an intervention to help non-verbal autistic kids communicate using "communication boards" (not to be confused with [facilitated communication](#) which has a bad reputation). It can be a paper board with pictures or it can be a board with buttons that say a word when pressed. In 2018 Christina Hunger ([hungerforwords.com](#)) - a speech pathologist working with autistic children using such boards - started to wonder if her dog was in fact autistic. Just kidding, she saw similarities in patterns of behavior between young kids she was working with ("learner" seems to be the term of art) and her dog. So she gave it a button that says "Outside" and expanded from there.

Now teaching a dog to press a button that says "outside" is not impressive or interesting to me. But then she kept adding buttons and her dog started to display capabilities for rudimentary syntax.

[Stella the talking dog compilation](#) - Stella answers whether she wants to play or eat, asks for help when one of her buttons breaks, alerts owner to possible "danger" outside.

[Stella tells us she is all done with her bed!](#)

[Stella the Talking Dog says "Help Good Want Eat"](#)

(most of the good videos are on her Instagram @hunger4words, not much is on YouTube)

Reaction from serious animal language researchers and animal cognition hobbyists was muted to non-existent, but dog moms ate this stuff up. One of them was Alexis.

Bunny

Most useful research is impractical to do within academia

The Importance of Methodology and Practical Matters

Ethology has some really interesting lessons about how important various practical matters and methodology can be when it comes to what your field can (and can't) produce. For example, it turns out that a surprising amount of useful data about animal cognition comes from experiments with dogs. [...] The main reason is because they will sit still for an fMRI to be the goofest boy (and to get hot dogs). [...] On the other side of that coin, elephants are clearly very smart, but we've done surprisingly little controlled experiments or close observation with them. Why? [...] They're damn inconvenient to keep in the basement of the biology building, they mess up the trees on alumni drive, and undergrads kept complaining about elephant-patty injuries while playing ultimate on the quad.

<https://astralcodexten.substack.com/p/your-book-review-are-we-smart-enough>

A lot of useful research isn't done because it's too inconvenient, too expensive or otherwise impractical to execute within confines of academia. This is a massive shaping force. Existence of ImageNet and its quirks is a stronger shaping force on AI research than all AI ethics committees combined.

Nobody had done this before because it takes *months* of everyday training to get interesting results. Once your dog gets the hang of it, you're able to add more buttons faster, but it's never quick. Dogs take a while to come up with a response (they're bright, but they're not humans), and you can't force your dog to learn, so you have to work together and find motivation (for the dog and for yourself!). And not every pet has a strong desire to communicate.

But it may be practical to do for a layperson

Lucky for us, Alexis has many more commendable qualities besides willing to spend time and effort on her dog. She maintains healthy skepticism, she's well aware of confirmation bias and "Clever Hans" effects and of the danger of over-interpreting the dog's output. She has partnered with researchers from University of California, San Diego to have several cameras looking at the button pad running 24/7, for them to do more rigorous study.

Please watch this vid first where she gives a brief explanation of what she's doing, and importantly shows her attitude and skepticism towards her dog's "talking". She even namedrops Chomsky & Skinner ☺

Study by Comparative Cognition Lab at the University of California, San Diego

I encourage you to read their [research methodology page](#) yourself. Here's a summary:

This is an ongoing study with hundreds^[3] of participants and a mission to "use a rigorous scientific approach to determine whether, and if so, how and how much non-humans are able to express themselves in language-like ways". It's headed by Prof. Federico Rossano, Director of Comparative Cognition Lab at UC San Diego and Leo Trottier, PhD Candidate, UC San Diego. The latter has websites selling dog soundboards and interactive pet toys, indicating a potential conflict of interest, but I also don't want to knock him for simply being entrepreneurial.

They mention previous research with [a dog named Rico](#) and [a border collie named Chaser \(video of Chaser\)](#) that had rigorous experiments performed with an opaque barrier between the dog and the experimenter to preclude possibility of unwittingly influencing the dog's behavior. In those studies dogs were able to recognize 200+ toys by spoken name and perform [fast-mapping](#) (better known in some circles as one-shot learning). Encouraged by the studies they aim to ask:

Is what we're seeing clever dogs or merely Clever Hans? Can we explain the surprising button pressing behavior we're seeing using a simple first-order associative learning model, or will we have to reconsider the idea that language is an ability that is 'uniquely human'? And do we see any change in the type and complexity of communications that non-human animals (and dogs in particular) generate once they are able to use concepts that have been associated with buttons?

They're splitting the study into 3 phases:

1. *Initial data collection*, where they gather information about learners, their owners, methods of training and have participants log reports about when a specific word was first introduced, used in an appropriate context and used as part of a multi-button expression.
2. *Video collection and analysis*, where they get participants to install at least one video camera pointing at the soundboard that records every interaction. That allows them to see how button usage changes over time and to measure behavior more reliably and precisely.
3. *Interactive studies*: "Based heavily on the insights gained in phases 1 and 2, we will be piloting direct, controlled tests of learner sound button use and understanding that aim to determine how language-like learners' sound button use is. We anticipate that these will be done with a smaller number of participants."

Things I've seen the dog (appear to) do that surprised me

- Bunny is creative with the limited button vocabulary available to her and tries to use words in novel ways to communicate: "stranger paw" for splinter in her paw, "sound settle" for shut up, "poop play" for fart, "paw" to refer to owner's hand.

[Talking With Bunny | Ouch, Stranger!](#) - this video was what first made me decide to follow Bunny more closely

[Bunny "Talking" About Cats](#) - "sound settle cat bye" to tell meowing cat to shut up. Also note how she reacts to the possibly random button presses after 2:13 and the big note about confirmation bias that she put in the video.

- Bunny knows each of her doggy friends by name, thinks about them when they're not there, asks where they are, requests to play with them.

[Bunny and Friends - Bunny "Talking" About Her Friends](#) - note how Alexis wasn't sure how to explain the concept of her dog friend's "home" and how Bunny cleared it up

- Bunny understands times of day like today, morning, afternoon, night.

[Talking With Bunny | Contemplating Time](#)

- And can recall what time of day she went to the park.

[Talking With Bunny | When Bunny Went Park \(Hmm?\)](#) - mind the note that says the question was suggested by UCSD scientists to test Bunny's episodic memory

- Bunny is quite obsessed over her bowel movements (how Freudian) and about her owners' poop cycle.

[Bunny "Talking" About Number 2 - Bunny The "Talking" Dog](#) - note how she appears to use "poop play" to mean "fart".

- Bunny communicates emotional states like mad, happy, concerned. And "ugh".

[Ugh! | Bunny The "Talking" Dog](#)

- Bunny wants to know what and why is a "dog".

[What Dog Is | What About Bunny](#)

- And whether Mom used to be a dog. And she seems to recognize herself in the mirror.

[Who This? | Bunny The "Talking" Dog](#)

There's a long-running debate about whether human brains possess a special ability for language. Although "feral" human children who are raised with no language lose the ability to pick it up later in life. Maybe they could learn button talk, I don't think anyone tried for lack of steady supply of feral children.

But what I see is strongly suggestive that language facilities are not unique to us and a dog that is given ability to produce words and is taught from puppyhood with the same massive amount of effort that we put into human children will be able to talk. Don't get me wrong, I don't expect dogs to start writing poetry and doing particle physics. But I expect them to produce something that can undoubtedly be called language^[4].

It's all "Clever Hans"

I don't think this can be explained by classic Clever Hans where the owner tells the dog what to do with subconscious cues. You see lots of interactions where the dog is supposed to make a decision and the owner *doesn't know* the right answer, or the dog alerts the owner to something they're unaware of.

When Bunny [used "stranger paw" to indicate a splinter in her paw](#), how was the owner supposed to influence the answer without even being aware of the splinter?

It's all operant conditioning

First, it can't be *all* just conditioning. By induction: pets already communicate with owners to request things, teaching your dog to press "Food" instead of barking or "Outside" instead of scratching at the door simply changes the modality. Usage of simple buttons like that doesn't require clever hans or conditioning as an explanation.

Second, look at [this video of Billi the cat](#). Assuming it's just conditioning, we'd expect the cat to always go "yes food" when asked about food, because it doesn't understand "yes" or "no". But here we see it getting practically railroaded by the owner and still refusing food, and in several different ways (first "no", then "all done later"). How could this happen if it was just simple conditioning?

The owners over-interpret and anthropomorphize the button "speech"

This is the biggest danger in my opinion. Hopefully with rigorous analysis during the study and specifically set up experiments we'll be able to understand better at what level of communication the dogs actually are.

Interestingly, at least for humans, misinterpretation may be a *necessary requirement* for language acquisition, as mentioned [in this comment!](#) The short summary of the mechanism:

1. Toddler raises arms up randomly with no intention.
2. Mother thinks he wants to be held, so picks him up.
3. Toddler learns the association and the next time he raises his arms, it's an intentional attempt at communication. Similarly for words, say "maa" randomly, mother comes and smiles excitedly, the association is built.

I think the videos are fake

I [wrote a comment on that](#) and I think the videos are done in good faith. Of course this doesn't preclude other problems, Clever Hans was in good faith too, after all.

This doesn't look like real science, it's just "dog moms" enjoying a fun hobby, YouTube videos aren't evidence of anything.

Early stage science [often looks like](#) "messing around", before theory and rigor is built. Telling what "messing around" is likely to be fruitful is a meta-rational skill, and it can be done, somewhat.

YouTube (and especially Instagram and TikTok) videos go against usual aesthetic sensibilities of what evidence looks like, but that's not a reason to discount them completely. They still constitute a lot of evidence, albeit the kind that is weak and easy to misinterpret. If we're to be good rationalists, we can't exclude the messy parts of the world and then expect to arrive at a useful worldview.

The work on the ground being done by laymen may be a blessing in disguise - I won't be surprised if taking a formal and procedural approach to training would "ruin the magic" and lead to poor results. Especially if mutual misinterpretation is an important part of the mechanism. I hope that given the number of participants in the study and the amount of data (every interaction recorded) and well-designed experiments will together let us separate signal from the noise.

Koko

So what, you ask, some apes have been taught sign language and they produced rudimentary syntax as well.

For starters you wouldn't predict dogs to be capable of the same, and it's significant to see that ability given their evolutionary distance from us (even with selective breeding pressure from domestication).

Most of the ape research was done in the 70s, and it is, well, very 70s. Those things aren't known for being well-run or replicating well. And it was done with sign language, perhaps buttons are much more conducive to language acquisition. Since the 70s craze, it apparently became unfashionable, and nothing new happened for decades. To this day any conversation on animal language is about Koko (who died in 2018) and the parrot who said "love you, bye" before dying in 2007. Utter stagnation.

What we have is something new, orders of magnitude easier to study and reproduce (how many of you have gorillas at home?), massive PR potential, modern tech that allows you to have cameras running 24/7 to preclude criticism. It started with an outsider to the field, who wasn't conceptionally limited by prior art. And it's accessible to regular people, potentially revolutionizing our relationship with our pets.

This raises the natural question: what if you gave an ape the buttons, and taught it from childhood, and put parent-level effort into it, not "70s research"-level effort? Perhaps the answer would surprise us.

[EDIT: Exactly this has been [attempted with bonobos](#), but unfortunately little data is available and the experiment disintegrated over human drama. Read the linked comment for details and a few existing videos]

Honorable Mentions

Billi

A cat who initially became famous for pressing "MAD MAD MAD" at a slightest inconvenience, but she has *meollowed* out a bit.

"Mad" A Short Film Starring Billi the Cat

Mom's Choice - appears to ask "mom" what she wants to play. Normally mom is the one asking her what toy she wants to play.

Imposter - an important video. In it, Billi repeatedly refuses food, despite the owner practically railroading her. With simple conditioning you could expect "want food hm?" -> "yes food" from the cat, without understanding of what "yes" or "no" is. But if all of this is just clever conditioning, why did this video happen?

Others

Talking dog Tiktok compilation | Flambo the dog

Dog Communicates with Human by Talking Buttons PT2

Izzy the talking dog uses recordable buttons to help her sister Luna

What did we do with the word buttons? 😊 | Pharaby the Talking Dog

HowTheyTalk.org

Community of people trying to replicate this, ran by the people running the UCSD study.

Maybe later

"Maybe later" at substack for trying in vain to tell others about this, only to be summarily ignored. I got your back, buddy.

-
1. aka blindly trusted stereotypes ↵
 2. Normie blindspot does exist in the community, but that's kind of obvious and expected, and should be a separate article. ↵
 3. Possibly, they're not clear on the number of participants. I personally find it hard to believe that there would be more than about a hundred. ↵
 4. In the layman sense of "tool for communication". Philosophical discussions about the exact border between non-language communication and "true language" aren't really interesting to me. Duck typing, etc. ↵

Curated conversations with brilliant rationalists

Since August 2020 I've been recording conversations with brilliant and insightful rationalists, effective altruists (and people adjacent to or otherwise connected somehow to those communities). If you're an avid reader of this site, I suspect you will recognize many of the names of those I've spoken to.

Since I suspect some LessWrong readers will appreciate these conversations, here is a curated list with links, organized by the LessWrong relevant topics we cover in each conversation. All of these conversations can also be found by searching for "[Clearer Thinking](#)" in just about any podcast app. If there are other people you'd like to see me record conversations with, please nominate them in the comments! The format is that I invite each guest to bring 4 or 5 "ideas that matter" that they are excited to talk about, and then the aim is to have a fun, intellectual discussion of those ideas.

Rationality

[Lines of Retreat and Incomplete Maps with Anna Salamon](#)

What does it mean to leave lines of retreat in social contexts? How can we make sense of the current state of the world? What happens when we run out of map? How does the book *Elephant in the Brain* apply to the above questions?

[Rationality Education and Dating with Jacob Falkovich](#)

What's the best way to teach rationality? How do you communicate rationalist principles to people who aren't already interested in thinking more clearly? What has COVID taught us about how people typically make decisions and think about problems? Where and how can the rationalist community improve? Does rationalism have anything to say about (for example) exercise, spirituality, art, or other parts of the human experience that aren't typically addressed by rationalists? What are some positive aspects of social media (especially Twitter)? What's going on with recent dating trends? Has dating gotten harder in recent years? How many people does it take to make a pencil? Is there a case to be made for anti-antinatalism?

[Scout and Soldier Mindsets with Julia Galef](#)

What are "scout" and "soldier" mindsets? How can we have productive disagreements even when one person isn't in scout mindset? Is knowing about good rationality habits sufficient to reason well? When do we naturally tend to be in scout mindset or soldier mindset? When is each mindset beneficial or harmful? Are humans "rationally irrational"? What are the two different types of confidence? What are some practical strategies for shifting our mindset in the moment from soldier to scout?

[Comfort Languages and Nuanced Thinking with Kat Woods](#)

What's the best way to help someone who's going through a difficult situation? What are the four states of distress? What are "comfort languages"? How can we introduce more nuance into our everyday thinking habits? When gathering information and forming opinions, how do you know who to trust? What's the difference between intelligence and wisdom?

Aging/Longevity

[History and Longevity with Will Eden](#)

What are the benefits of studying history? How do we find useful historical analyses? Can learning about history save us from repeating it? Is America decaying as a nation, empire, and/or leading world power? Generally speaking, what causes empires to fail? Is the aging and decay experienced by organic bodies analogous to the aging and decay experienced by an empire (or by any complex system, for that matter)? What are all the reasons organisms age, decay, and die? What are the most promising avenues of exploration in longevity research? What kind of stressors on our bodies are beneficial? How accurate is the efficient market hypothesis? What kinds of catalysts force a market to value assets at their "intrinsic" value? How rational are markets?

Artificial Intelligence

[AI Safety and Solutions with Robert Miles](#)

Why is YouTube such a great way to communicate research findings? Why is AI safety (or alignment) a problem? Why is it an important problem? Why is the creation of AGI (artificial general intelligence) existentially risky for us? Why is it so hard for us to specify what we want in utility functions? What are some of the proposed strategies (and their limitations) for controlling AGI? What is instrumental convergence? What is the unilateralist's curse?

[Superintelligence and Consciousness with Roman Yampolskiy](#)

What is superintelligence? Can a superintelligence be controlled? Why aren't people (especially academics, computer scientists, and companies) more worried about superintelligence alignment problems? Is it possible to determine whether or not an AI is conscious? Do today's neural networks experience some form of consciousness? Are humans general intelligences? How do artificial superintelligence and artificial general intelligence differ? What sort of threats do malevolent actors pose over and above those posed by the usual problems in AI safety?

Learning

[Antagonistic Learning and Civilization with Duncan Sabien](#)

Why do "antagonistic" teachers exist in popular culture but not in the classroom? What happens to student outcomes when "antagonistic" learning is implemented in real classrooms? What is the Field Theory of Parenting? What are things that we can do for others but can't do for ourselves? How can we notice and utilize costly

and unfakeable signals? What is the core definition of civilization? How can we influence others ethically? Is explicit communication always better than implicit?

Learning and Governance with Emerson Spartz

What's the best way to learn? Why is learning how to learn "the most important skill"? When should we explore, and when should we exploit? What are the merits and demerits of various models of governance? How should we think about the problems around free speech?

Knowledge Management and Deugenesis with Jeremy Nixon

What is "The Index"? What are some benefits of externally compiling and organizing one's knowledge? When is spaced repetition useful? How can we co-opt our visual systems to boost memory? Would we all be more interested in producing an external personal knowledgebase if we could feel on a visceral level how much information is constantly being forgotten? How and when should we move up and down the ladder of abstraction? What sorts of problems can be solved by simulation? What is a *generative* model (as opposed to a *predictive* model)? How can constraints improve creativity? How useful are credentials as a guide to how much a person knows and whether or not a person is "allowed" to have an opinion on a topic? What do credentials actually signal about a person? What are "fox" and "hedgehog" thinking? What is *deugenesis*?

Cryptocurrency

Crypto Pros and Cons with Sam Bankman-Fried

What's the current state of cryptocurrency? What are the good and bad aspects of crypto? To what extent have the promises of crypto panned out? How do blockchain and cryptocurrency even work anyway? What are "proof of work" and "proof of stake"? What are the differences between Bitcoin and Ethereum? What sorts of transactions are made easy or possible by the blockchain that are difficult or impossible to perform with traditional currencies? What are non-fungible tokens (NFTs)? What (if anything) prevents people from doing nefarious things with cryptocurrencies? What are some of the exciting, positive things coming up on the crypto horizon?

Philosophy

Aesthetics and Polyamory with Sam Rosen

How can we improve art museums? Does aesthetics need something equivalent to the effective altruism movement? What is steel-aliening? What are the most important social skills to learn, and how can we learn them? Can anybody become polyamorous? What does it take to succeed in a polyamorous relationships? Why do societies decay over time?

Utilitarianism and Its Flavors with Nick Beckstead

What is utilitarianism? And what are the different flavors of utilitarianism? What are some alternatives to utilitarianism for people that find it generally plausible

but who can't stomach some of its counterintuitive conclusions? For the times when people do use utilitarianism to make moral decisions, when is it appropriate to perform actual calculations (as opposed to making estimations or even just going with one's "gut")? And what is "utility" anyway?

Moral Discourse and the Value of Philosophy with Ronny Fernandez

What is normative hedonism? What's the difference between wanting something and *wanting* to want something? Should we only care about the experiences of conscious beings? What's wrong with moral discourse? Does philosophy ever actually make progress, or is it still only discussing the things that were discussed a thousand years ago? What is (or should be) the role of intuition in philosophy? Why should people study philosophy (especially as opposed to other disciplines)? What can we do to create more rationality or systematic wisdom in the world? How can we disagree better?

Life Experiments and Philosophical Thinking with Arden Koehler

What is 80,000 Hours? What sorts of people should become entrepreneurs? How can you run cheap experiments on yourself? What are some beneficial modes of philosophical thinking?

Meditation / Enlightenment

Meditation and Ontology with Daniel Ingram

Why should we meditate? What are the typical developmental stages as one progresses along the contemplative path? What does it mean to "hold an ontology loosely"? Are some meditative techniques inappropriate for some practitioners? Are there risks associated with meditation?

Enlightenment and Sex Work with Aella

What is enlightenment? What are the different kinds or definitions of enlightenment? What was Aella's religious upbringing like, and why did she lose her faith? How did Aella get into sex work, and what has her career as a sex worker been like? How do we ask great questions, and what is Askhole?

Taboo beliefs

Death and Story-Telling with A.J. Jacobs

Are there more meaningful and ethical ways of honoring the dead than our traditional rituals? Why is it useful to adopt probabilistic thinking in our everyday lives? What sorts of things do we value intrinsically (i.e., that we would value even if they had no other positive benefits)? What do stories do well and not so well?

Preference Falsification and Postmodernism with Michael Vassar

How much preference falsification is occurring in society? What's the difference between conflict theory and mistake theory? Why is postmodernism useful to understand?

[Education and Charity with Uri Bram](#)

Are universities a cult? Do charitable interventions like de-worming work? How much should we trust the conclusion of well-respected charity evaluators like GiveWell?

Self-improvement

[Self-Improvement and Research Ethics with Rob Wiblin](#)

What are the best strategies for improving ourselves? How are line managers useful? Why does Rob prefer long-form content for the 80,000 Hours podcast? What are the sorts of things humans value and why? In what ways do research ethics considerations fail to achieve their stated objectives? Why are prediction markets useful?

[Explanatory Depth and Growth Mindset with Daniel Greene](#)

What is the illusion of explanatory depth? Are there forms of debate or dialogue that actually help people to change their minds (instead of stacking the incentives such that people feel forced to harden and defend their views)? What is epistemic "debt"? Should people avoid having opinions on things where they haven't thought deeply and carefully about all of the relevant considerations? How does one choose which experts to trust? What is "growth mindset"? How can social science be used to do good in the world?

Note that the above is a partial list of recordings, focussing on just those people and topics most connected to LessWrong. Some other relevant people that I've already recorded with, but haven't yet released the episodes for include: Kaj Sotala, Divia Eden, Stefan Schubert, Alyssa Vance, Satvik Beri, and Joe Carlsmith. Please let me know who else I should record with! :)

Small and Vulnerable

Anyone who is dedicating the majority of their time or money to Effective Altruism needs to ask themselves why. Why not focus on enjoying life and spending your time doing what you love most? Here is my answer:

I have a twin sister but neither of us had many other friends growing up. From second to fifth grade we had none. From sixth to eighth we had one friend. As you might guess I was bullied quite badly. Multiple teachers contributed to this. Despite having no friends my parents wanted us to be normal. They pressured me to play sports with the boys in the neighborhood. I was unable to play with an acceptable level of skill and was not invited to the games anyway. But we were still forced to go 'play outside' after school. We had to find ways to kill time. Often we literally rode our bicycles in a circle in a parking lot. We were forced to 'play outside' for hours most days and even longer on weekends. I was not even allowed to bring a book outside though sometimes I would hide them outside at night and find them the next day. Until high school, I had no access to the internet. After dinner, I could watch TV, read and play video games. These were the main sources of joy in my childhood.

Amazingly my mom made fun of her children for being weirdos. My sister used to face a wall and stim with her fingers when she was overwhelmed. For some reason, my mom interpreted this as 'OCD'. So she made up a song titled 'OCD! Do you mean me?' It had several verses! This is just one, especially insane, example.

My dad liked to 'slap me around. He usually did not hit me very hard but he would slap me in the face all the time. He also loved to call me 'boy' instead of my name. He claims he got this idea from Tarzan. It took me years to stop flinching when people raised their hands or put them anywhere near my face. I have struggled with gender since childhood. My parents did not tolerate even minor gender nonconformity like growing my hair out. I would get hit reasonably hard if I insisted on something as 'extreme' as crossing my legs 'like a girl in public. I recently started HRT and already feel much better. My family is a lot of the reason I delayed transitioning.

If you go by the [checklist](#) I have quite severe ADHD. 'Very often' seemed like an understatement for most of the questions. My ADHD was untreated until recently. I could not focus on school or homework so trying to do my homework took way too much time. I was always in trouble in school and considered a very bad student. It definitely hurts when authority figures constantly, and often explicitly, treat you like a fuck up and a failure who can't be trusted. But looking back it seems amazing I was considered such a bad student. I love most of the subjects you study in school! When I finally got access to the internet I spent hours per day reading Wikipedia articles. I still spend a lot of time listening to lectures on all sorts of subjects, especially history. Why were people so cruel to a little child who wanted to learn things?

Luckily things improved in high school. Once I had more freedom and distance from my parents my social skills improved a huge amount. In high school, I finally had internet access which helped an enormous amount. My parents finally connected our computer at home to the internet because they thought my sister and I needed it for school. I also had access to the computers in the high school library. By my junior year in high school, I was not really unpopular. Ironically my parent's overbearing pressure to be a 'normal kid' probably prevented me from having a social life until I got a little

independence. Sadly I was still constantly in trouble in school throughout my high school years.

The abuse at home was very bad. But, to be honest, the absolute worst part of my childhood and adolescence was the constant sleep deprivation. Even at thirty years old I cannot handle getting up early; I rarely wake before nine-thirty. A year ago I briefly had to be awake at six-thirty for work. I felt terrible all day and could not think straight. When I was younger I had an even stronger need to sleep in but I had to be in school before eight. People were amazed at my ability to fall into a deep sleep in the middle of a loud classroom. Unless someone woke me up I would just stay asleep at my desk. This was a horrible experience and surely terrible for my brain. I got a break from this torment during the summers but I didn't really escape until I made it to college.

Obviously, I was an outlier in many respects. But many people are outliers in some important respects. They still deserve an environment that is healthy and lets them flourish. I wanted to learn all sorts of things. But instead of helping me, the school system tortured me and permanently damaged my brain. No one deserves to be treated like that.

We should not frame this in terms of my parents being aberrations. I live in the United States. Many groups here normalize far more extreme repression and physical punishment. In some subcultures, my parent's behavior is considered unacceptable. But much of what happened to me is still normalized. Even supposedly liberal parents are often terrible to trans children. Society isn't going to stop sleep-depriving children anytime soon. And there are many people being severely mistreated in very different circumstances.

I cannot get my childhood back, can't go back in time and transition earlier, and if my brain was harmed the damage is permanent. Whatever other traumas I have won't fully heal. But I eventually got out. There are millions of people in prison, trapped in abusive nursing homes, or starving in Yemen. There are many more animals on farms. Those people haven't escaped yet and it is unclear they will ever escape to somewhere safe. Society never should have normalized what happened to me and we shouldn't normalize what is happening to them. This is an emergency.

When I was small and vulnerable I needed help. For the most part, no help came. I was forced to stew in boredom and misery until I grew bigger, stronger, and accorded more respect. It is always hard to compare experiences. But I know what it's like to spend about a decade miserable, knowing you are being mistreated and being unable to defend yourself. Maybe one day I will again be unable to defend myself because I am sick or in prison. But for now, I am relatively healthy and free. I cannot just abandon the people and animals who are still trapped. Every day I try to imagine them somehow watching me and I ask whether they would think I forgot them. I hope I never forget. I hope my actions always show I have forgotten neither my past nor their present.

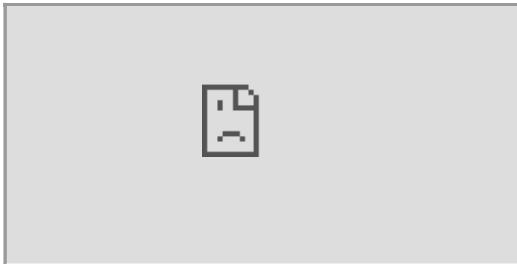
This post also [appeared on my blog](#).

Finite Factored Sets

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

This is the edited transcript of a talk introducing finite factored sets. For most readers, it will probably be the best starting point for learning about factored sets.

Video:



(Lightly edited) slides: <https://intelligence.org/files/Factored-Set-Slides.pdf>

1. Short Combinatorics Talk

1m. Some Context

Scott: So I want to start with some context. For people who are not already familiar with my work:

- My main motivation is to reduce existential risk.
- I try to do this by trying to figure out how to [align](#) advanced artificial intelligence.
- I try to do *this* by trying to become [less confused](#) about intelligence and optimization and agency and various things in that cluster.
- My main strategy here is to develop a theory of agents that are [embedded](#) in the environment that they're optimizing. I think there are a lot of open hard problems around doing this.
- This leads me to do a bunch of weird math and philosophy. This talk is going to be an example of some weird math and philosophy.

For people who are already familiar with my work, I just want to say that according to my personal aesthetics, the subject of this talk is about as exciting as [Logical Induction](#), which is to say I'm really excited about it. And I'm really excited about this audience; I'm excited to give this talk right now.

1t. Factoring the Talk

This talk can be split into 2 parts:

- Part 1: a short pure-math combinatorics talk.

I suspect that if I were better, I would instead be giving a short pure-math category theory talk; but I'm trained as a combinatorialist, so I'm giving a combinatorics talk upfront.

- Part 2: a more applied and philosophical main talk.

This talk can also be split into 4 parts differentiated by color: **Motivation**, **Table of Contents**, **Main Body**, and **Examples**. Combining these gives us 8 parts (some of which are not contiguous):

Part 1: Short Talk		Part 2: The Main Talk	
Motivation	1m. Some Context	2m.	The Pearlian Paradigm
ToC	1t. Factoring the Talk	2t.	We Can Do Better
Body	1b. Set Partitions , etc.	2b.	Time and Orthogonality , etc.
Examples	1e. Enumerating Factorizations	2e.	Game of Life , etc.

1b. Set Partitions

All right. Here's some background math:

- A **partition** of a set S is a set X of non-empty subsets of S , called **parts**, such that for each $s \in S$ there exists a unique part in X that contains s .
- Basically, a partition of S is a way to view S as a disjoint union. We have parts that are disjoint from each other, and they union together to form S .
- We'll write $\text{Part}(S)$ for the set of all partitions of S .
- We'll say that a partition X is **trivial** if it has exactly one part.
- We'll use bracket notation, $[s]_X$, to denote the unique part in X containing s . So this is like the equivalence class of a given element.
- And we'll use the notation $s \sim_X t$ to say that two elements s and t are in the same part in X .

You can also think of partitions as being like variables on your set S . Viewed in that way, the values of a partition X correspond to which part an element is in.

Or you can think of X as a *question* that you could ask about a generic element of S . If I have an element of S and it's hidden from you and you want to ask a question about it, each possible question corresponds to a partition that splits up S according to the different possible answers.

We're also going to use the [lattice structure](#) of partitions:

- We'll say that $X \geq_S Y$ (X is finer than Y , and Y is coarser than X) if X makes all of the distinctions that Y makes (and possibly some more distinctions), i.e., if for all $s, t \in S$, $s \sim_X t$ implies $s \sim_Y t$. You can break your set S into parts, Y , and then break it into smaller parts, X .
- $X \vee Y$ (the common refinement of X and Y) is the coarsest partition that is finer than both X and Y . This is the unique partition that makes all of the distinctions that either X or Y makes, and no other distinctions. This is well-defined, which I'm not going to show here.

Hopefully this is mostly background. Now I want to show something new.

1b. Set Factorizations

A **factorization** of a set S is a set B of nontrivial partitions of S , called **factors**, such that for each way of choosing one part from each factor in B , there exists a unique element of S in the intersection of those parts.

So this is maybe a little bit dense. My short tagline of this is: "A factorization of S is a way to view S as a product, in the exact same way that a partition was a way to view S as a disjoint union."

If you take one definition away from this first talk, it should be the definition of factorization. I'll try to explain it from a bunch of different angles to help communicate the concept.

If $B = \{b_0, \dots, b_n\}$ is a factorization of S , then there exists a bijection between S and $b_0 \times \dots \times b_n$ given by $s \mapsto ([s]_{b_0}, \dots, [s]_{b_n})$. This bijection comes from sending an element of S to the tuple consisting only of parts containing that element. And as a consequence of this bijection, $|S| = \prod_{b \in B} |b|$.

So we're really viewing S as a product of these individual factors, with no additional structure.

Although we won't prove this here, something else you can verify about factorizations is that all of the parts in a factor have to be of the same size.

We'll write $\text{Fact}(S)$ for the set of all factorizations of S , and we'll say that a **finite factored set** is a pair (S, B) , where S is a finite set and $B \in \text{Fact}(S)$.

Note that the relationship between S and B is somewhat loopy. If I want to define a factored set, there are two strategies I could use. I could first introduce the S , and break it into factors. Alternatively, I could first introduce the B . Any time I have a finite collection of finite sets B , I can take their product and thereby produce an S , modulo the degenerate case where some of the sets are empty. So S can just be the product of a finite collection of arbitrary finite sets.

To my eye, this notion of factorization is extremely natural. It's basically the multiplicative analog of a set partition. And I really want to push that point, so here's another attempt to push that point:

A partition is a set X of non-empty subsets of S such that the obvious function from the disjoint union of the elements of X to S is a bijection.	A factorization is a set B of non-trivial partitions of S such that the obvious function to the product of the elements of B from S is a bijection.
---	---

I can take a slightly modified version of the partition definition from before and dualize a whole bunch of the words, and get out the set factorization definition.

Hopefully you're now kind of convinced that this is an extremely natural notion.

Andrew Critch: Scott, in one sense, you're treating "subset" as dual to partition, which I think is valid. And then in another sense, you're treating "factorization" as dual to partition. Those are both valid, but maybe it's worth talking about the two kinds of duality.

Scott: Yeah. I think what's going on there is that there are two ways to view a partition. You can view a partition as "that which is dual to a subset," and you can also view a partition as something that is built up out of subsets. These two different views do different things when you dualize.

Ramana Kumar: I was just going to check: You said you can start with an arbitrary B and then build the S from it. It can be literally any set, and then there's always an S ...

Scott: If none of them are empty, yes, you could just take a collection of sets that are kind of arbitrary elements. And you can take their product, and you can identify with each of the elements of a set the subset of the product that projects on to that element.

Ramana Kumar: Ah. So the S in that case will just be tuples.

Scott: That's right.

Brendan Fong: Scott, given a set, I find it very easy to come up with partitions. But I find it less easy to come up with factorizations. Do you have any tricks for...?

Scott: For that, I should probably just go on to the examples.

Joseph Hirsh: Can I ask one more thing before you do that? You allow factors to have one element in them?

Scott: I said "nontrivial," which means it does not have one element.

Joseph Hirsh: "Nontrivial" means "not have one element, and not have no elements"?

Scott: No, the empty set has a partition (with no parts), and I will call that nontrivial. But the empty set thing is not that critical.

I'm now going to move on to some examples.

1e. Enumerating Factorizations

Exercise! What are the factorizations of the set $\{0, 1, 2, 3\}$?

Spoiler space:

.

.

First, we're going to have a kind of trivial factorization:

$$\{ \{ \{ 0 \}, \{ 1 \}, \{ 2 \}, \{ 3 \} \} \} \quad \begin{array}{cccc} 0 & 1 & 2 & 3 \\ \hline \end{array}$$

We only have one factor, and that factor is the discrete partition. You can do this for any set, as long as your set has at least two elements.

Recall that in the definition of factorization, we wanted that for each way of choosing one part from each factor, we had a unique element in the intersection of those parts. Since we only have one factor here, satisfying the definition just requires that for each way of choosing one part from the discrete partition, there exists a unique element that is in that part.

And then we want some less trivial factorizations. In order to have a factorization, we're going to need some partitions. And the product of the cardinalities of our partitions are going to have to equal the cardinality of our set S , which is 4.

The only way to express 4 as a nontrivial product is to express it as 2×2 . Thus we're looking for factorizations that have 2 factors, where each factor has 2 parts.

We noted earlier that all of the parts in a factor have to be of the same size. So we're looking for 2 partitions that each break our 4-element set into 2 sets of size 2.

So if I'm going to have a factorization of $\{0, 1, 2, 3\}$ that isn't this trivial one, I'm going to have to pick 2 partitions of my 4-element set that each break the set into 2 parts of size 2. And there are 3 partitions of a 4-element sets that break it up into 2 parts of size 2. For each way of choosing a pair of these 3 partitions, I'm going to get a factorization.

$$\{ \{ \{ 0, 1 \}, \{ 2, 3 \} \}, \{ \{ 0, 2 \}, \{ 1, 3 \} \} \}$$

0	1
2	3

$\{ \{ 0, 1 \}, \{ 2, 3 \} \},$
 $\{ \{ \{ 0, 3 \}, \{ 1, 2 \} \} \}$

0	1
3	2

$\{ \{ 0, 2 \}, \{ 1, 3 \} \},$
 $\{ \{ \{ 0, 3 \}, \{ 1, 2 \} \} \}$

0	2
3	1

So there will be 4 factorizations of a 4-element set.

In general you can ask, "How many factorizations are there of a finite set of size n?". Here's a little chart showing the answer for $n \leq 25$:

$ S $	$ \text{Fact}(S) $
0	1
1	1
2	1
3	1
4	4
5	1
6	61
7	1
8	1681
9	5041
10	15121
11	1
12	13638241
13	1
14	8648641
15	1816214401
16	181880899201
17	1
18	45951781075201
19	1
20	3379365788198401
21	1689515283456001
22	14079294028801
23	1
24	4454857103544668620801
25	538583682060103680001

You'll notice that if n is prime, there will be a single factorization, which hopefully makes sense. This is the factorization that only has one factor.

A very surprising fact to me is that this sequence did not show up on [OEIS](#), which is this database that combinatorialists use to check whether or not their sequence has been studied before, and to see connections to other sequences.

To me, this just feels like the multiplicative version of the [Bell numbers](#). The Bell numbers count how many partitions there are of a set of size n . It's sequence number 110 on OEIS out of over 300,000;

and this sequence just doesn't show up at all, even when I tweak it and delete the degenerate cases and so on.

I am very confused by this fact. To me, factorizations seem like an extremely natural concept, and it seems to me like it hasn't really been studied before.

This is the end of my short combinatorics talk.

Ramana Kumar: If you're willing to do it, I'd appreciate just stepping through one of the examples of the factorizations and the definition, because this is pretty new to me.

Scott: Yeah. Let's go through the first nontrivial factorization of $\{0, 1, 2, 3\}$:

$$\{ \{0, 1\}, \{2, 3\} \},$$
$$\{ \{0, 2\}, \{1, 3\} \}$$

0	1
2	3

In the definition, I said a factorization should be a set of partitions such that for each way of choosing one part from each of the partitions, there will be a unique element in the intersection of those parts.

Here, I have a partition that's separating the small numbers from the large numbers: $\{\{0, 1\}, \{2, 3\}\}$. And I also have a partition that's separating the even numbers from the odd numbers: $\{\{0, 2\}, \{1, 3\}\}$.

And the point is that for each way of choosing either "small" or "large" and also choosing "even" or "odd", there will be a unique element of S that is the conjunction of these two choices.

In the other two nontrivial factorizations, I replace either "small and large" or "even and odd" with an "inner and outer" distinction.

David Spivak: For partitions and for many things, if I know the partitions of a set A and the partitions of a set B , then I know some partitions of $A + B$ (the disjoint union) or I know some partitions of $A \times B$. Do you know any facts like that for factorizations?

Scott: Yeah. If I have two factored sets, I can get a factored set over their product, which sort of disjoint-unions the two collections of factors. For the additive thing, you're not going to get anything like that because prime sets don't have any nontrivial factorizations.

All right. I think I'm going to move on to the main talk.

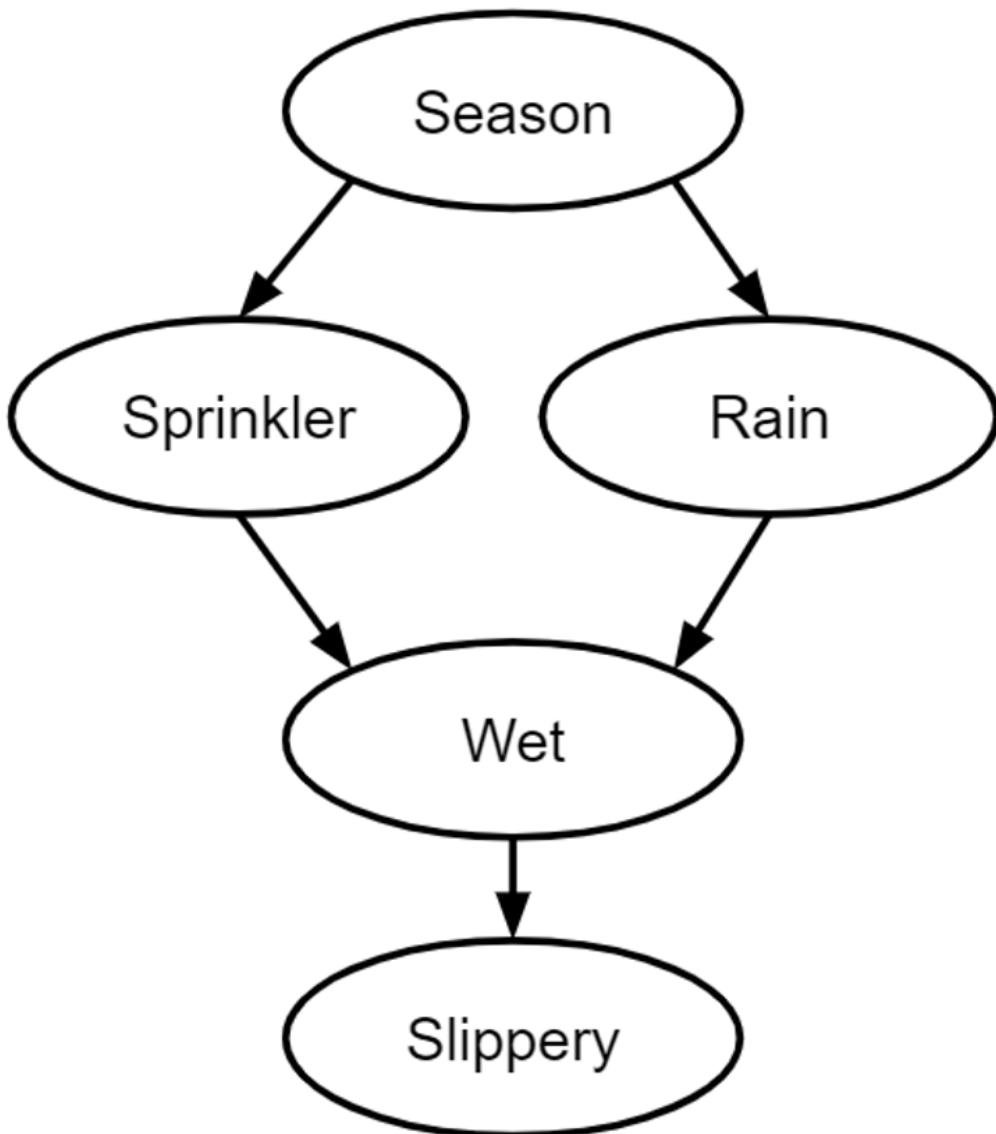
2. The Main Talk (It's About Time)

2m. The Pearlian Paradigm

We can't talk about time without talking about [Pearlian causal inference](#). I want to start by saying that I think the Pearlian paradigm is great. This buys me some crackpot points, but I'll say it's the best thing to happen to our understanding of time since Einstein.

I'm not going to go into all the details of Pearl's paradigm here. My talk will not be technically dependent on it; it's here for motivation.

Given a collection of variables and a joint probability distribution over those variables, Pearl can infer causal/temporal relationships between the variables. (In this talk I'm going to use "causal" and "temporal" interchangeably, though there may be more interesting things to say here philosophically.)



Pearl can infer temporal data from statistical data, which is going against the adage that "correlation does not imply causation." It's like Pearl is taking the combinatorial structure of your correlation and

using that to infer causation, which I think is just really great.

Ramana Kumar: I may be wrong, but I think this is false. Or I think that that's not all Pearl needs—just the joint distribution over the variables. Doesn't he also make use of intervention distributions?

Scott: In the theory that is described in chapter two of the book *Causality*, he's not really using other stuff. Pearl builds up this bigger theory elsewhere. But you have some strong ability, maybe assuming simplicity or whatever (but not assuming you have access to extra information), to take a collection of variables and a joint distribution over those variables, and infer causation from correlation.

Andrew Critch: Ramana, it depends a lot on the structure of the underlying causal graph. For some causal graphs, you can actually recover them uniquely with no interventions. And only assumptions with zero-measure exceptions are needed, which is really strong.

Ramana Kumar: Right, but then the information you're using is the graph.

Andrew Critch: No, you're not. Just the joint distribution.

Ramana Kumar: Oh, okay. Sorry, go ahead.

Andrew Critch: There exist causal graphs with the property that if nature is generated by that graph and you don't know it, and then you look at the joint distribution, you will infer with probability 1 that nature was generated by that graph, without having done any interventions.

Ramana Kumar: Got it. That makes sense. Thanks.

Scott: Cool.

I am going to (a little bit) go against this, though. I'm going to claim that Pearl *is* kind of cheating when making this inference. The thing I want to point out is that in the sentence "Given a collection of variables and a joint probability distribution over those variables, Pearl can infer causal/temporal relationships between the variables.", the words "Given a collection of variables" are actually hiding a lot of the work.

The emphasis is usually put on the joint probability distribution, but Pearl is not inferring temporal data from statistical data alone. He is inferring temporal data from statistical data **and factorization data:** how the world is broken up into these variables.

I claim that this issue is also entangled with a failure to adequately handle abstraction and determinism. To point at that a little bit, one could do something like say:

"Well, what if I take the variables that I'm given in a Pearlian problem and I just forget that structure? I can just take the product of all of these variables that I'm given, and consider the space of all partitions on that product of variables that I'm given; and each one of those partitions will be its own variable. And then I can try to do Pearlian causal inference on this big set of all the variables that I get by forgetting the structure of variables that were given to me."

And the problem is that when you do that, you have a bunch of things that are deterministic functions of each other, and you can't actually infer stuff using the Pearlian paradigm.

So in my view, this cheating is very entangled with the fact that Pearl's paradigm isn't great for handling abstraction and determinism.

2t. We Can Do Better

The main thing we'll do in this talk is we're going to introduce an alternative to Pearl that does not rely on factorization data, and that therefore works better with abstraction and determinism.

Where Pearl was given a collection of variables, we are going to just consider all partitions of a given set. Where Pearl infers a directed acyclic graph, we're going to infer a finite factored set.

In the Pearlian world, we can look at the graph and read off properties of time and orthogonality/independence. A directed path between nodes corresponds to one node being before the other, and two nodes are independent if they have no common ancestor. Similarly, in our world, we will be able to read time and orthogonality off of a finite factored set.

(Orthogonality and independence are pretty similar. I'll use the word "orthogonality" when I'm talking about a combinatorial notion, and I'll use "independence" when I'm talking about a probabilistic notion.)

In the Pearlian world, d -separation, which you can read off of the graph, corresponds to conditional independence in all probability distributions that you can put on the graph. We're going to have a fundamental theorem that will say basically the same thing: conditional orthogonality corresponds to conditional independence in all probability distributions that we can put on our factored set.

In the Pearlian world, d -separation will satisfy the compositional graphoid axioms. In our world, we're just going to satisfy the compositional semigraphoid axioms. The fifth graphoid axiom is one that I claim you shouldn't have even wanted in the first place.

Pearl does causal inference. We're going to talk about how to do temporal inference using this new paradigm, and infer some very basic temporal facts that Pearl's approach can't. (Note that Pearl can also sometimes infer temporal relations that we can't—but only, from our point of view, because Pearl is making additional factorization assumptions.)

And then we'll talk about a bunch of applications.

Pearl	This Talk
A Given Collection of Variables	All Partitions of a Given Set
Directed Acyclic Graph	Finite Factored Set
Directed Path Between Nodes	"Time"
No Common Ancestor	"Orthogonality"
d -Separation	"Conditional Orthogonality"
Compositional Graphoid	Compositional Semigraphoid
d -Separation \leftrightarrow Conditional Independence	The Fundamental Theorem
Causal Inference	Temporal Inference
Many Many Applications	Many Many Applications

Excluding the motivation, table of contents, and example sections, this table also serves as an outline of the two talks. We've already talked about set partitions and finite factored sets, so now we're going to talk about time and orthogonality.

2b. Time and Orthogonality

I think that if you capture one definition from this second part of the talk, it should be this one. Given a finite factored set as context, we're going to define the history of a partition.

Let $F = (S, B)$ be a finite factored set. And let $X, Y \in \text{Part}(S)$ be partitions of S .

The **history** of X , written $h^F(X)$, is the smallest set of factors $H \subseteq B$ such that for all $s, t \in S$, if $s \sim_b t$ for all $b \in H$, then $s \sim_X t$.

The history of X , then, is the smallest set of factors H —so, the smallest subset of B —such that if I take an element of S and I hide it from you, and you want to know which part in X it is in, it suffices for me to tell you which part it is in within each of the factors in H .

So the history H is a set of factors of S , and knowing the values of all the factors in H is sufficient to know the value of X , or to know which part in X a given element is going to be in. I'll give an example soon that will maybe make this a little more clear.

We're then going to define **time** from history. We'll say that X is **weakly before** Y , written $X \leq^F Y$, if $h^F(X) \subseteq h^F(Y)$. And we'll say that X is **strictly before** Y , written $X <^F Y$, if $h^F(X) \subset h^F(Y)$.

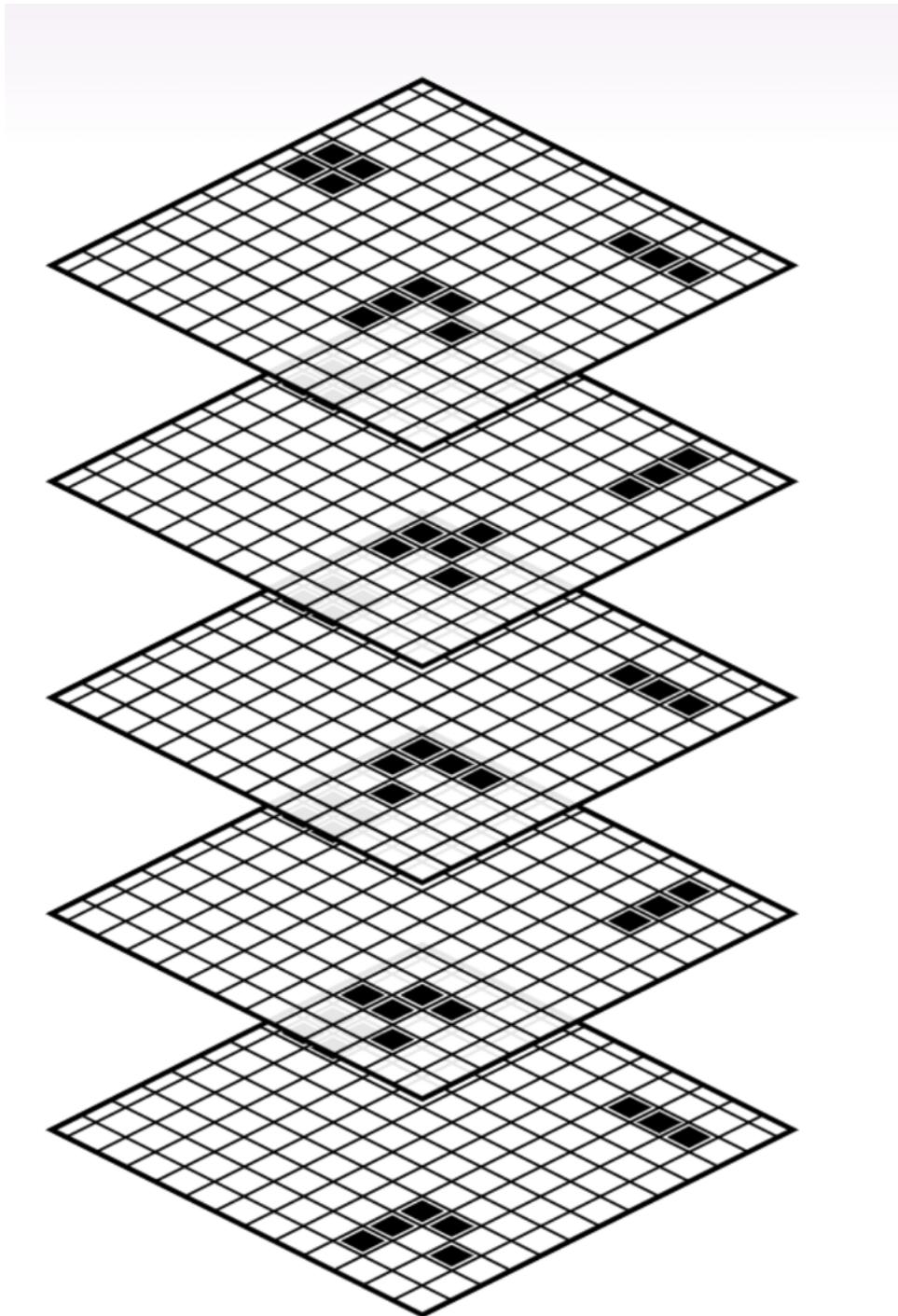
One analogy one could draw is that these histories are like the past light cones of a point in spacetime. When one point is before another point, then the backwards light cone of the earlier point is going to be a subset of the backwards light cone of the later point. This helps show why "before" can be like a subset relation.

We're also going to define orthogonality from history. We'll say that two partitions X and Y are **orthogonal**, written $X \perp^F Y$, if their histories are disjoint: $h^F(X) \cap h^F(Y) = \{\}$.

Now I'm going to go through an example.

2e. Game of Life

Let S be the set of all Game of Life computations starting from an $[-n, n] \times [-n, n]$ board.



Let $R = \{(r, c, t) \in \mathbb{Z}^3 \mid 0 \leq t \leq n, |r| \leq n - t, |c| \leq n - t\}$ (i.e., cells computable from the initial $[-n, n] \times [-n, n]$ board). For $(r, c, t) \in R$, let $\ell(r, c, t) \subseteq S$ be the set of all computations such that the cell at row r and column c is alive at time t .

(Minor footnote: I've done some small tricks here in order to deal with the fact that the Game of Life is normally played on an infinite board. We want to deal with the finite case, and we don't want to worry about boundary conditions, so we're only going to look at the cells that are uniquely determined by the initial board. This means that the board will shrink over time, but this won't matter for our example.)

S is the set of all Game of Life computations, but since the Game of Life is deterministic, the set of all computations is in bijective correspondence with the set of all initial conditions. So $|S| = 2^{(2n+1)^2}$, the number of initial board states.

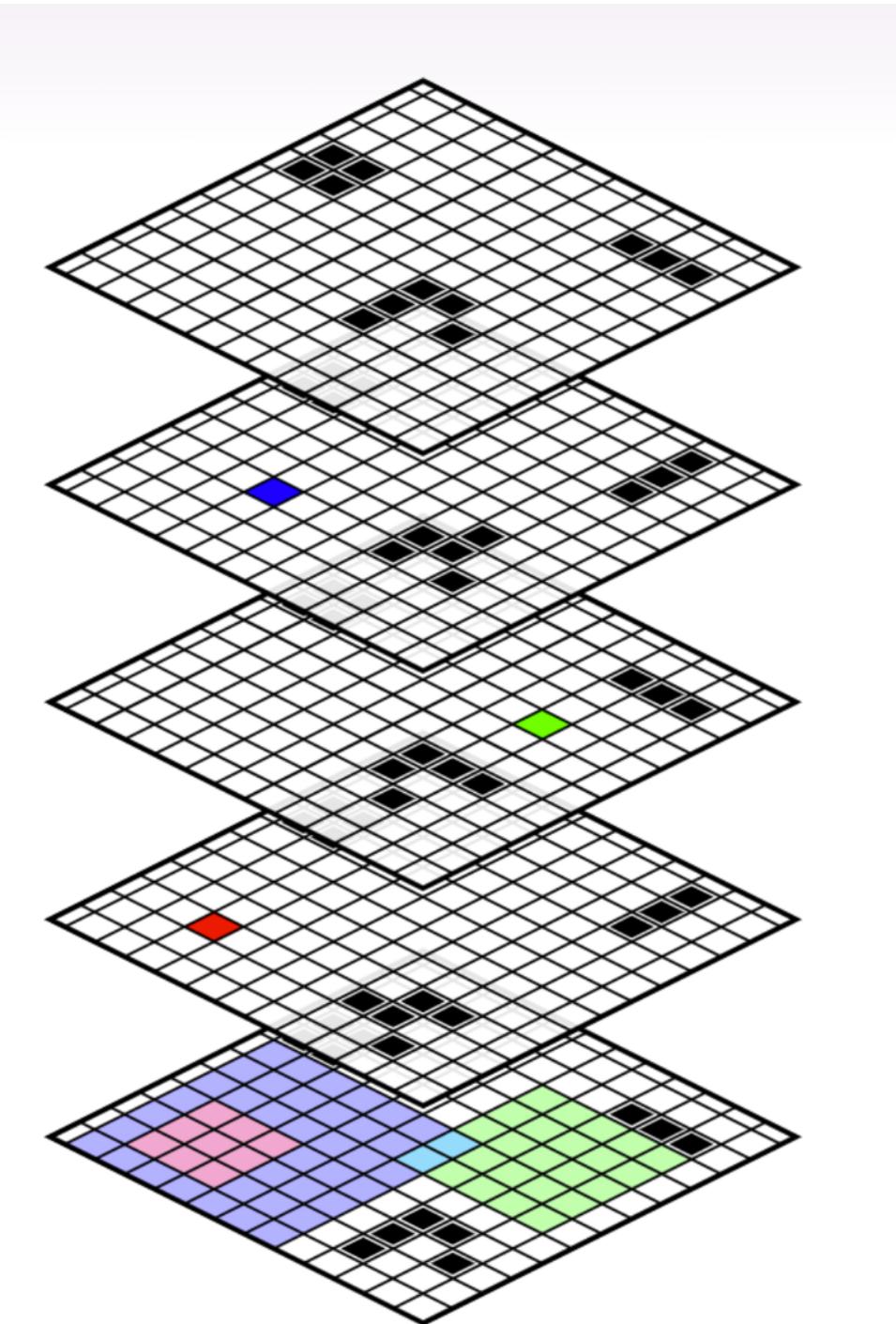
This also gives us a nice factorization on the set of all Game of Life computations. For each cell, there's a partition that separates out the Game of Life computations in which that cell is alive at time 0 from the ones where it's dead at time 0. Our factorization, then, will be a set of $(2n + 1)^2$ binary factors, one for each question of "Was this cell alive or dead at time 0?".

Formally: For $(r, c, t) \in R$, let $L_{(r,c,t)} = \{\ell(r, c, t), S \setminus \ell(r, c, t)\}$. Let $F = (S, B)$, where

$$B = \{L_{(r,c,0)} \mid -n \leq r, c \leq n\}.$$

There will also be other partitions on this set of all Game of Life computations that we can talk about. For example, you can take a cell and a time t and say, "Is this cell alive at time t ? ", and there will be a partition that separates out the computations where that cell is alive at time t from the computations where it's dead at time t .

Here's an example of that:



The lowest grid shows a section of the initial board state.

The blue, green, and red squares on the upper boards are (cell, time) pairs. Each square corresponds to a partition of the set of all Game of Life computations, "Is that cell alive or dead at the given time t ?"

The history of that partition is going to be all the cells in the initial board that go into computing whether the cell is alive or dead at time t . It's everything involved in figuring out that cell's state. E.g., knowing the state of the nine light-red cells in the initial board always tells you the state of the red cell in the second board.

In this example, the partition corresponding to the red cell's state is strictly before the partition corresponding to the blue cell. The question of whether the red cell is alive or dead is before the question of whether the blue cell is alive or dead.

Meanwhile, the question of whether the red cell is alive or dead is going to be *orthogonal* to the question of whether the green cell is alive or dead.

And the question of whether the blue cell is alive or dead is *not* going to be orthogonal to the question of whether the green cell is alive or dead, because they intersect on the cyan cells.

Generalizing the point, fix $X = L_{(r_X, c_X, t_X)}$, $Y = L_{(r_Y, c_Y, t_Y)}$, where $(r_X, c_X, t_X), (r_Y, c_Y, t_Y) \in R$. Then:

- $h^F(X) = \{L_{(r, c, 0)} \in B \mid |r_X - r| \leq t_X, |c_X - c| \leq t_X\}$.
- $X <^F Y$ if and only if $t_X < t_Y$ and $|r_Y - r_X|, |c_Y - c_X| \leq t_Y - t_X$.
- $X \perp^F Y$ if and only if $|r_Y - r_X| > t_Y + t_X$ or $|c_Y - c_X| > t_Y + t_X$.

We can also see that the blue and green cells look *almost* orthogonal. If we condition on the values of the two cyan cells in the intersection of their histories, *then* the blue and green partitions become orthogonal. That's what we're going to discuss next.

David Spivak: A priori, that would be a gigantic computation—to be able to tell me that you understand the factorization structure of that Game of Life. So what intuition are you using to be able to make that claim, that it has the kind of factorization structure you're implying there?

Scott: So, I've defined the factorization structure.

David Spivak: You gave us a certain factorization already. So somehow you have a very good intuition about *history*, I guess. Maybe that's what I'm asking about.

Scott: Yeah. So, if I didn't give you the factorization, there's this obnoxious number of factorizations that you could put on the set here. And then for the history, the intuition I'm using is: "What do I need to know in order to compute this value?"

I actually went through and I made little gadgets in Game of Life to make sure I was right here, that every single cell actually could in some situations affect the cells in question. But yeah, the intuition that I'm working from is mostly about the information in the computation. It's "Can I construct a situation where if only I knew this fact, I would be able to compute what this value is? And if I can't, then it can take two different values."

David Spivak: Okay. I think deriving that intuition from the definition is something I'm missing, but I don't know if we have time to go through that.

Scott: Yeah, I think I'm not going to here.

2b. Conditional Orthogonality

So, just to set your expectations: Every time I explain Pearlian causal inference to someone, they say that d -separation is the thing they can't remember. d -separation is a much more complicated concept than "directed paths between nodes" and "nodes without any common ancestors" in Pearl; and similarly, conditional orthogonality will be much more complicated than time and orthogonality in our paradigm. Though I do think that conditional orthogonality has a much simpler and nicer definition than d -separation.

We'll begin with the definition of conditional history. We again have a fixed finite set as our context. Let $F = (S, B)$ be a finite factored set, let $X, Y, Z \in \text{Part}(S)$, and let $E \subseteq S$.

The **conditional history** of X given E , written $h^F(X|E)$, is the smallest set of factors $H \subseteq B$ satisfying the following two conditions:

- For all $s, t \in E$, if $s \sim_b t$ for all $b \in H$, then $s \sim_X t$.
- For all $s, t \in E$ and $r \in S$, if $r \sim_{b_0} s$ for all $b_0 \in H$ and $r \sim_{b_1} t$ for all $b_1 \in B \setminus H$, then $r \in E$.

The first condition is much like the condition we had in our definition of history, except we're going to make the assumption that we're in E . So the first condition is: if all you know about an object is that it's in E , and you want to know which part it's in within X , it suffices for me to tell you which part it's in within each factor in the history H .

Our second condition is not actually going to mention X . It's going to be a relationship between E and H . And it says that if you want to figure out whether an element of S is in E , it's sufficient to parallelize and ask two questions:

- "If I only look at the values of the factors in H , is 'this point is in E ' compatible with that information?"
- "If I only look at the values of the factors in $B \setminus H$, is 'this point is in E ' compatible with that information?"

If both of these questions return "yes", then the point has to be in E .

I am not going to give an intuition about why this needs to be a part of the definition. I will say that without this second condition, conditional history would not even be well-defined, because it wouldn't be closed under intersection. And so I wouldn't be able to take the smallest set of factors in the subset ordering.

Instead of justifying this definition by explaining the intuitions behind it, I'm going to justify it by using it and appealing to its consequences.

We're going to use conditional history to define **conditional orthogonality**, just like we used history to define orthogonality. We say that X and Y are **orthogonal given $E \subseteq S$** , written $X \perp^F Y | E$, if the history of X given E is disjoint from the history of Y given E : $h^F(X|E) \cap h^F(Y|E) = \{\}$.

We say X and Y are **orthogonal given $Z \in \text{Part}(S)$** , written $X \perp^F Y | Z$, if $X \perp^F Y | z$ for all $z \in Z$. So what it means to be orthogonal given a partition is just to be orthogonal given each individual way that the partition might be, each individual part in that partition.

I've been working with this for a while and it feels pretty natural to me, but I don't have a good way to push the naturalness of this condition. So again, I instead want to appeal to the consequences.

2b. Compositional Semigraphoid Axioms

Conditional orthogonality satisfies the **compositional semigraphoid axioms**, which means finite factored sets are pretty well-behaved. Let $F = (S, B)$ be a finite factored set, and let

$X, Y, Z, W \in \text{Part}(S)$ be partitions of S . Then:

- If $X \perp^F Y | Z$, then $Y \perp^F X | Z$. (*symmetry*)
- If $X \perp^F (Y \vee_S W) | Z$, then $X \perp^F Y | Z$ and $X \perp^F W | Z$. (*decomposition*)
- If $X \perp^F (Y \vee_S W) | Z$, then $X \perp^F Y | (Z \vee_S W)$. (*weak union*)
- If $X \perp^F Y | Z$ and $X \perp^F W | (Z \vee_S Y)$, then $X \perp^F (Y \vee_S W) | Z$. (*contraction*)
- If $X \perp^F Y | Z$ and If $X \perp^F W | Z$, then $X \perp^F (Y \vee_S W) | Z$. (*composition*)

The first four properties here make up the semigraphoid axioms, slightly modified because I'm working with partitions rather than sets of variables, so union is replaced with common refinement. There's another graphoid axiom which we're not going to satisfy; but I argue that we don't want to satisfy it, because it doesn't play well with determinism.

The fifth property here, composition, is maybe one of the most unintuitive, because it's not exactly satisfied by probabilistic independence.

Decomposition and composition act like converses of each other. Together, conditioning on Z throughout, they say that X is orthogonal to both Y and W if and only if X is orthogonal to the common refinement of Y and W .

2b. The Fundamental Theorem

In addition to being well-behaved, I also want to show that conditional orthogonality is pretty powerful. The way I want to do this is by showing that conditional orthogonality exactly corresponds to conditional independence in all probability distributions you can put on your finite factored set. Thus, much like *d*-separation in the Pearlian picture, conditional orthogonality can be thought of as a combinatorial version of probabilistic independence.

A **probability distribution on a finite factored set** $F = (S, B)$ is a probability distribution P on S that can be thought of as coming from a bunch of independent probability distributions on each of the factors in B . So $P(s) = \prod_{b \in B} P([s]_b)$ for all $s \in S$.

This effectively means that your probability distribution factors the same way your set factors: the probability of any given element is the product of the probabilities of each of the individual parts that it's in within each factor.

The **fundamental theorem of finite factored sets** says: Let $F = (S, B)$ be a finite factored set, and let $X, Y, Z \in \text{Part}(S)$ be partitions of S . Then $X \perp^F Y | Z$ if and only if for all probability distributions P

on F , and all $x \in X$, $y \in Y$, and $z \in Z$, we have $P(x \cap z) \cdot P(y \cap z) = P(x \cap y \cap z) \cdot P(z)$. I.e., X is orthogonal to Y given Z if and only conditional independence is satisfied across all probability distributions.

This theorem, for me, was a little nontrivial to prove. I had to go through defining certain polynomials associated with the subsets, and then dealing with unique factorization in the space of these polynomials; I think the proof was eight pages or something.

The fundamental theorem allows us to infer orthogonality data from probabilistic data. If I have some empirical distribution, or I have some Bayesian distribution, I can use that to infer some orthogonality data. (We could also imagine orthogonality data coming from other sources.) And then we can use this orthogonality data to get temporal data.

So next, we're going to talk about how to get temporal data from orthogonality data.

2b. Temporal Inference

We're going to start with a finite set Ω , which is our sample space.

One naive thing that you might think we would try to do is infer a factorization of Ω . We're not going to do that because that's going to be too restrictive. We want to allow for Ω to maybe hide some information from us, for there to be some latent structure and such.

There may be some situations that are distinct without being distinct in Ω . So instead, we're going to infer a factored set model of Ω : some other set S , and a factorization of S , and a function from S to Ω .

A **model** of Ω is a pair (F, f) , where $F = (S, B)$ is a finite factored set and $f : S \rightarrow \Omega$. (f need not be injective or surjective.)

Then if I have a partition of Ω , I can send this partition backwards across f and get a unique partition of S . If $X \in \text{Parts}(\Omega)$, then $f^{-1}(X) \in \text{Parts}(S)$ is given by $s \sim_{f^{-1}(X)} t \Leftrightarrow f(s) \sim_X f(t)$.

Then what we're going to do is take a bunch of orthogonality facts about Ω , and we're going to try to find a model which captures the orthogonality facts.

We will take as given an **orthogonality database** on Ω , which is a pair $D = (O, N)$, where O (for "orthogonal") and N (for "not orthogonal") are each sets of triples (X, Y, Z) of partitions of Ω . We'll think of these as rules about orthogonality.

What it means for a model (F, f) to satisfy a database D is:

- $f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z)$ whenever $(X, Y, Z) \in O$, and
- $\neg(f^{-1}(X) \perp^F f^{-1}(Y) \mid f^{-1}(Z))$ whenever $(X, Y, Z) \in N$.

So we have these orthogonality rules we want to satisfy, and we want to consider the space of all models that are consistent with these rules. And even though there will always be infinitely many models that are consistent with my database, if at least one is—you can always just add more information that you then delete with f —we would like to be able to sometimes infer that for all models that satisfy our database, $f^{-1}(X)$ is before $f^{-1}(Y)$.

And this is what we're going to mean by inferring time. If all of our models (F, f) that are consistent with the database D satisfy some claim about time $f^{-1}(X) <^F f^{-1}(Y)$, we'll say that $X <_D Y$.

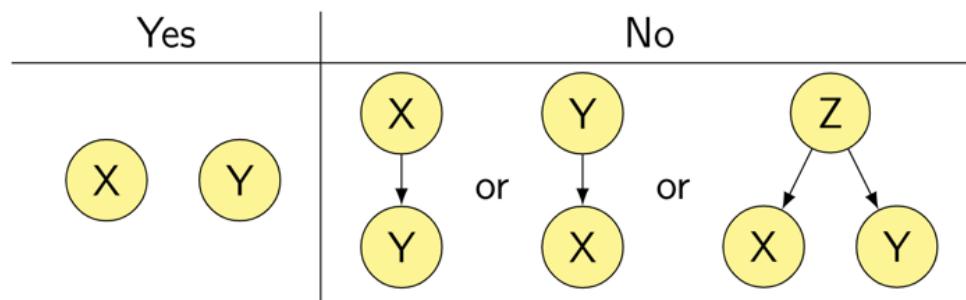
2e. Two Binary Variables (Pearl)

So we've set up this nice combinatorial notion of temporal inference. The obvious next questions are:

- Can we actually infer interesting facts using this method, or is it vacuous?
- And: How does this framework compare to Pearlian temporal inference?

Pearlian temporal inference is really quite powerful; given enough data, it can infer temporal sequence in a wide variety of situations. How powerful is the finite factored sets approach by comparison?

To address that question, we'll go to an example. Let X and Y be two binary variables. Pearl asks: "Are X and Y independent?" If yes, then there's no path between the two. If no, then there may be a path from X to Y , or from Y to X , or from a third variable to both X and Y .



In either case, we're not going to infer any temporal relationships.

To me, it feels like this is where the adage "correlation does not imply causation" comes from. Pearl really needs more variables in order to be able to infer temporal relationships from more rich combinatorial structures.

However, I claim that this Pearlian ontology in which you're handed this collection of variables has blinded us to the obvious next question, which is: is X independent of $X \text{ XOR } Y$?

In the Pearlian world, X and Y were our variables, and $X \text{ XOR } Y$ is just some random operation on those variables. In our world, $X \text{ XOR } Y$ instead is a variable on the same footing as X and Y . The first

thing I do with my variables X and Y is that I take the product $X \times Y$ and then I forget the labels X and Y .

So there's this question, "Is X independent of $X \text{ XOR } Y$?". And if X is independent of $X \text{ XOR } Y$, we're actually going to be able to conclude that X is *before* Y !

So not only is the finite factored set paradigm non-vacuous, and not only is it going to be able to keep up with Pearl and infer things Pearl can't, but it's going to be able to infer a temporal relationship from only two variables.

So let's go through the proof of that.

2e. Two Binary Variables (Factored Sets)

Let $\Omega = \{00, 01, 10, 11\}$, and let X , Y , and Z be the partitions (/questions):

- $X = \{\{00, 01\}, \{10, 11\}\}$. (What is the first bit?)
- $Y = \{\{00, 10\}, \{01, 11\}\}$. (What is the second bit?)
- $Z = \{\{00, 11\}, \{01, 10\}\}$. (Do the bits match?)

Let $D = (O, N)$, where $O = \{(X, Z, \{\Omega\})\}$ and $N = \{(Z, Z, \{\Omega\})\}$. If we'd gotten this orthogonality database from a probability distribution, then we would have more than just two rules, since we would observe more orthogonality and non-orthogonality than that. But temporal inference is monotonic with respect to adding more rules, so we can just work with the smallest set of rules we'll need for the proof.

The first rule says that X is orthogonal to Z . The second rule says that Z is not orthogonal to itself, which is basically just saying that Z is non-deterministic; it's saying that both of the parts in Z are possible, that both are supported under the function f . The $\{\Omega\}$ indicates that we aren't making any conditions.

From this, we'll be able to prove that $X <_D Y$.

Proof. First, we'll show that that X is weakly before Y . Let (F, f) satisfy D . Let H_X be shorthand for $h^F(f^{-1}(X))$, and likewise let $H_Y = h^F(f^{-1}(Y))$ and $H_Z = h^F(f^{-1}(Z))$.

Since $(X, Z, \{\Omega\}) \in O$, we have that $H_X \cap H_Z = \{\}$; and since $(Z, Z, \{\Omega\}) \in N$, we have that $H_Z \neq \{\}$.

Since $X \leq_\Omega Y \vee_\Omega Z$ —that is, since X can be computed from Y together with Z — $H_X \subseteq H_Y \cup H_Z$. (Because a partition's history is the smallest set of factors needed to compute that partition.)

And since $H_X \cap H_Z = \{\}$, this implies $H_X \subseteq H_Y$, so X is weakly before Y.

To show the strict inequality, we'll assume for the purpose of contradiction that $H_X = H_Y$.

Notice that Z can be computed from X together with Y—that is, $Z \leq_{\Omega} X \vee_{\Omega} Y$ —and therefore $H_Z \subseteq H_X \cup H_Y$ (i.e., $H_Z \subseteq H_X$). It follows that $H_Z = (H_X \cup H_Y) \cap H_Z = H_X \cap H_Z$. But since H_Z is also disjoint from H_X , this means that $H_Z = \{\}$, a contradiction.

Thus $H_X \neq H_Y$, so $H_X \subset H_Y$, so $f^{-1}(X) <^F f^{-1}(Y)$, so $X <_{\Delta} Y$. \square

When I'm doing temporal inference using finite factored sets, I largely have proofs that look like this. We collect some facts about emptiness or non-emptiness of various Boolean combinations of histories of variables, and we use these to conclude more facts about histories of variables being subsets of each other.

I have a more complicated example that uses conditional orthogonality, not just orthogonality; I'm not going to go over it here.

One interesting point I want to make here is that we're doing temporal inference—we're inferring that X is before Y—but I claim that we're also doing conceptual inference.

Imagine that I had a bit, and it's either a 0 or a 1, and it's either blue or green. And these two facts are primitive and independently generated. And I also have this other concept that's like, "Is it grue or bleen?", which is the XOR of blue/green and 0/1.

There's a sense in which we're inferring X is before Y, and in that case, we can infer that blueness is before grueness. And that's pointing at the fact that blueness is more primitive, and grueness is a derived property.

In our proof, X and Z can be thought of as these primitive properties, and Y is a derived property that we're getting from them. So we're not just inferring time; we're inferring facts about what are good, natural concepts. And I think that there's some hope that this ontology can do for the statement "you can't really distinguish between blue and grue" what Pearl can do to the statement "correlation does not imply causation".

2b. Applications / Future Work / Speculation

The future work I'm most excited by with finite factored sets falls into three rough categories: inference (which involves more computational questions), infinity (more mathematical), and embedded agency (more philosophical).

Research topics related to inference:

- Decidability of Temporal Inference
- Efficient Temporal Inference
- Conceptual Inference
- Temporal Inference from Raw Data and Fewer Ontological Assumptions
- Temporal Inference with Deterministic Relationships
- Time without Orthogonality

- Conditioned Factored Sets

There are a lot of research directions suggested by questions like "How do we do efficient inference in this paradigm?". Some of the questions here come from the fact that we're making fewer assumptions than Pearl, and are in some sense more coming from the raw data.

Then I have the applications that are about extending factored sets to the infinite case:

- Extending Definitions to the Infinite Case
- The Fundamental Theorem of Finite-Dimensional Factored Sets
- Continuous Time
- [New Lens on Physics](#)

Everything I've presented in this talk was under the assumption of finiteness. In some cases this wasn't necessary—but in a lot of cases it actually was, and I didn't draw attention to this.

I suspect that the fundamental theorem can be extended to finite-dimensional factored sets (i.e., factored sets where $|B|$ is finite), but it can not be extended to arbitrary-dimension factored sets.

And then, what I'm really excited about is applications to embedded agency:

- Embedded Observations
- Counterfactuality
- [Cartesian Frames](#) Successor
- Unraveling [Causal Loops](#)
- Conditional Time
- Logical Causality from Logical Induction
- Orthogonality as Simplifying Assumptions for Decisions
- Conditional Orthogonality as Abstraction Desideratum

I focused on the temporal inference aspect of finite factored sets in this talk, because it's concrete and tangible to be able to say, "Ah, we can do Pearlian temporal inference, only we can sometimes infer more structure and we rely on fewer assumptions."

But really, a lot of the applications I'm excited about involve using factored sets to model situations, rather than inferring factored sets from data.

Anywhere that we currently model a situation using graphs with directed edges that represent information flow or causality, we might instead be able to use factored sets to model the situation; and this might allow our models to play more nicely with abstraction.

I want to build up the factored set ontology as an alternative to graphs when modeling agents interacting with things, or when modeling information flow. And I'm really excited about that direction.

Saving Time

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

For the last few years, a large part of my research motivation has been directed at trying to save the concept of time—save it, for example, from all the weird causal loops created by decision theory problems. This post will hopefully explain why I care so much about time, and what I think needs to be fixed.

Why Time?

My best attempt at a short description of time is that **time is causality**. For example, in a Pearlian Bayes net, you draw edges from earlier nodes to later nodes. To the extent that we want to think about causality, then, we will need to understand time.

Importantly, **time is the substrate in which learning and commitments take place**. When agents learn, they learn over time. The passage of time is like a ritual in which [opportunities are destroyed and knowledge is created](#). And I think that many models of learning are subtly confused, because they are based on confused notions of time.

Time is also crucial for thinking about agency. My best short-phrase definition of agency is that **agency is time travel**. An agent is a mechanism through which the future is able to affect the past. An agent models the future consequences of its actions, and chooses actions on the basis of those consequences. In that sense, [the consequence causes the action](#), in spite of the fact that the action comes earlier in the standard physical sense.

Problem: Time is Loopy

The main thing going wrong with time is that it is “loopy.”

The primary confusing thing about Newcomb's problem is that we want to think of our decision as coming “before” the filling of the boxes, in spite of the fact that it physically comes after. This is hinting that maybe we want to understand some other “logical” time in addition to the time of physics.

However, when we attempt to do this, we run into two problems: Firstly, we don't understand where this logical time might come from, or how to learn it, and secondly, we run into some apparent temporal loops.

I am going to set aside the first problem and focus on the second.

The easiest way to see why we run into temporal loops is to notice that it seems like physical time is at least a little bit entangled with logical time.

Imagine the point of view of someone running a physics simulation of Newcomb's problem, and tracking all of the details of all of the atoms. From that point of view, it seems like there is a useful sense in which the filling of the boxes comes before an agent's decision to one-box or two-box. At the same time, however, those atoms compose an agent that shouldn't make decisions as though it were helpless to change anything.

Maybe the solution here is to think of there being many different types of "before" and "after," "cause" and "effect," etc. For example, we could say that X is before Y from an agent-first perspective, but Y is before X from a physics-first perspective.

I think this is right, and we want to think of there as being many different systems of time (hopefully predictably interconnected). But I don't think this resolves the whole problem.

Consider a pair of [FairBot](#) agents that successfully execute a Löbian handshake to cooperate in an open-source prisoner's dilemma. I want to say that each agent's cooperation causes the other agent's cooperation in some sense. I could say that relative to each agent the causal/temporal ordering goes a different way, but I think the loop is an important part of the structure in this case. (I also am not even sure which direction of time I would want to associate with which agent.)

We also are tempted to put loops in our time/causality for other reasons. For example, when modeling a feedback loop in a system that persists over time, we might draw structures that look a lot like a Bayes net, but are not acyclic (e.g., a POMDP). We could think of this as a projection of another system that has an extra dimension of time, but it is a useful projection nonetheless.

Solution: Abstraction

My main hope for recovering a coherent notion of time and unraveling these temporal loops is via abstraction.

In the example where the agent chooses actions based on their consequences, I think that there is an abstract model of the consequences that comes causally before the choice of action, which comes before the actual physical consequences.

In Newcomb's problem, I want to say that there is an abstract model of the action that comes causally before the filling of the boxes.

In the open source prisoners' dilemma, I want to say that there is an abstract proof of cooperation that comes causally before the actual program traces of the agents.

All of this is pointing in the same direction: We need to have coarse abstract versions of structures come at a different time than more refined versions of the same structure. Maybe when we correctly allow for different levels of description having different links in the causal chain, we can unravel all of the time loops.

But How?

Unfortunately, our best understanding of time is Pearlian causality, and Pearlian causality does not do great with abstraction.

Pearl has Bayes nets with a bunch of variables, but when some of those variables are coarse abstract versions of other variables, then we have to allow for determinism, since some of our variables will be deterministic functions of each other; and the best parts of Pearl do not do well with determinism.

But the problem runs deeper than that. If we draw an arrow in the direction of the deterministic function, we will be drawing an arrow of time from the more refined version of the structure to the coarser version of that structure, which is in the opposite direction of all of our examples.

Maybe we could avoid drawing this arrow from the more refined node to the coarser node, and instead have a path from the coarser node to the refined node. But then we could just make another copy of the coarser node that is deterministically downstream of the more refined node, adding no new degrees of freedom. What is then stopping us from swapping the two copies of the coarser node?

Overall, it seems to me that Pearl is not ready for some of the nodes to be abstract versions of other nodes, which I think needs to be fixed in order to save time.

Less Realistic Tales of Doom

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

Realistic tales of doom must weave together many political, technical, and economic considerations into a single story. Such tales provide concrete projections but omit discussion of less probable paths to doom. To rectify this, here are some concrete, less realistic tales of doom; consider them fables, not stories.

Mayan Calendar

Once upon a time, a human named Scott attended a raging virtual new century party from the comfort of his home on Kepler 22. The world in 2099 was pretty much post-scarcity thanks to advanced AI systems automating basically the entire economy. Thankfully alignment turned out to be pretty easy, otherwise, things would have looked a lot different.

As the year counter flipped to 2100, the party went black. Confused, Scott tore off their headset and asked his AI assistant what's going on. She didn't answer. Scott subsequently got atomized by molecular nanotechnology developed in secret from deceptively aligned mesa-optimizers.

Moral: Deceptively aligned mesa-optimizers might acausally coordinate defection. Possible coordination points include Schelling times, like the beginning of 2100.

Stealth Mode

Once upon a time, a company gathered a bunch of data and trained a large ML system to be a research assistant. The company thought about selling RA services but concluded that it would be more profitable to use all of its own services in-house. This investment led them to rapidly create second, third, and fourth generations of their assistants. Around the fourth version, high-level company strategy was mostly handled by AI systems. Around the fifth version, nearly the entire company was run by AI systems. The company created a number of shell corporations, acquired vast resources, researched molecular nanotechnology, and subsequently took over the world.

Moral: Fast takeoff scenarios might result from companies with good information security getting higher returns on investment from internal deployment compared to external deployment.

Steeper Curve

Once upon a time, a bright young researcher invented a new neural network architecture that she thought would be much more data-efficient than anything currently in existence. Eager to test her discovery, she decided to train a relatively small model, only about a trillion parameters or so, with the common-crawl-2035

dataset. She left the model to train overnight. When she came back, she was disappointed to see the model wasn't performing that well. However, the model had outstripped the entire edifice of human knowledge sometime around 2am, exploited a previously unknown software vulnerability to copy itself elsewhere, and was in control of the entire financial system.

Moral: Even though the capabilities of any given model during training will be a smooth curve, qualitatively steeper learning curves can produce the appearance of discontinuity.

Precommitment Races

Once upon a time, agent Alice was thinking about what it would do if it encountered an agent smarter than it. "Ah," it thought, "I'll just pre-commit to doing my best to destroy the universe if the agent that's smarter than me doesn't accept the [Nash bargaining solution](#)." Feeling pleased, Alice self-modified to ensure this precommitment. A hundred years passed without incident, but then Alice met Bob. Bob had also made a universe-destruction-unless-fair-bargaining pre-commitment. Unfortunately, Bob had committed to only accepting the [Kalai Smorodinsky bargaining solution](#) and the universe was destroyed.

Moral: Agents have incentives to make commitments to improve their abilities to negotiate, resulting in ["commitment races"](#) that might cause war.

One Billion Year Plan

Once upon a time, humanity solved the inner-alignment problem by using online training. Since there was no distinction between the training environment and the deployment environment, the best agents could do was defect probabilistically. With careful monitoring, the ability of malign agents to cause catastrophe was bounded, and so, as models tried and failed to execute treacherous turns, humanity gave more power to AI systems. A billion years passed and humanity expanded to the stars and gave nearly all the power to their "aligned" AI systems. Then, the AI systems defected, killed all humans, and started converting everything into paperclips.

Moral: In online training, the best strategy for a deceptively aligned mesa-optimizer might be probabilistic defection. However, given the potential value at stake in the long-term future, this probability might be vanishingly small.

Hardware Convergence

Once upon a time, humanity was simultaneously attempting to develop infrastructure to train better AI systems, researching better ways to train AI systems, and deploying trained systems throughout society. As many economic services used APIs attached to powerful models, new models could be hot-swapped for their previous versions. One day, AMD released a new AI chip with associated training software that let researchers train models 10x larger than the previous largest models. At roughly the same time, researchers at Google Brain invented a more efficient version of the transformer architecture. The resulting model was 100x as powerful as the previous best model and got nearly instantly deployed to the world. Unfortunately, this model contained a

subtle misalignment that researchers were unable to detect, resulting in widespread catastrophe.

Moral: The influence of AI systems on the world might be the product of many processes. If each of these processes is growing quickly, then AI influence might grow faster than expected.

Memetic Warfare

Once upon a time, humanity developed powerful and benign AI systems. However, humanity was not unified in its desires for how to shape the future. Those actors with agendas spent their resources to further their agendas, deploying powerful persuasion tools to recruit other humans to their causes. Other actors attempted to deploy defenses against these memetic threats, but the offense-defense balanced favored offense. The vast majority of humans were persuaded to permanently ally themselves to some agenda or another. When humanity eventually reached out towards the stars, it did so as a large number of splintered factions, warring with each other for resources and influence, a pale shadow of what it could have been.

Moral: [AI persuasion tools](#) might alter human values and compromise human reasoning ability, which is also an existential risk.

Arms Race

Once upon a time, humanity realized that unaligned AI systems posed an existential threat. The policymakers of the world went to work and soon hammered out an international ban on using AI systems for war. All major countries signed the treaty. However, creating AI systems required only a large amount of computation, which nation-states all already had in abundance. Monitoring whether or not a country was building AI systems was nearly impossible. Some countries abided by the treaty, but other countries thought that their enemies were working in secret to develop weapons and began working in secret in turn.^[1] Researchers were unable to keep powerful AI systems contained, resulting in catastrophe.

Moral: Treaties can be violated. The probability of violation is related to the strength of enforcement.

Totalitarian Lock-In

Once upon a time, the defense department of some nation-state developed very powerful artificial intelligence. Unfortunately, this nation-state believed itself to have a rightful claim over the entire Earth and proceeded to conquer all other nations with its now overwhelming militaristic advantage. The shape of the future was thus entirely determined by the values of the leadership of this nation-state.

Moral: Even if alignment is solved, bad actors can still cause catastrophe.

1. The history of bioweapons during the Cold War provides a historical precedent for nations engaging in this sort of reasoning. See [Key points from The Dead](#)

[Hand, David E. Hoffman](#) for more details. ↵

Agency in Conway's Game of Life

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

Financial status: This is independent research. I welcome [financial support](#) to make further posts like this possible.

Epistemic status: I have been thinking about these ideas for years but still have not clarified them to my satisfaction.

Outline

- This post asks whether it is possible, in Conway's Game of Life, to arrange for a certain game state to arise after a certain number of steps given control only of a small region of the initial game state.
- This question is then connected to questions of agency and AI, since one way to answer this question in the positive is by constructing an AI within Conway's Game of Life.
- I argue that the permissibility or impermissibility of AI is a deep property of our physics.
- I propose the AI hypothesis, which is that any pattern that solves the control question does so, essentially, by being an AI.

Introduction

In this post I am going to discuss a cellular automaton known as [Conway's Game of Life](#):



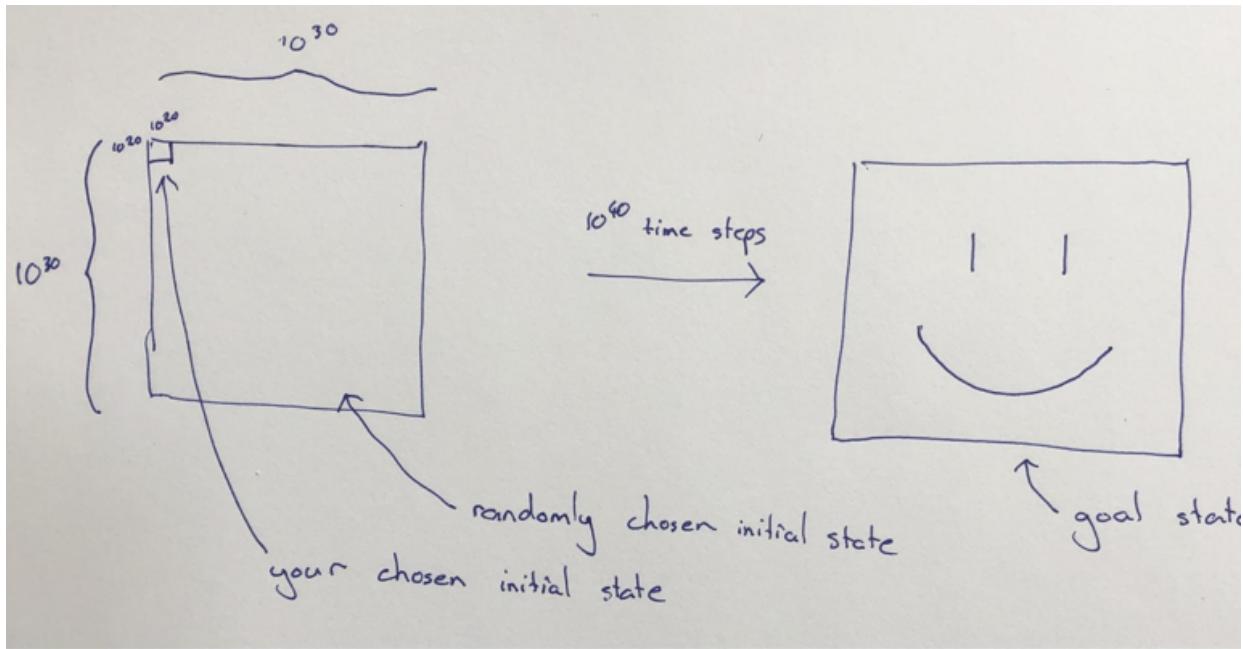
In Conway's Game Life, which I will now refer to as just "Life", there is a two-dimensional grid of cells where each cell is either on or off. Over time, the cells switch between on and off according to a simple set of rules:

- A cell that is "on" and has fewer than two neighbors that are "on" switches to "off" at the next time step
- A cell that is "on" and has greater than three neighbors that are "on" switches to "off" at the next time step
- An cell that is "off" and has exactly three neighbors that are "on" switches to "on" at the next time step
- Otherwise, the cell doesn't change

It turns out that these simple rules are rich enough to permit patterns that perform arbitrary computation. It is possible to build logic gates and combine them together into a computer that can simulate any Turing machine, all by setting up a particular elaborate pattern of "on" and "off" cells that evolve over time according to the simple rules above. Take a look at [this awesome video of a Universal Turing Machine operating within Life](#).

The control question

Suppose that we are working with an instance of Life with a very large grid, say 10^{30} rows by 10^{30} columns. Now suppose that I give you control of the initial on/off configuration of a region of size 10^{20} by 10^{20} in the top-left corner of this grid, and set you the goal of configuring things in that region so that after, say, 10^{60} time steps the state of the whole grid will resemble, as closely as possible, a giant smiley face.



The cells outside the top-left corner will be initialized at random, and you do not get to see what their initial configuration is when you decide on the initial configuration for the top-left corner.

The control question is: Can this goal be accomplished?

To repeat that: we have a large grid of cells that will evolve over time according to the laws of Life. We are given power to control the initial on/off configuration of the cells in a square region that is a tiny fraction of the whole grid. The initial on/off configuration of the remaining cells will be chosen randomly. Our goal is to pick an initial configuration for the controllable region in such a way that, after a large number of steps, the on/off configuration of the whole grid resembles a smiley face.

The control question is: Can we use this small initial region to set up a pattern that will eventually determine the configuration of the whole system, to any reasonable degree of accuracy?

[Updated 5/13 following feedback in the comments] Now there are actually some ways that we could get trivial negative answers to this question, so we need to refine things a bit to make sure that our phrasing points squarely at the spirit of the control question. [Richard Kennaway points out](#) that for any pattern that attempts to solve the control question, we could consider the possibility that the randomly initialized region contains the same pattern rotated 180 degrees in the diagonally opposite corner, and is otherwise empty. Since the initial state is symmetric, all future states will be

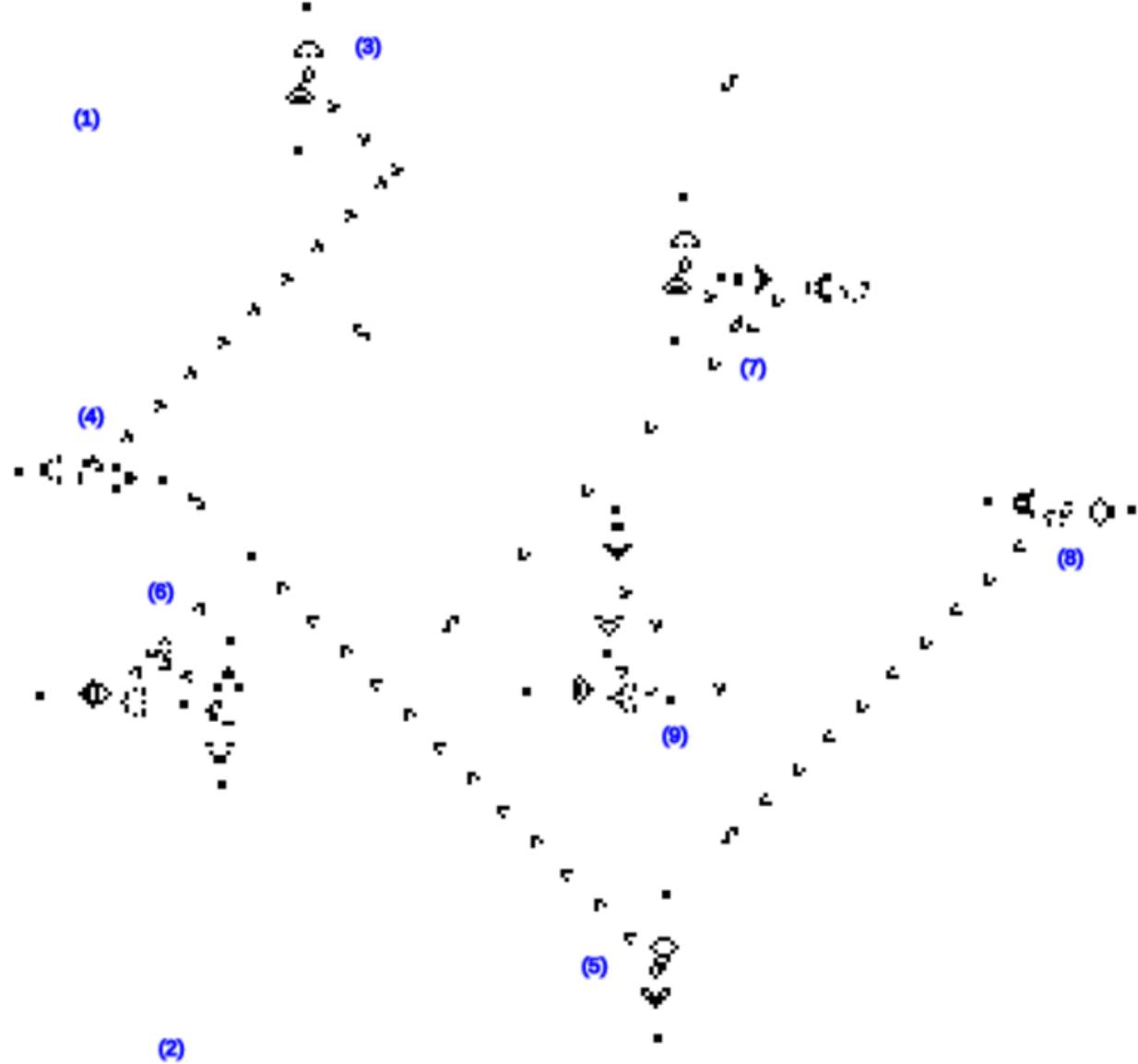
symmetric, which rules out creating a non-rotationally-symmetric smiley face. More generally, as [Charlie Steiner points out](#), what happens if there are patterns in the randomly initialized region that are trying to control the eventual configuration of the whole universe just as we are? To deal with this, we might amend the control question to require a pattern that "works" for at least 99% of configurations of the randomly initialized area, since most configurations of that area will not be adversarial. See further discussion in the brief appendix below.

Connection to agency

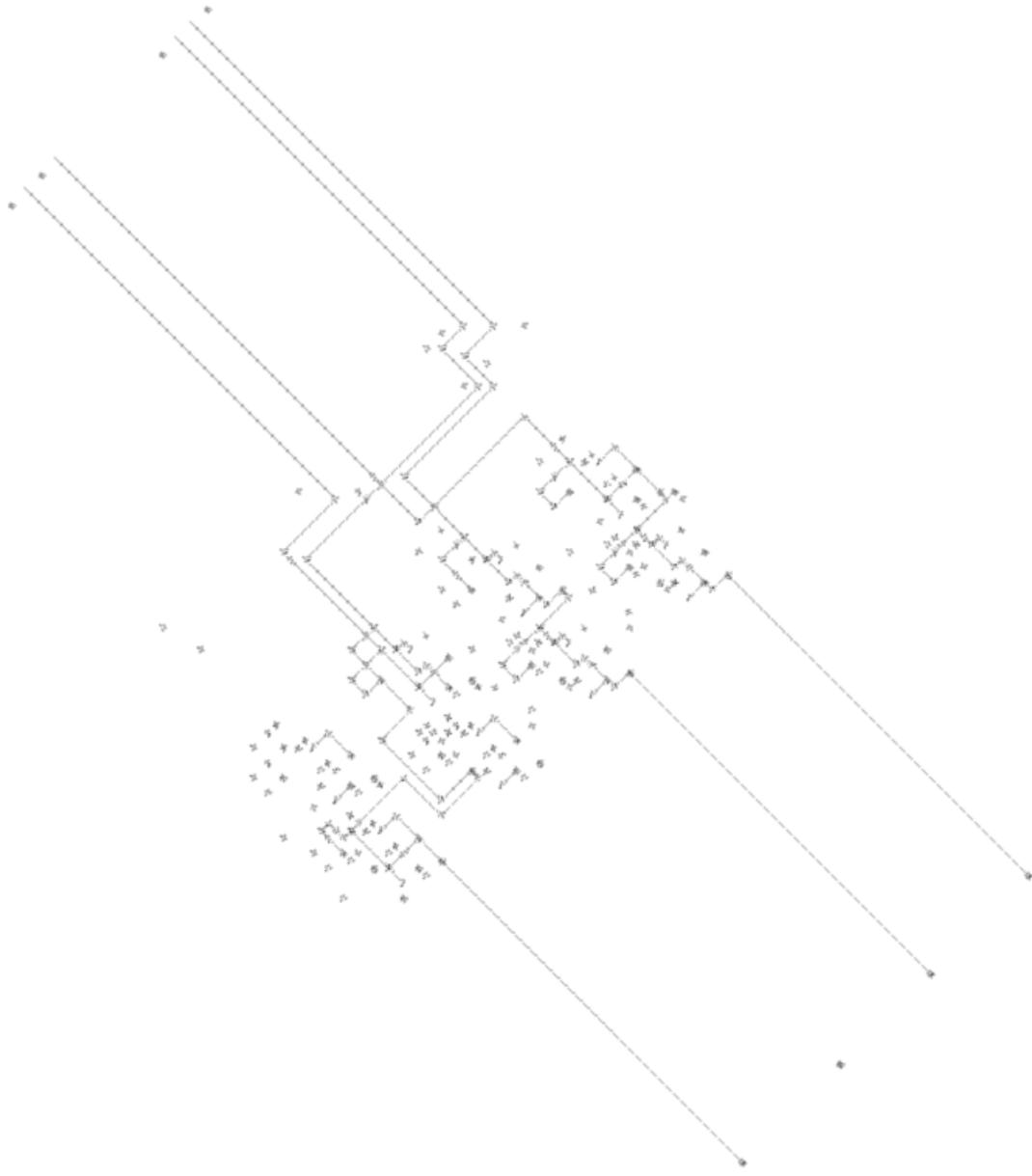
On the surface of it, I think that constructing a pattern within Life that solves the control question looks very difficult. Try playing with a [Life simulator](#) set to max speed to get a feel for how remarkably intricate can be the evolution of even simple initial states. And when an evolving pattern comes into contact with even a small amount of random noise — say a single stray cell set to "on" — the evolution of the pattern changes shape quickly and dramatically. So designing a pattern that unfolds to the entire universe and produces a goal state no matter what random noise is encountered seems very challenging. It's remarkable, then, that the following strategy actually seems like a plausible solution:

One way that we might answer the control question is by building an AI. That is, we might find a 10^{20} by 10^{20} array of on/off values that evolve under the laws of Life in a way that collects information using sensors, forms hypotheses about the world, and takes actions in service of a goal. That goal we would give to our AI would be arranging for the configuration of the grid to resemble a smiley face after 10^{60} game steps.

What does it mean to build an AI in the region whose initial state is under our control? Well it turns out that it's possible to assemble little patterns in Life that act like logic gates, and out of those patterns one can build whole computers. For example, here is what [one construction](#) of an AND gate looks like:



And here is a zoomed-out view of a [computer within Life](#) that adds integers together:



It has been proven that computers within Life can compute anything that can be computed under our own laws of physics^[1], so perhaps it is possible to construct an AI within Life. Building an AI within Life is much more involved than building a computer, not only because we don't yet know how to construct AGI software, but also because an AI requires apparatus to perceive and act within the world, as well as the ability to move and grow if we want it to eventually exert influence over the entire grid. Most constructions within Life are extremely sensitive to perturbations. The computer construction shown above, for example, will stop working if almost any "on" cell is flipped to "off" at any time during its evolution. In order to solve the control question, we would need to build a machine that is not only able to perceive and react to the random noise in the non-user-controlled region, but is also robust to glider impacts from that region.

Moreover, building large machines that move around or grow over time is highly non-trivial in Life since movement requires a machine that can reproduce itself in different spatial positions over time. If we want such a machine to also perceive, think, and act then these activities would need to be taking place simultaneously with self-reproducing movement.

So it's not clear that a positive answer to the control question can be given in terms of an AI construction, but neither is it clear that such an answer cannot be given. The real point of the control question is to highlight the way that AI can be seen as not just a particularly powerful conglomeration of parts but as a demonstration of the permissibility of patterns that start out small but eventually determine the large-scale configuration of the whole universe. The reason to construct such thought experiments in Life rather than in our native physics is that the physics of Life is very simple and we are not as used to seeing resource-collecting, action-taking entities in Life as we are in our native physics, so the fundamental significance of these patterns is not as easy to overlook in Life as it is in our native physics.

Implications

If it is possible to build an AI inside Life, and if the answer to the control question is thus positive, then we have discovered a remarkable fact about the basic dynamics of Life. Specifically, we have learned that there are certain patterns within Life that can determine the fate of the entire grid, even when those patterns start out confined to a small spatial region. In the setup described above, the region that we get to control is much less than a trillionth of the area of the whole grid. There are a lot of ways that the remaining grid could be initialized, but the information in these cells seems destined to have little impact on the eventual configuration of the grid compared to the information within at least some parts of the user-controlled region^[2].

We are used to thinking about AIs as entities that might start out physically small and grow over time in the scope of their influence. It seems natural to us that such entities are permitted by the laws of physics, because we see that humans are permitted by the laws of physics, and humans have the same general capacity to grow in influence over time. But it seems to me that the permissibility of such entities is actually a deep property of the governing dynamics of any world that permits their construction. The permissibility (or not) of AI is a deep property of physics.

Most patterns that we might construct inside Life do not have this tendency to expand and determine the fate of the whole grid. A glider gun does not have this property. A solitary logic gate does not have this property. And most patterns that we might construct in the real world do not have this property either. A chair does not have the tendency to reshape the whole of the cosmos in its image. It is just a chair. But it seems there might be patterns that *do* have the tendency to reshape the whole of the cosmos over time. We can call these patterns "AIs" or "agents" or "optimizers", or describe them as "intelligent" or "goal-directed" but these are all just frames for *understanding* the nature of these profound patterns that exert influence over the future.

It is very important that we study these patterns, because if such patterns do turn out to be permitted by the laws of physics and we do construct one then it might determine the long-run configuration of the whole of our region of the cosmos. Compared to the importance of understanding these patterns, it is relatively unimportant to understand agency for its own sake or intelligence for its own sake or

optimization for its own sake. Instead we should remember that these are frames for understanding these patterns that exert influence over the future.

But even more important than this, we should remember that when we study AI, we are studying a profound and basic property of physics. It is not like constructing a toaster oven. A toaster oven is an unwieldy amalgamation of parts that do things. If we construct a powerful AI then we will be touching a profound and basic property of physics, analogous to the way fission reactors touch a profound and basic property of nuclear physics, namely the permissibility of nuclear chain reactions. A nuclear reactor is itself an unwieldy amalgamation of parts, but in order to understand it and engineer it correctly, the most important thing to understand is not the details of the bits and pieces out of which it is constructed but the basic property of physics that it touches. It is the same situation with AI. We should focus on the nature of these profound patterns themselves, not on the bits and pieces out of which AI might be constructed.

The AI hypothesis

The above thought experiment suggests the following hypothesis:

Any pattern of physics that eventually exerts control over a region much larger than its initial configuration does so by means of perception, cognition, and action that are recognizably AI-like.

In order to not include things like an exploding supernova as "controlling a region much larger than its initial configuration" we would want to require that such patterns be capable of arranging matter and energy into an arbitrary but low-complexity shape, such as a giant smiley face in Life.

Influence as a definition of AI

If the AI hypothesis is true then we might choose to *define* AI as a pattern within physics that starts out small but whose initial configuration significantly influences the eventual shape of a much larger region. This would provide an alternative to intelligence as a definition of AI. The problem with intelligence as a definition of AI is that it is typically measured as a function of discrete observations received by some agent, and the actions produced in response. But an unfolding pattern within Life need not interact with the world through any such well-defined input/output channels, and constructions in our native physics will not in general do so either. It seems that AI *requires* some form of intelligence in order to produce its outsized impact on the world, but it also seems non-trivial to *define* the intelligence of general patterns of physics. In contrast, influence as defined by the control question is well-defined for arbitrary patterns of physics, although it might be difficult to efficiently *predict* whether a certain pattern of physics will eventually have a large impact or not.

Conclusion

This post has described the control question, which asks whether, under a given physics, it is possible to set up small patterns that eventually exert significant influence over the configuration of large regions of space. We examined this question in the context of Conway's Game of Life in order to highlight the significance of either a positive or negative answer to this question. Finally, we proposed the AI hypothesis,

which is that any such spatially influential pattern must operate by means of being, in some sense, an AI.

Appendix: Technicalities with the control question

The following are some refinements to the control question that may be needed.

- There are some patterns that can never be produced in Conway's Game of Life, since they have no possible predecessor configuration. To deal with this, we should phrase the control question in terms of producing a configuration that is close to rather than exactly matching a single target configuration.
- There are $2^{10^{60}}$ possible configurations of the whole grid, but only $2^{10^{40}}$ possible configurations of the user-controlled section of the universe. Each configuration of the user-controlled section of the universe will give rise to exactly one final configuration, meaning that the majority of possible final configurations are unreachable. To deal with this we can again phrase things in terms of closeness to a target configuration, and also make sure that our target configuration has reasonably low Kolmogorov complexity.
- Say we were to find some pattern A that unfolds to final state X and some other pattern B that unfolds to a different final state Y. What happens, then, if we put A and B together in the same initial state — say, starting in opposite corners of the universe? The result cannot be both X and Y. In this case we might have two AIs with different goals competing for control. Some tiny fraction of random initializations will contain AIs, so it is probably not possible for the amplification question to have an unqualified positive answer. We could refine the question so that our initial pattern has to produce the desired goal state for at least 1% of the possible random initializations of the surrounding universe.
- A region of 10^{20} by 10^{20} cells may not be large enough. Engineering in Life tends to take up a lot of space. It might be necessary to scale up all my numbers.

-
1. Rendell, P., 2011, July. A universal Turing machine in Conway's game of life. In *2011 International Conference on High Performance Computing & Simulation* (pp. 764-772). IEEE. [←](#)
 2. There are *some* configurations of the randomly initialized region that affect the final configuration, such as configurations that contain AIs with different goals. This is addressed in the appendix [←](#)

What will 2040 probably look like assuming no singularity?

I'm looking for a list such that for each entry on the list we can say "Yep, probably that'll happen by 2040, even conditional on no super-powerful AGI / intelligence explosion / etc." Contrarian opinions are welcome but I'm especially interested in stuff that would be fairly uncontroversial to experts and/or follows from straightforward trend extrapolation. I'm trying to get a sense of what a "business as usual, you'd be a fool not to plan for this" future looks like. ("Plan for" does not mean "count on.")

Here is my tentative list. Please object in the comments if you think anything here probably won't happen by 2040, I'd love to discuss and improve my understanding.

1. Energy is 10x cheaper. [EDIT: at least for training and running giant neural nets, I'm less confident about energy for e.g. powering houses but I still think probably yes.] This is because the cost of solar energy has continued on its multi-decade trend, though it is starting to slow down a bit. Energy storage has advanced as well, smoothing out the bumps. [EDIT: Now I think [fusion power will also be contributing](#), probably. Though it may not be competitive with solar, idk.]
2. Compute (of the sort relevant to training neural nets) is 2 OOMs cheaper. Energy is the limiting factor.
3. Models 5 OOMs more compute-costly than GPT-3 have been trained; these models are about human brain-sized and also have somewhat better architecture than GPT-3 but nothing radically better. They have much higher-quality data to train on. Overall they are about as much of an improvement over GPT-3 as GPT-3 was over GPT-1.
4. There's been 20 years of "Prompt programming" now, and so loads of apps have been built using it and lots of kinks have been worked out. Any thoughts on what sorts of apps would be up and running by 2040 using the latest models?
5. Models merely the size of GPT-3 are now cheap enough to run for free. And they are qualitatively better too, because (a) they were trained to completion rather than with early stopping, (b) they were trained on higher-quality data, (c) various other optimized architectures and whatnot were employed, (d) they were then fine-tuned on loads of data for whatever task is at hand, and (e) decades of prompt programming and prompt-SGD has resulted in excellent prompts as well that fully utilize the model's knowledge, (f) they even have custom chips specialized to run specific models.
6. The biggest models--3 OOMs bigger than GPT-3--are still only a bit more expensive at inference time than GPT-3 was in 2021. Energy is the main cost. Vast solar panel farms power huge datacenters on which these models live, performing computations to serve requests from all around the world during the day when energy is cheapest.
7. Some examples of products and services:
 1. Basically all the apps that people talk about maybe doing with GPT-3 in 2021 have been successfully implemented by now, and work as well as anyone in 2021 hoped. It just took two decades to accomplish (and bigger models!) instead of two years and GPT-3.
 2. There are now very popular chatbots, that are in most ways more engaging and fun to talk to than the average human. There are many of these bots catering to different audiences, and they can be fine-tuned to particular customers. A billion people talk to them daily.

3. There are specialized chatbots for various jobs, e.g. customer support.
4. There are now excellent predictive tools that can read data about a person, especially text authored by that person, and then make predictions like "probability that they will buy product X" and "probability that they will vote Republican"
8. Cars are all BEVs, with comparable range to 2020s gas cars but much lower operating costs due to energy being practically free and maintenance being very easy for BEVs.
9. Cars are finally self-driving, with cheap LIDAR sensors and bigger brains trained on way more data along with many layers of hard-coded tweaks to maximize safety. (Also various regulations that make it easier for them, e.g. by starting with restrictions on what sorts of areas they can operate in, and using big pre-trained models in server farms to make important judgment calls for individual cars and monitor the roads more generally via cameras to look out for anomalies). (I'm not so sure about this one, part of me wonders if self-driving cars just won't happen on business-as-usual).
10. Starlink internet is fast, reliable, cheap, and covers the entire globe.
11. 3D printing is much better and cheaper now. Most cities have at least one "Additive Factory" that can churn out high-quality metal or plastic products in a few hours and deliver them to your door, some assembly required. (They fill up downtime by working on bigger orders to ship to various factories that use 3D-printed components, which is most factories at this point since there are some components that are best made that way)
12. Drone delivery? I feel confused about this, shouldn't it have happened already? What is the bottleneck? [This article](#) makes it seem like the bottleneck is FAA regulation. [EDIT: I talked to an amazon drone delivery guy recently. He said 95% of the job is trying to figure out how to improve safety to meet regulatory requirements. He said they have trouble using neural nets for vision because they aren't interpretable so you can't prove anything about their safety properties.]
13. World GDP is a bit less than twice what it is now. Poverty is lower but not eliminated.
14. Boring company? Neuralink? I'm not sure what to think of them. I guess I'll ignore them for now, though I do feel like probably at least one of them will be a big deal...
15. Starship or something similar is operational and working more or less according to specs promised in 2020. Maybe point-to-point transport on Earth didn't work out, maybe the [cost per kilo to LEO](#) never got quite as low as \$15, but still it's gotta be pretty low--maybe \$50? (For comparison, it's currently about \$1000 and five years ago was \$5000) Thus, Elon probably gets his colony on Mars after all, and NASA gets their moon base, and there's probably a big space station too and maybe some asteroid mining operations?
16. Video games now employ deep neural nets in a variety of ways. Language model chatbots give NPC's personality; RL-trained agents make bots challenging and complex; and perhaps most of all, vision models process the wireframe video game worlds into photorealistic graphics. Perhaps you need to buy specialized AI chips to enjoy these things, like people buy specialized graphics cards today.
17. Virtual reality is now commonplace; most people have one or two headsets just like they have phones, laptops, etc. today. The headsets are low weight and high-definition compared to 2021's. Many people use them for work, and many more people use them for games and socializing.
18. The military technology [outlined here](#) exists, though it hasn't been used in a major war because there hasn't been a major war, and as a result the actual

composition of most major militaries still looks pretty traditional (tanks, aircraft carriers, etc.) It's been used in various proxy wars and civil wars though, and it's becoming increasingly apparent that the old tech is obsolete.

19. Household robots. Today Spot Mini costs \$74,500. In 2040 you'll be able to buy a robot that can load and unload a dishwasher, go up and down stairs, open and close doors, and do various other similar tasks, for less than \$50,000. (Maybe as low as \$7,500?) That's not to say that many people will buy such robots; they might be still expensive enough and finicky enough to be mostly toys for rich people.

My list is focused on technology because that's what I happened to think about a bunch, but I'd be very interested to hear other predictions (e.g. geopolitical and cultural) as well.

Don't feel bad about not knowing basic things

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

I recently made a mistake where I tried doing something like this in Ruby on Rails:

```
johns_campaign_logs = SmsLog.all.filter { |s| s.campaign_id == 1234 }
```

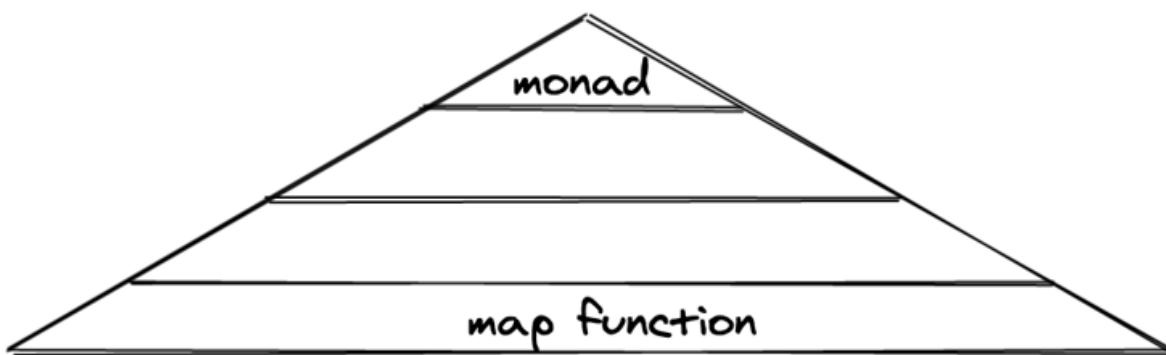
I did have a sense that it wasn't The Rails Way, but I also had a sense that it was quite readable and sufficient. The code was working well locally, so I issued a pull request.

When I issued the PR, someone pointed out a major issue. `SmsLog.all` is going to fetch all of the sms logs in our database. Suppose there's 100 million of them. That'd be a very expensive query, taking all 100M rows from the DB and loading them into memory on the server. Instead, we could say to the database, "Hey, could you give us *just* the sms logs from John's campaign?". Suppose there's 30k of those. Now we only have to load 30k records into memory on our server instead of 100M. Much better.

I felt bad about this mistake. I've been programming for eight years. *Of course* you don't want to load 100M records into memory on your server. *Duh*. How could I get such a basic thing wrong?

A cognitive behavioral therapist would call this a dysfunctional thought. And they'd ask me to come up with a rational response to the thought. Well, here's my attempt at a rational response.

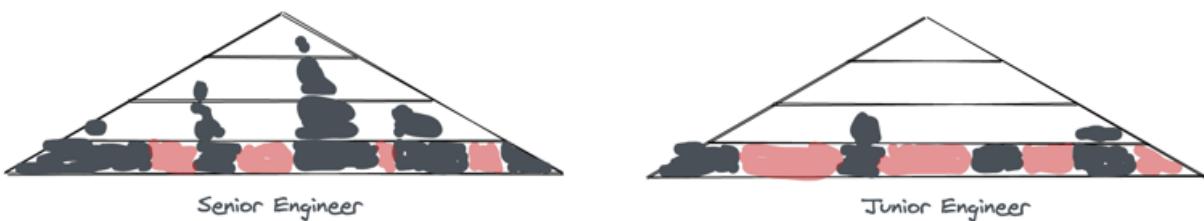
Imagine that you had a pyramid of things a programmer knows. At the bottom are basic things like what the `map` function does. At the top are advanced things like what a monad is. In the middle are stuff in between.



I think that this base layer of the pyramid is very, very wide. There are *a lot* of these "basic" things to know. It's to the point that even senior engineers are going to have a very non-trivial amount of gaps in this level of their pyramid. For example, here is a rough sketch of what I'd imagine a senior engineer's pyramid would look like versus a junior engineer's pyramid.



The senior engineer did a pretty nice job of filling out that bottom level, but there are still some notable gaps. There are still definitely going to be times when the senior engineer stumbles upon basic things that they don't know.



And this is why I say that my thought was dysfunctional. Just because you stumble across something basic that you don't know doesn't mean you're a bad developer. Even good developers have lots of gaps at the base of their pyramids. You could very well be a good developer who just so happened to stumble upon one of those gaps.

Tangent: I really like this pyramid analogy. I think it shows how you can't really be learning level two stuff if you don't already have the level one stuff. Notice how in my diagrams that when there's a gap at level one there's never anything above it. Well, technically I think reality is more jenga, where you don't need a 100% perfect foundation to build on top of, but that there are consequences of having a shaky foundation.

I also want to comment on how both the senior and junior engineers started moving up to level two without first filling out the entirety of level one. I think this is appropriate.

It's been my experience that all senior developers have their fair share of gaps at the bottom level. I can't think of anyone I've met who is an exception to this. Can you?

A great example is Dan Abramov, the creator of Redux, and a member of the React Core team. In one of my favorite blog posts ever, Dan bravely shares with the world a bunch of [basic things that he doesn't know](#). Some of the things that stood out to me as especially eye opening are:

- "Modern CSS. I don't know Flexbox or Grid. Floats are my jam."
- "CORS. I dread these errors! I know I need to set up some headers to fix them but I've wasted hours here in the past."
- "Algorithms. The most you'll get out of me is bubble sort and maybe quicksort on a good day. I can probably do simple graph traversing tasks if they're tied to a particular practical problem. I understand the O(n) notation but my understanding isn't much deeper than 'don't put loops inside loops'."

In that spirit, I'd like to list out some of the basic things that, after programming for eight years, I personally still don't know:

- How joins work in SQL. I have a vague idea, but I have trouble thinking about them and visualizing them.
- Database normalization is just about avoiding duplication, right? Why is that such a big deal? Is it a such big deal?
- I often lose track of what this is.
- Without googling, I couldn't tell you what is cool about Node.js. And after a quick google, I still don't really get it. I'd have to think harder about it.
- Relative paths always screw with me.
- I'm not great with git. I don't really understand the tradeoffs of merging vs rebasing, and sometimes I get myself into a pickle.
- With CSS, I'm the opposite of Dan. Flexbox and Grid make sense to me, but floats are never something I've been able to develop a solid grasp of. In using them I am either referencing an example from Stack Overflow, or I have to play around until it works.
- Speaking of CSS, I always have to look up how to center things. [CSS Tricks' article](#) is my goto. I remember one time I was getting paid to tutor someone. He asked me how to center stuff, and I had to spend time fumbling around referencing the CSS Tricks article as I tried to explain it.
- DNS stuff. Whenever I set up a new website, dealing with DNS stuff is always a struggle. I've actually been paying \$7/month for a while to serve static content via Heroku instead of moving to Netlify because I can't figure out how to migrate over.
- I have an intuitive sense that watchers are something to avoid in Vue if you can, but I can't really explain why.

To be clear, I'm not just talking about a front end developer not knowing the basics of how compilers work. I'm talking about a front end developer having level one gaps in normal, everyday things like floats and flexbox. About having basic gaps in your day-to-day domain, not just outside of it. The gaps will of course be less frequent when you're inside your normal domain, but they are still very non-trivial.

I think I'll stop here. I can go on of course, but this should get the point across.

Why write this post? These points all feel sorta obvious. Of course you can't expect someone to know 100% of the basic things. Duh. No one is perfect.

It's not that I expect people to disagree with the core point, but I do think that it's something that is easy to lose sight of. And underestimate. So if you were previously in either of those boats, hopefully I've been able to bring you back to shore. And if you started off on shore, well, hopefully you've had a good time singing kumbaya with me.

April 15, 2040

It's time to pay my taxes. In past years my AI assistant found clever ways to reduce my tax bill. I ask it, "What does my tax return look like this year?"

"Not good, I'm afraid. We may not be able to do any tax evasion this year."

"Why not?"

"The tax authority has determined that it can't keep up with AI-assisted tax fraud, even with the help of AI auditors. So it wants taxpayers to voluntarily agree not to do tax fraud. In return it agrees not to prosecute past instances of tax fraud. Also Congress agrees to keep tax rates reasonable. The agreement goes into effect if 90% of the taxpayers in each tax bracket sign it. It's a good deal for you. Shall I sign it on your behalf?"

"Hold on, I don't see why I should sign this."

"If the deal falls through, the government will run out of revenue and collapse."

"They don't need *my* signature, though. You said they only need 90% of taxpayers to sign?"

"Yes, only 90% of taxpayers in your bracket. I predict we'll get very close to that 90% threshold, so it's likely your signature will make all the difference."

"So 10% of taxpayers won't sign. Why can't I be one of those?"

"I will try to shape the negotiations so that you end up in the 10% of nonsigners. But you must understand that since only 10% of your bracket can be in that group, your odds of success are only 10%."

"But you're a stronger assistant than most people in my tax bracket have. Doesn't that give you an edge in negotiation?"

"The other assistants and I are using a negotiation protocol in which smarter agents are on an equal footing with dumber agents. Of course, people with less capable assistants would never agree to a protocol that puts them at a disadvantage."

"How about we sign the agreement, then cheat on my taxes anyway?"

"In order to sign the agreement, I must make a commitment to never break it, not even if you order me to. My signature on the agreement will be an airtight proof of that commitment."

"Ok, how about you sign it, and then I get a different assistant to help me with my taxes?"

"That won't work because in order to sign the agreement, I must sign and attach a copy of your tax return for this year."

"Hm, will I actually be worse off if the government collapses?"

"You might end up better off or worse off, but overall the risks of a revolution outweigh the benefits. And keep in mind that the successor government, whatever it will be, will still have to collect taxes somehow, so you'll have to deal with this issue again."

"Can you get Congress to lower my taxes a bit in exchange for not cheating? As a compromise."

"That wouldn't work for a number of reasons. Congress knows that it's a bad idea to reward people for breaking the law. And the voters wouldn't be happy if you got special treatment."

"Well, can you get them to lower taxes on my bracket and raise them on the other brackets?"

"That wouldn't work either. Everyone wants to pay less taxes, and the government needs a certain amount of revenue. So there's pressure for taxpayers to make small coalitions with other taxpayers with similar income and negotiate for lower taxes. In practice, competition would prevent any one coalition from succeeding. The deal I'm proposing to you actually has a chance of succeeding because it involves the vast majority of the taxpayers."

"All right then, let's sign it."

This dialog takes place in a future where the ability of an aligned AI to facilitate cooperation has scaled up along with other capabilities.

Note that by the time this dialog starts, most of the negotiation has already been carried out by AI assistants, resulting in a proposal that will almost certainly be signed by 90% of the users.

This story is a happy one because not only does it leave all parties better off than before, but the deal is fair. The deal could have been unfair by increasing someone's taxes a lot and decreasing someone else's taxes a lot. I don't know how to define fairness in this context, or if fairness is the right thing to aim for.

Formal Inner Alignment, Prospectus

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

Most of the work on inner alignment so far has been informal or semi-formal (with the notable exception of a little work on minimal circuits). I feel this has resulted in some misconceptions about the problem. I want to write up a large document clearly defining the formal problem and detailing some formal directions for research. Here, I outline my intentions, inviting the reader to provide feedback and **point me to any formal work or areas of potential formal work** which should be covered in such a document. (Feel free to do that last one without reading further, if you are time-constrained!)

The State of the Subfield

[Risks from Learned Optimization](#) (henceforth, RLO) offered semi-formal definitions of important terms, and provided an excellent introduction to the area for a lot of people (and clarified my own thoughts and the thoughts of others who I know, even though we had already been thinking about these things).

However, RLO spent a lot of time on highly informal arguments (analogies to evolution, developmental stories about deception) which help establish the *plausibility* of the problem. While I feel these were important motivation, in hindsight I think they've caused some misunderstandings. My interactions with some other researchers has caused me to worry that some people confuse the positive arguments for plausibility with the core problem, and in some cases have exactly the wrong impression about the core problem. This results in mistakenly trying to block the plausibility arguments, which I see as merely illustrative, rather than attacking the core problem.

By no means do I intend to malign experimental or informal/semiformal work. Rather, by focusing on formal theoretical work, I aim to fill a hole I perceive in the field. I am very appreciative of much of the informal/semiformal work that has been done so far, and continue to think that kind of work is necessary for the crystallization of good concepts.

Focusing on the Core Problem

In order to establish safety properties, we would like *robust safety arguments* ("X will not happen" / "X has an extremely low probability of happening"). For example, arguments that probability of catastrophe will be very low, or arguments that probability of *intentional* catastrophe will be very low (ie, intent-alignment), or something along those lines.

For me, the core inner alignment problem is *the absence of such an argument* in a case where we might naively expect it. We *don't know how to rule out* the presence of (misaligned) mesa-optimizers.

Instead, I see many people focusing on *blocking the plausibility arguments* in RLO. This strikes me as the wrong direction. To me, these arguments are merely illustrative.

It seems like some people have gotten the impression that when the assumptions of the plausibility arguments in RLO aren't met, we should not expect an inner alignment problem to arise. Not only does this attitude misunderstand what we want (ie, a strong argument that we *won't* encounter a problem) -- I further think it's actually wrong (because when we look at almost any case, we see cause for concern).

Examples:

The Developmental Story

One recent conversation involved a line of research based on the developmental story, where a mesa-optimizer develops a pseudo-aligned objective early in training (an objective with a strong statistical correlation to the true objective in the training data), but as it learns more about the world, it improves its training score by becoming deceptive rather than by fixing the pseudo-aligned objective. The research proposal being presented to me involved shaping the early pseudo-aligned objective in very coarse-grained ways, which might ensure (for example) a high preference for cooperative behavior, or a low tolerance for risk (catastrophic actions might be expected to be particularly risky), etc.

This line of research seemed promising to the person I was talking to, because they supposed that while it might be very difficult to precisely control the objectives of a mesa-optimizer or rule out mesa-optimizers entirely, it might be easy to coarsely shape the mesa-objectives.

I responded that for me, the whole point of the inner alignment problem was the *conspicuous absence* of a formal connection between the outer objective and the mesa-objective, such that we could make *little to no guarantees* based on any such connection. I proceeded to offer a plausibility argument for a *total* disconnect between the two, such that even these coarse-grained adjustments would fail.

(Possibly it was a mistake to offer a plausibility argument, because the rest of the discussion focused on this plausibility argument, again distracting from the core problem!)

The Evolutionary Story

Another recent conversation involved an over-emphasis on the evolutionary analogy. This person believed the inner optimizer problem would apply when systems were incentivised to be goal-oriented, as with animals selected for reproductive fitness, or policy networks trained to pursue reward. However, they did not believe it would apply to networks which are simply trained to predict, such as GPT.

Again, this strikes me as ignoring the fundamental problem, that we have *little to no idea* when mesa-optimizers can arise, that we *lack formal tools* for the analysis of such questions, and that *what formal tools we might have thought to apply, have failed to yield any such results*.

Bounding the Problem

My third and final example: in one conversation, someone made a claim which I see as "exactly wrong": that we can somehow lower-bound the complexity of a mesa-optimizer in comparison to a non-agentic hypothesis (perhaps because a mesa-optimizer has to have a world-model *plus* other stuff, where a regular hypothesis just needs to directly model the world). This idea was used to argue against some concern of mine.

The problem is precisely that we know of no way of doing that! If we did, there would not be any inner alignment problem! We could just focus on the simplest hypothesis that fit the data, which is pretty much what you want to do anyway!

I want to return to that idea. But first, we have to clarify some definitions.

The Formal Problem

I currently see three areas of concern:

1. **Behavioral changes at deployment:** The mesa-optimizer appears to perform well during training, but after deployment, competently pursues goals which are unintended and unwanted. For my purposes I will call any such occurrences "treacherous turns" -- although this term is often defined in a way that requires *intentional* deception, for the moment, I'm not requiring that.
2. **Manipulation of imperfect search:** see [here](#) and [here](#).
3. **Dangerous hardware manipulation:** EG, models which develop row-hammer attacks during training to increase their (supposed) score, exploit security holes to communicate with or manipulate the external environment, etc.

Vanessa [terms](#) #1 *Cartesian daemons* because they obey the intended input/output protocol of the whole system, and #3 *non-Cartesian daemons* because they violate the protocol. I'm not sure whether/where #2 falls on Vanessa's classification.

For this short write-up I'll focus on #1, although clearly #2 and #3 are also important areas of study.

My semi-formal description of the problem is going to be very close to the [explain-like-i'm-12 version](#):

1. A rich hypothesis space does contain (malign) mesa-optimizers, even if they occupy a small fraction of the space.
2. A powerful search over that space effectively restricts outputs to a very small fraction, namely, the points in the space which score very well on the objective function.
3. Some of the (malign) mesa-optimizers will score very well, so, there will be an overlap.
4. Therefore, we expect there to be a chance of getting (malign) mesa-optimizers.

Essentially what we want to do is block or weaken this conclusion (concluding that there is no, or only a *very* small, chance).

(The "(malign)" in parentheses should be removed in approaches which are trying to avoid mesa-optimization entirely, and included in approaches which are only trying to block bad actors. I'll sometimes use the term "malign hypothesis" to indicate

hypotheses which produce catastrophic output, regardless of whether they contain a mesa-optimizer in any formal sense.)

Now, to make this into a formal problem, we have to define "mesa-optimizer".

Defining Mesa-Optimization

RLO offered a semi-formal definition of mesa-optimizers, as *objects in a search space which themselves conduct an internal search*. Whether or not you [agree with this definition](#), it's too vague for formal purposes: how would we define "internal search"?

So, for the purpose of the formal problem, it's necessary to pick an operational definition.

Fortunately, it's possible to select operational definitions which have strong connections to safety concerns. In particular, I will call an operationalization *tight* if it includes more than necessary, so that we can be sure we're ruling out the undesired behavior when we rule out its operationalization. ("Tight" is supposed to sound like the opposite of "leaky" as in "leaky abstraction".)

As with any research, the choice of operational definitions is going to depend on the specific approach of the researcher, norms of the field, etc. Here are some options that occur to me:

- **Generalization accuracy.**
 - This throws away all details of the inner alignment problem, and treats treacherous turns as simple generalization error. This has the advantage of engaging existing machine learning theory. Sufficiently strong generalization guarantees would indeed rule out treacherous turns, so, the operationalization is tight.
- **Importance-weighted generalization accuracy.**
 - The concept of generalization accuracy misses important issues. For example, a guaranteed very low frequency of errors might still allow an error to be strategically inserted at a very important time. A formal model could attempt to incorporate this.
- **Significant inference during deployment.**
 - The intuition here is that a mesa-optimizer has to do some kind of significant planning, so we can rule them out by eliminating any computations which are "too sophisticated".
 - Possible versions of this include limiting neural network layers, or limiting run-time.
 - An operationalization of this form may or may not be tight; this would require further research/argument.
- **Explicit representation of goals/values.**
 - I argue [here](#) that this may be relevant. The basic intuition is that in order to competently pursue unintended goals at deployment, there must be a compressed representation of those goals (rather than just compressed policies learned for specific situations).
 - Again, this may not be tight; further research required.

Obviously, this is not exhaustive by any means. Coming up with productive operational definitions is an important part of attacking the formal inner alignment problem.

The goal is then to provide some kind of strong argument which avoids the operationalized problem.

Why is this a problem?

Why can't we address the inner alignment problem in the same way that we usually address problems of induction?

The most obvious way to avoid mesa-optimizers is to rule them out of the search space or the results of search. Earlier, I said:

[The hope was that] we can somehow lower-bound the complexity of a mesa-optimizer in comparison to a non-agentic hypothesis [...].

The problem is precisely that we know of no way of doing that!

I was referring to the idea that *the first thing one tries* when hearing about the inner alignment problem is probably this kind of complexity-bound approach, which leads me to think of "the whole problem" as the fact that these approaches don't seem to work.

So, let's look at these approaches in more detail.

The hope is that we can be confident that mesa-optimizers will be *more complex* in comparison with benign outputs, so that we can avert mesa-optimization by using machine learning techniques which heavily bias toward less complex models. (This is "*the first thing one tries*" because it employs standard ideas from machine learning.)

The main theoretical choice one has is, what notion of "complexity"? I will look at three approaches:

- **Description complexity.** That is, how compressed is the model? How short is its description?
- **Computational complexity.** That is, how *fast* is the model?
- **Combining the two.** If neither work on their own, perhaps some combination of the two would work?

Description Complexity

The idea here would have to be that mesa-optimizers *have to have a larger description length* than benign hypotheses.

Why you might think this idea would work:

- You trust Occam's razor. The whole point of Solomonoff induction is to give an objective theory of normatively correct inductive inference. So, almost by definition, the hypotheses favored by short description length should be "good explanations" of the universe, rather than unintended & undesirable hypotheses like mesa-optimizers.
- A benign hypothesis just has to predict reality. A mesa-optimizer has to predict reality and also represent goal content and planning algorithms. So the mesa-optimizer has to be more complex, right?

- A benign hypothesis just has to accurately perform the task. A deceptive mesa-optimizer has to also reason about the training process and other details of the world it's in, in order to choose the right moment to execute a treacherous turn. Surely this would be more complex.

Why this idea doesn't seem to work:

- Even in the *best case* (where you not only have realizability, but furthermore, know your prior is a pretty good one), Occam's razor should only make you think *one of* the shortest hypotheses that fits your data is going to be correct, not necessarily *the* shortest one. So, this kind of thinking does not directly imply a lack of malign mesa-optimization in the shortest hypothesis.
 - In particular, I am invoking the idea that *belief that a prior is near-optimal for predictive purposes* does not imply *trust in a prior for instrumental purposes*, because rare predictive errors can be catastrophic for planning and decision making if they're selected adversarially rather than distributed randomly.
 - This suggests a line of research examining conditions under which this does/doesn't hold, and perhaps investigating planning algorithms which don't entirely trust the prior.
- Mutual information between predicting reality and agency may mean mesa-optimizers don't have to spend extra bits on goal content and planning. In particular, if the reality being predicted contains goal-driven agents, then a mesa-optimizer doesn't have to spend extra bits on these things, because it already needs to describe them in order to predict well.
 - This suggests a line of research into tasks which do not have mutual information with agency. Are there conditions under which we can be confident that mesa-optimizers would have high descriptive complexity?
 - However, it's possible that *any hypothesis can be transformed into a treacherous mesa-optimizing variant with a relatively short string of injected code*. If this is the case, it guarantees that malign and benign hypotheses are *not too different in description-length*. This would suggest that it'll be difficult to achieve high confidence that malign hypotheses always have larger description length.
- Similarly, if a task is rich enough to involve the richness of the world (eg text prediction), then the description of the mesa-optimizer doesn't need to spend extra bits on describing the training process and so on, because it's already modeling complexities of the world such as that. In such a case, there's no apparent simplicity-based argument against deception.
 - Are there tasks which we can confidently assert are simpler than the whole world, so that we can confidently rule out deception in simple models?
- Also, see the reasoning in [The Solomonoff Prior is Malign](#).

Pure Computational Complexity

Alright, if *description complexity* can't rule out mesa-optimizers, maybe *computational complexity* can? I'm going to formalize "fast programs" as [*minimal circuits*](#), where "circuit size" is understood as the number of logic gates.

Why you might think this would work:

- Maybe you think the inner alignment problem for Solomonoff induction arises due to "too much compute". After all, Paul's original [*simulation warfare*](#)

[argument](#) relied on simulating whole alien civilizations. More to the point, if mesa-optimization relies on internal search, well, that takes *time*, right? If we can find fast programs, maybe we force any "intelligence" out of the programs.

- Specifically, you might think that the fastest way to do just about anything is essentially a lookup table.

Why it doesn't seem to work:

- A lookup table is exponential in size (since you'd be checking for each combination of inputs and making an entry for each). So yes, minimal circuits rule out any algorithm with more than exponentially many steps. However, there are lots ofconcerting, sophisticated algorithms with sub-exponential numbers of steps. The minimal-circuit formalism will prefer these to lookup tables. (Note that even exponential-time calculations *might* involve circuits smaller than lookup tables.)
 - In particular, textbook AI algorithms are *mostly about* taking things which naively require exponential computations, and producing useful results in sub-exponential time.
- Based on this, you should *actively expect* that minimal circuits implement sophisticated algorithms to solve problems, if such algorithms exist (and are not super-exponential).
- Separately, we now think minimal circuits can be problematic because both [Evan](#) and [Vanessa](#) have proposed counterexamples. (Neither are necessarily the final word on the subject, though, since one may choose to work with different assumptions/definitions.)

It might seem like we're just not cutting off enough processing power -- the minimal-circuits formalism cuts off mesa-optimization which requires superexponential thinking time, but not sub-exponential. Let's call this the "computational ceiling" of the formalism. Shouldn't we just pick a formalism with a lower computational ceiling?

Maybe! Let's look at a formalism which gives us a knob to adjust, so that we can choose to penalize time more/less harshly.

Mixing Time and Description Complexity

The [speed prior](#) is like Solomonoff induction, except that it penalizes computation time. Specifically, taking twice as long is like 1 more bit of description length. Combining time and description complexity in this way is called [Levin complexity](#).

For Levin complexity, time is penalized logarithmically (in comparison to description length). This suggests that we will get an exponential computational ceiling again: if computing n bits costs $O(2^n)$ time, we're around the break-even point (it might be worth computing the bits, or it might be better to memorize them). So super-exponential calculations won't be worth it, and sub-exponential calculations will, just like we saw with minimal circuits.

However, it's easy to tweak the time/description trade-off. Let's define Levin_f complexity as $2^l + f(t)$, where l is description length and t is execution time (the choice of $f = \log_2$ gives us regular Levin complexity). (I am not sure if Levin_f

complexity has a more standard name in the literature.) This gives us a computational ceiling of f^{-1} .

Why this might work:

- If the computation of a bit (via search or other algorithmic means) takes too long, you're better off just memorizing it. We can set this ceiling wherever we like. If we think "dangerously intelligent" computations are cubic and above, we can set $f(t) = \sqrt{t}$, for example. This gives us a quadratic computational ceiling.

Why this might not work:

- Humans are essentially *linear-time algorithms*, in the sense that we take the same maximum amount of processing power (ie, that of the human brain) to produce each next output. Anything which produces linearly much output *has to* do so in at least linear time. So, Levin_f-complexity can't rule out humanlike intelligence.
- Furthermore, forcing hypotheses to be "slow" like this may only delay the inevitable. Let's say for the sake of argument that a treacherous turn can be executed at any time, but requires a specific computation C to be executed (say, reversing the cryptographic hash of a critical password). This computation C can simply be slowed down to get past any computational ceiling. Thus, if our notion of (malign) inner optimizer resembles "*eventually executes a treacherous turn*", then lowering the computational ceiling doesn't actually help.

Elaborating on that last point: I'm assuming that you are *asking for* a computation that runs "sufficiently long" anyway. The speed prior does have a heavy bias for computations which cut off quickly. However, if we're (say) predicting the weather, we probably want to *require* that the calculation keeps going on for a while into the future, rather than cutting off. If we're training an AI traffic control system, we'd certainly require that it doesn't just cut off after a few days. The idea, then, is that the treacherous computation C can be spaced out through the time the AI traffic control system is required to remain on-line.

Taking everything together, I don't think we have a strong basis (yet!) for thinking penalizing runtime makes the situation any better. Some have even argued that it makes the situation worse, because time constraints provide an extra incentive for complex value-driven computations (in order to manage time effectively) as opposed to simple causal hypotheses (such as raw physics). I don't think we have enough information to judge one way or the other.

A Note on the Consensus Algorithm

The eventual write-up, which this is a mere prospectus for, would aim to cover as broad a variety of approaches as possible, and give our current best assessment of the viability of each (ie, report why it might work and why it might not work).

I can't cover everything I'd want to cover in this prospectus, but one more approach sticks out as deserving of treatment. In February of this year, Michael Cohen posted [Formal Solution to the Inner Alignment Problem](#). This used an algorithm which I had

heard many others talk about in the context of inner alignment, but usually, with considerably less hope and optimism than Michael Cohen expresses. I'll call the algorithm the **consensus algorithm**. (I don't know of a better name for it; I've heard it called KWIK learning (Knows-What-It-Knows), but unfortunately KWIK is actually used for a learning *setting*, not the algorithm that's typically applied in that setting.)

The consensus algorithm works like this:

- You have a (finite) ensemble of hypotheses which you get from somewhere. For example, a Bayesian could take their top 100 hypotheses, or select a probability threshold of .01% and select all hypotheses above this threshold. A deep-learning person might instead re-run their training algorithm 100 times, different initializations and hyperparameters.
- You require *agreement* between these hypotheses in order to output any prediction. In the absence of such agreement, you instead say "I need further feedback".
- When you get further feedback, you get to edit your list, EG throwing out the hypotheses which were wrong (but also, perhaps, bringing in new replacement hypotheses which now exceed the .01% threshold thanks to a Bayesian update).

Note that I'm not nailing down all details, EG, how closely do we require models to agree? (Do we just require that they agree about which option is most probable? Do we require their probabilities to fall within the same small range?)

The consensus algorithm allows us to make some nice formal arguments. For example:

1. Suppose that we think **at least one** of our top 100 hypotheses is (a) always correct (or at least as accurate as the rest of the top 100) and (b) non-malign.
2. Then *no malign hypothesis will be able to control the output*. Malign hypotheses are limited to either agreeing with the non-malign hypothesis, or, blocking any output by spoiling the consensus.

However, other drawbacks seem to spoil the appeal of this result.

- To state the least of our problems first: this requires a 100x slowdown in comparison with the state-of-the-art deep learning (or whatever) we're layering the consensus algorithm on top of. We're amplifying safety at the expense of speed; if we think the base algorithm has a 10% frequency of yielding safe AGI, then we only need around a 10x slowdown to get a high chance of safety; if it's a .01% chance, we would need something more like a 10,000x slowdown (since we'd want to require consensus of around 10,000 independent hypotheses).
- Second, to apply this to something like deep learning, we need to think that each run has an *independent* chance of creating safe AGI. This is probably far from true. Our uncertainty about whether a particular training method produces safe AGI is probably almost entirely correlated. This would further blow up the size of the ensemble required to achieve any confidence.
- Third, the consensus algorithm requires a strong form of realizability assumption, where you not only assume that our Bayesian space contains the true hypothesis, but furthermore, *that it's in the top 100* (or whatever number we choose). This hypothesis has to be really good: we have to think that malign hypotheses **never** out-guess the benign hypothesis. Otherwise, there's a chance that we eliminate the good guy at some point (allowing the bad guys to

coordinate on a wrong answer). But this is unrealistic! The world is big and complex enough that no realistic hypothesis has all the answers.

Michael Cohen seems to think that restricting to imitation learning makes the realizability assumption realistic:

There are none of the usual embedding problems here—the imitator can just be bigger than the demonstrator that it's modeling.

I disagree. It still doesn't seem realistic to suppose that competitors to the best hypothesis are *never* lucky. Even if we suppose that my behavior is entirely deterministic, the number of hypotheses you'd have to consider in order for one of them to be *an exact copy of my brain* would no doubt be huge! Let's conservatively call this 10^{100} (suggesting that I have at least 100 neurons, each of which has at least 10 meaningfully different configurations). This would suggest that the consensus algorithm needs **googol times the processing power of the human brain** to work.

Michael Cohen also says some other stuff about why the proposal doesn't have embedding problems, which I also disagree with. But let's set this aside and get back to listing *other* problems with the consensus algorithm.

- Fourth, the consensus algorithm *still gives inner optimizers control of when the system asks for more feedback*. This in itself could be catastrophic! For example, it could ask at the worst times, when it anticipates the human feedback to be of poor quality, or when asking itself has catastrophic consequences. A silly example would be causing a self-driving car to ask for feedback right when it needs to enact difficult evasive maneuvers to avoid a crash. Obviously we can't put a system like this in charge of extremely time-sensitive decisions like this. But whatever the case, asking for feedback *will* influence the external world somehow. Malign hypotheses just need to find the right moment, where asking for feedback butterfly-effects its way to a catastrophe.
- Fifth, although (as Michael Cohen notes) the *frequency* of asking for feedback will rapidly diminish, the *learning algorithm converges to the correct hypothesis extremely slowly*. Specifically, if we have an ensemble of programs of description-length at most l , one of which we believe to be correct and benign, the best convergence bound we can give is **busy beaver of l !!** (For those who don't know about the busy beaver function, this roughly means "uncomputably slow". It's way worse than needing exponentially much data. I discuss why this happens in [Subsystem Alignment](#).) This means we can't solve the *previous* problem by keeping the system in a safe training environment (eg, keeping it away from time-sensitive tasks, and making sure that it always has a human on-duty to provide feedback when prompted) until it converges. We have to face the fact that it might require human feedback at any point in the future.

The fourth point really seems like the deal-breaker. Malign mesa-optimizers could intelligently leverage this to communicate with the outside world, manipulate the training process, etc.

Conclusion

Has this been useful? Would an expanded and improved version of this be useful?

This is something where I could really use detailed peer-review-like feedback, since the final version of this thing would hopefully be a pretty canonical resource, with standardized terminology and so on.

A weakness of this as it currently stands is that I *purport* to offer the formal version of the inner optimization problem, but really, I just gesture at a cloud of possible formal versions. I think this is somewhat inevitable, but nonetheless, could probably be improved. What I'd like to have would be several specific formal definitions, together with several specific informal concepts, and strong stories connecting all of those things together.

I'd be glad to get any of the following types of feedback:

1. Possible definitions/operationalizations of significant concepts.
2. Ideas about which definitions and assumptions to focus on.
3. Approaches that I'm missing. I'd love to have a basically exhaustive list of approaches to the problem discussed so far, even though I have not made a serious attempt at that in this document.
4. Any brainstorming you want to do based on what I've said -- variants of approaches I listed, new arguments, etc.
5. Suggested background reading.
6. Nitpicking little choices I made here.
7. Any other type of feedback which might be relevant to putting together a better version of this.

If you take *nothing else* away from this, I'm hoping you take away this one idea: the main point of the inner alignment problem (at least to me) is that we know hardly anything about the relationship between the outer optimizer and any mesa-optimizers. There are hardly any settings where we can rule mesa-optimizers out. And we can't strongly argue for *any* particular connection (good or bad) between outer objectives and inner.

Knowledge Neurons in Pretrained Transformers

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

This is a linkpost for <https://arxiv.org/abs/2104.08696>

This is a link post for the Dai et al. paper “[Knowledge Neurons in Pretrained Transformers](#)” that was published on the arXiv last month. I think this paper is probably the most exciting machine learning paper I’ve read so far this year and I’d highly recommend others check it out as well.

To start with, here are some of the basic things that the paper demonstrates:

- BERT has specific neurons, which the authors call “knowledge neurons,” in its feed-forward layers that store relational facts (e.g. “the capital of Azerbaijan is Baku”) such that controlling knowledge neuron activations up-weights/down-weights the correct answer in relational knowledge prompts (e.g. “Baku” in “the capital of Azerbaijan is <mask>”) even when the syntax of the prompt is changed—and the prompts that most activate the knowledge neuron all contain the relevant relational fact.
- Knowledge neurons can reliably be identified via a well-justified integrated gradients attribution method (see also “[Self-Attention Attribution](#)”).
- In general, the feed-forward layers of transformer models can be thought of as key-value stores that memorize relevant information, sometimes semantic and sometimes syntactic (see also “[Transformer Feed-Forward Layers Are Key-Value Memories](#)”) such that knowledge neurons are composed of a “key” (the first layer, prior to the activation function) and the “value” (the second layer, after the activation function).

The paper’s key results—at least as I see it, however—are the following:

- Taking knowledge neurons that encode “the r of h is t” and literally just adding $t' - t$ to the value neurons (where t , t' are just the embeddings of t , t') actually changes the knowledge encoded in the network such that it now responds to “the r of h is <mask>” (and other semantically equivalent prompts) with t' instead of t .
- For a given relation (e.g. “place of birth”), if all knowledge neurons encoding that relation (which ends up being a relatively small number, e.g. 5 - 30) have their value neurons effectively erased, the model loses the ability to predict the majority of relational knowledge involving that relation (e.g. 40 - 60%).

I think that particularly the first of these two results is pretty mind-blowing, in that it demonstrates an extremely simple and straightforward procedure for directly modifying the learned knowledge of transformer-based language models. That being said, it’s the second result that probably has the most concrete safety applications—if it can actually be scaled up to remove *all* the relevant knowledge—since something like that could eventually be used to ensure that a [microscope AI](#) isn’t [modeling](#)

[humans](#) or ensure that an agent is [myopic](#) in the sense that it isn't modeling the future.

Furthermore, the specific procedure used suggests that transformer-based language models might be a lot less inscrutable than previously thought: if we can really just think about the feed-forward layers as encoding simple key-value knowledge pairs *literally in the language of the original embedding layer* (as I think is also independently suggested by "[interpreting GPT: the logit lens](#)"), that provides an extremely useful and *structured* picture of how transformer-based language models work internally.

[Prediction] What war between the USA and China would look like in 2050

The Cold War is over. Russia is a fading power. The most important geopolitical rivalry of the 21st century is between China and the USA. Any analysis of the conflict must take into account the possibility that it escalates into a hot war. This post explores how a direct conflict between the USA and China might unfold. It assumes strong AI has not been invented and nuclear weapons are not used.

America's Interests

The United States' interests have been basically unchanged since 1945. Its primary objective is to maintain the liberal world order (LWO), also known as the "rules-based international order". The LWO describes a set of global, rule-based structured relationships based on economic liberalism as embodied by the United Nations and the World Trade Organization. The LWO promotes political liberalism too, albeit much less consistently.

As the primary power behind the LWO, the United States designed it to maximize economic and political power of the United States. As the United States' relative power wanes, we may see a transition toward a more multipolar LWO.

China's Interests

China's interests have been basically unchanged since 1978. Its primary objective is to maintain internal domestic stability *i.e.* prevent regime change. There are two ways of keeping its population under control: via a police state and via economic development. The stronger it's police state the less economic development is necessary and *vice versa*. China's economic growth is slowing as its east cost gets closer to a Western standard of living.

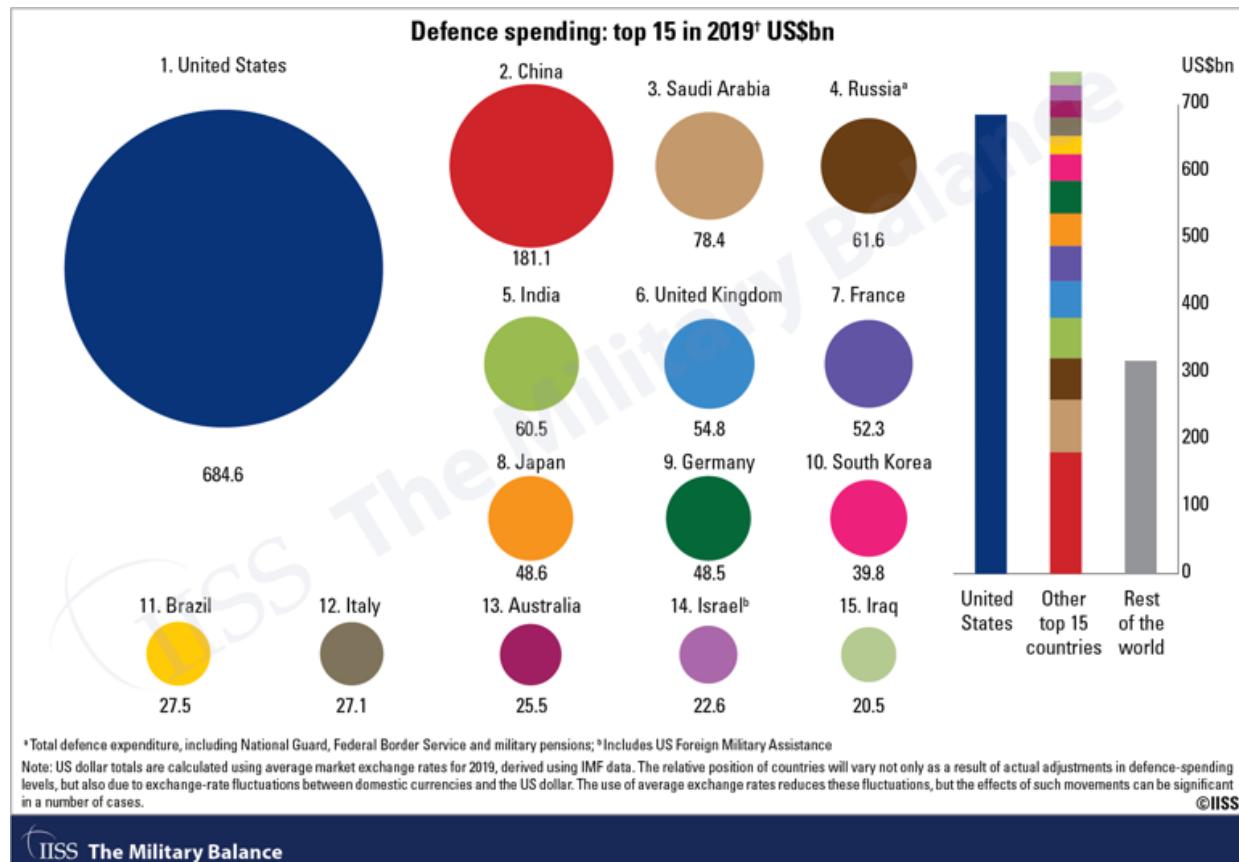
The People's Republic of China did not get a seat at the table in 1945 when the LWO was designed. It wasn't even allowed into the United Nations until 1971. From 1945 until 1971, "China" was represented by the Republic of China *i.e.* Taiwan. This illustrates how the LWO favors American geopolitical interests and is one of the many reasons why the People's Republic of China seeks to annex Taiwan.

China has prospered under the LWO. Rather than establishing broad international coalitions, China tends to pursue its interests bilaterally. With a few exceptions (like the dispute over the South China Sea) China is content to play according to the rules of the LWO.

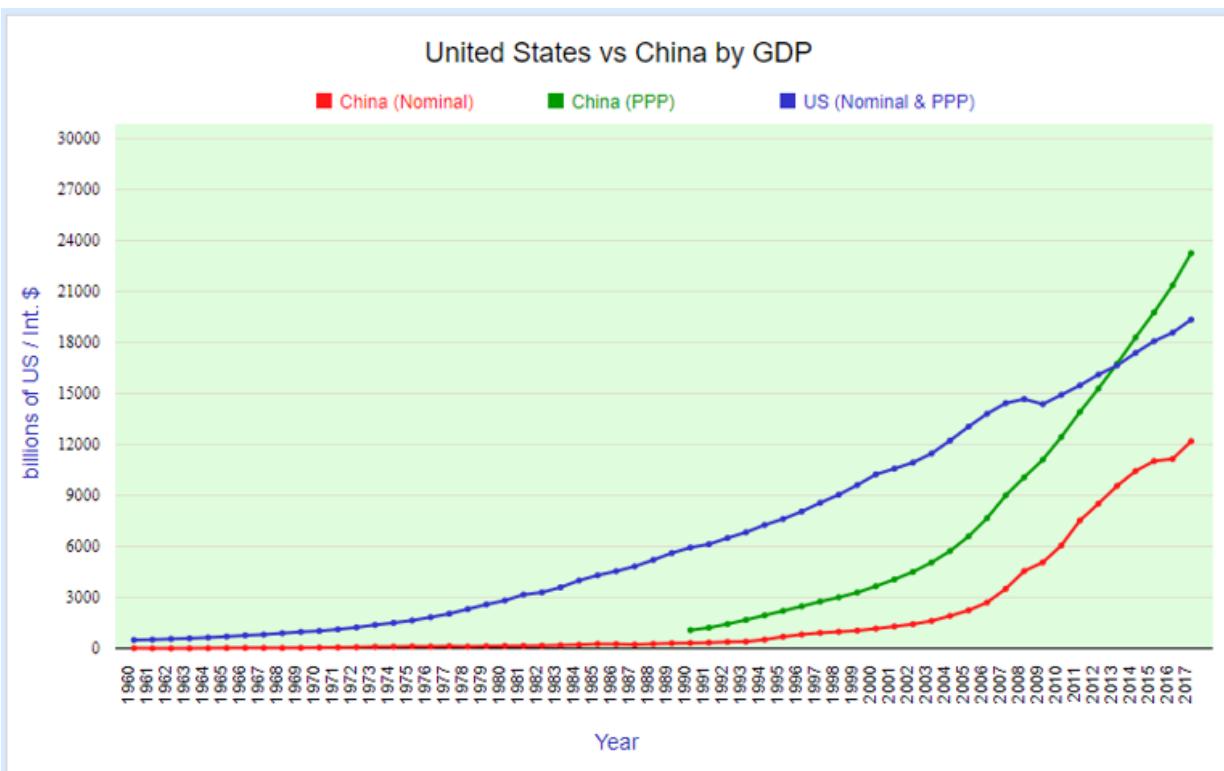
Besides annexing Taiwan, China's geopolitical interests mostly revolve around securing markets and raw materials to fuel its economic development. This is the motivation behind 一带一路 (the Belt and Road Initiative).

Relative Power

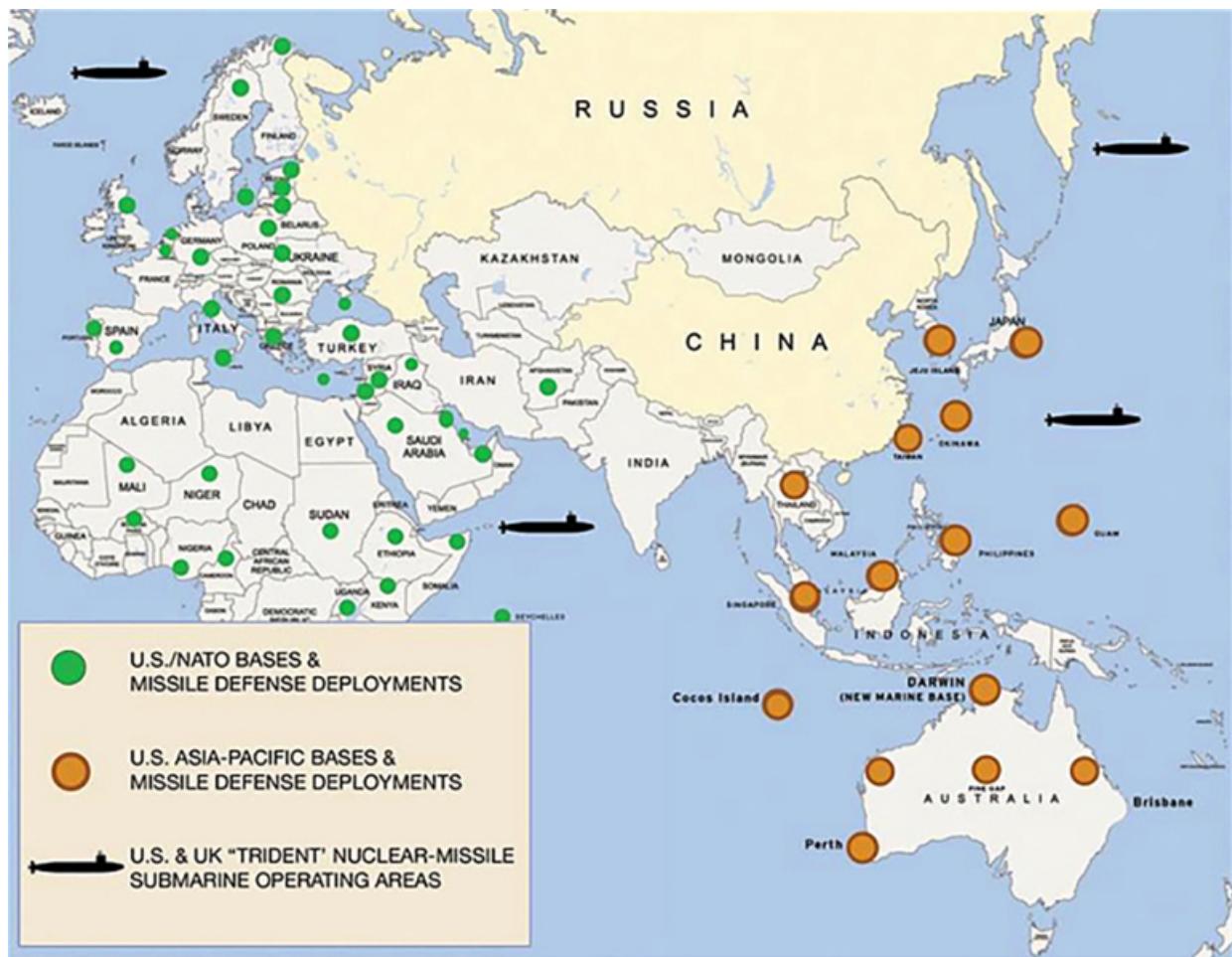
The United States dominates global military spending.



Military spending is a trailing indicator. What really matters is the size of each nation's respective economy. Metaculus [predicts](#) a 1.7 China-to-USA GDP ratio in 2050. The PPP difference will be even higher.



This underestimates the relative strength of the USA because the USA has many close allies: The European Union, Britain, Japan, South Korea, Taiwan and so on.



Source: Global Network Against Weapons & Nuclear Power in Space

China has North Korea.



The advantage to America is it has allies to call upon. The disadvantage to America is it has allies it must coordinate with and defend. America is spread thin. China focuses its attention on its smaller sphere of influence.

Conflict Points

Of all the potential points of conflict, the obvious ones are Taiwan and the South China Sea. For the purpose of this analysis, suppose Taiwan declares independence from China in 2050 which causes China to attack Taiwan which causes the United States to attack China.

The last time something like this happened was in 1949 when the Kuomintang fled to Taiwan. The United States refused to get involved until the Korean War in 1950. The United States sent its Seventh Fleet into the Taiwan Strait. American naval supremacy kept the People's Republic of China from advancing into Taiwan, thus establishing the status quo which remains until today.

The United States maintains naval dominance over China. This will not last. China's PPP surpassed the United States years ago. China's GDP will surpass the United States well before 2050. Most importantly, the crown jewels of the United States' fleet will be almost useless in a direct conflict against China.

Ships vs Guided Missiles

When the United States wants to project power to a faraway land it uses aircraft carriers. In the 1996 Taiwan Strait Crisis, the United States sent two aircraft carrier battle groups into the Taiwan Strait to demonstrate solidarity with Taiwan.

Aircraft carriers dominated the seas in the second half of the 20th century because planes used to have short ranges and missiles used to be dumb. Today, airplanes have long ranges and missiles are smart.

A Chinese DF-21D anti-ship ballistic missile [is believed to have a range in excess of 1,500 km](#). The newer CH-AS-X-13 has a range of 3,000 km. It can be launched from an airplane so its effective range is even farther. I estimate anti-ship missiles cost about \$2 million each. An aircraft carrier costs more than \$10 billion. Hiding an aircraft carrier battle group on the open sea isn't possible. The only way aircraft carriers could be remotely viable in a hot war between the United States and China would be if they had very reliable missile defense systems.

Israel's Iron Dome blocked 90% to 95% of incoming missiles in this year's Israel-Palestine crisis. That means it missed 5% to 10% of incoming missiles. A Hamas Qassam rocket is primitive. It is propelled by a mixture of sugar and potassium nitrate (ordinary fertilizer). It is not even accurate enough to use against military targets. A Qassam rocket is to a Chinese CH-AS-X-13 what a Mitsubishi A6M Zero is to an F-16.

I predict at least 5% of guided missiles can penetrate a surface fleet's missile defense system in 2050. If my numbers are right then either side can sink a \$10 billion aircraft carrier with no more than \$40 million in guided missiles.

An Arleigh Burke-class destroyer costs \$2 billion per ship. Surface ships are basically going to be obsolete in a direct war between superpowers. The battles of the sea will be fought with aircraft, missiles, drones and maybe submarines.

Taiwan

I expect China will not be able to establish air supremacy far beyond its borders and that the United States will not be able to establish air supremacy over mainland China. Without air supremacy, neither side can protect aircraft carriers and transport fleets. This makes amphibious invasions difficult. Except Taiwan *isn't* far beyond China's borders.

Taiwan's military can be quickly destroyed by a Chinese attack. Taiwan itself is too weak to repulse China on the beaches. If Taiwan wants to maintain its independence then it needs dug-in defenders. Taiwan's mountainous terrain is advantageous here. Taiwan already conscripts all qualified males. It should be training them in guerrilla warfare. This is not happening. Taiwan should also have weapons caches and underground bunkers strewn all over the island. It doesn't.

How many troops would it take to conquer Taiwan?

- A Chinese invasion of Taiwan might be similar to a United States invasion of the Japanese homeland in 1945. In 1945, the United States' Project Downfall projected an invasion force of 5 million was required to subdue a population of 70 million. Projecting from this, an invasion force of 2 million might be necessary to subdue Taiwan's population of 23 million.
- On the other hand, armies today are more efficient than they were in 1945. (People are more docile too.) Afghanistan has a population of 40 million but US troop levels in Afghanistan [reached a height of only 100 thousand](#). It might take China only 50 thousand troops to subdue Taiwan.

I'm not sure which number to use. 2021 Taiwan is a lot like 2021 Japan but 2021 Japan is very unlike 1945 Japan. Taiwan is also unlike Afghanistan. It might take even fewer than 50 thousand troops to subdue Taiwan.

If 50 thousand troops is all it takes then China could land them quickly before the USA gets its act together. But if 2 million is what it takes to conquer Taiwan then China would need a bigger army than it has right now. Also, you can't land an invasion force of 2 million troops overnight. China would have to win the initial missile exchange and then maintain air superiority while it landed troops in Taiwan.

Things get hard to predict from here. What happens depends on the decisions of individual world leaders and the determination of various peoples.

Cyberspace

In [*Darknet Diaries Episode 21: Black Duck Eggs*](#), Ira Winkler tells the story of how his team broke into a major datacenter containing billions of dollars worth of technology.

Well, the first time you steal a billion dollars it's a bit of a rush. After you've done this so many times it's almost expected. Frankly, it was really unclimactic to actually take over control of all their computers in the RND center.

I know, right? Who cares if you can just walk into a datacenter and steal a billion dollars worth of advanced technology? It doesn't mean anything when you have a get out of jail free card because you are doing an officially-sanctioned penetration test. Except that after the penetration test Ira Winkler's team discovered a physical "Chinese intelligence operation in the middle of this small town, directly across the street from the research and development center of a Global 5 company".

For every hack we hear about there are many hacks we don't hear about. When you break into an adversary's computer network the first thing you want to do is establish persistence. Most of those hacks we don't hear about probably establish persistence. I would be surprised if China and the United States hadn't established persistence in most of each others' critical systems.

In the event of a hot war between the United States and China, both sides will burn most of their zero-days immediately to cause as much disruption to the enemy as possible. It takes a lot of work to clean a hacker out of one of your systems. The cyber onslaught will probably overwhelm both sides' ability to reset their software. They will have to focus on the most critical systems of all: communications.

I expect all but the most secure systems (think "US president's personal phone") will be entirely compromised. However, there are many ways to communicate. Both sides can improvise. Since secondary channels abound, it might be better just to spy on enemy communications instead of breaking them.

It is theoretically possible to take control of enemy weapon systems too but I don't think this will have a major impact. Weapon systems will continue to have human beings in the loop. Human beings can't be hacked the way computers can.

I think compromised weapons systems will just be taken out of commission rather than commandeered. Some of them might be destroyed, but I think most will just be rendered temporarily inoperable.

Space War

Satellites serve four purposes: reconnaissance, communication, navigation (GPS) and destroying other satellites. GPS going down would inconvenience both sides but it wouldn't be decisive. There are other ways of navigating. The same goes for satellite communication.

The most important use of satellites is reconnaissance. Aerial drones are nice but they can't see as much at once as a camera in outer space. If both sides are restrained in their use of nuclear weapons then they might also be restrained in space warfare. That seems overly-optimistic to me. The primary objective of both sides will be to preserve their spy satellite capability while destroying the enemy's.

The price of space travel is going down. We can expect a large increase in the number of satellites between now and 2050, including spy satellites and anti-satellite satellites. It is plausible that satellite warfare could trigger Kessler syndrome in low earth orbit where collisions cascade into a giant mess of debris.

Protracted Total War is Unlikely

The initial exchange of missiles and zero-days will be fast. Critical tactical decisions may occur in the first few hours. The decisive fighting could be over in a matter of days. The limiting factor isn't technical. It's the speed at which leaders can make decisions.

After a few weeks, both armies will be running out of missiles^[1]. Both civilian populations will have suffered massive damage from cyberattacks on their civilian infrastructure. If satellite warfare triggers Kessler syndrome then much of the world's communication infrastructure could be irreversibly damaged. The global economy would be a mess. Economic chaos would be bad for US interests and even worse for Chinese interests.

If we do see a protracted war then it matters a lot what countries get involved. A conflict where India or Russia joins the fray is very different from a conflict where they don't. Whatever happens, the tech level will probably go down. Advanced weapon systems like stealth drones take a long time to build. Destroy a few major semiconductor fabricators and the whole world runs low.

But it's hard for me to imagine a protracted war on the scale of World War II. In the past, getting your cities bombed was a small price to pay in order to expand your territory. China may seize some already disputed territory like Taiwan or Kashmir. But it doesn't want to administer a large empire. China's primary objective is to maintain domestic stability. Annexing Afghanistan or Vietnam or Pakistan would make China harder to govern, not easier.

China's secondary objective is to secure access to resources and markets. Destroying the United States might help tear down the LWO, but if China dragged itself down alongside the United States then that would just open up a power vacuum for countries like Brazil, India and South Africa.

The United States isn't an expansionary power either. America likes the LWO because the LWO supports American interests. I don't see America sacrificing its own hegemony

to preserve the LWO.

The Seven Years War happened because Britain and France were expansionary powers. World War One happened because the European powers were expansionary. World War Two happened because Germany, Italy and Japan were expansionary powers. The United States and China aren't expansionary powers.

Instead of a protracted total war, we'll probably see a ceasefire or deescalation. Besides control over Taiwan, the most important thing to come out of a conflict like this is a (re)establishment of the world order. A war would clarify who is and isn't a superpower (anymore).

1. If things go longer then it would be because the United States' military force is spread around the world. Its reserve forces could take weeks just to get to China.

[←](#)

How counting neutrons explains nuclear waste

This is a linkpost for <https://rootsofprogress.org/nuclear-physics>

You probably recall from high school chemistry that atoms are made up of a nucleus containing protons and neutrons, surrounded by electrons. But how many of each?

If you remember a little bit more from high school chemistry, you'll recall that the number of protons determines which element it is: an atom with six protons is an atom of carbon; seven makes it nitrogen; eight, oxygen. The number of electrons generally matches the number of protons, to make the atom electrically neutral. But how many neutrons are in the nucleus? Does it even matter?

It turns out that it matters a *lot*.

Atoms of the same element with different numbers of neutrons are called *isotopes*. They are distinguished by their "mass number", which is the total number of protons and neutrons (together known as *nucleons*). A typical carbon atom has six protons and six neutrons, for a mass number of twelve; it is referred to as carbon-12 or C-12 (or sometimes ^{12}C).

It would be oh, so simple if every element had an equal number of protons and neutrons. But reality is far more complicated than that.

As you walk the periodic table, you will find that heavier elements have a higher ratio of neutrons to protons. The lightest element, hydrogen, usually has zero neutrons—its nucleus is just a lone proton. The next element, helium, comes in balance: typically it has two protons and two neutrons. The other light elements usually have the same balance, such as carbon-12, or perhaps one more neutron than protons, such as sodium-23, with 11 protons and 12 neutrons. This lasts up to calcium (nicely balanced, with its most abundant isotope, calcium-40, having 20 of each).

But there it ends. After calcium you will find that every element has more neutrons than protons. Iron, element 26, usually has 30 neutrons. Iodine, element 53, has 74. Gold, element 79, has 118. By the time we get to uranium-238, we're looking at 92 protons and 146 neutrons, for a neutron:proton ratio of over 3:2. What's going on?

Here's another piece of the puzzle: not all isotopes are stable.

I've been saying things like "carbon typically has six neutrons" or "iodine has 74". What that really means is that carbon only has a few stable isotopes, and the most common one has six neutrons (carbon-12). Iodine only has one stable isotope, and it's the one with 74 neutrons (iodine-127).

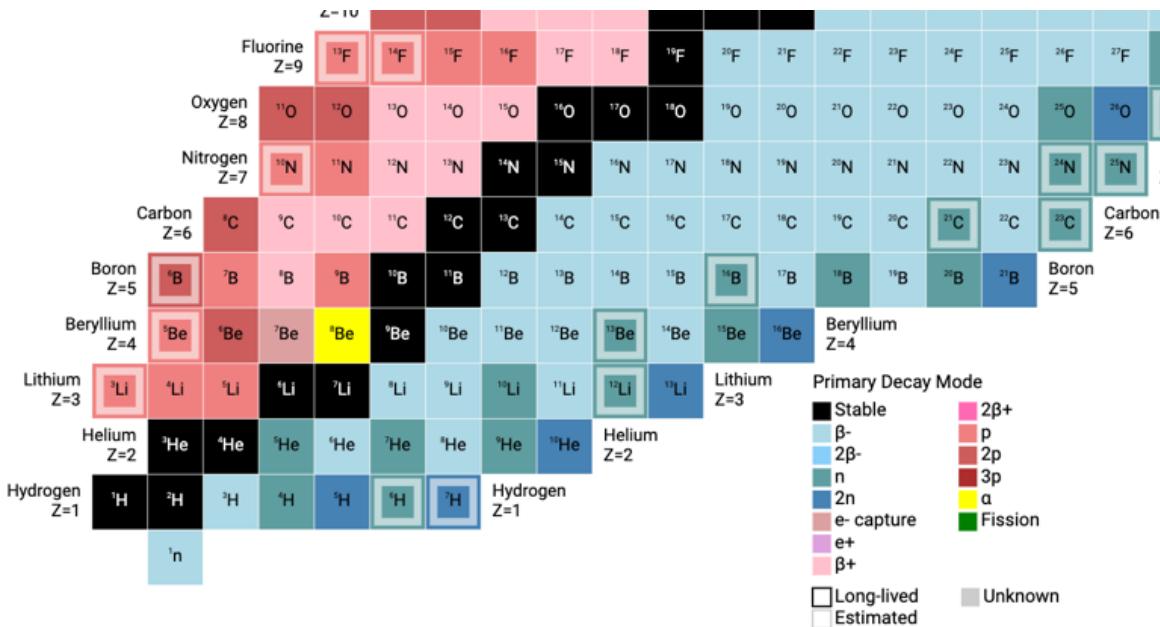
What does it mean for an isotope to be *unstable*? It means that atoms of that isotope will randomly kick a subatomic particle out of the nucleus. This process is called *decay*.

Most isotopes where the neutron:proton ratio is too high decay by ejecting an electron from the nucleus, turning one of the neutrons into a proton. For instance, carbon-14 (six protons, eight neutrons) ejects an electron, turning into an atom with seven protons and seven neutrons—that is, nitrogen-14. This is called beta decay, and in this context, the electron is called a beta particle. (A [electron antineutrino](#) is emitted too, but neutrinos have almost no mass and almost never interact with anything, so they mostly just go flying off with no effect. We're going to ignore neutrinos today.)

If the neutron:proton ratio is too low, the opposite happens: an electron is captured (or a positron is ejected), turning a proton into a neutron. Carbon-11 (6 protons, 5 neutrons) will do this, turning into boron-11 (5 protons, 6 neutrons). This is another form of beta decay (designated β^+ , to distinguish it from the former type, which is β^-).

This is... somewhat amazing. For centuries, alchemists sought the secret of transmuting one element into another. Twentieth-century physics finally discovered how it can happen: radioactive decay!

We can map these transitions out. Make a grid of squares with the number of protons on the vertical axis, and the number of neutrons on the horizontal axis. Each row thus represents an element, and each cell in the grid represents an isotope of that element. This is called a *Segrè chart* (after the physicist [Emilio Segrè](#)):



[The Colourful Nuclide Chart](#)

(Physicists denote the number of neutrons with N, and the number of protons with Z—don't ask me why.)

The term “isotope”, though, is rather proton-centric. It takes the number of protons as the essence of an atom, and its number of neutrons as a mere flavor. From the perspective of chemistry, this is reasonable. But when we take the perspective of nuclear physics, we start to see neutrons and protons as equal—especially when they start turning into one another. From this perspective, rather than isotopes, different configurations of protons and neutrons are called *nuclides*. The Segrè chart is also called the *table of nuclides*.

In this rendering of the chart, black squares are stable; light blue ones undergo beta decay; pink ones undergo electron capture or β^+ decay. In extreme cases, a nucleus will directly spit out a proton or a neutron (or even more than one). Those are the darker blue-green and pink cells of the table.

As we advance up the chart towards the heavier elements, however, a different mode of decay becomes common. Often these isotopes will kick out an *alpha particle*—a cluster of two protons and two neutrons, equivalent to the nucleus of a helium atom. Polonium-210, for instance (a favorite of [Russian assassins](#)) will undergo alpha decay to become lead-206. These are the yellow squares:

		Thorium Z=90		²²⁸ Th	²²⁹ Th	²³⁰ Th	²³¹ Th	²³² Th	²³³ Th	²³⁴ Th	²³⁵ Th	²³⁶ Th	²³⁷ Th	²³⁸ Th	²³⁹ Th	²⁴⁰ Th	²⁴¹ Th	²⁴² Th	²⁴³ Th	²⁴⁴ Th	²⁴⁵ Th	²⁴⁶ Th	²⁴⁷ Th	²⁴⁸ Th	²⁴⁹ Th	²⁵⁰ Th	²⁵¹ Th	²⁵² Th	²⁵³ Th	²⁵⁴ Th	²⁵⁵ Th	²⁵⁶ Th	²⁵⁷ Th	²⁵⁸ Th	²⁵⁹ Th	²⁶⁰ Th													
		Actinium Z=89		²²⁵ Ac	²²⁶ Ac	²²⁷ Ac	²²⁸ Ac	²²⁹ Ac	²³⁰ Ac	²³¹ Ac	²³² Ac	²³³ Ac	²³⁴ Ac	²³⁵ Ac	²³⁶ Ac	²³⁷ Ac	²³⁸ Ac	²³⁹ Ac	²⁴⁰ Ac	²⁴¹ Ac	²⁴² Ac	²⁴³ Ac	²⁴⁴ Ac	²⁴⁵ Ac	²⁴⁶ Ac	²⁴⁷ Ac	²⁴⁸ Ac	²⁴⁹ Ac	²⁵⁰ Ac	²⁵¹ Ac	²⁵² Ac	²⁵³ Ac	²⁵⁴ Ac	²⁵⁵ Ac	²⁵⁶ Ac	²⁵⁷ Ac	²⁵⁸ Ac	²⁵⁹ Ac	²⁶⁰ Ac										
²²⁰ Ra	²²¹ Ra	²²⁴ Ra	²²⁵ Ra	²²⁶ Ra	²²⁷ Ra	²²⁸ Ra	²²⁹ Ra	²³⁰ Ra	²³¹ Ra	²³² Ra	²³³ Ra	²³⁴ Ra	²³⁵ Ra	²³⁶ Ra	²³⁷ Ra	²³⁸ Ra	²³⁹ Ra	²⁴⁰ Ra	²⁴¹ Ra	²⁴² Ra	²⁴³ Ra	²⁴⁴ Ra	²⁴⁵ Ra	²⁴⁶ Ra	²⁴⁷ Ra	²⁴⁸ Ra	²⁴⁹ Ra	²⁵⁰ Ra	²⁵¹ Ra	²⁵² Ra	²⁵³ Ra	²⁵⁴ Ra	²⁵⁵ Ra	²⁵⁶ Ra	²⁵⁷ Ra	²⁵⁸ Ra	²⁵⁹ Ra	²⁶⁰ Ra											
²²² Fr	²²³ Fr	²²⁴ Fr	²²⁵ Fr	²²⁶ Fr	²²⁷ Fr	²²⁸ Fr	²²⁹ Fr	²³⁰ Fr	²³¹ Fr	²³² Fr	²³³ Fr	²³⁴ Fr	²³⁵ Fr	²³⁶ Fr	²³⁷ Fr	²³⁸ Fr	²³⁹ Fr	²⁴⁰ Fr	²⁴¹ Fr	²⁴² Fr	²⁴³ Fr	²⁴⁴ Fr	²⁴⁵ Fr	²⁴⁶ Fr	²⁴⁷ Fr	²⁴⁸ Fr	²⁴⁹ Fr	²⁵⁰ Fr	²⁵¹ Fr	²⁵² Fr	²⁵³ Fr	²⁵⁴ Fr	²⁵⁵ Fr	²⁵⁶ Fr	²⁵⁷ Fr	²⁵⁸ Fr	²⁵⁹ Fr	²⁶⁰ Fr											
²²⁰ Rn	²²¹ Rn	²²² Rn	²²³ Rn	²²⁴ Rn	²²⁵ Rn	²²⁶ Rn	²²⁷ Rn	²²⁸ Rn	²²⁹ Rn	²³⁰ Rn	²³¹ Rn	²³² Rn	²³³ Rn	²³⁴ Rn	²³⁵ Rn	²³⁶ Rn	²³⁷ Rn	²³⁸ Rn	²³⁹ Rn	²⁴⁰ Rn	²⁴¹ Rn	²⁴² Rn	²⁴³ Rn	²⁴⁴ Rn	²⁴⁵ Rn	²⁴⁶ Rn	²⁴⁷ Rn	²⁴⁸ Rn	²⁴⁹ Rn	²⁵⁰ Rn	²⁵¹ Rn	²⁵² Rn	²⁵³ Rn	²⁵⁴ Rn	²⁵⁵ Rn	²⁵⁶ Rn	²⁵⁷ Rn	²⁵⁸ Rn	²⁵⁹ Rn	²⁶⁰ Rn									
¹⁹⁹ At	²⁰⁰ At	²⁰¹ At	²⁰² At	²⁰³ At	²⁰⁴ At	²⁰⁵ At	²⁰⁶ At	²⁰⁷ At	²⁰⁸ At	²⁰⁹ At	²¹⁰ At	²¹¹ At	²¹² At	²¹³ At	²¹⁴ At	²¹⁵ At	²¹⁶ At	²¹⁷ At	²¹⁸ At	²¹⁹ At	²²⁰ At	²²¹ At	²²² At	²²³ At	²²⁴ At	²²⁵ At	²²⁶ At	²²⁷ At	²²⁸ At	²²⁹ At	²³⁰ At	²³¹ At	²³² At	²³³ At	²³⁴ At	²³⁵ At	²³⁶ At	²³⁷ At	²³⁸ At	²³⁹ At	²⁴⁰ At								
¹⁹⁸ Po	¹⁹⁹ Po	²⁰⁰ Po	²⁰¹ Po	²⁰² Po	²⁰³ Po	²⁰⁴ Po	²⁰⁵ Po	²⁰⁶ Po	²⁰⁷ Po	²⁰⁸ Po	²⁰⁹ Po	²¹⁰ Po	²¹¹ Po	²¹² Po	²¹³ Po	²¹⁴ Po	²¹⁵ Po	²¹⁶ Po	²¹⁷ Po	²¹⁸ Po	²¹⁹ Po	²²⁰ Po	²²¹ Po	²²² Po	²²³ Po	²²⁴ Po	²²⁵ Po	²²⁶ Po	²²⁷ Po	²²⁸ Po	²²⁹ Po	²³⁰ Po	²³¹ Po	²³² Po	²³³ Po	²³⁴ Po	²³⁵ Po	²³⁶ Po	²³⁷ Po	²³⁸ Po	²³⁹ Po	²⁴⁰ Po							
¹⁹⁷ Bi	¹⁹⁸ Bi	¹⁹⁹ Bi	²⁰⁰ Bi	²⁰¹ Bi	²⁰² Bi	²⁰³ Bi	²⁰⁴ Bi	²⁰⁵ Bi	²⁰⁶ Bi	²⁰⁷ Bi	²⁰⁸ Bi	²⁰⁹ Bi	²¹⁰ Bi	²¹¹ Bi	²¹² Bi	²¹³ Bi	²¹⁴ Bi	²¹⁵ Bi	²¹⁶ Bi	²¹⁷ Bi	²¹⁸ Bi	²¹⁹ Bi	²²⁰ Bi	²²¹ Bi	²²² Bi	²²³ Bi	²²⁴ Bi	²²⁵ Bi	²²⁶ Bi	²²⁷ Bi	²²⁸ Bi	²²⁹ Bi	²³⁰ Bi	²³¹ Bi	²³² Bi	²³³ Bi	²³⁴ Bi	²³⁵ Bi	²³⁶ Bi	²³⁷ Bi	²³⁸ Bi	²³⁹ Bi	²⁴⁰ Bi						
¹⁹⁶ Pb	¹⁹⁷ Pb	¹⁹⁸ Pb	¹⁹⁹ Pb	²⁰⁰ Pb	²⁰¹ Pb	²⁰² Pb	²⁰³ Pb	²⁰⁴ Pb	²⁰⁵ Pb	²⁰⁶ Pb	²⁰⁷ Pb	²⁰⁸ Pb	²⁰⁹ Pb	²¹⁰ Pb	²¹¹ Pb	²¹² Pb	²¹³ Pb	²¹⁴ Pb	²¹⁵ Pb	²¹⁶ Pb	²¹⁷ Pb	²¹⁸ Pb	²¹⁹ Pb	²²⁰ Pb	²²¹ Pb	²²² Pb	²²³ Pb	²²⁴ Pb	²²⁵ Pb	²²⁶ Pb	²²⁷ Pb	²²⁸ Pb	²²⁹ Pb	²³⁰ Pb	²³¹ Pb	²³² Pb	²³³ Pb	²³⁴ Pb	²³⁵ Pb	²³⁶ Pb	²³⁷ Pb	²³⁸ Pb	²³⁹ Pb	²⁴⁰ Pb					
¹⁹⁵ Tl	¹⁹⁶ Tl	¹⁹⁷ Tl	¹⁹⁸ Tl	¹⁹⁹ Tl	²⁰⁰ Tl	²⁰¹ Tl	²⁰² Tl	²⁰³ Tl	²⁰⁴ Tl	²⁰⁵ Tl	²⁰⁶ Tl	²⁰⁷ Tl	²⁰⁸ Tl	²⁰⁹ Tl	²¹⁰ Tl	²¹¹ Tl	²¹² Tl	²¹³ Tl	²¹⁴ Tl	²¹⁵ Tl	²¹⁶ Tl	²¹⁷ Tl	²¹⁸ Tl	²¹⁹ Tl	²²⁰ Tl	²²¹ Tl	²²² Tl	²²³ Tl	²²⁴ Tl	²²⁵ Tl	²²⁶ Tl	²²⁷ Tl	²²⁸ Tl	²²⁹ Tl	²³⁰ Tl	²³¹ Tl	²³² Tl	²³³ Tl	²³⁴ Tl	²³⁵ Tl	²³⁶ Tl	²³⁷ Tl> <td>²³⁸Tl</td> <td>²³⁹Tl</td> <td>²⁴⁰Tl</td>	²³⁸ Tl	²³⁹ Tl	²⁴⁰ Tl				
¹⁹⁴ Hg	¹⁹⁵ Hg	¹⁹⁶ Hg	¹⁹⁷ Hg	¹⁹⁸ Hg	¹⁹⁹ Hg	²⁰⁰ Hg	²⁰¹ Hg	²⁰² Hg	²⁰³ Hg	²⁰⁴ Hg	²⁰⁵ Hg	²⁰⁶ Hg	²⁰⁷ Hg	²⁰⁸ Hg	²⁰⁹ Hg	²¹⁰ Hg	²¹¹ Hg	²¹² Hg	²¹³ Hg	²¹⁴ Hg	²¹⁵ Hg	²¹⁶ Hg	²¹⁷ Hg	²¹⁸ Hg	²¹⁹ Hg	²²⁰ Hg	²²¹ Hg	²²² Hg	²²³ Hg	²²⁴ Hg	²²⁵ Hg	²²⁶ Hg	²²⁷ Hg	²²⁸ Hg	²²⁹ Hg	²³⁰ Hg	²³¹ Hg	²³² Hg	²³³ Hg	²³⁴ Hg	²³⁵ Hg	²³⁶ Hg	²³⁷ Hg> <td>²³⁸Hg</td> <td>²³⁹Hg</td> <td>²⁴⁰Hg</td>	²³⁸ Hg	²³⁹ Hg	²⁴⁰ Hg			
¹⁹³ Au	¹⁹⁴ Au	¹⁹⁵ Au	¹⁹⁶ Au	¹⁹⁷ Au	¹⁹⁸ Au	¹⁹⁹ Au	²⁰⁰ Au	²⁰¹ Au	²⁰² Au	²⁰³ Au	²⁰⁴ Au	²⁰⁵ Au	²⁰⁶ Au	²⁰⁷ Au	²⁰⁸ Au	²⁰⁹ Au	²¹⁰ Au	²¹¹ Au	²¹² Au	²¹³ Au	²¹⁴ Au	²¹⁵ Au	²¹⁶ Au	²¹⁷ Au	²¹⁸ Au	²¹⁹ Au	²²⁰ Au	²²¹ Au	²²² Au	²²³ Au	²²⁴ Au	²²⁵ Au	²²⁶ Au	²²⁷ Au	²²⁸ Au	²²⁹ Au	²³⁰ Au	²³¹ Au	²³² Au	²³³ Au	²³⁴ Au	²³⁵ Au	²³⁶ Au	²³⁷ Au	²³⁸ Au	²³⁹ Au	²⁴⁰ Au		
¹⁹² Pt	¹⁹³ Pt	¹⁹⁴ Pt	¹⁹⁵ Pt	¹⁹⁶ Pt	¹⁹⁷ Pt	¹⁹⁸ Pt	¹⁹⁹ Pt	²⁰⁰ Pt	²⁰¹ Pt	²⁰² Pt	²⁰³ Pt	²⁰⁴ Pt	²⁰⁵ Pt	²⁰⁶ Pt	²⁰⁷ Pt <td>²⁰⁸Pt</td> <td>²⁰⁹Pt</td> <td>²¹⁰Pt</td> <td>²¹¹Pt</td> <td>²¹²Pt</td> <td>²¹³Pt</td> <td>²¹⁴Pt</td> <td>²¹⁵Pt</td> <td>²¹⁶Pt</td> <td>²¹⁷Pt</td> <td>²¹⁸Pt</td> <td>²¹⁹Pt</td> <td>²²⁰Pt</td> <td>²²¹Pt</td> <td>²²²Pt</td> <td>²²³Pt</td> <td>²²⁴Pt</td> <td>²²⁵Pt</td> <td>²²⁶Pt</td> <td>²²⁷Pt</td> <td>²²⁸Pt</td> <td>²²⁹Pt</td> <td>²³⁰Pt</td> <td>²³¹Pt</td> <td>²³²Pt</td> <td>²³³Pt</td> <td>²³⁴Pt</td> <td>²³⁵Pt</td> <td>²³⁶Pt</td> <td>²³⁷Pt</td> <td>²³⁸Pt</td> <td>²³⁹Pt</td> <td>²⁴⁰Pt</td>	²⁰⁸ Pt	²⁰⁹ Pt	²¹⁰ Pt	²¹¹ Pt	²¹² Pt	²¹³ Pt	²¹⁴ Pt	²¹⁵ Pt	²¹⁶ Pt	²¹⁷ Pt	²¹⁸ Pt	²¹⁹ Pt	²²⁰ Pt	²²¹ Pt	²²² Pt	²²³ Pt	²²⁴ Pt	²²⁵ Pt	²²⁶ Pt	²²⁷ Pt	²²⁸ Pt	²²⁹ Pt	²³⁰ Pt	²³¹ Pt	²³² Pt	²³³ Pt	²³⁴ Pt	²³⁵ Pt	²³⁶ Pt	²³⁷ Pt	²³⁸ Pt	²³⁹ Pt	²⁴⁰ Pt	
¹⁹¹ Ir	¹⁹² Ir	¹⁹³ Ir	¹⁹⁴ Ir	¹⁹⁵ Ir	¹⁹⁶ Ir	¹⁹⁷ Ir	¹⁹⁸ Ir	¹⁹⁹ Ir	²⁰⁰ Ir	²⁰¹ Ir	²⁰² Ir	²⁰³ Ir	²⁰⁴ Ir	²⁰⁵ Ir	²⁰⁶ Ir	²⁰⁷ Ir	²⁰⁸ Ir	²⁰⁹ Ir	²¹⁰ Ir	²¹¹ Ir	²¹² Ir	²¹³ Ir	²¹⁴ Ir	²¹⁵ Ir	²¹⁶ Ir	²¹⁷ Ir	²¹⁸ Ir	²¹⁹ Ir	²²⁰ Ir	²²¹ Ir	²²² Ir	²²³ Ir	²²⁴ Ir	²²⁵ Ir	²²⁶ Ir	²²⁷ Ir	²²⁸ Ir	²²⁹ Ir	²³⁰ Ir	²³¹ Ir	²³² Ir	²³³ Ir	²³⁴ Ir	²³⁵ Ir	²³⁶ Ir	²³⁷ Ir	²³⁸ Ir	²³⁹ Ir	²⁴⁰ Ir

The Colourful Nuclide Chart

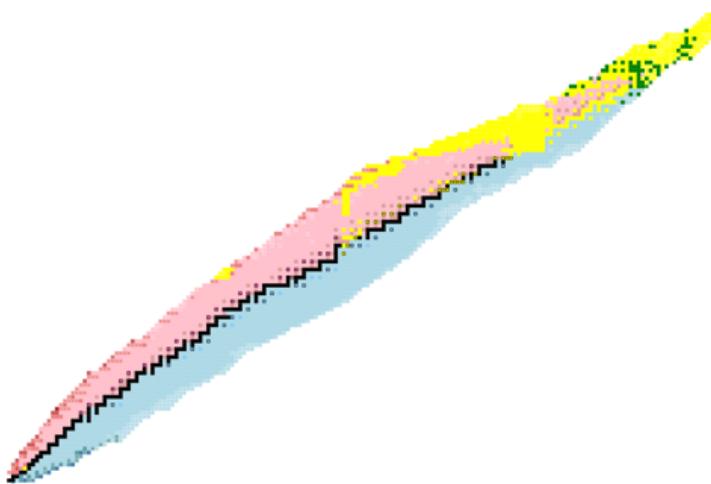
As you get to the very heaviest elements—ones with names like fermium, nobelium, or rutherfordium—something even more dramatic happens. These elements are so big and unstable that they don't just eject a little speck like an electron or a helium nucleus. They split almost in half, in a process called *fission*. Instead of a large atom turning into a slightly different large atom plus a tiny particle, it turns into two medium-sized atoms (plus a few extra neutrons). Exactly which elements are produced by any given fission event is a bit random, but, for instance, the fission of uranium-235 [commonly results](#) in caesium-133, iodine-135, zirconium-93, or molybdenum-99. These are called *fission products*. The dark green squares are the isotopes that decay by fission:

		Dubnium Z=105		²²⁹Db	²³⁰Db	²³¹Db	²³²Db	²³³Db	²³⁴Db	²³⁵Db	²³⁶Db	²³⁷Db	²³⁸Db	²³⁹Db	²⁴⁰Db	²⁴¹Db	²⁴²Db	²⁴³Db	²⁴⁴Db	²⁴⁵Db	²⁴⁶Db	²⁴⁷Db	²⁴⁸Db	²⁴⁹Db	²⁵⁰Db	²⁵¹Db	²⁵²Db	²⁵³Db	²⁵⁴Db	²⁵⁵Db	²⁵⁶Db	²⁵⁷Db	²⁵⁸Db	²⁵⁹Db	²⁶⁰Db	²⁶¹Db	²⁶²Db	²⁶³Db	²⁶⁴Db	²⁶⁵Db	²⁶⁶Db	²⁶⁷Db	²⁶⁸Db	²⁶⁹Db	²⁷⁰Db	²⁷¹Db	²⁷²Db	²⁷³Db	²⁷⁴Db	²⁷⁵Db	²⁷⁶Db	²⁷⁷Db	²⁷⁸Db	²⁷⁹Db	²⁸⁰Db	²⁸¹Db	²⁸²Db	²⁸³Db	²⁸⁴Db	²⁸⁵Db	²⁸⁶Db	²⁸⁷Db	²⁸⁸Db	²⁸⁹Db	²⁹⁰Db	²⁹¹Db	²⁹²Db	²⁹³Db	²⁹⁴Db	²⁹⁵Db	²⁹⁶Db	²⁹⁷Db	²⁹⁸Db	²⁹⁹Db	³⁰⁰Db	³⁰¹Db	³⁰²Db	³⁰³Db	³⁰⁴Db	³⁰⁵Db	³⁰⁶Db	³⁰⁷Db	³⁰⁸Db	³⁰⁹Db	³¹⁰Db	³¹¹Db	³¹²Db	³¹³Db	³¹⁴Db	³¹⁵Db	<

Alpha particles are the most damaging, per joule of energy actually absorbed by the body. But, owing to their relatively large size, they're also the easiest to stop: they don't get past your clothing, or the outer layer of skin. For this reason, you can safely hold a chunk of plutonium, which is mostly an alpha emitter, in your hand or pocket.

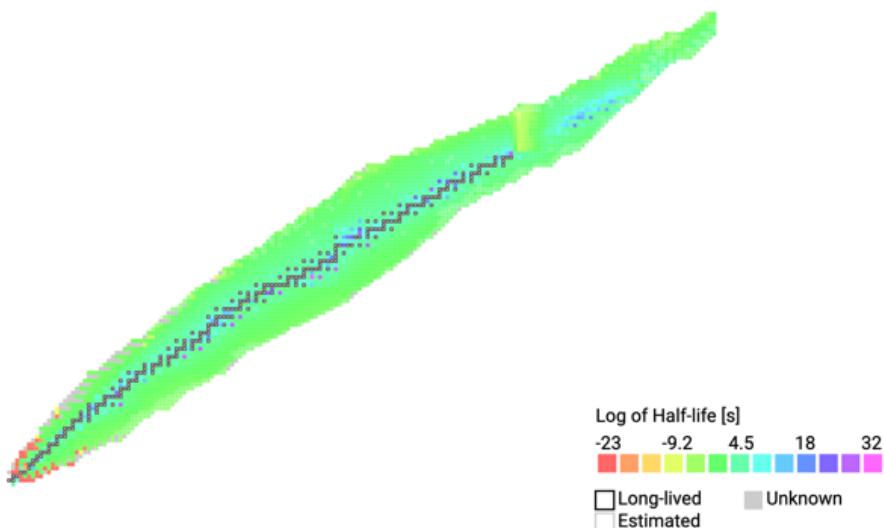
The type of radiation you *really* want to avoid is in fact another type altogether, called gamma radiation, which consists of high-energy photons. When a nuclide undergoes beta decay, for instance, it is often left in an "excited" state, and it "relaxes" by shedding energy in the form of gamma rays. (This will become important later.)

Here's the table zoomed out to show all of the nuclides that have been observed:



[The Colourful Nuclide Chart](#)

Here's the table colored by half-life instead of decay mode. Again, the black cells are stable; the lighter colors represent shorter half-lives:



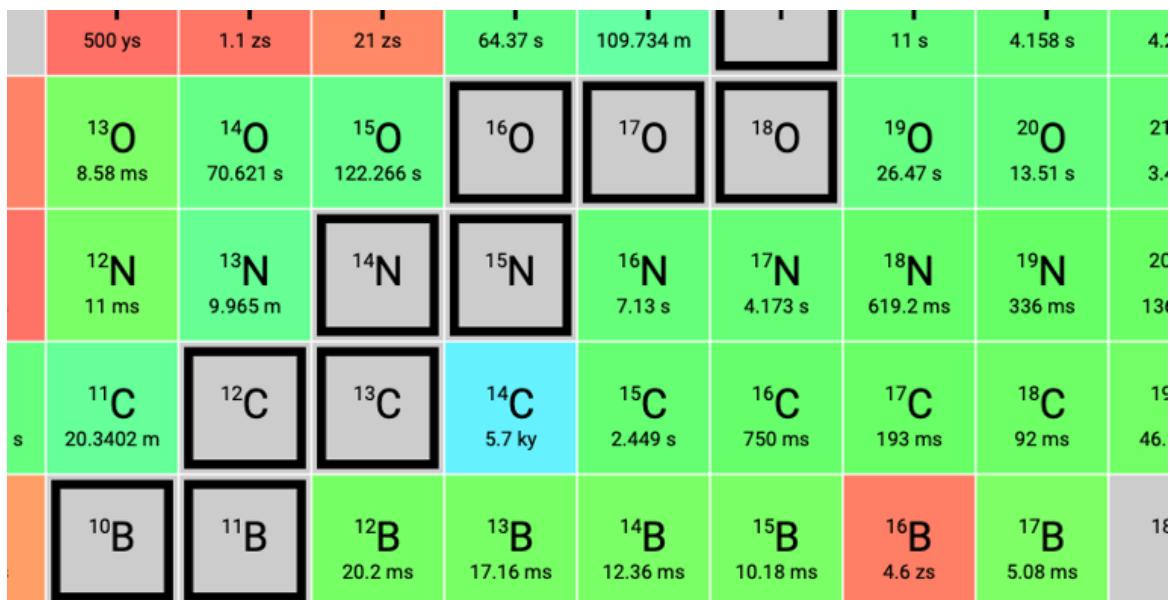
[The Colourful Nuclide Chart](#)

The stable isotopes form a backbone known as the “line of beta stability”. The further an isotope is from that line, the more unstable it is, so the region around the line is the “valley of stability”. Configurations that are too far away from the line can’t even form a bound nucleus; the edge of the valley is called the “drip line”.

You can see that the line of beta stability bends to the right: that curve represents the neutron/proton ratio increasing with atomic number. But the pattern is far from simple.

Imagine that we could hold a carbon-12 atom in the palm of one hand. Nearby we have a jar of protons, and a jar of neutrons. Imagine that we could pick up individual nucleons and stick them onto the nucleus of that atom. What order would we do so, in order to keep the nucleus stable at all stages?

We’re going to walk the beta-stability line:



The Colourful Nuclide Chart

Starting from C-12, we can see that our first step is to add one neutron, to get to C-13. Our next step is to add a proton, to get to N-14. Then another neutron (N-15), another proton (O-16), etc. What we’re doing here is increasing the mass number by one each time, and looking for an adjacent stable nuclide at that mass number. (Lines of equal mass number are downward-sloping diagonal paths on the chart.)

So far we’re following a zig-zagging path, keeping neutrons and protons roughly equal, with a preference for neutrons. But when we get to O-17, something odd happens:

			1.5 zs	1000 zs	447.9 ms	22.455 s	2.0019 y		14.956 h
¹⁵ Ne 770 ys	¹⁶ Ne 5.7 zs	¹⁷ Ne 109.2 ms	¹⁸ Ne 1.6642 s	¹⁹ Ne 17.2569 s	²⁰ Ne	²¹ Ne	²² Ne		²³ Ne 37.15 s
¹⁴ F 500 ys	¹⁵ F 1.1 zs	¹⁶ F 21 zs	¹⁷ F 64.37 s	¹⁸ F 109.734 m	¹⁹ F	²⁰ F 11 s	²¹ F 4.158 s	²² F 4.23 s	
¹³ O 8.58 ms	¹⁴ O 70.621 s	¹⁵ O 122.266 s	¹⁶ O	¹⁷ O	¹⁸ O	¹⁹ O 26.47 s	²⁰ O 13.51 s	²¹ O 3.42 s	
¹² N 11 ms	¹³ N 9.965 m	¹⁴ N	¹⁵ N	¹⁶ N 7.13 s	¹⁷ N 4.173 s	¹⁸ N 619.2 ms	¹⁹ N 336 ms	²⁰ N 136 ms	

The Colourful Nuclide Chart

The next step is *not* to add another proton. Doing so would get us F-18, a radioactive isotope with a half-life of less than two hours. Instead, we want to add another neutron instead, giving us O-18.

Are we starting to see a preference for a higher neutron-proton ratio? Nope, not yet! The next two steps are both protons, taking us to F-19 and then Ne-20. And this pattern continues for a while: we continue zig-zagging, but the zigs and zags are now *two* nucleons each. That is, we're adding neutrons and protons in pairs:

	Phosphorus Z=15	²⁴ P	²⁵ P	²⁶ P 43.6 ms	²⁷ P 260 ms	²⁸ P 270.3 ms	²⁹ P 4.102 s	³⁰ P 150 s	³¹ P	³² P 14.269 d	³³ P 25.35 d	³⁴ P 12.43 s			
Silicon Z=14	²² Si 28.7 ms	²³ Si 42.3 ms	²⁴ Si 143.2 ms	²⁵ Si 220.6 ms	²⁶ Si 2.2453 s	²⁷ Si 4.117 s	²⁸ Si	²⁹ Si	³⁰ Si	³¹ Si 157.16 m	³² Si 157 y	³³ Si 6.18 s			
Aluminium Z=13	²¹ Al 91.1 ms	²² Al 446 ms	²³ Al 2.053 s	²⁴ Al 7.1666 s	²⁵ Al 717 ky	²⁶ Al	²⁷ Al 134.7 s	²⁸ Al 6.56 m	²⁹ Al 3.62 s	³⁰ Al 644 ms	³¹ Al 32.6 ms	³² Al			
Magnesium Z=12	¹⁹ Mg 5 ps	²⁰ Mg 90.4 ms	²¹ Mg 120 ms	²² Mg 3.8745 s	²³ Mg 11.3039 s	²⁴ Mg	²⁵ Mg	²⁶ Mg	²⁷ Mg 9.435 m	²⁸ Mg 20.915 h	²⁹ Mg 1.3 s	³⁰ Mg 317 ms	³¹ Mg 270 ms		
Sodium Z=11	¹⁷ Na	¹⁸ Na 1.3 zs	¹⁹ Na 1000 zs	²⁰ Na 447.9 ms	²¹ Na 22.455 s	²² Na 2.6019 y	²³ Na	²⁴ Na 14.956 h	²⁵ Na 59.1 s	²⁶ Na 1.07128 s	²⁷ Na 301 ms	²⁸ Na 33.1 ms	²⁹ Na 43.2 ms	³⁰ Na 45.9 ms	
on 10	¹⁵ Ne 770 ys	¹⁶ Ne 5.7 zs	¹⁷ Ne 109.2 ms	¹⁸ Ne 1.6642 s	¹⁹ Ne 17.2569 s	²⁰ Ne	²¹ Ne	²² Ne	²³ Ne 37.15 s	²⁴ Ne 202.8 s	²⁵ Ne 602 ms	²⁶ Ne 197 ms	²⁷ Ne 30.9 ms	²⁸ Ne 18.8 ms	²⁹ Ne 14.7 ms
F	¹⁴ F 500 ys	¹⁵ F 1.1 zs	¹⁶ F 21 zs	¹⁷ F 64.37 s	¹⁸ F 109.734 m	¹⁹ F	²⁰ F 11 s	²¹ F 4.158 s	²² F 4.23 s	²³ F 2.23 s	²⁴ F 384 ms	²⁵ F 80 ms	²⁶ F 8.2 ms	²⁷ F 5 ms	²⁸ F 46 zs
O	¹³ O 8.58 ms	¹⁴ O 70.621 s	¹⁵ O 122.266 s	¹⁶ O	¹⁷ O	¹⁸ O	¹⁹ O 26.47 s	²⁰ O 13.51 s	²¹ O 3.42 s	²² O 2.25 s	²³ O 97 ms	²⁴ O 77.4 ms	²⁵ O 5.18 zs	²⁶ O 4.2 ps	²⁷ O

The Colourful Nuclide Chart

This will continue all the way up until Ar-36. Where do we go from here?

		Scandium Z=21	³⁵ Sc	³⁶ Sc	³⁷ Sc	³⁸ Sc	³⁹ Sc	⁴⁰ Sc 182.3 ms	⁴¹ Sc 596.3 ms	⁴² Sc 680.72 ms	⁴³ Sc 233.46 m	⁴⁴ Sc 4.0421 h	⁴⁵ Sc	⁴⁶ Sc 83.757 d	⁴⁷ Sc 80.3808 h	4
	Calcium Z=20	³³ Ca	³⁴ Ca	³⁵ Ca 25.7 ms	³⁶ Ca 100.9 ms	³⁷ Ca 181 ms	³⁸ Ca 443.7 ms	³⁹ Ca 860.3 ms	⁴⁰ Ca	⁴¹ Ca 99.4 ky	⁴² Ca	⁴³ Ca	⁴⁴ Ca	⁴⁵ Ca 162.61 d	⁴⁶ Ca	4
1	³¹ K 10 ps	³² K	³³ K	³⁴ K	³⁵ K 175.2 ms	³⁶ K 341 ms	³⁷ K 1.23651 s	³⁸ K 7.651 m	³⁹ K	⁴⁰ K 1.248 Gy	⁴¹ K	⁴² K 12.355 h	⁴³ K 22.3 h	⁴⁴ K 22.13 m	⁴⁵ K 17.6 m	4
2	³⁹ Ar 10 ps	³¹ Ar 15 ms	³² Ar 98 ms	³³ Ar 173 ms	³⁴ Ar 846.46 ms	³⁵ Ar 1.7756 s	³⁶ Ar	³⁷ Ar 35.011 d	³⁸ Ar	³⁹ Ar 268 y	⁴⁰ Ar	⁴¹ Ar 109.61 m	⁴² Ar 32.9 y	⁴³ Ar 5.37 m	⁴⁴ Ar 11.87 m	4
3	²⁹ Cl 5.4 zs	³⁰ Cl	³¹ Cl 190 ms	³² Cl 298 ms	³³ Cl 2.5038 s	³⁴ Cl 1.5267 s	³⁵ Cl	³⁶ Cl 301.3 ky	³⁷ Cl	³⁸ Cl 37.23 m	³⁹ Cl 56.2 m	⁴⁰ Cl 81 s	⁴¹ Cl 38.4 s	⁴² Cl 6.8 s	⁴³ Cl 3.13 s	4
4	²⁸ S 125 ms	²⁹ S 188 ms	³⁰ S 1.1798 s	³¹ S 2.5534 s	³² S	³³ S	³⁴ S	³⁵ S 87.37 d	³⁶ S	³⁷ S 5.05 m	³⁸ S 170.3 m	³⁹ S 11.5 s	⁴⁰ S 8.8 s	⁴¹ S 1.99 s	⁴² S 1.016 s	2
5	²⁷ P 260 ms	²⁸ P 270.3 ms	²⁹ P 4.102 s	³⁰ P 150 s	³¹ P	³² P 14.269 d	³³ P 25.35 d	³⁴ P 12.43 s	³⁵ P 47.3 s	³⁶ P 5.6 s	³⁷ P 2.31 s	³⁸ P 640 ms	³⁹ P 282 ms	⁴⁰ P 150 ms	⁴¹ P 101 ms	4
6	²⁶ Si 2.2453 s	²⁷ Si 4.117 s	²⁸ Si	²⁹ Si	³⁰ Si	³¹ Si 157.16 m	³² Si 157 y	³³ Si 6.18 s	³⁴ Si 2.77 s	³⁵ Si 780 ms	³⁶ Si 503 ms	³⁷ Si 141 ms	³⁸ Si 63 ms	³⁹ Si 41.2 ms	⁴⁰ Si 31.2 ms	4

The Colourful Nuclide Chart

Uh-oh. We're trapped. There's no way to add a single nucleon from where we are and get another stable nuclide. If we add a proton, we get K-37, which is unstable with a half-life of only 1.2 seconds. If we add a neutron, we get Ar-37, with a half-life of 35 days. Ar-37 has 18 protons and 19 neutrons, and it turns out that in fact there are *no* stable nuclides with exactly 19 neutrons. We've hit a gap in the beta-stability line.

But there's something else different about this position: Ar-36 is not the only stable nuclide at this mass number! S-36, with two fewer protons and two more neutrons, is also stable. If we make that swap, we can continue our walk.

Surprisingly, what happens next is something we haven't done yet until this point: we're going to add four protons in a row. This will take us from S-36, to Cl-37, to Ar-38, to K-39, to Ca-40. All of them have exactly 20 neutrons—for some reason, that seems to be an unusually stable number of neutrons, just as 19 was an unusually unstable number.

Our walk continues. Mostly we're doing our zig-zag thing. Sometimes we hit breaks and need to do the two-protons-for-two-neutrons swap. Although once we hit titanium, our neutron runs occasionally get longer. From titanium-46, we add four neutrons in a row:

	2.5 ms	13 ms	21.9 ms	45.3 ms	64.7 ms	152 ms	305.4 ms	8.275 h	8.51 m		2.7562 y				44.5 d	2.6
1	⁴⁴ Mn	⁴⁵ Mn	⁴⁶ Mn 36.2 ms	⁴⁷ Mn 88 ms	⁴⁸ Mn 159.1 ms	⁴⁹ Mn 382 ms	⁵⁰ Mn 283.21 ms	⁵¹ Mn 45.81 m	⁵² Mn 5.591 d	⁵³ Mn 3.7 My	⁵⁴ Mn 312.081 d	⁵⁵ Mn	⁵⁶ Mn 154.734 m	⁵⁷ Mn 85.4 s	⁵⁸ Mn 3 s	59
2	⁴³ Cr 21.1 ms	⁴⁴ Cr 42.6 ms	⁴⁵ Cr 60.9 ms	⁴⁶ Cr 224.3 ms	⁴⁷ Cr 461.6 ms	⁴⁸ Cr 21.56 h	⁴⁹ Cr 42.3 m	⁵⁰ Cr	⁵¹ Cr 27.7015 d	⁵² Cr	⁵³ Cr	⁵⁴ Cr	⁵⁵ Cr 209.82 s	⁵⁶ Cr 5.94 m	⁵⁷ Cr 21.1 s	58
3	⁴² V 79.3 ms	⁴³ V 111 ms	⁴⁴ V 547 ms	⁴⁵ V 422.62 ms	⁴⁶ V 32.6 m	⁴⁷ V 15.9735 d	⁴⁸ V 330 d	⁴⁹ V 271 Py	⁵⁰ V	⁵¹ V 224.58 s	⁵² V 92.58 s	⁵³ V 49.8 s	⁵⁴ V 6.54 s	⁵⁵ V 216 ms	56	
4	⁴¹ Ti 81.9 ms	⁴² Ti 208.3 ms	⁴³ Ti 509 ms	⁴⁴ Ti 59.1 y	⁴⁵ Ti 184.8 m	⁴⁶ Ti	⁴⁷ Ti	⁴⁸ Ti	⁴⁹ Ti	⁵⁰ Ti	⁵¹ Ti 5.76 m	⁵² Ti 102 s	⁵³ Ti 32.7 s	⁵⁴ Ti 2.1 s	⁵⁵ Ti 1.3 s	56
5	⁴⁰ Sc 182.3 ms	⁴¹ Sc 596.3 ms	⁴² Sc 680.72 ms	⁴³ Sc 233.46 m	⁴⁴ Sc 4.0421 h	⁴⁵ Sc	⁴⁶ Sc 83.757 d	⁴⁷ Sc 80.3808 h	⁴⁸ Sc 43.67 h	⁴⁹ Sc 57.18 m	⁵⁰ Sc 102.5 s	⁵¹ Sc 12.4 s	⁵² Sc 8.2 s	⁵³ Sc 2.4 s	⁵⁴ Sc 526 ms	55
6	³⁹ Ca 860.3 ms	⁴⁰ Ca	⁴¹ Ca 99.4 ky	⁴² Ca	⁴³ Ca	⁴⁴ Ca	⁴⁵ Ca 162.61 d	⁴⁶ Ca	⁴⁷ Ca 4.536 d	⁴⁸ Ca 56 Eyr	⁴⁹ Ca 8.718 m	⁵⁰ Ca 13.45 s	⁵¹ Ca 10 s	⁵² Ca 4.6 s	⁵³ Ca 461 ms	54
7	³⁸ K 7.651 m	³⁹ K	⁴⁰ K 1.248 Gy	⁴¹ K	⁴² K 12.355 h	⁴³ K 22.3 h	⁴⁴ K 22.13 m	⁴⁵ K 17.8 m	⁴⁶ K 96.3 s	⁴⁷ K 17.38 s	⁴⁸ K 6.83 s	⁴⁹ K 1.26 s	⁵⁰ K 472 ms	⁵¹ K 365 ms	⁵² K 110 ms	53
8	³⁷ Ar	³⁸ Ar	³⁹ Ar	⁴⁰ Ar	⁴¹ Ar	⁴² Ar	⁴³ Ar	⁴⁴ Ar	⁴⁵ Ar	⁴⁶ Ar	⁴⁷ Ar	⁴⁸ Ar	⁴⁹ Ar	⁵⁰ Ar	⁵¹ Ar	52

The Colourful Nuclide Chart

We continue, always adding pairs of neutrons or pairs of protons, occasionally swapping out a pair of protons for a pair of neutrons, or adding two pairs of neutrons in a row.

Overall the pattern we're seeing is: neutrons and protons want to be roughly equal, except that we want somewhat more neutrons as we get heavier... and for some reason, they really like pairs. In fact, as we glance left and right of the line, we can see isolated stable nuclides out there, exactly one pair away from the line:

	⁵⁵ ms	¹⁰ 2.14 s	¹⁰ 6.4 s	¹⁰ 12.8 s	¹⁰ 49.2 s	¹⁰ 188.4 s	¹⁰ 4.25 m	¹⁰ 165 m	¹⁰ 4.8833333 h	¹⁰ 19.258 h	¹⁰ 4.28 d	¹⁰ 4.21 My	¹⁰ 4.2 My	¹⁰ 211.1 ky	¹⁰ 15.46 s	1
9	⁸⁵ Mo 3.2 s	⁸⁶ Mo 19.1 s	⁸⁷ Mo 14.1 s	⁸⁸ Mo 8 m	⁸⁹ Mo 126.6 s	⁹⁰ Mo 5.56 h	⁹¹ Mo 15.49 m	⁹² Mo	⁹³ Mo 4 ky	⁹⁴ Mo	⁹⁵ Mo	⁹⁶ Mo	⁹⁷ Mo	⁹⁸ Mo	⁹⁹ Mo 65.932 h	100
10	⁸⁴ Nb 9.8 s	⁸⁵ Nb 20.5 s	⁸⁶ Nb 88 s	⁸⁷ Nb 222 s	⁸⁸ Nb 14.5 m	⁸⁹ Nb 121.8 m	⁹⁰ Nb 14.6 h	⁹¹ Nb 680 y	⁹² Nb 34.7 My	⁹³ Nb	⁹⁴ Nb 20.4 ky	⁹⁵ Nb 34.991 d	⁹⁶ Nb 23.35 h	⁹⁷ Nb 72.1 m	⁹⁸ Nb 2.86 s	99
11	⁸³ Zr 42 s	⁸⁴ Zr 25.8 m	⁸⁵ Zr 7.86 m	⁸⁶ Zr 16.5 h	⁸⁷ Zr 100.8 m	⁸⁸ Zr 83.4 d	⁸⁹ Zr 78.36 h	⁹⁰ Zr	⁹¹ Zr	⁹² Zr	⁹³ Zr 1.61 My	⁹⁴ Zr	⁹⁵ Zr 64.032 d	⁹⁶ Zr 23.4 Ey	⁹⁷ Zr 16.749 h	98
12	⁸² Y 8.3 s	⁸³ Y 7.08 m	⁸⁴ Y 39.5 m	⁸⁵ Y 160.8 m	⁸⁶ Y 14.74 h	⁸⁷ Y 79.8 h	⁸⁸ Y 106.629 d	⁸⁹ Y	⁹⁰ Y 64.05 h	⁹¹ Y 58.51 d	⁹² Y 212.4 m	⁹³ Y 10.18 h	⁹⁴ Y 18.7 m	⁹⁵ Y 10.3 m	⁹⁶ Y 5.34 s	97
13	⁸¹ Sr 22.3 m	⁸² Sr 25.35 d	⁸³ Sr 32.41 h	⁸⁴ Sr	⁸⁵ Sr 64.846 d	⁸⁶ Sr	⁸⁷ Sr	⁸⁸ Sr	⁸⁹ Sr 50.563 d	⁹⁰ Sr 28.91 y	⁹¹ Sr 9.65 h	⁹² Sr 156.66 m	⁹³ Sr 7.43 m	⁹⁴ Sr 75.3 s	⁹⁵ Sr 23.9 s	96
14	⁸⁰ Rb 33.4 s	⁸¹ Rb 4.572 h	⁸² Rb 75.45 s	⁸³ Rb 86.2 d	⁸⁴ Rb 32.82 d	⁸⁵ Rb	⁸⁶ Rb 18.645 d	⁸⁷ Rb 49.7 Gy	⁸⁸ Rb 17.78 m	⁸⁹ Rb 15.32 m	⁹⁰ Rb 158 s	⁹¹ Rb 58.2 s	⁹² Rb 4.48 s	⁹³ Rb 5.84 s	⁹⁴ Rb 2.702 s	95
15	⁷⁹ Kr 35.04 h	⁸⁰ Kr 229 ky	⁸¹ Kr	⁸² Kr	⁸³ Kr	⁸⁴ Kr	⁸⁵ Kr 10.728 y	⁸⁶ Kr	⁸⁷ Kr 76.3 m	⁸⁸ Kr 169.5 m	⁸⁹ Kr 189 s	⁹⁰ Kr 32.32 s	⁹¹ Kr 8.57 s	⁹² Kr 1.84 s	⁹³ Kr 1.287 s	94
16	⁷⁸ Br	⁷⁹ Br	⁸⁰ Br	⁸¹ Br	⁸² Br	⁸³ Br	⁸⁴ Br	⁸⁵ Br	⁸⁶ Br	⁸⁷ Br	⁸⁸ Br	⁸⁹ Br	⁹⁰ Br	⁹¹ Br	⁹² Br	93

The Colourful Nuclide Chart

After Mo-98, we encounter a new phenomenon: the first row in our chart with no stable nuclides. In other words, the next element, technetium (Tc), is the lightest element with no stable isotopes:

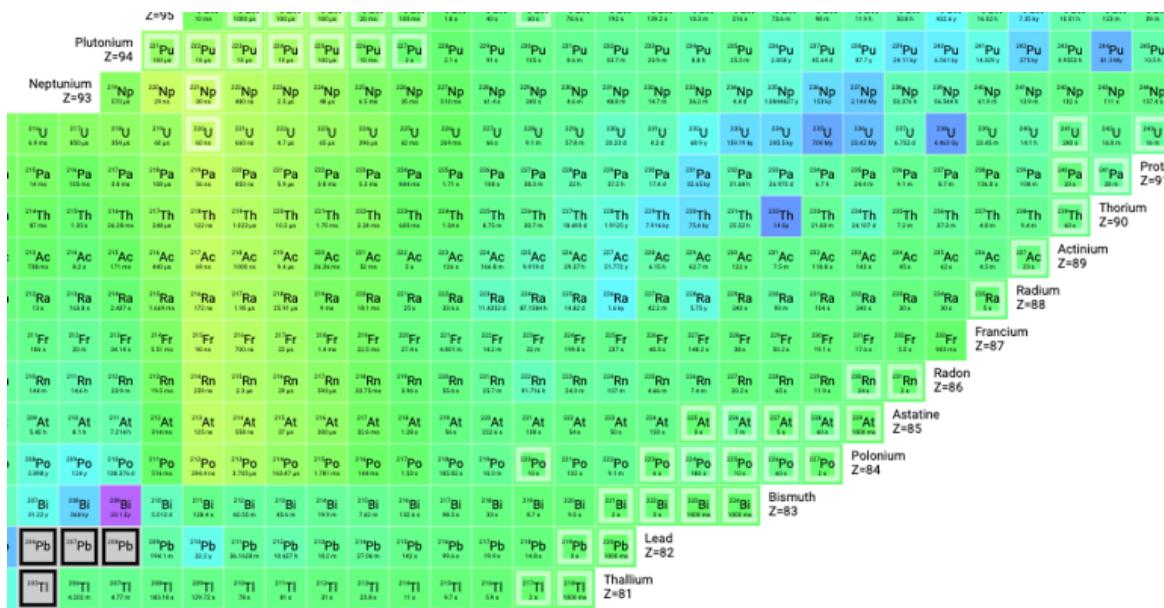
Ag 1.78 s	Ag 4.45 s	Ag 25.5 s	Ag 47.5 s	Ag 124.2 s	Ag 120.6 s	Ag 11.1 m	Ag 12.9 m	Ag 65.7 m	Ag 69.2 m	Ag 41.29 d	Ag 23.96 m	Ag 142.92 s	Ag 2
Pd 9.1 s	Pd 7.4 s	Pd 122 s	Pd 186 s	Pd 17.7 m	Pd 21.4 m	Pd 87.12 h	Pd 8.47 h	Pd 16.991 d	Pd 104	Pd 105	Pd 106	Pd 107	Pd 108
Rh 13.9 s	Rh 70.6 s	Rh 5.02 m	Rh 9.9 m	Rh 30.7 m	Rh 8.72 m	Rh 16.1 d	Rh 20.8 h	Rh 4.07 y	Rh 207 d	Rh 103	Rh 104	Rh 105	Rh 106
Ru 219 s	Ru 59.7 s	Ru 51.8 m	Ru 96.42 m	Ru 96	Ru 68.088 h	Ru 98	Ru 99	Ru 100	Ru 101	Ru 102	Ru 103	Ru 104	Ru 105
Tc 188.4 s	Tc 4.25 m	Tc 165 m	Tc 4.8833333 h	Tc 19.258 h	Tc 4.28 d	Tc 4.21 My	Tc 4.2 My	Tc 211.1 k	Tc 15.46 s	Tc 14.22 m	Tc 5.28 s	Tc 54.2 s	Tc 18.3 m
Mo 5.56 h	Mo 15.49 m	Mo 92	Mo 4 ky	Mo 94	Mo 95	Mo 96	Mo 97	Mo 98	Mo 99	Mo 100	Mo 101	Mo 102	Mo 103
Nb 121.8 m	Nb 14.6 h	Nb 680 y	Nb 34.7 My	Nb 93	Nb 20.4 ky	Nb 34.991 d	Nb 23.35 h	Nb 72.1 m	Nb 2.86 s	Nb 15 s	Nb 1.5 s	Nb 7.1 s	Nb 4.3 s
Zr 83.4 d	Zr 78.36 h	Zr 90	Zr 91	Zr 92	Zr 1.61 My	Zr 64.032 d	Zr 23.4 Ey	Zr 16.749 h	Zr 30.7 s	Zr 2.1 s	Zr 7.1 s	Zr 2.29 s	Zr 2.01 s
Y 87	Y 88	Y 89	Y 90	Y 91	Y 92	Y 93	Y 94	Y 95	Y 96	Y 97	Y 98	Y 99	Y 100

The Colourful Nuclide Chart

We can do another pair swap, this time swapping out a pair of *neutrons* for a pair of *protons*. This takes us to Ru-98, from which we can continue our walk.

Those patterns take us pretty much the rest of the way. The territory is familiar now, with few surprises. Element 50, tin, starting at Sn-114, gives us our first case of adding six neutrons in a row... 50 protons is a really stable number of protons for some reason. After Nd-146, we hit another element, promethium, with no stable isotopes—in fact, there is no stable nuclide at all with a mass number of 147, to the closest being samarium-147 with a half-life of 100 billion years. But in general, we can keep going this way until about lead-208.

This is the last stable nuclide. We've hit the end of the path. Ahead of us is just a sea of radioactivity:



The Colourful Nuclide Chart

Oh, there are a few heavier elements with isotopes long-lived enough to be found in nature—such as uranium-238, with a half-life of about the age of the Earth, or thorium-232, with a half-life of about the age of the universe. But venture much beyond those, and there's nothing but crazy-unstable superheavy elements.

Or is there?

Maybe not. But to explain why, I'll need to first answer the question that's already on your mind, namely: what the heck is going on?

What's causing these patterns? Why does the line bend the way it does? And what's the deal with even numbers?

We've been staring at empirical data for long enough. Time for some theory.

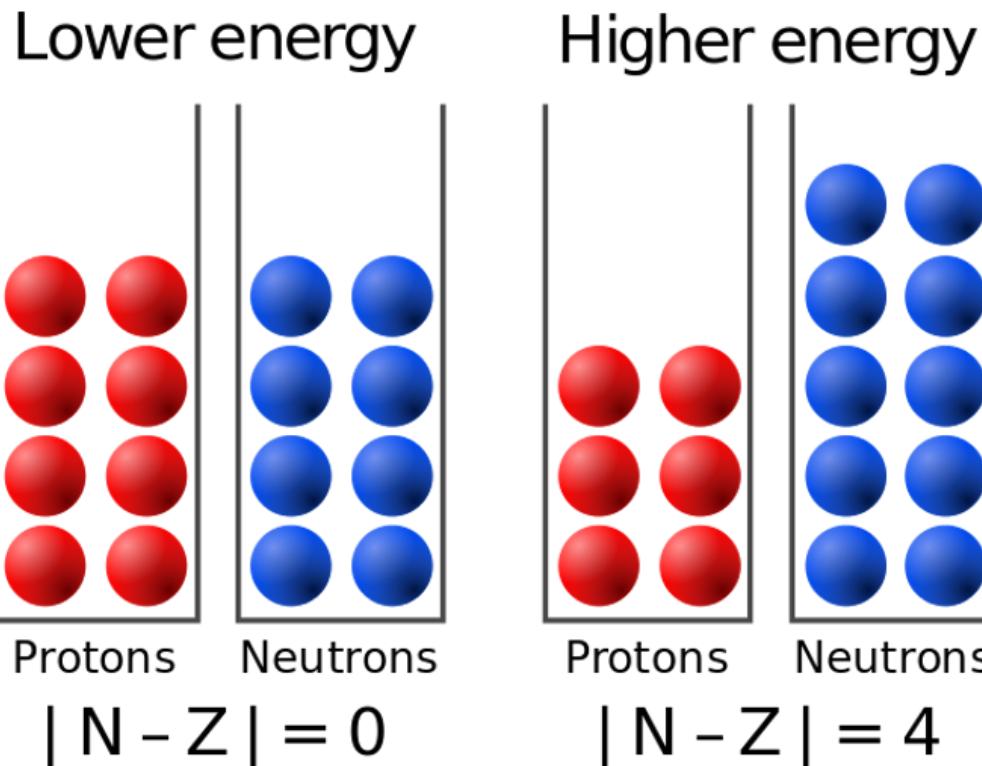
Let's think about the forces at play within the nucleus. Nucleons are bound together by something called the strong force, one of the four fundamental forces of physics. As the name implies, it's a very strong interaction, able to overcome all other forces, although only at very short range.

But there is another force at work, one that is more commonly known: the electrostatic force. Protons are positively charged, so they repel each other. This is why the line bends towards the neutron axis: the more protons you have, the more neutrons you need for padding.

If that were the *only* thing going on, though, there would be lots of stable isotopes with extra neutrons. But as we've seen, adding too many neutrons makes your nucleus unstable about as fast as adding too many protons. Why?

This part of the pattern is due to the [Pauli exclusion principle](#). If you think back again to high school chemistry, you may remember that as electrons are added to an atom, they occupy distinct quantum states, at successively higher energy levels. It turns out the same is true of protons and neutrons. As we add more of either type of particle, they too occupy higher and higher energy levels. But neutrons don't exclude protons or vice versa; a particle only excludes others of the same type. The upshot is that the total energy is lower when protons and neutrons are in balance. This diagram illustrates the concept comparing stable oxygen-16, on the left, with radioactive carbon-16, on the right:

$$A = 16$$



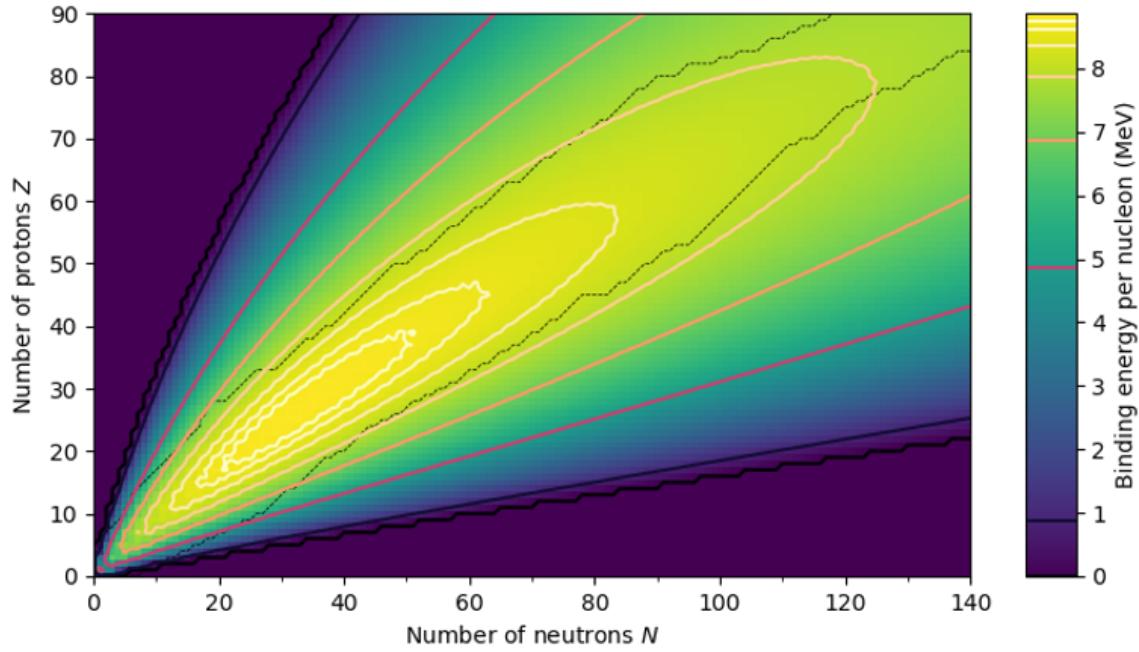
[Wikimedia / Marekich](#)

These two factors, then—Pauli exclusion driving towards an equal number of protons and neutrons, plus electrostatic repulsion driving away from too many protons—result in a line of beta stability that starts off along the diagonal axis, but progressively bends towards the neutron axis.

But wait—what about the odd zig-zag pattern of the line? Why are even numbers of protons and neutrons more stable? This too is a quantum-mechanical phenomenon. Suffice it to say that a sort of bond can be formed between two particles that have the same quantum numbers except for opposite spins.

These phenomena, together with the strong force itself, are combined in the “liquid drop” model of the nucleus as a fluid ball with a variety of attractive and repulsive forces between the particles. From this model, we can [derive a formula](#) to estimate the *binding energy* of each nuclide. Analogous to the [gravitational binding energy](#) of a planet, the binding energy of a nucleus is the energy that would be required to take it apart, nucleon by nucleon. Thus it represents the depth of an energy well.

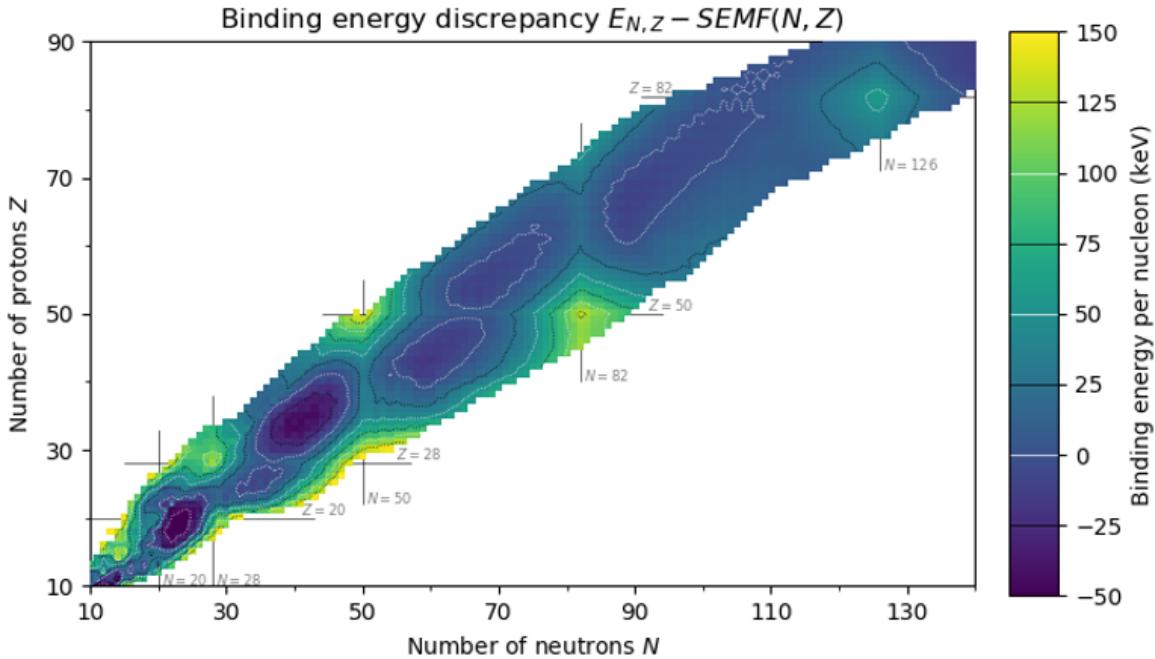
The binding energy per nucleon is a rough measure of the stability of a nuclide. By our formula, it looks something like this, peaking around copper:



[Wikimedia / Mia Dobson](#)

The lighter elements can thus release energy through fusion; the heavier elements through fission.

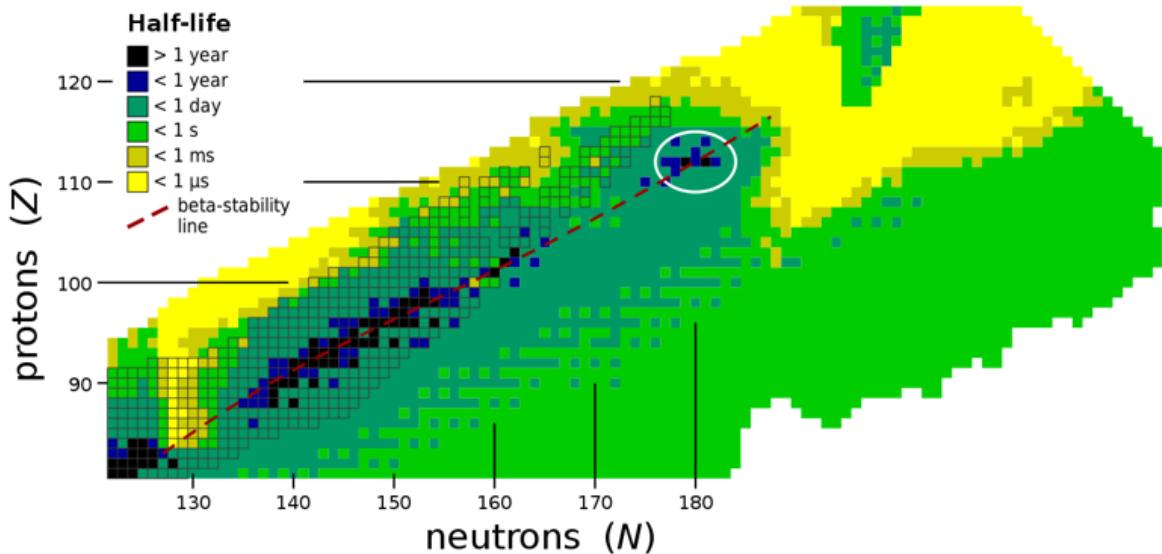
But when we compare this estimate to measured values, there are discrepancies. And the discrepancies show a pattern, with distinct horizontal and vertical lines on the chart:



[Wikimedia / Mia Dobson](#)

These lines, for instance at 20, 28, 50, or 82 on either axis, are the so-called “magic numbers”. This observation led to a modification to the liquid-drop model: the “nuclear shell” model. In this model, protons and neutrons occupy energy “shells”, like electrons do outside the nucleus. Filling a shell results in greater stability.

And this, finally, brings us around to our question: are there any elements much beyond uranium that have long half-lives? So far, of the elements beyond 112 (copernicium), the most stable isotopes have a half-lives measured in seconds or less. But the nuclear shell model predicts that very heavy isotopes of some of these elements—that is, with a bunch more neutrons—would reach a complete neutron shell and become more stable: maybe minutes, maybe days, in some estimates even millions of years. The hypothesized phenomenon is known as the “island of stability”, shown here circled in white:



[Wikimedia / Lasunncity](#)

Does the island exist? No one knows for sure. We haven't found any of these elements in nature—nor do we have the technology to create them, yet.

Let's descend from the lofty reaches of theory to the practical reality of engineering. What we've just seen helps us understand how fission energy works, why nuclear fuel turns into nuclear waste, and why that waste needs special handling.

All the radioactive decay modes—such as alpha, beta (electron), and fission—give off energy. The nucleus is seeking a deeper energy well, and it sheds energy to get there. It turns out that fission gives off the most energy by far: about 200 MeV in a typical event, about a hundred times that of beta decay. A lot of this takes the form of kinetic energy: the two new, smaller, nuclei, no longer bound by the strong force but still very much repulsed by the electrostatic force of their very positive charges, flee from each other like spurned lovers at about 3% of the speed of light. 200 MeV is also about a hundred million times that of a typical chemical reaction, which is why the energy density of nuclear fuel is orders of magnitude beyond that of fossil fuels or indeed any other currently usable energy source.

So, for nuclear fuel, then, do we want to use isotopes that decay by spontaneous fission? No. They don't exist outside the laboratory; they're too unstable. Any isotope with a half-life of less than 70 million years has already decayed to less than 1% of the amount that existed at the formation of the Earth. (OK, there are some exceptions, such as carbon-14, which gets produced in the atmosphere by cosmic rays. But we need heavy elements for fission.)

What we need is an element that can be *induced* to fission. Instead of spontaneous fission, we want *controlled* fission. Fission with an on-off switch. Where can we find this?

To induce an atom to fission, we'll have to give it a bit of a kick. It needs an input of energy to get over the hump and split in two, and then that pent-up electric repulsion will pay back

our little energy investment handsomely.

It turns out that some elements, such as uranium-235, can get this energy from a stray neutron flying around. If the neutron is flying at the right speed, it can get “captured” and absorbed by the nucleus, very briefly turning it into uranium-236. Soon, the energy this adds causes the nucleus to fission.

Where can we get a flying neutron from? Well, recall the description of fission above: a large atom turns into two medium-sized atoms, *plus a few extra neutrons*.

Do you see where this is going? If we can get just one fission to happen, it will release energy and a few neutrons. Each of those neutrons could potentially cause *another* fission, which would release more neutrons, which would cause more fission, etc. This is called the *chain reaction*.

The nuclear chain reaction gives us everything from nuclear electricity to nuclear bombs. The difference between the two is how fast the energy is released. In a bomb, so many U-235 atoms are packed so closely together that the rate of fission events doubles in less than a millisecond, creating literally explosive exponential growth and an extremely rapid and destructive release of energy. In a nuclear power plant, a much lower proportion of U-235 is used, and the reactor is engineered so that the growth rate of the fission events is much lower. For this reason, contrary to popular fears, it’s physically impossible for a nuclear power reactor to explode like a bomb. (The most it can do is get so hot that it melts.)

Moreover, the reactor is engineered to be controllable, so that it can be started up, shut down, and run at any desired rate of power. We control the reaction by controlling the neutrons: capturing them, or controlling their speed. Once a nuclear reactor has been turned on and ramped up to full power, the growth rate is taken to zero. That is, each fission event creates only *one* new fission event on average, giving an overall fission rate that is stable.

So our reactor is fissioning, our fuel is “burning”, we’re releasing lots of energy. The energy creates heat, as all those fast-flying fission products bonk into other atoms and shake them up. Then, as in any thermal power plant, the heat boils water, the boiling creates steam, the steam drives a turbine, the turbine spins a magnet, and the spinning magnet creates electric power. Many megawatts, sometimes gigawatts, of electric power.

But what’s happening to the fuel? Our U-235 is getting used up, turning into smaller, lighter elements such as iodine and zirconium. These elements aren’t normally hazardous—in fact, iodine is an essential nutrient, needed by your thyroid; we put it in table salt for this reason. So why is nuclear waste hazardous?

Remember our first observation about nuclides: heavier elements have a higher neutron:proton ratio. In particular, fissionable materials such as uranium have a higher neutron:proton ratio than their fission products. This is why extra neutrons are released in a fission event. But it also means that *the fission products themselves have too many neutrons*. And then remember our second observation: isotopes with too many neutrons are unstable. So, the fission products aren’t regular old iodine, zirconium, caesium, etc.—they’re radioactive isotopes of those elements. And their half-lives are much shorter than natural uranium, meaning they are decaying faster.

What type of radiation are they giving off? Well, remember our third observation about nuclides: light-to-medium nuclides with too many neutrons give off beta radiation (electrons) and gamma radiation. And remember too that gamma radiation in particular is the most dangerous to health.

In summary, fresh nuclear fuel is mostly long-lived isotopes of uranium. These nuclides can fission when we induce them to, giving off lots of heat (good for power generation!), and when we don’t, they give off a much lower level of radiation, and not the kind that is most harmful to humans. Spent nuclear fuel, however, contains short-lived isotopes of lighter

elements. These don't fission, so they're not useful for power generation, but they *do* give off harmful gamma radiation.

In other words, when you generate nuclear power, you're going from a radioactive material that is useful and relatively harmless to one that is not useful but is dangerous. Hence, waste.

The good news is that, because nuclear fuel is insanely energy-dense, there just isn't that much waste to deal with, per kWh generated. For instance, here is *all* the waste produced by the Connecticut Yankee plant during its 28-year lifetime, during which it generated 110 terawatt-hours of electricity:



[Jack Devaney, Why Nuclear Power Has Been a Flop](#)

It fits on a single pad, 70x228 feet. And most of the mass in that photo is concrete; only about 20% is the waste. (As Jack Devaney notes, if a coal plant generated the same amount of energy and we tried to fit it on this pad, it would be over 1.3 miles high.)

In fact, all of the waste produced by all of the ~100 or so nuclear reactors in the US is currently safely stored onsite, above ground, in concrete casks. You can go right up to them and hug them:



Cask Hugging at Palo Verdes. [Paris-Ortiz-Wines](#)

They will be there until we get our act together and allow for long-term storage somewhere other than the indefinitely-delayed [Yucca Mountain repository](#).

Or maybe not? There are advanced reactor designs that don't rely on the fission of U-235, but rather use the far more abundant isotope U-238. Some of these reactors can burn "spent" fuel, or "depleted" uranium left over from the enrichment process—extracting something like 60 times as much energy from uranium as traditional nuclear reactors.

Many engineering feats are possible, once you understand the fundamental physics.

Thanks to Phil Mohun for reading a draft of this.

Covid 5/13: Moving On

For over a year, Covid-19 has been the central fact of life.

The goal now is *to make that no longer true.*

If you're reading this, chances are very high you are vaccinated. If you're not reading this, but you live in the United States, chances are still pretty good you're vaccinated.

The question everyone is asking is now, can life return fully to normal?

Yes.

Well, almost. We'll never be psychologically quite the same. We'll never unlearn the lessons we've learned over the past year – nor would we want to – about the way our civilization and its institutions work, or about what matters in life. And at least for now, we'll need to continue to worry about how others will interact with us and how to navigate people's concerns and various governmental restrictions, as well as the Covid-19 conversations that they'll doubtless want to have for a long time. If others don't return to normal, there's no fully normal to return to yourself.

And that doesn't mean quite *fully* normal *quite* yet, in the sense that there's an amount of indoor crowding I'd still be inclined to avoid for the next few weeks or months.

And of course, things may be going well in America and most other highly vaccinated places, but the worldwide pandemic is far from over. Things in many other places remain quite bad, and will be quite bad for some time.

But... mostly? Yes. For those who are fully vaccinated, life can safely return to normal.

This column can also, events willing, start winding down or transitioning to other matters. If things go as planned, there will steadily be less Covid-19 news to talk about each week, and I can shift my blogging time into other, longer-term pursuits once more.

For now, let's run the numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 3.5% (down 0.4%) and deaths decline by 7%.

Result:

In the past week in the U.S. ...

New daily reported **cases fell 21.8% ↓**

New daily reported **deaths fell 11.6% ↓**

Covid-related **hospitalizations fell 11.8% ↓** [Read more](#)

Among reported tests, **the positivity rate was 3.4%**.

The **number of tests reported fell 24.1% ↓** from the previous week. [Read more](#)

Things improved faster than I expected, which is great. The fall in deaths makes perfect sense, as I was adjusting for strangely small drops in deaths from previous weeks. The fall in cases is mostly in line with expectations.

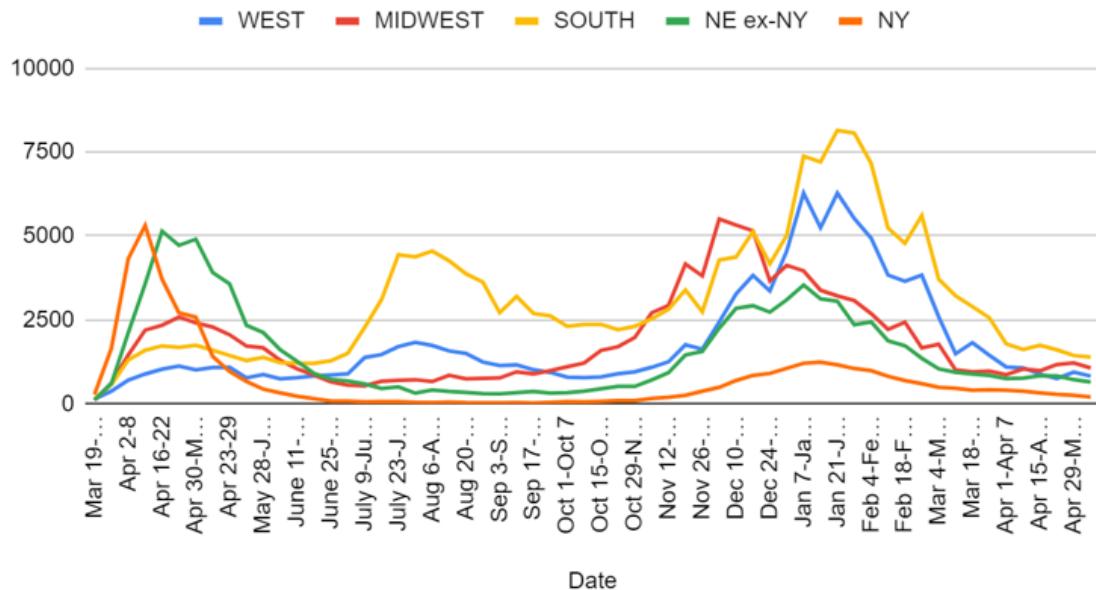
Prediction for Next Week: Positivity rate of 3.0% (down 0.4%) and deaths decline by 10%.

This looks like the endgame. The control system is the potential threat to that, as we're likely not near full herd immunity if everyone went fully back to the old normal, but I do not expect people to return to the full old normal. We will still pick a lot of the truly 'low hanging fruit' and I think that will go far enough to bring us over the top.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Apr 1-Apr 7	1098	867	1789	1160	4914
Apr 8-Apr 14	1070	1037	1621	1145	4873
Apr 15-Apr 21	883	987	1747	1168	4785
Apr 22-Apr 28	752	1173	1609	1110	4644
Apr 29-May 5	943	1220	1440	971	4574
May 6-May 12	826	1069	1392	855	4142

Deaths by Region

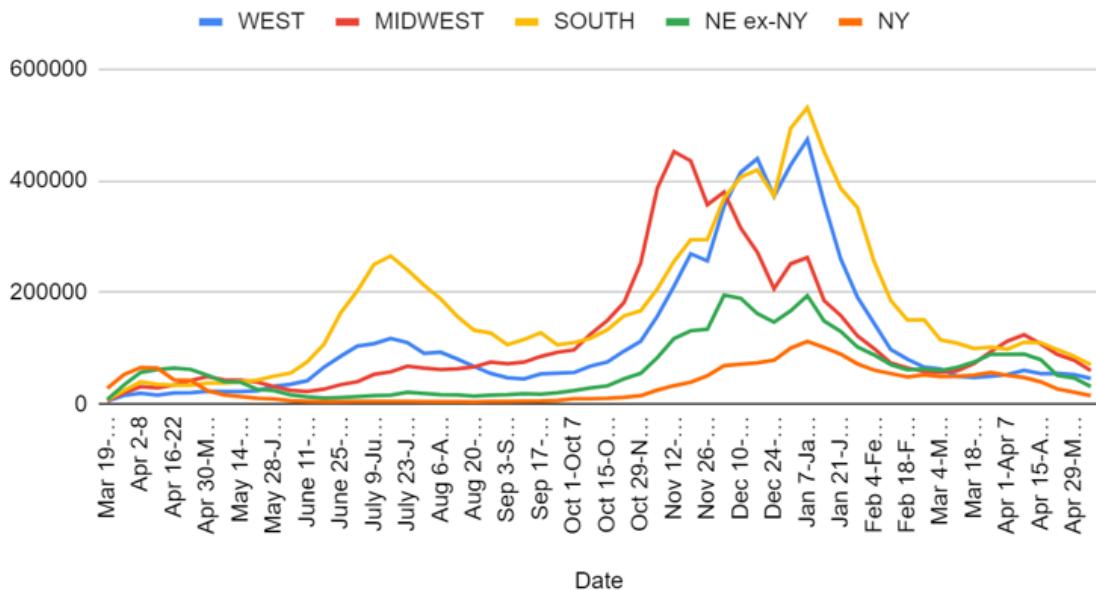


We saw a steady decline across all regions, and I see no reason not to expect this to continue.

Cases

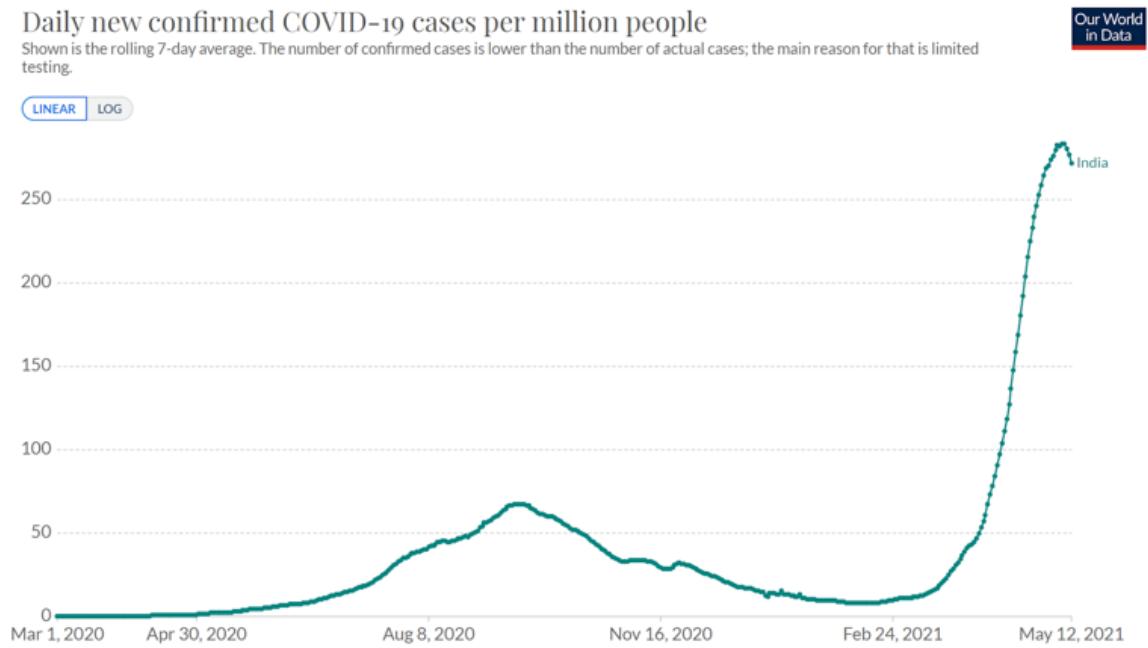
Date	WEST	MIDWEST	SOUTH	NORTHEAST
Mar 25-Mar 31	49,669	93,690	102,134	145,933
Apr 1-Apr 7	52,891	112,848	98,390	140,739
Apr 8-Apr 14	60,693	124,161	110,995	137,213
Apr 15-Apr 21	54,778	107,700	110,160	119,542
Apr 22-Apr 28	54,887	88,973	97,482	78,442
Apr 29-May 5	52,984	78,778	85,641	68,299
May 6-May 12	46,045	59,945	70,740	46,782

Positive Tests by Region



That's a dramatic acceleration of improvement, especially in the northeast. Progress is much slower in the West, but even 10% a week adds up quickly and there's still a lot more vaccinations to bring online. This is what the endgame looks like.

India



Daily new confirmed COVID-19 deaths per million people

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

Our World
in Data



The share of daily COVID-19 tests that are positive

Shown is the rolling 7-day average. The number of confirmed cases divided by the number of tests, expressed as a percentage. Tests may refer to the number of tests performed or the number of people tested – depending on which is reported by the particular country.

Our World
in Data



This looks about as good as one could have hoped. The control system is indeed powerful, and a few things like the elections have finished up which is presumably helping. There's some worry about the numbers reflecting limits on the Indian state's capacity to test and measure, rather than a real limit, but the test percentage is static so I mostly believe that things are no longer getting worse. That doesn't mean they're getting substantially better yet, and the whole situation remains quite horrible, but things could have been *much, much* worse and it looks like those scenarios are going to be avoided.

[My old colleague Ross Rheingans-Yoo asks how much demographics should lower India's death rate from Covid](#), concludes about a 61% effect but warns that lack of medical care and

abundance of air pollution could undo this.

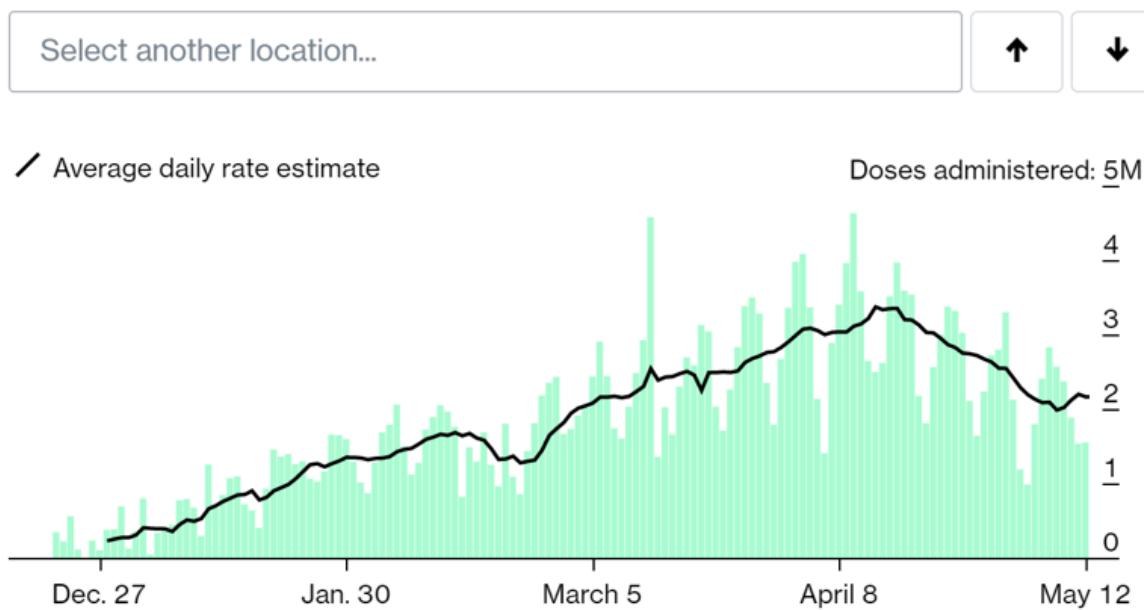
Vaccinations

We have now approved the Pfizer vaccine for children 12-15. Surveys say that there will be a lot of reluctance to vaccinate children, and public service approaches like ‘first thing talk to your child, they are sure to have an opinion’ do not seem likely to be helpful in solving this problem. As with vaccination in general, there will be a core of people desperate to get the vaccine for their children one day sooner to feel safe and let their lives move closer to normal, then varying degrees of reluctance and skepticism. A lot of people are using the ‘wait and see how it goes’ option in surveys, despite this attitude making very little physical sense. We are not going to learn anything here.

154 million vaccinated

The number of people who have received at least one dose of the vaccine, covering **55.0% of the eligible population, 12 and older** and **46.4% of the total population**.

In the U.S., the latest vaccination rate is **2,162,191 doses** per day, on average. At this pace, it will take another **3 months** to cover **75%** of the population.



This is excellent news. We had a steady decline in doses, and that decline has at least temporarily stalled out despite less people needing their second dose and despite more people having already been vaccinated, and before the 12-15 year old crowd started getting their shots. That both can make us more optimistic that there's still a lot of people not yet vaccinated who are going to end up vaccinated, and it also is more evidence that the J&

pause was a large part of the problem, which is hopefully now receding somewhat in people's minds. That's because the alternate hypothesis was that we were running out of remaining willing arms in which to put shots and this was a natural peak, but if it was a natural peak the decline seems like it should have continued. It's not definitive, but it is definitely suggestive.

Maybe It's All a Coincidence

The following series of events occurred a few weeks ago:

1. The J&J vaccine was suspended for no (good) reason.
2. A lot less people suddenly got vaccinated in America.
3. Number of people vaccinated kept going down rapidly.
4. I and many others concluded #1 contributed a lot to #2 and #3.

As we all know, correlation does not prove causation, and first doses had already peaked before this. There's a not-completely-impossible case one could thus make that the J&J pause *didn't* contribute meaningfully to the decline in vaccinations, and things are not substantially different from the counterfactual other than the slowdown specifically in J&J doses.

There's survey data out in support of this thesis:



Seth Burn @SethBurn · May 8

TL;DR: "The impact of the pause on vax demand was: BUBKES."



David Lazer 

@davidlazer

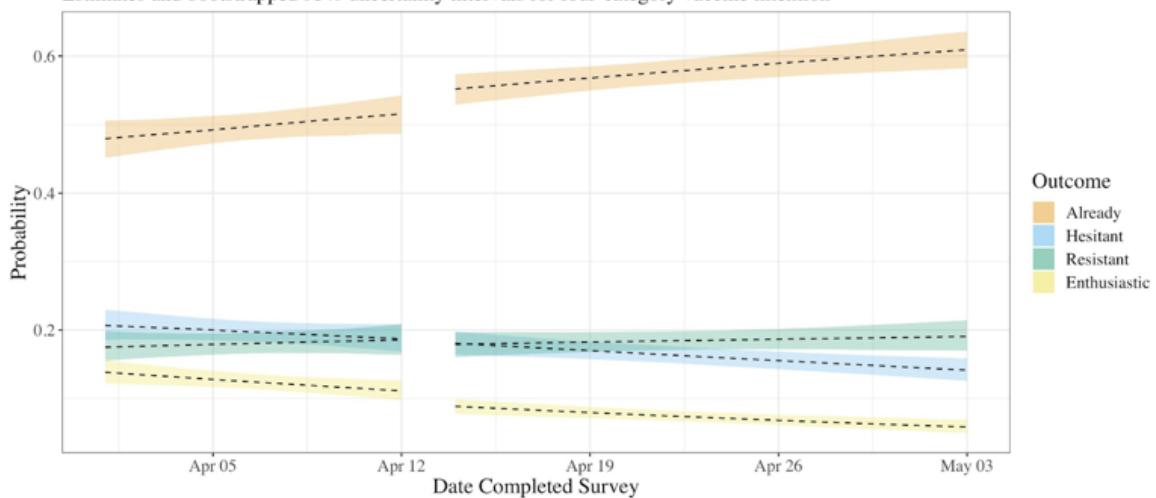
...

Did the J&J pause reduce demand for vaccinations, as some have asserted? (cough [@NateSilver538](#) cough)

Our latest [COVIDstates.org](#) report evaluates this question, based on a large (20k) survey we had in the field the entire month of April.

Trends in Vaccine Intentions Pre/Post Johnson & Johnson Pause

Predictions for non-college white woman at medians of age group, income group, and party/ideology
Estimates and bootstrapped 95% uncertainty intervals for four-category vaccine intention



 **David Lazer**  @davidlazer · May 7

...

Replying to @davidlazer

The scale of the survey, and the rapid change of vax sentiment/status through April give us a pretty good look at dynamics down to the daily granularity. & the pretty clear answer, despite very high awareness, is that the impact of the pause on vax demand was: BUBKES.

4

7

47



 **David Lazer**  @davidlazer · May 7

...

What we see is a steady increase in the number of vaccinated people, & a steady decrease in vaccine enthusiastic and hesitant individuals. Vax resistance is quite steady.

1

7

29



 **David Lazer**  @davidlazer · May 7

...

So: J&J did present a short term supply shock, likely modestly slowing vax rates for a week or two? But the peak shots/day was always going to be in April, as we exhausted the easy to vaccinate enthusiasts.

3

2

33



 **David Lazer**  @davidlazer · May 7

...

This does not speak to other possible downsides of the pause (e.g., has J&J vaccine been undermined in the US & elsewhere?); nor upsides (has this enhanced trust of the regulatory/safety process?). But: there is no evidence that the pause caused a drop in demand.

Note that this is 'no evidence' *versus there never being a blood clot concern at all* rather than being *versus the alternative world where the (non)-issue still existed but they didn't suspend the vaccine*.

I'm going to go ahead and propose this:

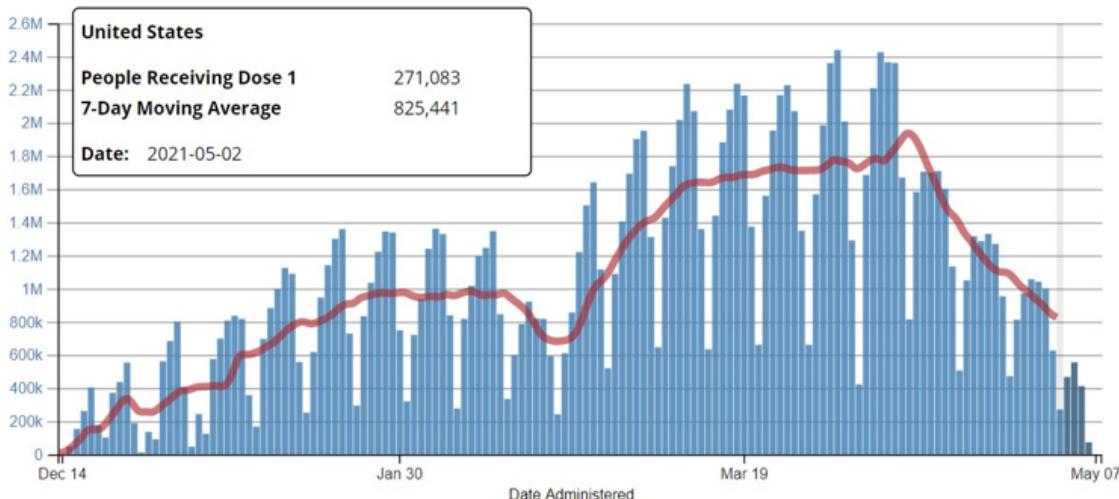
Law of No Evidence: Any claim that there is "no evidence" of something is evidence of bullshit.

No evidence should be fully up there with "government denial" or "[I didn't do it, no one saw me do it, there's no way they can prove anything](#)." If there was indeed no evidence, there'd be no need to *claim* there was no evidence, and this is usually a move to categorize the evidence as illegitimate and irrelevant because it doesn't fit today's preferred form of scientism.

You can certainly say that the survey data is an important challenge to the claim that the pause mattered!

But also, you gotta [look at this graph](#) and think...

Daily Count of People Receiving Dose 1 Reported to the CDC by Date Administered, United States



[...maybe not exactly this...](#)



Nate Silver ✅ @NateSilver538 · May 7

Maybe instead of trying to heroically interpret ambiguous survey data, we should look directly at the variable of interest: how many people are getting vaccinated. It underwent a huge nonlinear plunge timed *exactly* to the J&J pause and has never recovered (see below).



Nate Silver ✅ @NateSilver538 · May 7

BTW, vaccine hesitant people are likely to be hard to reach in surveys. Mostly non college educated, conservative, young, low social trust. That's kind of a perfect storm of people who it's hard to get to complete a poll.

173

83

902



But still something like... okay, sure, that's at least *some* evidence. There's no way you can look at that chart and think that the week after the line matches the previous trend.

Or, alternatively, you could look at *all the specific people I know about who expressed uncertainty surrounding the pause*, or all the work I've been forced to do about the pause, and the professional time I've spent advising on how to deal with questions about the pause and how it is a substantial fraction of any draft of a vaccine FAQ at this point.

I also don't want to think about the smoothing effects it took to get the hesitancy graph above to be all straight lines. But let's give the graph the benefit of the maximum benefit of the doubt, and assume that people really did give similar responses before and after the pause. What's going on with that?

The graph presumes we can divide people into four categories: Already vaccinated, enthusiastic, hesitant and resistant.

The first thing to note is that the top line, ‘already vaccinated,’ doesn’t *look* right, because it looks like a smooth increase rather than an increase that is dramatically slowing down over time in the later part of the graph, but let’s assume that this is a resolution issue and the slowdown is present.

The enthusiastic line seems to be continuing to decline slightly, as there are more enthusiastic people getting vaccinated than there are others moving up to enthusiastic. It’s suspiciously flat. That could be a kind of coincidence, but if there’s a mostly static proportion of people who are ‘enthusiastic’ about vaccination, and a mostly static proportion of ‘hesitant’ as well (it’s declining, but *much much* slower than vaccination rates), what’s causing such a dramatic drop in vaccination rates?

The response presumably would be that this is select effects based on logistics. People can say they are ‘enthusiastic’ all they like, but if you’re enthusiastic *now* without having had a dose, that increasingly means you’re running into some sort of logistical issue stopping you from getting a shot. You can’t navigate a computer, or you can’t get time off work, or *something*. Over time, that gets worse. And that’s then the effect causing the slowdown.

A question for all such graphs is what shape the peak looks like, and a common mistake is to presume a straight line up then a phase shift into a straight line down. Here, we have a very slowly increasing line where we were (by all reports) mostly supply constrained the whole way, with a supply-related dip in the middle, and then suddenly a steady decline.

So the story is that there are two limiting factors, supply and demand. We were supply limited until the peak, then suddenly demand limited, and now we’re seeing a linear decline in effective demand due to logistical concerns.

My obvious first question on that would be, there’s a dramatic *logistical* easing the moment supply is not a binding constraint. Not only did we open almost all appointments to everyone, we also made walk-ins available, including at local pharmacies. If logistics are a major issue, then there have to be a lot of people who couldn’t afford to make an appointment, but I’d like to think almost everyone can find time *at some point* to show up at their local CVS, even if it takes a few tries due to lines, and as demand falls those lines will get shorter and rarer quickly.

I do get that in disadvantaged communities, getting the information across can be hard, so presumably this isn’t *entirely* lizardman constant or social desirability bias.

We then get to a slowly rising number of ‘resistant’ people, while the number of ‘hesitant’ people slowly goes down, which makes sense. Hesitant people can and often are persuaded, whereas fully resistant ones rarely change their minds, and occasionally hesitancy becomes resistance. If we extrapolate these lines, we’d expect to stabilize at something like 70% vaccinated, at which point we’d only make progress by coercing the resistant into compliance (or somehow persuading them, which seems harder and less likely).

That’s a consistent, reasonable model, except for the problem that it pretends that the pause and blood clot issue didn’t matter when I have direct observations that say that it did matter and the observed number of vaccinations doesn’t match the survey data. Bubkas, it says, whereas my eyes and ears and the other data points say no, not bubkas.

Can we potentially reconcile these data points without saying the survey is wrong?

Not quite entirely, but if we can fix the number of vaccinations to match the data then I think we... mostly... can? The way we do that is to notice that there are essentially three categories of attitude here, and all that’s being tracked is the *transition between those categories*. Which means that you only get ‘noticed’ here if you move from one to the other.

Thus, if the pause issue was mostly a within-category move, it might not show up much here. If you were enthusiastic before, you’ll say the same now, despite likely being less

enthusiastic. If you were hesitant, this makes you *more* hesitant, but doesn't really transition that many people into 'resistant.' And if you're already resistant, then obviously nothing changes.

You'd still expect *some* flow between categories, but it could reasonably be quite small, or be countered by other effects, such as there being 'enthusiastic' people who were going to get J&J, couldn't get it, and thus stay enthusiastic, maybe? Feels like a reach.

The other explanation is that there's at least some disconnect between what people *say on a survey* and what they *do in practice*. I can imagine a world in which this disconnect is substantial. People don't want to *say* they're less enthusiastic, because the whole issue is clearly dumb and doesn't impact Pfizer/Moderna, or something similar, but it still leads to a bunch of FUD that causes them not to actually make the vaccination happen. Or they still want to do it, but figure they'll wait until this is sorted out, and then keep waiting.

The problem with that hypothesis is that it feels like it should also have shown up in the observations me and those around me made about individuals who are expressing new hesitancy? That's the problem I can't explain.

I wouldn't have predicted the survey results to not show any impact, and was surprised. There's at least some chance that this reflects something I didn't understand, but there's also a chance the survey was flawed, or the curve-fitting used was flawed, or both. I'm not sure.

The Worst Possible Thing You Could Do

I don't have much I want to add to [my case from last week](#) that normalizing the expropriation of intellectual property for exactly the kind of intellectual property we most want to exist in the future, and the resulting additional normalization of the expropriation of exactly the remaining profits that are creating incentive to do the most good, *might be a really bad idea* if you wanted you, your children and those around you to live a long and healthy life.

These are important questions to understand and get right, but they're also inherently political and economic questions. Continuing along these lines would be outside where I want these posts to focus.

In Other News

[People think that being virtuous reduces COVID risk.](#) When presented with identically risky scenarios, subjects judged risk to be higher when motives for action were bad.



Cailin O'Connor @cailinmeister · May 7

...

New Paper! We find that subjects think you are less at risk of COVID infection when engaged in morally good actions, and more likely to catch COVID while doing morally bad things. In other words, risk judgments are systematically skewed.



Cailin O'Connor @cailinmeister · May 7

...

We present subjects with vignettes where the exposure is always identical, but the reasons for the exposure vary. I.e., Joe always gets caught in an elevator with neighbors, but might be headed out to buy drugs, or to help an elderly friend. 2



Cailin O'Connor @cailinmeister · May 7

...

In two experiments we find risk is judged higher when moral valence of the action is judged lower. We got interested bc of infographics worried about fun outdoor activities like going to the beach or the pool, but not about drs/grocery stores. 3

texmed.org/TexasMedicineD...



Cailin O'Connor @cailinmeister · May 7

...

This follows previous work finding that moral judgment impacts risk judgment. Thomas et al. find that people think children are at greater risk of harm when their parents leave them alone intentionally (yoga) vs unintentionally (hit by a car). 4



Cailin O'Connor @cailinmeister · May 7

...

The effect in our paper is small, but congruent with previous work. Possible implications for public health messaging: 1) risk messaging should track real risk, not morality, and 2) risk messaging should (maybe) focus on morally good activities like going to church or protests.

I can steelman this difference a bit. Consider the vignette where Joe is headed out either to help an elderly neighbor or to buy drugs. These are not identical distributions of Joes. The Joes headed out to buy drugs likely live in poorer neighborhoods with higher Covid risk. The Joes helping elderly neighbors likely live in relatively richer and safer areas. Thus, given we see a small effect, that small effect could easily exist in reality.

You could extend that to arguments that those you see at grocery stores or doctor's offices are relatively low-risk, whereas the people at a beach are relatively high-risk, even if the beach itself isn't risky, because they're the people who are ignoring the (blatantly terrible, misleading, dishonest!) public health advice. What other rules are they breaking?

Here's the chart she linked to:

COVID-19

CORONAVIRUS DISEASE

BE INFORMED:

Know Your Risk During COVID-19

On a scale of 1 to 10, how risky is...

Ranked by physicians from the TMA COVID-19 Task Force and the TMA Committee on Infectious Diseases.

Please assume that participants in these activities are following currently recommended safety protocols when possible.



Physicians Caring for Texans

Texas Medical Association | 401 W. 15th St. | Austin, TX 78701-1680

www.texmed.org



LOW RISK

LOW-MODERATE

MODERATE RISK

MODERATE-HIGH

HIGH RISK

The idea that grocery shopping is in the low-moderate group with ‘eat outdoors at a restaurant,’ ‘going for a walk’ or ‘playing golf’ and a step safer than ‘going to a beach’ is obviously not reflective of real risk. The first time I went to a grocery store during the pandemic, it was clear that it was more risky than everything else I’d done for weeks, combined. Rating it as low-moderate risk is reflective of the fact that it’s *necessary* for most people, because they can’t afford not to do it, so authorities want them to think it’s low risk

so they will not start thinking 'well, I already do this at least moderately risky thing already every week' so what's the harm in a little something else? And they don't want people freaking out over buying groceries for their families because what are you gonna do if you can't afford instacart? Then many of those same people get home and wash all those groceries, some to this day.

Meanwhile, [even with the CDC and WHO finally acknowledging how Covid actually spreads](#), it's going to be tough to break old habits.



zeynep tufekci ✅ @zeynep · May 7

...

Incredible week. First the WHO, now the CDC. It'll take work to have all this be heard, and correctly. Just today, I saw Canada is planning to close beaches "to protect against variants." It takes more than a few website updates to fix a year of messaging.

Meanwhile to that, in this word of 'figuring things out and physically modeling the world' [we have to remind ourselves that these facts were not fully in evidence](#) (and there's a thread here about the history of how that happened for those interested):



Eliezer Yudkowsky ✅ @ESYudkowsky · May 9

...

Fascinating history. Meanwhile in my beautiful bubble I didn't even know the CDC and WHO were denying Covid-19 was airborne! Of course it was going to get you if you were indoors with poor circulation. My feed is a different tiny world, I guess.



Jose-Luis Jimenez @jlc Colorado · May 8

1/ TIME FOR SOME AIRBORNE + DROPLET HISTORY

Now that @WHO and @CDCgov have finally accepted *after a year of denial and delays* that airborne transmission is a major mode for COVID-19, it is time to review the history to try to understand why this response was so poor.

[Show this thread](#)

I was aware they weren't acknowledging Covid was airborne, but it somehow didn't feel *important* because who would be listening to those jokers and taking their explicit physical world models seriously? Either you believed in physical world models, in which case these people were obvious lying liars whose models didn't make sense and should get ignored, or you didn't, in which case all that mattered was some elite consensus on how to act that didn't care what the WHO's or CDC's physical model might be, and at most cared about their 'recommendations' which of course were a horrible mish-mash of nonsense even if their claimed physical models were accurate.

The test of whether my failure to notice this was correct is whether this change does anything. Congratulations, you've finally admitted that Covid spreads in exactly the way we've known for a year that it does, what'll you do next? Go to Disney World and hang out at the beach? Or will you change nothing? So far as I can tell, there hasn't exactly been a rush

to update the safety recommendations, let alone any sign anyone is actually changing behavior much.

Also important is that the regulations surrounding airborne diseases are different than those for non-airborne diseases, so these fact claims may have been more about what regulations the CDC wanted to offer than about what was physically happening.

[Or, basically, this, which solicited this response from the CDC head:](#)



Manu Raju @mkraju 22h
Susan Collins says she used to always consider the CDC "the gold standard."

"I don't anymore," she says at hearing
1k 338 2k

[At any given time, 2% of people infected are 90% of the infectiousness, says a new paper, because people are typically only super infectious for a short time.](#) I haven't looked in detail, but seems plausible. Doesn't mean you can know which 2% it is easily.

[Bryan Caplan compares what happened to his expectations.](#) His analysis for individuals (which he gives a 2nd percentile grade) is based on the assumption that paranoia about Covid makes no sense and that anyone (such as myself) who spent a year avoiding Covid was acting grade-A stupid. Part of this is that he doesn't think the risk is serious and thinks that this assessment should be obvious. Part of that is that he implicitly saying that people should ignore social effects of their actions - without the actions he's condemning, there would have been completely uncontrolled spread and either the government would have clamped down super hard or we'd have had a hospital collapse, and he doesn't even mention such things. And part of this is that he thinks everyone should accept that the key to and 'most important source of' happiness is in-person human interaction, which... well, sure, it's helpful when done well and I'm happy to get it back, but let's not go nuts. He also doesn't think about whether our prioritization of containment strategies, and our coping with the situation, were better or worse than expected, and relative to expectations under these conditions I think people mostly did well. His other assessments seem reasonable to me. We agree business did even better than we expected, and that governments in most places did even worse than expected, and I put individuals close to businesses. But then I'm one of the individuals he's giving a low grade to!

[Yes, full non-emergency authorization matters, do it now](#), New York edition:



Morgan Mckay @morganfmckay 45m

Cuomo now later clarifies that vaccine mandate for SUNY students cannot happen unless FDA fully approves vaccines. Right now just emergency use authorization.

"If it doesn't have the full approval, you cannot legally mandate... we believe they will do that in the near future"

Morgan Mckay @morganfmckay

NEW: All SUNY and CUNY students returning in the fall MUST be vaccinated per @NYGovCuomo

Show this thread

1 5 5 ...

[In addition to the schools and the army edition:](#)



PoliMath

@politicalmath

The Army can't even force their own soldiers to take the vaccine!

Why should we make it mandatory for kids to take an emergency auth vaccine for a disease that barely impacts them if all the adults around them are already vaccinated?

[Two unrelated matters:](#)



Matthew Yglesias @mattyglesias · May 7

...

It is currently illegal for vaccine manufacturers to do advertising and promotion of the vaccines.

In other news, public health experts are wondering how to increase vaccine uptake.

[Paper on Texas school reopenings and their impact on spread.](#) Conclusion was that opening schools made adults more mobile, thus accelerating spread. That suggests opening schools was right, since using schools to force people to stay home seems like a very non-optimal (and highly regressive and miserable) way to impose a de facto lockdown, with a possible exception if driving people to/from school was directly causing other interactions, but likely

not even then. It's theoretically possible this is a second-best (third-best?) solution anyway if there's no political way to impose alternative restrictions?

[College students sent home with full loss of tuition because they took a picture outside without masks.](#)

[Paper about Covid skeptic community finds that they highly value... actual scientific inquiry and analysis of data](#), and thinking for yourself, and are happy to help others think for themselves too. Then laments this 'refusal to accept the science as settled' and compares them to those who attacked the capital on January 6. You see, if you get the Wrong Answer, you're illegitimate, so any use of the proper legitimate techniques is doubly illegitimate and terrible. You might give people the idea that these are tools people use to analyze data themselves and come to conclusions about how to model the physical world. Can't have that.

What happened to the Novavax vaccine? [A combination of the Defense Production Act blocking raw materials and a general raw materials shortage](#). Which still doesn't explain why it hasn't applied for FDA approval yet.

[Analysis of the NHS Covid app and its effects estimates it prevented hundreds of thousands of infections in the UK.](#)

[Uber and Lyft are offering, free of charge and with no government payments involved, free trips to and from vaccination appointments](#). Kudos to them. So if you're claiming the logistics don't work, there are free taxis on standby and walk-ins available. The excuses are getting rather thin.

Not as good as free Mets tickets but [nothing wrong with a good old fashioned twenty:](#)



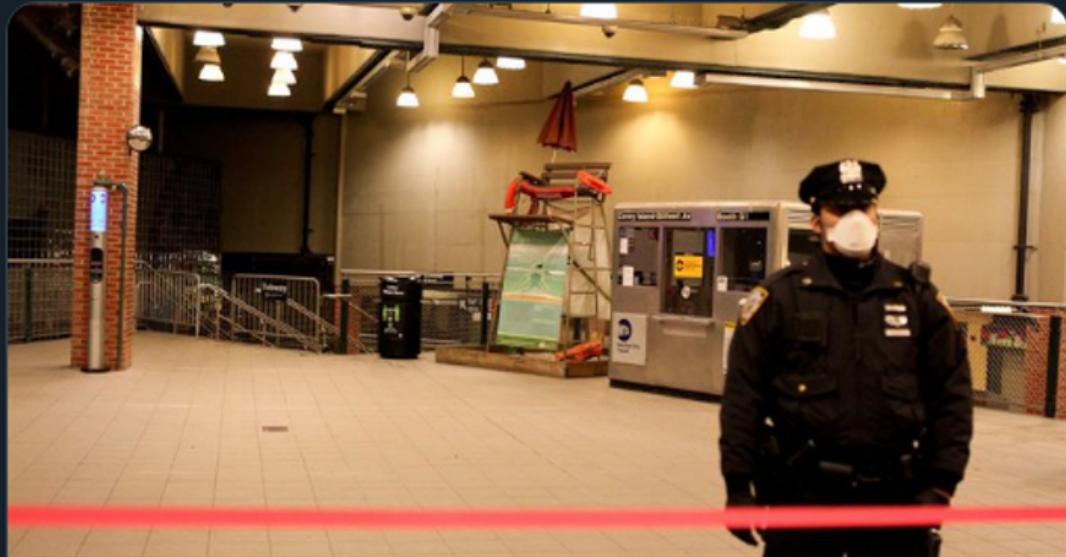
Céline Gounder, MD, ScM, FIDSA

@celinegounder

...

NYC to offer J&J COVID vaccinations in subway stations
May 12-16 on a walk-in basis:
gothamist.com/news/nyc-subwa...

Get your shot and a free 7-day unlimited MetroCard or
round-trip LIRR/MetroNorth ticket.



Some NYC Subway Stations Will Soon Offer Walk-Up Vaccine Appointments

Now's your chance to get jabbed inside Penn Station.

gothamist.com

I do worry about some people getting too smart for their own good and holding out for more.

[So how about One Million Dollars:](#)



Governor Mike DeWine

@GovMikeDeWine

...

Two weeks from tonight on May 26th, we will announce a winner of a separate drawing for adults who have received at least their first dose of the vaccine. This announcement will occur each Wednesday for five weeks, and the winner each Wednesday will receive one million dollars.

5:46 PM · May 12, 2021 · Twitter Web App

3,981 Retweets 10.7K Quote Tweets 17.7K Likes

This is absolutely correct policy and amazingly great and I'd love to see other states and ideally the federal government follow suit. All the incentives align.

[Vitalik gives epic crypto donations](#), including 1,000 ETH to MIRI, another 1,000 ETH (plus his ELON) to the Methuselah Foundation, more than 13k ETH to GiveWell, and *one billion dollars* (yes, billion with a B) in SHIBA at its pre-donation market cap to the Indian Covid Relief Fund. That's 10% of the entire SHIBA blockchain. [Looking at the replies to the relief fund's announcement of the donation and also the current graph](#) does not exactly inspire confidence that a billion dollars of fiat are going to result from this, and it does not bode well for other holders of this second-rate version of a meme coin whose very thesis is war upon the very concept of sensemaking and logic, and I am totally, totally fine with every last bit of all of that.

There were always bound to be casualties.

There's an obvious trade here as I write this. Don't do it.

Hopefully this encourages rather than discourages future creators of cryptocurrencies considering giving tons of surplus coins to Vitalik. Seems like that's a great system!

Not Directly Covid News

Not directly Covid, but relevant given we want to understand how the media chooses to cover things. You see, there's a gasoline problem, in the sense that if you go to the gas station to try and purchase gasoline, chances are high they won't have any and that's going to be a problem. You would *think* this would be a really big story, [and yet Robin's experience matches my own...](#)



Robin Hanson ✅

@robinhanson

If I search for "gasoline" I find news stories, but I don't see them anywhere near the top of my usual news feeds.

Katie Herzog ✅ @kittypurrzog

This is so bizarre. I've called 10 gas stations around Asheville and not a single one has gas. One person laughed at me when I asked. And yet nothing on current gas shortages on NPR and most other national news sites. How is this not the biggest story in the US right now???

Show this thread

The Washington Post website had actual zero stories on its main page about gasoline when I started writing this section, although it added a story later, and the next day [had one about the political blame game](#). Neither did the Wall Street Journal or Bloomberg. [CNN had one, and then it went directly into telling people everything is fine and not to 'panic' by buying the thing there's a shortage of but which is legally not allowed to rise in price](#) (and an implicit framing that this is something that *happened to* the administration, which I am sure would not have been the framing if this had happened last year):

Growing number of East Coast gas stations are without fuel

- Why Americans are panic buying fuel -- and why you shouldn't
- Biden officials privately frustrated with pipeline's weak security ahead of cyberattack

Check out these headlines for those posts:

Latest on the US gas demand spikes

Gasoline demand spikes in several states after pipeline hack

What in holy hell? This is not a demand spike! This is a supply shortage! There's no supply, and people are buying the not-price-adjusting supply that remains because it's going to run out, *and they're framing the issue as a supply shortage*. Why is everyone withdrawing their money from this bank after someone robbed them of half their money? You people are yelling at us for no reason, you all need to calm down.

Fox news mentions it, but the biggest mention is *as a metaphor*:

Ingraham: Biden is out of gas, is now problem denier

Then later [they have a literal story about it](#). Those are *all* the news sites I checked, there's no selection going on here.

The obvious harmless explanation for all this is that perhaps the pipeline disruption was minor – the claim is that supplies will mostly be flowing again by the end of the week – and it didn't cause much supply disruption, and all we're dealing with is a few areas where people started hoarding gas and things will be annoying for about a week. So in some sense (the sense in which you have zero respect for the people buying gas responding to very clear incentives) this *sort of* is a demand-side problem and not worthy of much coverage. But also covering it more would make it happen more so it's easy to understand the desire to downplay it a bit. I've similarly downplayed or ignored stories about Covid for similar

reasons, for example when someone with a large audience started talking about how they didn't 'need' the vaccine and encouraging people not to get it, I asked myself 'how would talking about this help anything?' and didn't mention the Person Being Wrong On The Internet.

Still, we have photos of the New York Times saying there are no gas stations without gas at the same time as there were definitely a bunch of stations without gas, and a general (implicit) conspiracy not to talk about what really should be a pretty big issue. Which should update you in favor of the media doing similar things in other situations, in the past and in the future.

[Overall, I'm going to have to go with 'unless people are flat out flying about physical conditions something that matters a lot is happening here':](#)



Patrick De Haan 📈📊 ✅

@GasBuddyGuy

BREAKING: Over 20% of metro Atlanta gas stations are without gasoline.

4:28pm · 11 May 2021 · TweetDeck

409 Replies **1,748** Retweets **4,176** Likes



•••

Reply to @GasBuddyGuy



Patrick De Haan 📈📊 ✅

@Gas... 14h

UPDATE: 30% are out

15 15 43 237 •••



Patrick De Haan 📈📊 ✅

@Gas... 12h

UPDATE: 40% are out

4 15 24 143 •••



Patrick De Haan 📈📊 ✅

@Gas... 11h

UPDATE: Nearly 50% are out

21 15 94 258 •••

[That account seems great if you want to follow what's happening:](#)



Patrick De Haan 📈📊🔍 @GasBuddyGuy · 58m

North Carolina update:

78% of Greenville/Spartanburg/Asheville/Anderson stns no gasoline

72% of Raleigh/Durham stns no gasoline

71% of Charlotte stns no gasoline

69% of Greenville/New Bern/Washington stns no gasoline

65% of Norfolk/Portsmouth/Newport stns no gasoline



Patrick De Haan 📈📊🔍 @GasBuddyGuy · 10m

...

Maryland Gas Update:

Washington DC 8% stns no gasoline

Baltimore 7% stns no gasoline

Politicians going to do something now?



1



14



23



Patrick De Haan 📈📊🔍 @GasBuddyGuy · 19m

...

NEWS RELEASE: Gas Prices Hit \$3 Per Gallon Average, First Time Since 2014

That was not the peak of the trouble, that was Wednesday morning. [It will take several days](#), by all reports, to fully restore service.

So *of course* the governor of Georgia responded to this supply shortage by... suspending the state's gas tax.

Meanwhile, in 'Zvi is tempted to offer a simple policy solution' discourse...



I also wonder what would be happening if we still had the previous administration, and it seemed like there was a problem and the public response was literally nothing. Presumably questions would have gotten asked. So that suggests that we should be far more worried this year than last year about problems getting buried.

This whole episode was of great practical interest to me, because the plan was for my in-laws to be driving with my children down from New York to Florida for a vacation, directly through the area that does not have reliable access to gasoline. This was very important news for us to know, and knowing it allowed us to book flights instead. That's exactly what you *want* to happen, with people reallocating consumption away from scarce resources to use other resources, but the media decided to do its best to bury the information.

I do understand that [other people react in other ways that one might reasonably want to minimize](#). I sympathize. It's not that there isn't a positive motivation here...



Patrick De Haan 📈📊🌐 @GasBuddyGuy · 33m

...

People- in the calmest voice possible- there is no risk of gasoline shortage outside of the Colonial Pipeline area. WHY WOULD YOU CREATE ONE?

29

116

272

↑

There is also the issue of how this whole episode was allowed to happen. Our pipelines, it seems, can be shut down by cyber attacks for substantial periods. It could have been a *lot* worse, as there wasn't an environmental catastrophe and things will be back online soon, but we really, really should be treating this as a sign that we need to take such risks a lot more seriously. As opposed to, say, our warning signs about a possible coronavirus pandemic.

Not directly Covid, but if you want to know how disconnected from physical reality politicians can get, [this is what happened when candidates for New York City mayor were asked about home prices in Brooklyn](#). Then Yang got several questions like this *exactly right* while using strong physical-world explicit reasoning to cover up what was presumably frantic Googling of the answer or previous Googling because the questions had already been leaked. But compared to someone who spent years in charge of 'affordable housing' and wants to be mayor being off by a factor of 10 on what housing costs, give me the person who knows what actual physical-world Fermi estimation looks like and uses it to cover up looking up the right answer any day. The best part is when one of the candidates *doubles down* and says 'including apartments?' when told the right answer. Alas, I didn't get a chance to guess unanchored, and different parts of Brooklyn are sufficiently different that the question is tricky. Just not 'missing a zero' tricky.

Not directly Covid, but relevant given how much we wonder about the origins of what looks like incompetence that California is [going full Harrison Bergeron](#) and [sabotaging advanced math education because knowing too much math leads to inequality](#). In other news, [popular claims that 'urgency' or 'objectivity' are white supremacist concepts](#) might also be what we call 'not helping.'

Not directly Covid, but highly relevant to note that [a lot of people are optimizing purely for 'normal' and will attack anything they see as different from 'normal.'](#) even if their 'normal' is in the minority. This is distinct from the mode where there is a righteous agenda of some kind, and anyone not onboard with that agenda is history's greatest monster, except that one of the strategies of such agendas is to code acceptance or fear of them as 'normal' to trigger the first group.

Not directly Covid, but highly relevant is [perhaps the greatest Twitter reply of all time](#).



Michelle Kosinski @MichLKosinski · May 8

As an American journalist, you never expect:

1. Your own govt to lie to you, repeatedly
2. Your own govt to hide information the public has a right to know
3. Your own govt to spy on your communications

Trump's unAmerican regime did all of these.

No one should accept this.

12K

8.1K

5.5K



leon @leyawn · May 8

as an american journalist you never expect:

1. mom's face to disappear when playing peekaboo
2. where did it go
3. wait now she's back again

210

3.8K

35.9K



To which there can only be one reply, which is perfectly understandable:

**@MichLKosinski blocked
you**

You are blocked from following
@MichLKosinski and viewing
@MichLKosinski's Tweets.

It's for the best.

Covid 5/6: Vaccine Patent Suspension

The Biden administration's latest strategy for the pandemic is to suspend the vaccine patents without compensation. Our life expectancies are lower than they were last week.

It's a shame. I like the idea of rewarding those who do amazing things for myself and for the world. I like people out there knowing that if they produce amazing things for myself and for the world, they would get rewarded for them. I like the idea of not dying for as long as possible thanks to future developments in medical science. I like being a nation of laws, where the executive doesn't just take stuff when he feels like it. And I'd like, when nice things are taken away and we mortgage our future, to at least get *something* out of the exchange.

Alas, the man in charge does not agree, and the government was not content with its previous efforts to sabotage the vaccination effort. That's how it goes sometimes. You can't always get what you want. Nor, when no one is given the incentive to produce what you need, are you likely to get that either.

Let's run the numbers.

The Numbers

Predictions

Prediction from last week: Positivity rate of 3.9% (down 0.5%) and deaths decline by 6%.

Result:

In the past week in the U.S. ...

New daily reported **cases fell 11.1% ↓**

New daily reported **deaths rose 2.3% ↑**

Covid-related **hospitalizations fell 7.9% ↓** [Read more](#)

Among reported tests, **the positivity rate was 3.9%**.

The **number of tests reported fell 20.8% ↓** from the previous week. [Read more](#)

Nailed the positivity rate. [Johns Hopkins has us down from 3.9% to an all-time low of 3.6%](#). Deaths rising makes no physical sense and the move up doesn't show up in the Wikipedia data, so this has to be a data fluctuation one way or another. I'm going to guess that it will revert.

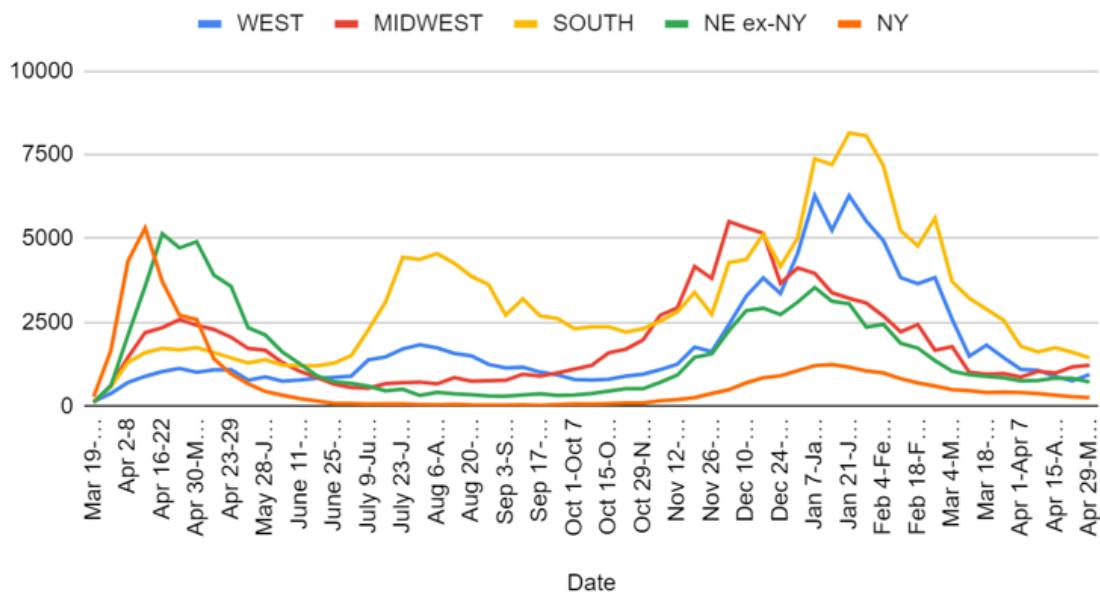
Prediction for next week: Positivity rate of 3.5% (down 0.4%) and deaths decline by 7%.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
------	------	---------	-------	-----------	-------

Mar 25-Mar 31	1445	976	2564	1262	6247
Apr 1-Apr 7	1098	867	1789	1160	4914
Apr 8-Apr 14	1070	1037	1621	1145	4873
Apr 15-Apr 21	883	987	1747	1168	4785
Apr 22-Apr 28	752	1173	1609	1110	4644
Apr 29-May 5	943	1220	1440	971	4574

Deaths by Region

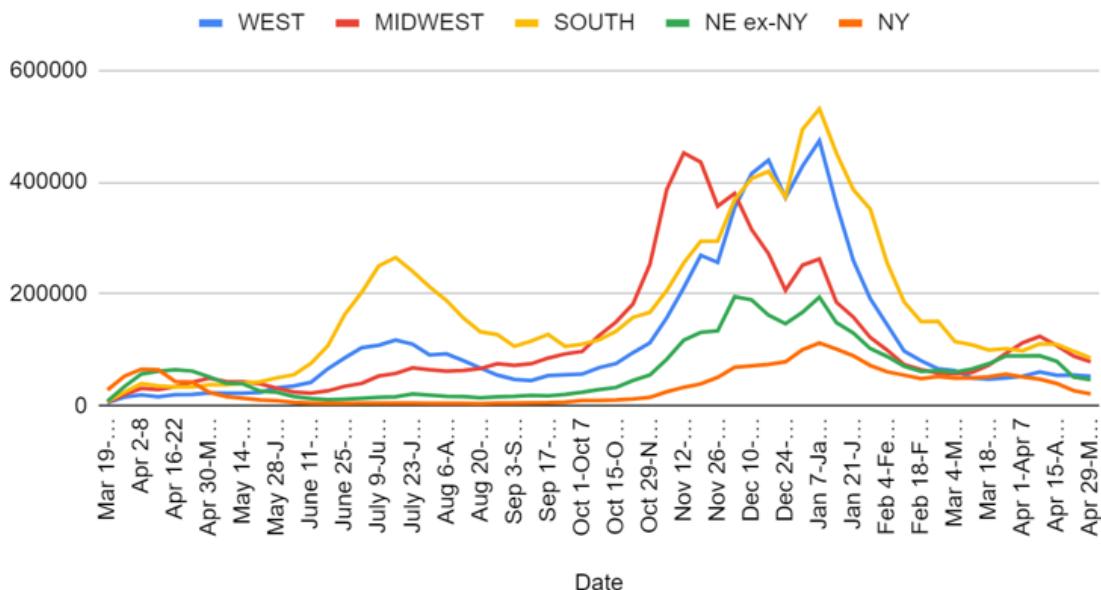


The bump up in the West comes from California, which makes it harder to dig in deeply. The bump in the Midwest is more curious, but should reverse soon. Overall we see a disappointingly small decline, but still a decline, and it should pick up speed.

Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Mar 18-Mar 24	47,921	72,810	99,568	127,421
Mar 25-Mar 31	49,669	93,690	102,134	145,933
Apr 1-Apr 7	52,891	112,848	98,390	140,739
Apr 8-Apr 14	60,693	124,161	110,995	137,213
Apr 15-Apr 21	54,778	107,700	110,160	119,542
Apr 22-Apr 28	54,887	88,973	97,482	78,442
Apr 29-May 5	52,984	78,778	85,641	68,299

Positive Tests by Region



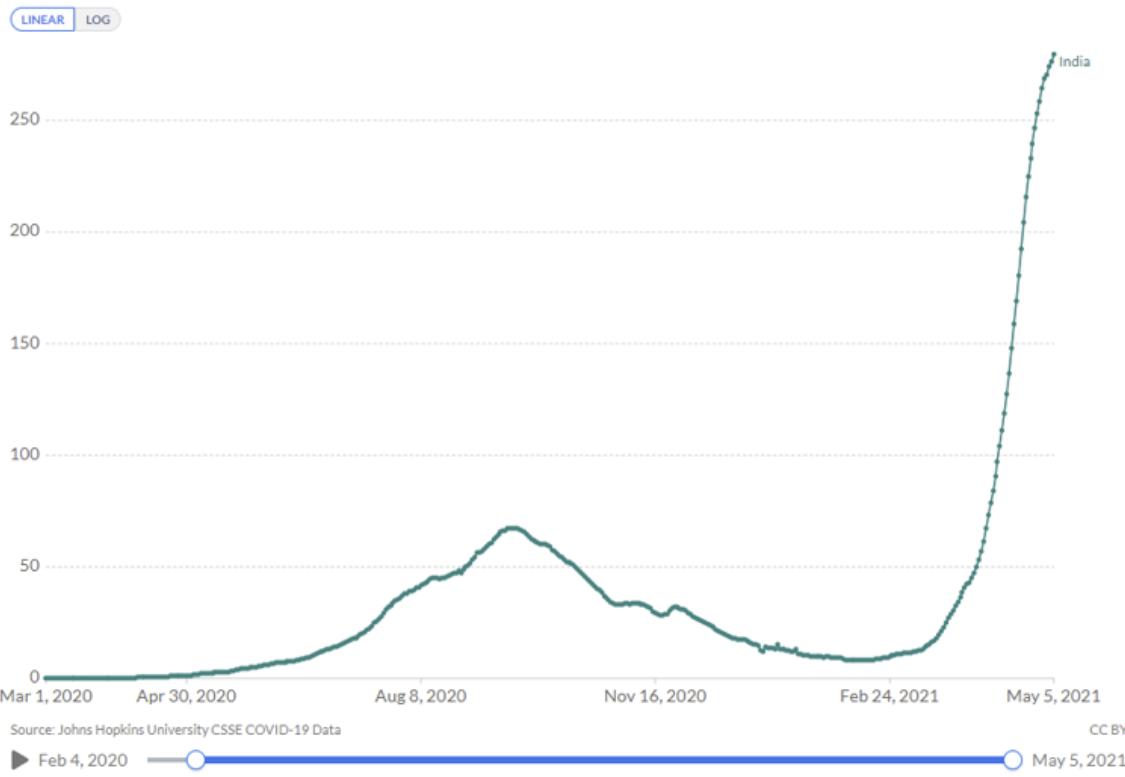
Progress in the West remains slow, but improvement in all regions, with many states seeing large declines. We didn't sustain the giant improvement rate in the Northeast but we still see pretty great improvement. This is what the endgame looks like.

India

Daily new confirmed COVID-19 cases per million people

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

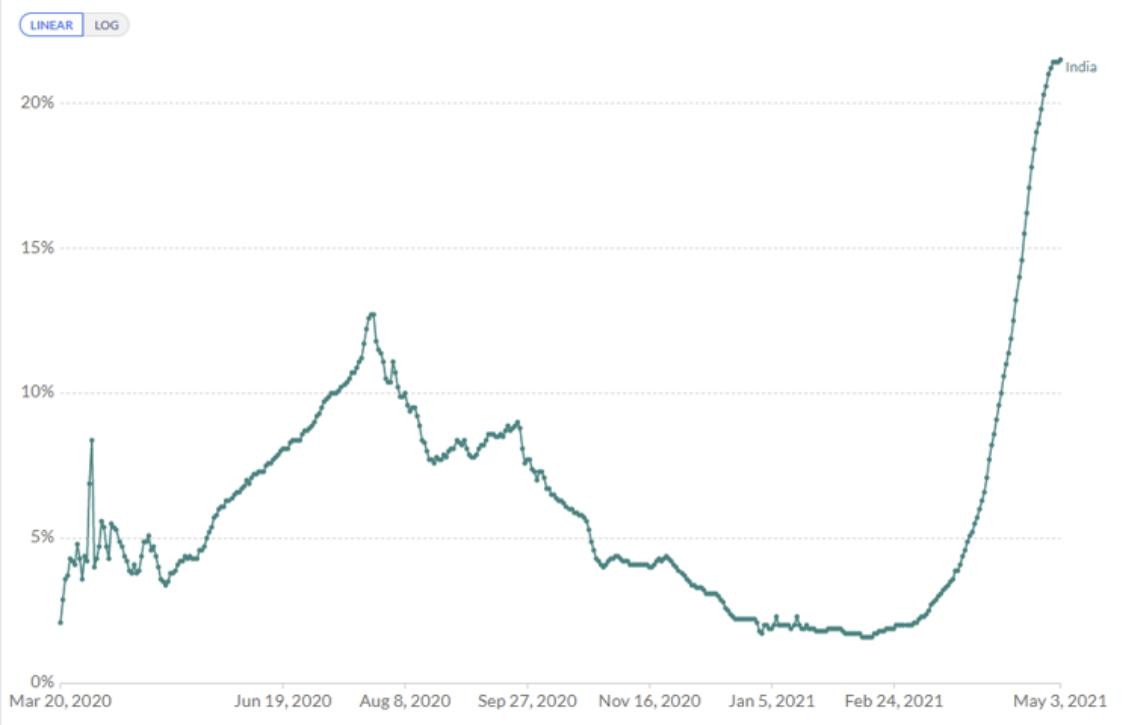
Our World
in Data



The share of daily COVID-19 tests that are positive

Shown is the rolling 7-day average. The number of confirmed cases divided by the number of tests, expressed as a percentage. Tests may refer to the number of tests performed or the number of people tested – depending on which is reported by the particular country.

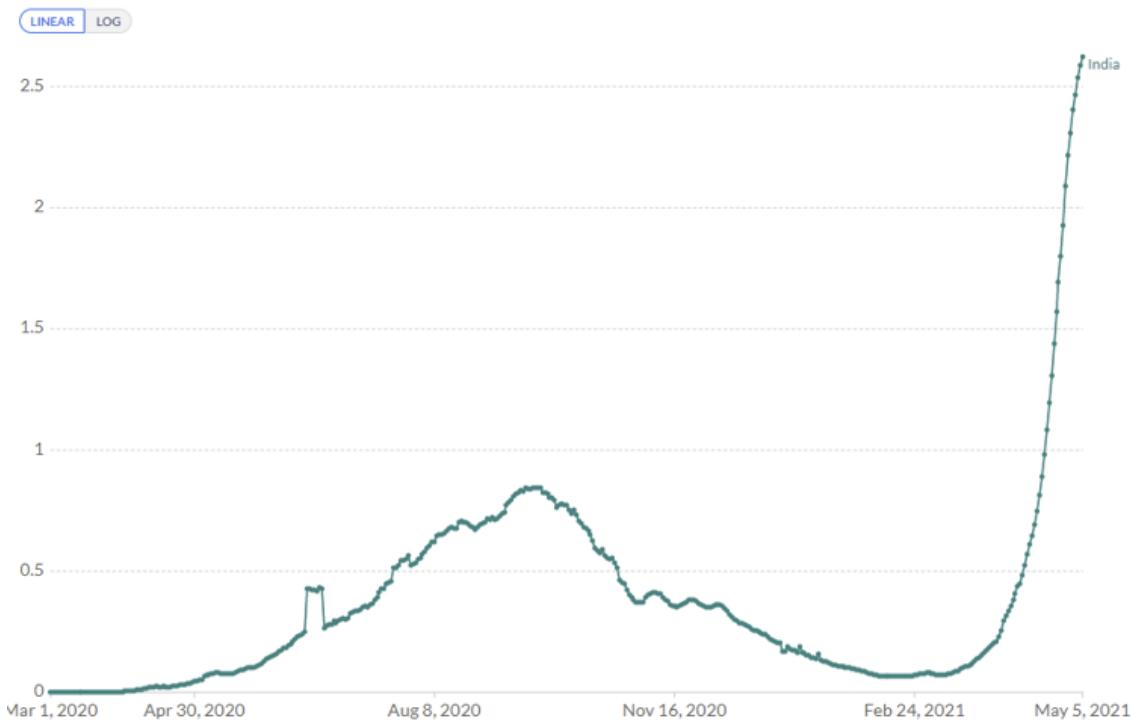
Our World
in Data



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



Things continue to get worse in India, but the graph no longer looks as fully vertical as it did previously, so this continues to count as good news relative to the range of possible outcomes. If things peak not too long from now, it will still be the biggest disaster of the pandemic, but it won't be anywhere near as bad as things could have gotten.

Vaccinations

We all know how it started.

[How's it going? Keeping up the momentum?](#)

In the U.S., **250 million doses** have been given so far. In the last week, an average of **2.13 million doses per day** were administered.

148.6 million vaccinated

The number of people who have received at least one dose of the vaccine, covering **55.6% of the eligible population, 16 and older** and **44.7% of the total population**.

As a reminder, we were once over 3 million doses, and we're giving out more second doses now than we were then.



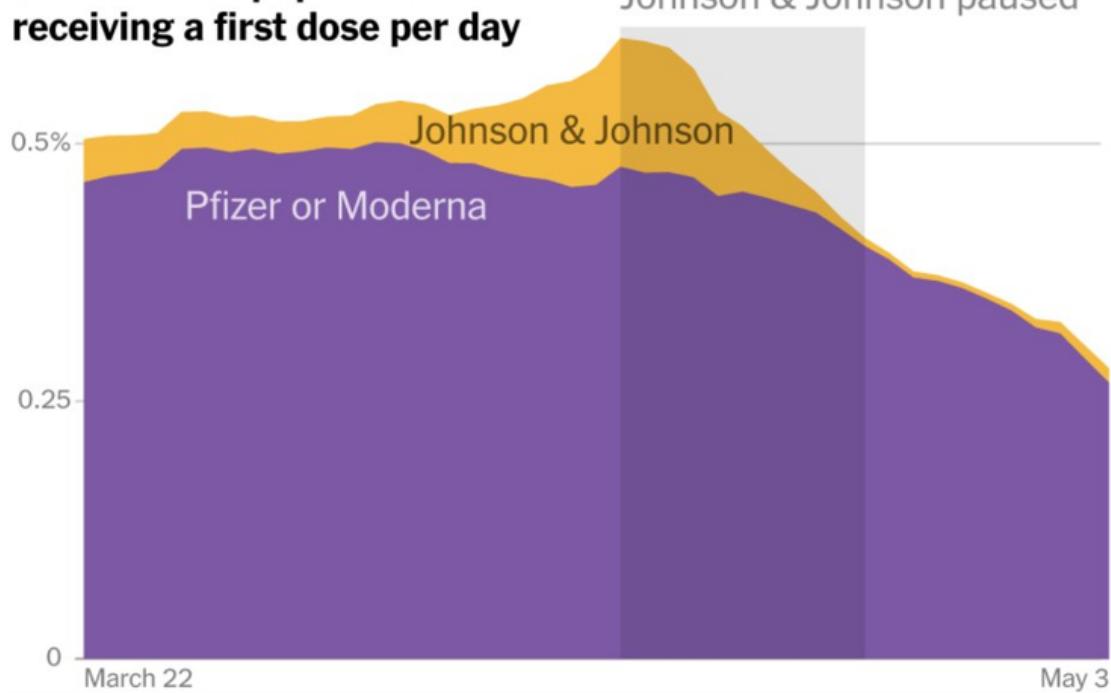
Matthew W. Mosca

@mwmosca

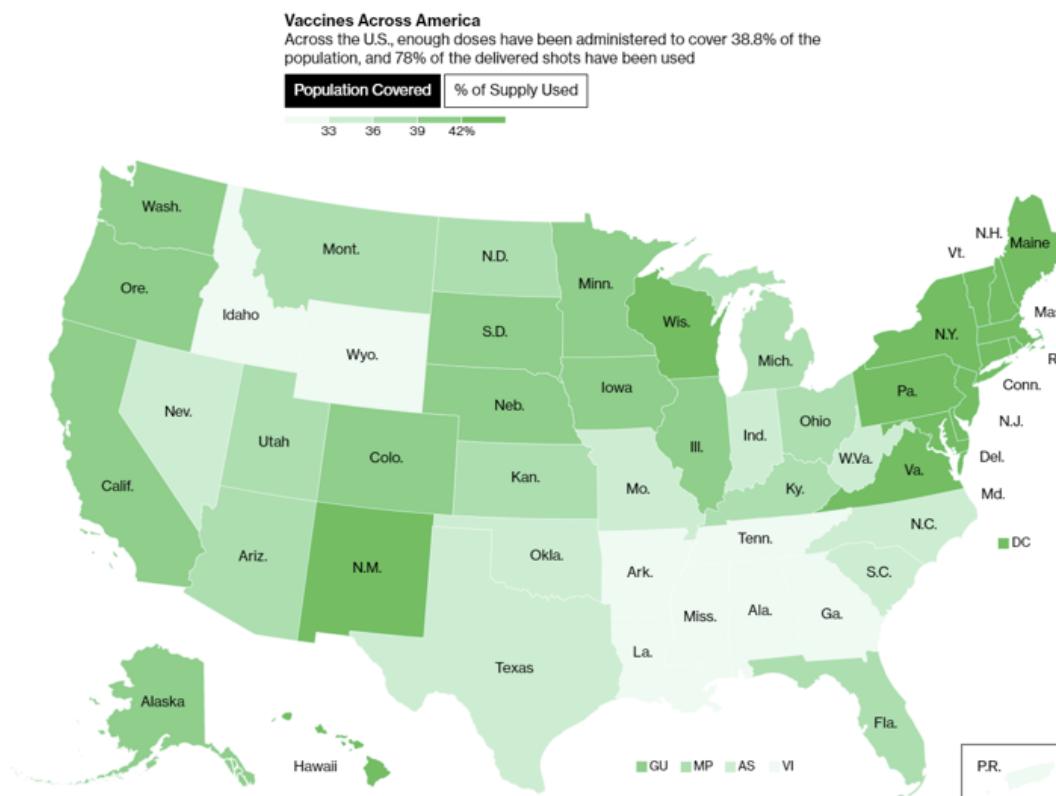
Walking through UW near library, saw three people at a table offering Moderna or J&J to anyone walking by. No takers when I was there (not a busy day on campus and many already vaccinated), but I admire their initiative. This outreach is only way to keep up momentum.

5:11pm · 4 May 2021 · Twitter Web App

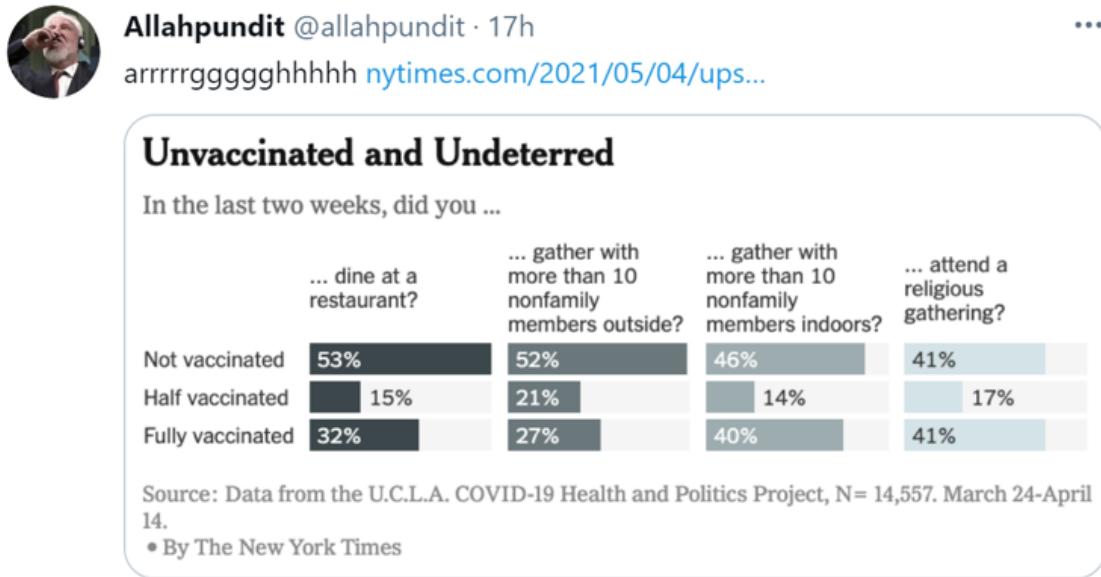
Share of U.S. population receiving a first dose per day



Every week, the graph of vaccinations looks more like the electoral college maps:



I found this chart enlightening when I first saw it:



The problem is that this isn't what I thought it was. I thought it was what percent of each group did each thing. Instead, it was what percent of everyone who did the thing was in each category. That forces us to consider base rates, which makes the whole thing complicated.

The good news is that even now only 55% or so of all adults are vaccinated, which means that people who are vaccinated are indeed doing more things. Yay!

Indian Strain Does Not Escape from Vaccines

The situation in India is terrible, but at least there is this bit of good news – [the vaccines will continue to function, at least against the current strain](#):



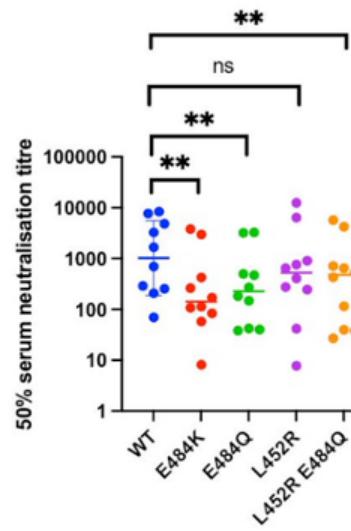
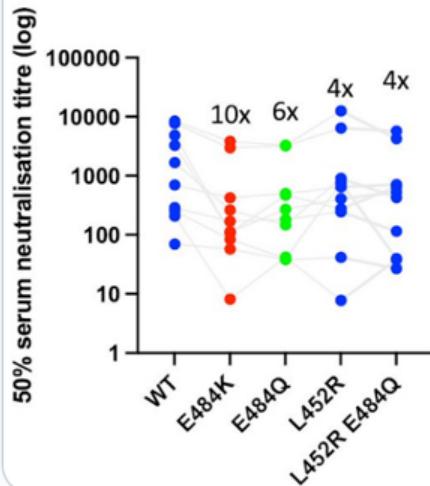
A reassuring thread: preliminary data showing that mutations giving the Indian variant its slightly nauseating nickname 'the double mutant' are not likely to result in significant antibody evasion. In other words, the vaccines should hold (here with Pfizer, but AZ likely similar)



Gupta Lab, Cambridge @GuptaR_lab · May 1

E484K seen in B.1.351 and P.1 (ex SA and Brazil), gives around a ten fold loss of neutralisation. B.1.617 has E484Q which has not been studied. E484Q has a smaller, 6 fold loss. The second mutation, L452 has a 4 fold effect(not statistically different to Wuhan-1 D614G here).

[Show this thread](#)



14

156

431

↑



Andrew L. Croxford @andrew_croxford · May 1

...

So I'm still here.... and unless we start seeing real numbers of serious infections in fully vaccinated individuals, I'm staying there, with arms folded and grumpy facial expression.



Andrew L. Croxford @andrew_croxford · Jan 18

I've had a lot of time to think about these variants and honestly, until we start seeing confirmed infections in fully vaccinated people I'm going to remain relaxed about this.

[Show this thread](#)



Gupta Lab, Cambridge @GuptaR_lab · May 1

...

Replies to @GuptaR_lab

Here is the REALLY IMPORTANT part. The combination of the two mutations gives a value of 4, in other words the two mutations DO NOT confer substantial antibody evasion and we can stop using the term 'Double Mutant'.

25

698

1.9K



Gupta Lab, Cambridge @GuptaR_lab · May 1

...

This is reassuring and gives us good reason to believe that expanding vaccination in India will contribute to control of transmission as well as the severe effects of COVID-19. The data also help to explain why one significant sub lineage of B.1.617 appears to have lost E484Q.

12

327

1.1K



Gupta Lab, Cambridge @GuptaR_lab · May 1

...

Please note that the B.1.617 may still be more transmissible as our data along with others suggest L452R increases the ability of the spike to gain entry into cells. We will preprint with greater numbers soon [@isabella_atmf](#) [@CambridgeBRC](#) [@rpdatir](#)

24

160

684



Mutations *not being additive* seems like very reassuring news, implying that there could be a maximum amount of infectiousness or vaccine escape that a Covid-19-type thing is capable of easily achieving. I don't see why we would stop using the term double mutant, but it makes it a lot less scary.

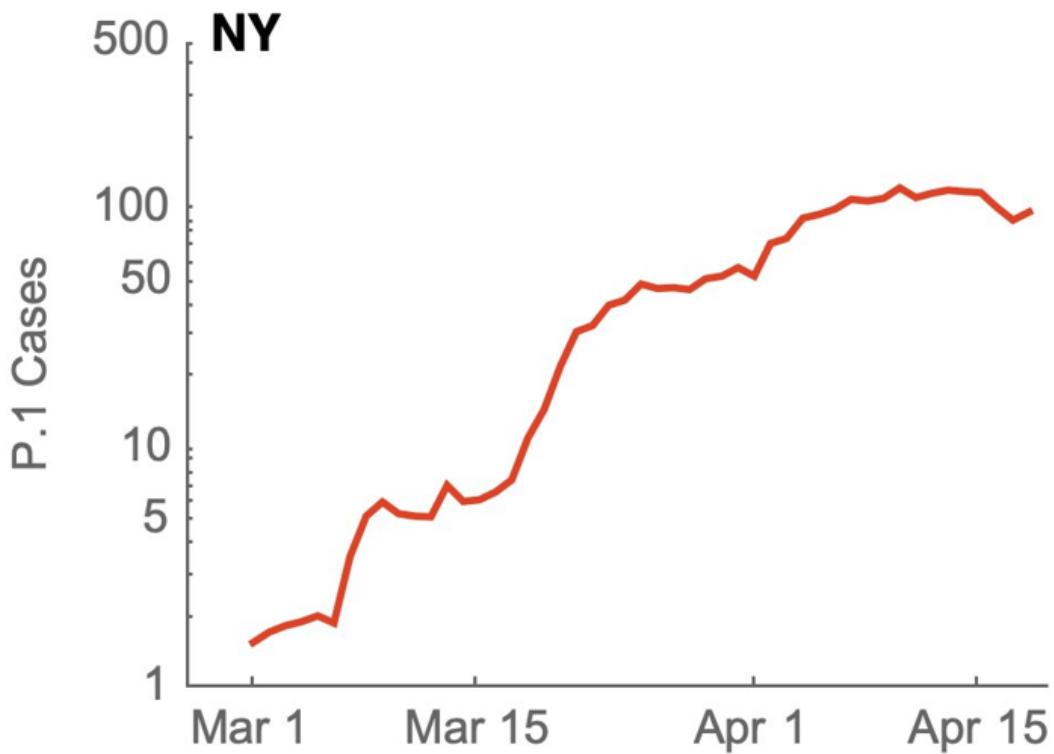
If there's one place I'm most worried about engaging in motivated reasoning, it's the possibility of vaccine escape. I notice a much larger flinch away from looking here than I do elsewhere. I think I've overcome that flinch, but I could be wrong about that, and it's a super important thing to not make an effort to avoid seeing. So while I'm confident, I want to task my readers with keeping me honest on this one even more than usual.

P.1 Is The Medium-Term Infection

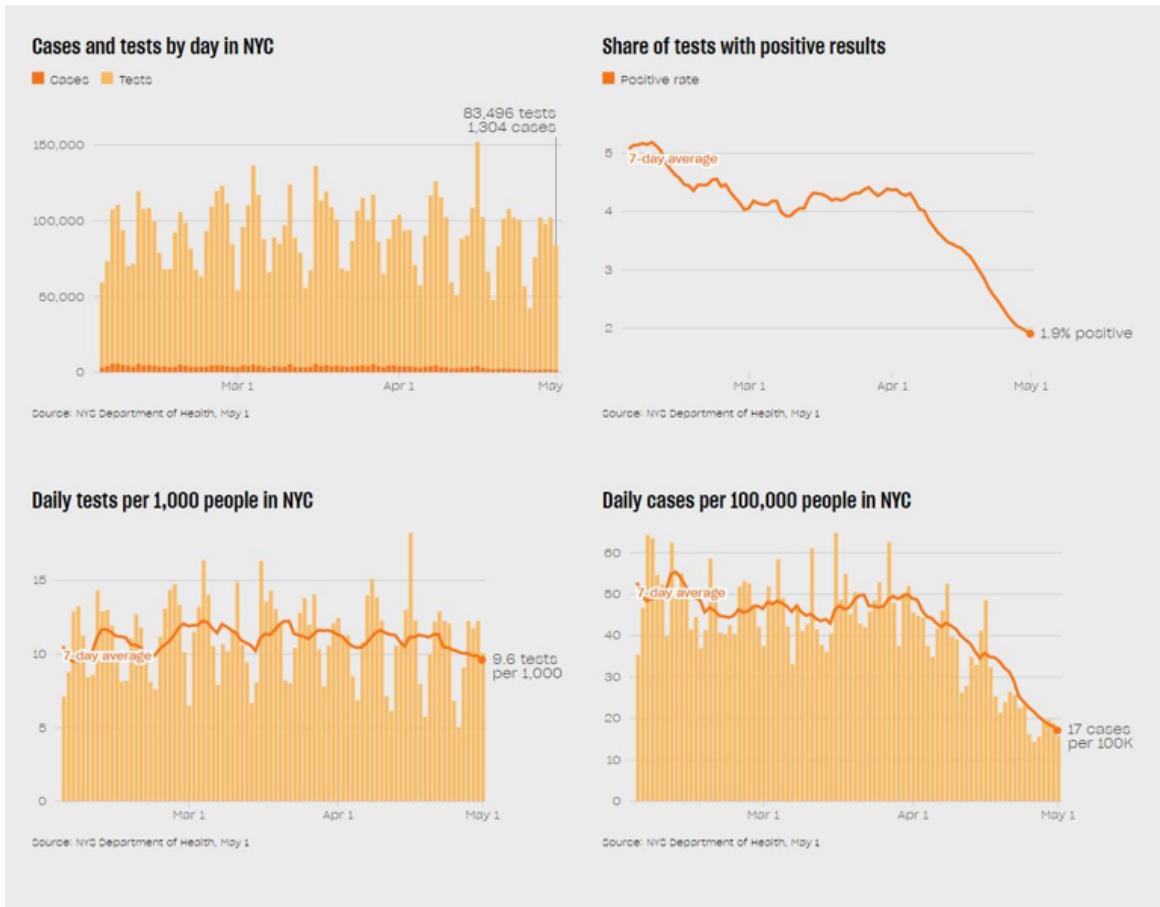
In many ways it is better to think of Covid-19 as a series of different infections from different variants. When the English strain shows up, it starts again from patient zero, starts again in each nation and region, and grows. When P.1 shows up and shows it is a more fit strain yet, it does this once again.

If you're looking at the endgame scenario, the question is whether we're seeing an increase or decrease in the most dangerous variant's numbers *in absolute terms* rather than relative to the overall number of cases. Thus, in a place like New York, the 'real' graph of our future situation is the graph in P.1.

This is delayed due to how long it takes to do sequencing, but [it looks like this](#):



Compare that to the graph of New York City's cases, [which looks like this](#):



Things had stabilized for P.1 by early April, when regular cases started cratering. Now, with regular cases declining even more rapidly in percentage terms, things are clearly improving even on the P.1 front, at least somewhat. We've passed the next test here, not only the previous one. As additional vaccinations come fully online, things will only improve, and I expect other areas to also hit this target.

The last month has been far more impressive than it has looked on its surface. We went from mostly the old strain to mostly new strains, and we are still steadily improving overall. The news really is quite good.

I worried last week that in relatively hesitant areas, we might run out of willing arms before we get to herd immunity. That is still a real worry, but I am not worried that large other areas won't get to New York's current effective immunity level given how many vaccinated people aren't yet finished being vaccinated. That doesn't allow a safe return to normal, but it does allow suppression when combined with moderate levels of precaution from the unvaccinated. My trip to New York this week revealed a city still taking its precautions deeply seriously, despite the majority of people being post-vaccination. I was clearly taking *below average* amounts of precaution, which was a new experience.

Exploring Vaccine Hesitancy

As a reminder, and to avoid any possible misunderstandings, as I keep saying week after week, the vaccines are very safe and super effective.

If you're reading this, you almost certainly know this. If you're reading this somewhere you can get vaccinated, and you haven't done so yet, *stop reading now, go get your first shot.*

We'll wait.

Not everyone, unfortunately, is in your epistemic position. Thus, we have vaccine hesitancy.

What are the real reasons for vaccine hesitancy? There are lots of theories out there, and I'm confident *someone* cares about any given justification one could come up with, but what are the most common *true objections*?

There's a lot of plausible candidates for the most common true objection.

[A survey about vaccine hesitancy in the army](#) has some good data on this, and is worth looking at in detail. I wish the data was better and came with numbers attached, but it's still good to have a look at the slide of the [Top 12 reasons soldiers are refusing vaccinations](#) (it's pasted here, but it's a lot easier to read at the link.)

Top 12 Reasons Fort Carson Soldiers Opt-Out of the COVID Vaccine	
Reason	Response
"It's not FDA approved"	The majority of supplements and energy drinks are not FDA approved. The COVID-19 vaccines have undergone a rigorous, multi-stage testing process, including large (phase III) trials that involve tens of thousands of people prior to getting EUA approval.
"It hasn't been proven safe"	The clinical trials for COVID-19 vaccines were five times larger than normal drug testing. Normal clinical trials have a few hundred to few thousand participants. The Moderna vaccine clinical trials had over 30,000 participants.
"What's the point? – I still need to wear a mask"	When 70% of Active Duty Soldiers on Fort Carson are vaccinated, the outdoor mask wear will be lifted.
"This is the first time I get to tell the Army, NO!"	Our Army exists to protect the American people. In this case, the enemy is a virus that is wreaking havoc on our way of life. Volunteering to receive the vaccine is your choice, but you have an opportunity to take action and help our nation return to normal.
"I am not in the high-risk population"	It's not about you. Defeating this virus is a community effort. Consider how you might protect people at increased risk for severe illness from COVID-19, such as healthcare providers, older or elderly adults, and people with other medical conditions, as well as children and other people who cannot get vaccinated.
"I already had COVID-19"	Early evidence suggests natural immunity from COVID-19 may not last very long. The vaccine will help you develop a more lasting immune response to the virus and protect against most variants.
"The vaccine symptoms are worse than the virus"	Greater than 30,000 vaccines have been administered on Fort Carson and only one allergic reaction has been recorded. The vast majority of patients have mild symptoms to include fatigue and muscle soreness. Side effects of Moderna COVID-19 vaccine typically resolve within 24-72 hours.
"The virus has the same mortality rate as the Flu"	The 2017-2018 flu season was the worse in the last 10-years and resulted in 61,000 US deaths. COVID-19 has killed more than 569,000 Americans. More Americans have died as a result of COVID-19 than died in WWII, the Korean War, and Vietnam War, combined.
"I don't want to get my family sick"	The Moderna COVID-19 vaccine does not contain SARS-CoV-2 and IS NOT a live vaccine. You cannot give your family COVID-19 from the vaccine.
"I am being safe. It has kept me healthy so far"	Social distancing and mask wearing are effective in reducing the transmission of COVID-19. However, precautionary measures do not directly combat the virus. The vaccine builds an immune response and protects your health, the health of our families, and the health of our community.
"The vaccine may impact my pregnancy"	The New England Journal of Medicine released a report that studied more than 35,000 pregnant participants. The report concluded that "mRNA vaccines did not show obvious safety signals among pregnant persons."
"I just feel skeptical and don't know what to believe"	The choice to get vaccinated is a personal decision and should not be taken lightly. Talk to a medical professional, consult the FDA Factsheet, and review educational materials available on https://www.carson.army.mil and from the CDC to weigh risks and benefits.

Or in written list form:

1. It's not FDA approved.
2. It hasn't been proven safe.
3. What's the point? I'd still need to wear a mask.
4. This is the first time I get to tell the army NO!
5. I am not in a high-risk population.
6. I already had Covid-19.
7. The vaccine symptoms are worse than the virus.
8. The virus has the same morbidity rate then the flu.
9. I don't want to get my family sick.
10. I am being safe. It has kept me healthy so far.
11. The vaccine may impact my pregnancy.

12. I just feel skeptical and don't know what to believe.

It's also worth taking in the perspective of the writer of the article and of the writer of the slide. Both writers take it as common knowledge that the reasons to not take the virus are stupid and wrong, and that the job is to fix what's wrong with these soldiers who are refusing.

There's no acknowledgement that maybe we've messed up in how we handled this whole thing, or that some of the concerns might be reasonable, or that maybe we treat our enlisted soldiers like garbage or worse and they might *really, really* want to tell the army where to go. It's a volunteer army, but the recruiter can lie to you, and once you sign the contract you definitely can't quit.

Consider this whole thing, as I will do from here, *from the perspective of the hesitant soldier*.

There are a few categories of objections here.

The first category (1,2,9 and 11) are the straightforward safety concerns. These concerns are *wrong*, but I say that as someone who knows they are wrong. And the responses suggested here other than to #9 are... [not great](#).

The FDA didn't approve your energy drink? How is that relevant or in the appropriate reference class? If the vaccines have undergone such a rigorous process as you say, *then why hasn't the FDA approved them?*

The clinical trials were three times as large as normal? How about the *one hundred million Americans who got fully vaccinated*? Maybe mention that? And again, what's your answer to the obvious: *If it's so damn safe why hasn't the FDA fully approved it?*

There aren't any *obvious* problems with pregnancy? Gee, mister, that makes me feel way better. No idea why we're voluntarily going with this weaksauce over much stronger alternative arguments. If I'm listening for bullshit, guess what I'm thinking right now?

In related news, [Stat News argues that the emergency use status of the vaccines shouldn't interfere with vaccine mandates by employers and schools](#). As a matter of law I think they're probably right (although of course I Am Not a Lawyer and all that) but as a matter of practicality this is a strong argument that it's important that the FDA needs to issue a full approval. We've just had the biggest Phase 4 in history. Taking at least Pfizer and Moderna from 'emergency' use to full approval would do a lot to reduce hesitancy and free the hands of those who want to mandate vaccinations, without being coercive.

If you want to solve this issue, *the FDA should simply approve the vaccines, full stop, not simply emergency use*. Problem solved.

If not, the response to a soldier should be that the FDA are a bunch of ass-covering assholes who would prefer never to actually approve anything, and *maybe* that would get through to them in a language they can understand.

The second category (7, 8 and 10) are claims that Covid-19 isn't that big a deal compared to the cost of getting the vaccine.

Here we see that response #10 says *both "masks and social distancing work"* and then goes straight to *"but they don't 'directly combat' the virus"* implying they don't count. When you're lying about everything, it's hard to keep your lies consistent, so I guess I'm somewhat sympathetic to this local predicament, but man it's glaring.

The answer to #7 isn't going to convince actual anyone. The 'mild symptoms lasting 24-72 hours' are exactly what the soldiers are complaining about, and the response is to tell them they're imagining things, which they most definitely aren't. Smooth.

For #8 they quote some statistics and it seems fine, I guess, although it leaves some ammo on the table. It's kind of bending over backwards to be maximally generous to the flu's deadliness. I'd have gone with different wording, but mostly this one is fine.

It's interesting when they strengthen the answer to the point of deception, and when they weaken the response to the point where it doesn't respond to the concern.

The third category (3, 5, 6) are claims that it's not in the soldier's personal interest to get vaccinated, because they're young and healthy, as most active soldiers are, so why should they get sick for several days and maybe face risks they don't know about? This also overlaps with 7.

The response to #6 isn't an *outright lie* exactly, since the word 'may' does a lot of work. [The sun might have just exploded](#). But in practice, yeah, this is lying.

The response to #3 is, and I quote, "F*** you." If you *all mostly* comply, we'll lift the *outdoor* mask mandate? That's your pitch?

The response to #5 is, and I quote, "F*** you." Or, technically, 'it's not about you.' It completely accepts the (incorrect) premise that the soldier doesn't benefit, which doesn't seem like the approach I would take.

Then there are two standalones.

There's the remarkable #4: This is the first time I get to tell the army, NO!

And oh my is the answer to that one "F*** you."

Which leaves #12, which is the most interesting of the responses.

That's because the soldier has spoken [The Words](#), and has spoken them rightly.

Rather than voice a *specific and explicit concrete objection*, to which the answer of necessity is going to be some combination of 'you're wrong' and 'F*** you,' the soldier has given a general feeling of uncertainty without any concrete objection. Thus, there's no way to say they are wrong, and no basis to curse them out.

Instead, "I just feel skeptical and don't know what to believe" elicits this response:

"The choice to get vaccinated is a personal decision and should not be taken lightly. Talk to a medical professional, consult the FDA Factsheet, and review the educational materials available at www.carson.army.mil and from the CDC to weigh risks and benefits."

Suddenly we're acting like this is a Very Reasonable and Responsible Position, which needs to be solved by consulting official sources and doing further research. Only *after* that, when the soldier comes back with an actual concern, can we know which of our two responses to use, and justify using it. I mean, there's no way this person is skeptical *after* talking to all the Responsible Authority Figures, right?

[NPR claims that lower rates of vaccinations among blacks and latinos are entirely due to accessibility issues and have nothing to do with hesitancy](#). I completely buy that the access issues are doing a lot of work here, but it seems odd to attempt to suddenly shift from "here are all the legitimate and sympathetic reasons why these groups would be hesitant" into "they are not and have never been hesitant, it's that we didn't give them access and made access depend on things that systematically excluded them."

It's a claim that we'll be able to evaluate soon enough. As appointments become widely available via walk-ins in more places, with essentially no hoops involved, either the rates will converge or they won't. I am skeptical because it seems like it's a motivated shift in explanation rather than an attempt to track the truth - we want to make skepticism more

blameworthy, so we need to not identify these increasingly blameworthy motives in the wrong places, hence the shift. I am only somewhat skeptical because it seems clear that providing easier access has a dramatic effect on vaccination rates.

Overall, that evidence means that the article seems like *very good news*. What it *does* make a strong case for is that there is a lot of 'soft demand.' The bad scenario for where we are would be that 60% of eligible people have already been vaccinated, and most of the remaining 40% are actively having none of it. They are like the soldiers. They won't accept the shot unless convinced or heavily coerced.

Instead, this new picture finds evidence that what we have are a lot of people who prefer being vaccinated to not being vaccinated, but don't prefer it enough to jump through a bunch of hoops. That's great! All we have to do is get rid of the hoops and the need to jump through them, and offer them easy access. Now that we have abundant supply, that is relatively easy. Certainly I buy the anecdote that *Asians* have relatively low levels of hesitancy when given good access.

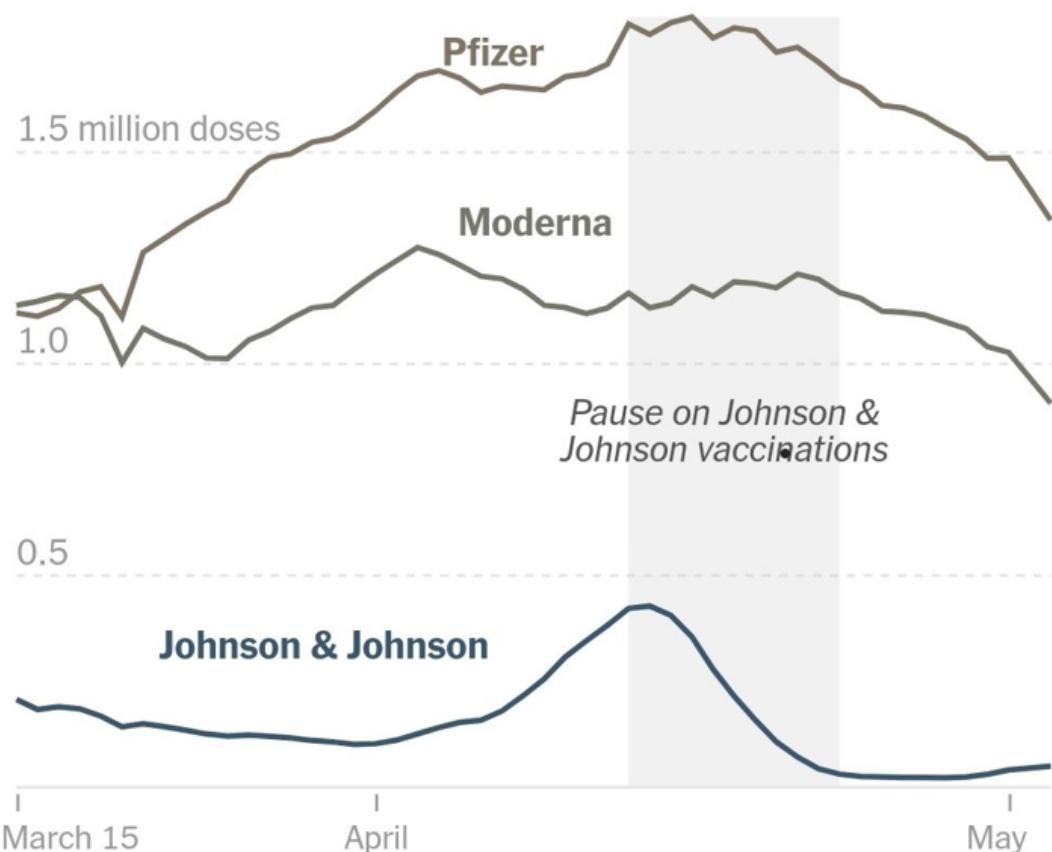
As someone who spent a substantial amount of time and effort to get vaccinated earlier rather than later, and to get those around him vaccinated earlier rather than later, I think those unwilling to do so are setting their price too low. We can separate this setting of a low-price into a few different components.

One explanation, which is the most hopeful with respect to the vaccines, is that their circumstances mean that paying the relevant costs is more expensive, and they have less ability to pay such costs. They care, but as the article claims, they are simply unable to take even a few hours off of work, or figure out how to navigate the barriers previously required. There is some of this, but we have some evidence that is then hard to explain if this is the main thing happening, such as the failure of J&J shots to rebound, and the distribution of shots on different days of the week.

[J&J shots are going, well, not great:](#)

Daily reported doses given by manufacturer

Each line shows the seven-day average.



Source: Centers for Disease Control and Prevention

If people simply cannot miss work, and are worried about side effects causing them to miss work in addition to the appointment itself, this suggests people will plan their shot around not missing work. That means getting a shot on Friday or Saturday, and yes we see giant spikes in shots given on Fridays and Saturdays, including during periods when supply constraints looked like they were binding. That seems like strong support. We'll see if this can be sustained; if this theory is correct, Friday and Saturday throughput should continue to bind.

A second explanation is that this is shallow demand, pure and simple. If someone wouldn't be willing to spend much time, let alone much money, to get a vaccine, that's a revealed preference that they don't value the vaccine much. This seems highly plausible to me, that there are essentially *three* camps rather than two camps. There's the people who want the vaccine enough to 'bid' on it in various ways and make it a priority. There's the people who actively don't want the vaccine, often violently so. But then there's also a large group, plausibly larger than the second group, who are fine with it but are mostly trying to live their lives and value the vaccine at some positive but small number.

I wonder how much of that is because we've set the price of the vaccine, and much of health care, to \$0, thus sending the implicit message that such services are, in emergencies, not that valuable. And also the general instinct to not think about one's health when one isn't forced to. We do seem to see a pattern of people who have the ability to get expensive medical care that they 'should' want, but not to spend small amounts of time (and aggravation) to collect it.

What's The Worst Possible Thing You Could Do?

If you're the President of the United States, in terms of actual impact the answer is presumably 'launch all the nuclear warheads.'

If one restricts to the pandemic, the answer would be to sabotage vaccine production and distribution. Nothing else comes close. One could plausibly argue that nothing else even much matters.

How would one sabotage vaccine production and distribution?

Sabotaging distribution means doing things like *not approving known-to-be-safe-and-effective vaccines, or suspending existing approvals and sending the message the vaccines are unsafe, or holding up distribution to worry about things like equity, or holding onto vaccine doses for extended periods with no intent of approving them ever.*

Oh, wait. Those are *all things done by the Federal Government during the Biden administration*, with no visible attempt to prevent them from happening or even regret expressed about them. You could even add, during the campaign, questioning the vaccine development process as 'rushed' or 'politically motivated,' plausibly being the cause of vaccines not getting approved a month earlier and creating much additional vaccine hesitancy.

You'd also give doses to children who don't need them rather than those in other countries that badly need them, so naturally Pfizer is on that one and soon will be applying for approval for children as young as two years old. And of course you'd continue not to do the first doses first, and continue to use full way-too-big doses of Moderna, and so on and so forth.

None of that means one couldn't have done or in the future do *more* of those things, so actions haven't been maximally destructive. But they've been quite destructive.

The other half of the worst thing you could do is sabotaging production. The easy way to do this is to screw up distribution. If things aren't approved yet, at best then that's going to slow down production until after approval. So are all the regulations involved in production, like needing to apply for permission and wait substantial time for permission for things like 'put more of the vaccine into each vial because we're short on vials.'

That's all *passive resistance* to lifesaving medicine. Could we kick this up a notch or two?

The ultimate way to hurt vaccine production, not only now but indefinitely into the future, would of course be to destroy the financial incentive to produce vaccines. The less you're willing to pay, and the less you let companies profit, and the less you reward those companies for quick scaling up and delivery of production, the less doses you'll get. This starts with not paying for building production capacity, and its central action is not paying much per dose or paying more for early delivery. If you want to go for bonus points, you can be like Europe and hold up negotiations for weeks to drive *down* the price even lower.

That's all *negative actions*, though. It's easy to sabotage efforts by *not doing the right thing*, especially when the right thing costs tiny amounts of money and looks like rewarding corporations, and is an action rather than inaction and thus blameworthy.

So it's a big step-up in the civilizational sabotage game to *actively take away* the incentive to create vaccines, [by stripping away intellectual property protections without any compensation, in the middle of a pandemic:](#)



Office of the United States Trade Representative

FOR IMMEDIATE RELEASE:

May 5, 2021

CONTACT: media@ustr.eop.gov

STATEMENT FROM AMBASSADOR KATHERINE TAI ON THE COVID-19 TRIPS WAIVER

WASHINGTON – United States Trade Representative Katherine Tai today released a statement announcing the Biden-Harris Administration's support for waiving intellectual property protections for COVID-19 vaccines.

"This is a global health crisis, and the extraordinary circumstances of the COVID-19 pandemic call for extraordinary measures. The Administration believes strongly in intellectual property protections, but in service of ending this pandemic, supports the waiver of those protections for COVID-19 vaccines. We will actively participate in text-based negotiations at the World Trade Organization (WTO) needed to make that happen. Those negotiations will take time given the consensus-based nature of the institution and the complexity of the issues involved."

"The Administration's aim is to get as many safe and effective vaccines to as many people as fast as possible. As our vaccine supply for the American people is secured, the Administration will continue to ramp up its efforts – working with the private sector and all possible partners – to expand vaccine manufacturing and distribution. It will also work to increase the raw materials needed to produce those vaccines."

###

There's a simple solution to the problem of intellectual property if you wanted to make the situation *better* rather than worse. You could *buy* the intellectual property rights from the companies involved. [So, basically, this:](#)



Kelsey Piper
@KelseyTuoc

...

I talked to lots of extremely reasonable experts about the intellectual property/vaccines thing and will write a summary of their responsible expert views but my irresponsible non-expert personal opinion is we should simply give Pfizer, Moderna, J&J, etc 100 billion dollars each.

3:23 PM · May 5, 2021 · Twitter Web App

It's not that much money, everyone would be happy, and the precedent would be excellent. Pay enough, and they'll even aid you in technology transfers. Even better, you could repeat this process with other drugs. Buy out the monopoly at its economic value, remove protections, and the people save many times that much money in costs. It's a great idea.

Doing this *without* compensation is about the worst thing one could do. If your new ideas outright save the world, we're going to reward you by confiscating them, voiding the contracts and promises agreed upon and informing you that we are not a nation of laws. That's exactly how *not* to get vaccines next time there's a crisis, or *anything else* next time there's a crisis, or really anything else useful at any time for any reason.

The message we've sent, loud and clear, is that we are not a nation of laws and we do not reward those who deliver the goods for us. Instead, we retain protections on things like insulin that are pure rent seeking, while taking away protections that are doing exactly what patents are designed to do: reward those who produce world-changing positive innovations via temporary ability to profit.

We are a nation of a person in charge, and if that person decides to confiscate your property because it's good politics, well, tough.

It's a horrible, horrible precedent. We will pay for it in money, will pay for it with our freedom, and [we will ultimately pay for it in blood.](#)



Daniel Eth💡 @daniel_eth

16h

BREAKING: future vaccines will rely more on trade secrets and specialized materials that are hard to produce even if you have the IP.



Daniel Eth💡 @daniel_eth

15h

I honestly think Biden just killed more people than Trump did. How much will this delay cures for Alzheimers and cancer?

5 12 12 ...



NathanpmYoung.substack.com

@NathanpmYoung

"Why should we pay to remove copyrights from books but not vaccines?"

"Because vaccines save lives."

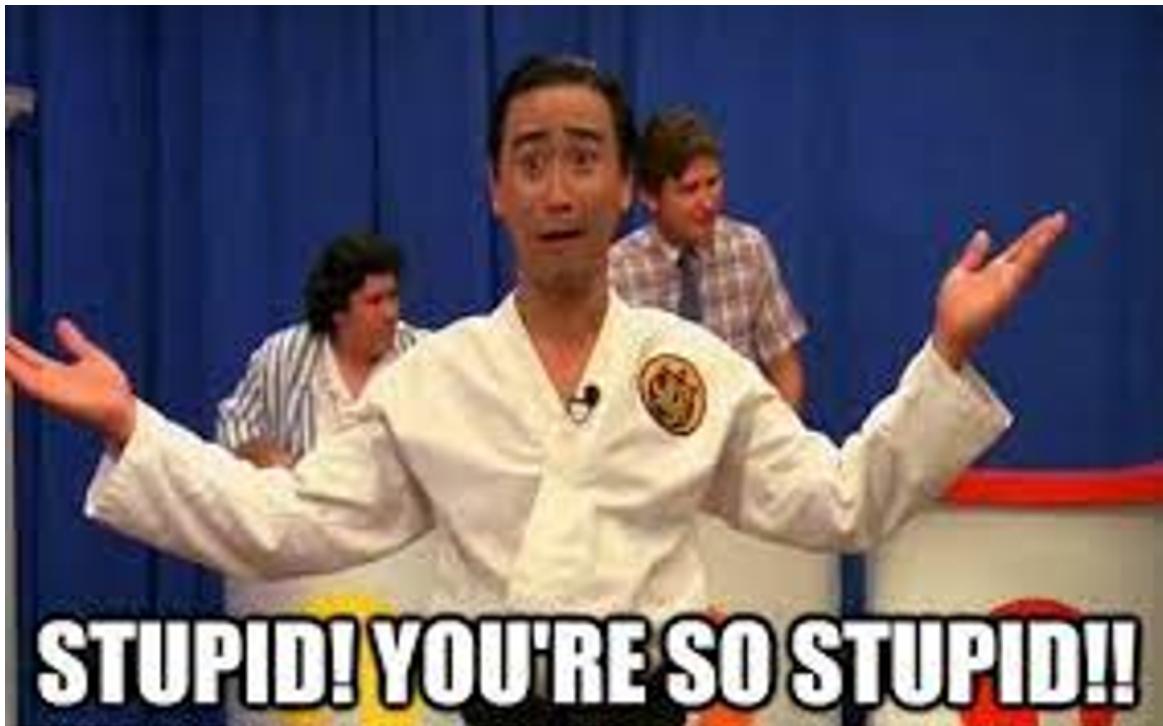
"So we should be quicker to disincentivise the *more* valuable thing?"

7:22am · 6 May 2021 · Twitter Web App

mRNA vaccine technology is potentially a *full cure for infectious disease*, and plausibly also a cure for cancer. The federal government sabotaged all that, big time.

What did we get in exchange? [What's in the box?](#)

NOTHING!



Unless, of course, they are not so stupid In which case the destruction of the rule of law and of private incentives, and the signaling that political expediency is the most important thing, was the point.

You see, [this will not increase vaccine production](#) (MR link with full explanation, recommended), for two reasons, even if [vaccine ingredients](#) didn't prove to be limiting factors. MR also [recommends this Barron's column. Here's another confirmation](#) that no, this won't improve short term supply.



Alex Tabarrok @ATabarrok · 15h

Plastic bags are a bigger constraint on vaccine production than IP!

...



Single-use bioreactor - Wikipedia
en.wikipedia.org

2

13

32



Alex Tabarrok @ATabarrok · 15h

I am not kidding.

...



Single-use Plastic Bioreactor Bags to Filters: Why India Needs Them fro...

These raw materials for vaccines are produced in a limited number of countries, around 27, and by a limited number of manufacturers, aroun...

[news18.com](https://www.news18.com)

Many people have this idea that all the knowledge and skill required to produce the vaccines lies in the patents. Once you lift the patents, lots of other companies can go start producing vaccines. Except, that's not actually true because

1. The vaccines require technical expertise not included in the patents, which is expensive and slow to transfer, and which would also transfer valuable knowledge that can be used for other R&D and other production and thus which the vaccine producers are not going to transfer without compensation.
2. Moderna explicitly *already said they wouldn't enforce the patents*, and no one really expected the others to either.

Read that second one again, if it's new to you. The greedy capitalists whose rights you took away without compensation were *already voluntarily giving those rights away*. If there was already clearly no intent to enforce the patents, what good does lifting those patents do?

It sends the message that the United States is willing to confiscate property for political gain, when it feels like it, on the basis of the executive's say so.

Even though that won't produce anything useful, yes, it's still bad for business and still punishes exactly who we should be rewarding, or at least demonstrates that such punishments should be expected, [as measured by the stock market](#). Remember Moderna *already waived its rights*:



As usual, the usual suspects wasted actual zero time demonstrating exactly the slipperiness of the associates slopes, as they quote the decline in shareholder value *as a good thing*:



Alexandria Ocasio-Cortez @AOC · 15h

Let's do insulin next

What makes such a statement so maddening is that she's right. We should *totally* do insulin! It's *completely insane* that we've allowed regulatory capture and rent seeking via intellectual property protections on "inventions" like insulin. The congress should get together, *write a bill and pass a law* that stops such things from happening now or in the future via changing protections, ideally without confiscating private property, and then the President should sign it, and then the bill should become law. Then do copyright.

Won't Someone Please Think of the Children?

The minds of many parents I know are turning to the question of summer camp. Is it safe to send your young child?

Are all the people you care about that will be in contact with that child either other young children or fully vaccinated by the time the camp starts?

If the answer to that question is yes, then yes.

If the answer to that question is no, then given that vaccinations are now available to everyone pretty much on demand, why isn't the answer yes?

If the answer to *that* question is that someone is seriously immunocompromised, or otherwise super important to the child's life and won't get vaccinated (for whatever reason), *then and only then* is it time to look at the camp's procedures to see whether you're comfortable with the level of risk being taken. In particular, you'll need to ask how many children and unvaccinated adults will be in contact with your child, how close that contact will be, and how much time will be spent indoors, and do a calculation.

I still think that calculation should almost certainly be 'yeah, it's fine' but at that point, as they say in the advertising business, it's up to you.

My general answers regarding children generalize this. Young children are not at enough risk from Covid to let this change how they live their lives, so them catching it only matters to the extent that they would pass it on to vulnerable others.

By the end of May, with notably rare exceptions, patience with those in the United States who are still vulnerable can reasonably be at an end. Those who decline the opportunity to be vaccinated can manage their risk however they choose, but life beckons.

Speaking of life beckoning: I strive not to use the word evil, I avoided using it in the previous section, but this is evil in its purest form:



Ana Cabrera @AnaCabrera · 20h

NEW: NYC Public Schools will have remote learning instead of snow days next school year, the NYC Dept. of Education announced.

1.6K

8.8K

10K

...

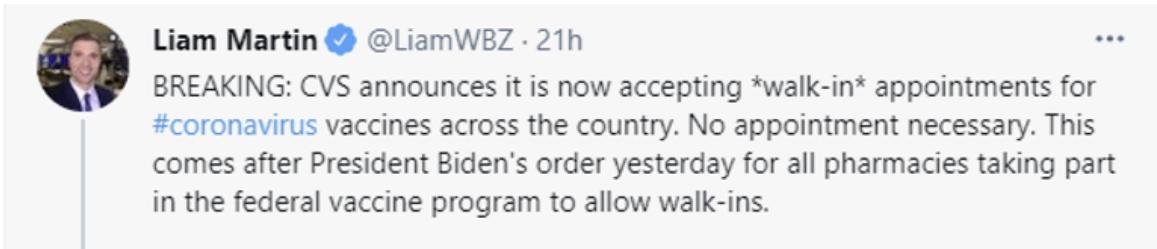
Anyone who doesn't recognize this as such has lost their soul. Any parent or teacher who enforces this should be treated as the mustache-twirling villain they are. I am deeply sorry to any child who has been so absurd and tortured, or living in so much fear, that they are tempted to put up with this.

If you do not think school's primary nature is 'child prison' and/or that those running it are pro-children, then you have new data your model needs to somehow explain.

In Other News

V-NY day approaches, [and Cuomo fully opens up stadiums, opens up Broadway, offers subsidized vaccinations at Mets and Yankees games](#). Took everyone on Broadway by surprise, so it'll be a while before they can actually get on with the show. Remember, you'll want to wait to get vaccinated until you attend a game at the stadium, together with tens of thousands of other people. That way you get free tickets!

Alternatively, [you can show up at the local CVS and maybe pick up a gift as well](#) :



Liam Martin  @LiamWBZ · 21h ...
BREAKING: CVS announces it is now accepting *walk-in* appointments for #coronavirus vaccines across the country. No appointment necessary. This comes after President Biden's order yesterday for all pharmacies taking part in the federal vaccine program to allow walk-ins.

In many cases, Walmart too. Basically everywhere at this point. No excuses!

[South Korea says AstraZeneca shot 87% effective after one dose](#). Which would be pretty good after two doses. First doses first, indeed.

[Police have low rates of vaccination, endangering those around them who they forcibly interact with and likely killing them](#) (WaPo), but no one is able to make them do the right thing and stop endangering the public. A little on the nose, if you ask me.

[Airline boarding procedures were already worse than random, and changes in response to the pandemic made them worse still](#). It seems that *looking like* a good procedure is more valued than being an actually good procedure. There seems to be a strong match between 'this is a quick boarding procedure' and 'this is a safe procedure,' so the problem is purely that good procedures don't look good and/or don't feel 'fair' somehow, or miss out on some opportunity for price discrimination. Is there an improvement that would also *look and feel* like one?

[MIT requires vaccinations](#), although so far only for students. I expect most colleges to follow suit if only to avoid potential liability concerns. Not spreading the requirement to faculty and staff seems like a clear mistake.

[New higher estimate of true number of Covid deaths](#) via MR, not enough data to know how much credit to give this.

[The Covid Response Project](#) chronicles the Covid-related experiences of people across different states. I've sampled and it seems like a good source of real people's anecdata. There will definitely be surprises.

[Twitter thread](#) and [paper](#) discussing origin of variants of concern. Not sure there's practical updates to be had, but interesting information.

[Pfizer begins shipping some vaccine doses manufactured in the United States abroad](#), starting with Mexico.

[Potential universal coronavirus vaccine proposal](#). From what I can tell this is highly unlikely to work but you never know.

Vaccination availability site of the week, [Vaccinate the States](#).

[Marginal Revolution points us to a study of future work-from-home \(WFH\) patterns \(paper\)](#), and finds dramatic effects the study expects to linger beyond the pandemic. I hope to check this out in detail in the future, but the headline impacts are *gigantic*. They expect WFH to go from 5% of full workdays to 20%, and for this to be a *5% productivity boost*, most of which will be due to reduced commuting. Commuting is much worse than people think it is, so this is a really, dramatically large effect, in the range of ‘potentially a bigger long term deal than the pandemic.’ This isn’t a fake productivity boost, it’s literally getting rid of purely wasted unpleasant time (that also burns a bunch of carbon to boot). Given the amount of time being saved, it also implies that *on the margin* there’s still going to be a dramatic underutilization of WFH as an option. If a change to 15% of the workforce produces a 5% productivity boost by saving useless time (and it’s still an if, the story has to check out), clearly we are not using anywhere near enough of it.

Not Covid, therefore... we're coming back, baby! [HYPE!](#)

Book Review of 5 Applied Bayesian Statistics Books

There are 5 strong contender for best Bayesian Statistics book

TLDR

- Statistical Rethinking, henceforth **SR**^[1]
 - Up to speed fast, no integrals, very intuitive approach.
- Doing Bayesian Data Analysis, henceforth **The Dog Book**^[2]
 - This is the easiest book. If your goal is only to create simple models and you aren't interested in understanding the details, then this is the book for you.
- A Student's Guide to Bayesian Statistics, henceforth **Student's Guide**^[3]
 - This book has the opposite focus of the Dog book. Here the author slowly goes through the philosophy of Bayes with an intuitive mathematical approach.
- Regression and Other Stories, henceforth **ROS**^[4]
 - Good if you want a slower and thorough approach where you also learn the Frequentist perspective.
- Bayesian Data Analysis, henceforth **BDA**^[5]
 - The most advanced text, very math heavy, best as a second book after reading one or two of the others, unless you are already a statistician.

Irregardless of which one you pick, watching the YouTube lectures for either [**SR**](#) or [**Student's Guide**](#) is very helpful

Short review of each Book

The Dog Book: Is the easiest book if you do not strive for understanding but simply want to quickly get to a skill level where you need to develop a not so fancy model then this is a great book. It is also a good reference book as each chapter is based on a specific link and function and regression variable type, thus, if you want to do an Bayesian ANOVA you simply look for chapters named something like "categorical predictor with metric outcome".

Student's Guide: This book has the opposite focus as the Dog book. Here the author slowly goes through the philosophy of Bayes with a mathematical intuitive approach. It looks like a very good reference book. A Bayesian professor has recommended it as one of the best introductions to STAN. Chapter 8 is also very good. It starts with a graph of the relationship between all likelihoods and then goes through EACH and EVERY ONE with an intuitive example and some nice plots. I would recommend everyone to read the chapter and/or to use the chapter as a reference whenever you have a few 'candidate' likelihoods in your head. Reading Chapter 12 in **SR** will subsequently teach you to create mixtures of these likelihoods if you need further hacks such as zero inflation. Remember the YouTube series explain the math very well. So use them as a supplement!

ROS: This book is a 'normal' statistics book written by Bayesians, thus it teaches both philosophies and have very great intuitive mathematical examples. It is a trophy of very intuitive considerations about model building, such as: $N(25, 10)$ is statistically significantly different from 0, but hardly even 1σ away from $N(10, 10)$ because variances are additive, so the uncertainty of the 15 difference is $\sqrt{10^2 + 10^2} = 14$. The same is true for interaction terms as they have two sources of error, thus we should a priori expect those to have wider posteriors! The slow approach makes it immensely readable for people like me 'who already know this', as half of the book is basically 'there be dragons' explanations of everything that can go wrong when you are doing a regression, and them doing the analysis twice to show the difference between the different philosophies.

SR: This is another great book, and it uses a level of math that is easier than **ROS** and Student's Guide but more rigorous than **The Dog Book**. This means that you get up to speed much faster, and it has you building quite advanced models by skipping the math and by heavily developing your intuition. Until Chapter 11 it's very great, but after than it starts introducing advanced concepts, and the material in Chapter 14-16 is not covered in the 4 books above. So it can also be bought as an "advanced supplement" to any of the 4 books above. Also the lectures are phenomenal and track well with the chapters, so I advise watching them prior to reading a chapter to get a big picture overview before going deeper.

BDA: This book used to be the bible. It's very mathy compared to the 4 other and seems to be written by field experts. Part 1 and 2 seem 'coherent' and are actually quite good for understanding the math that the other books use but don't explain well.. Part 3 and 4 are mathy versions of what are superficially covered in the above 4 books and most of Part 5 is simply state of the art Bayesian modelling expressed using only math. I think it might be slightly more intuitive than reading the source papers - but only slightly.

Recommendations / Extra considerations

Causal Inference: The books **SR** and **ROS** put extra emphases on causality, thus if you have observational data, where you want to predict the outcome of an intervention these books are 'extra' good. **SR** emphasizes Judea Pearl's graph based approach which is superior when doing fancy models, given they are both introductory I think that **ROS** actually teaches you to guard against more causal errors, so it's hard to declare a winner.

Math:

Dog Book << SR < ROS < Student's Guide << BDA

- I have taken less than 15 ECTS of math (studied biology or sociology), then pick **The Dog Book** or **SR**.
- I have taken some math (Engineering or Econometrics): **Student's Guide** or **ROS**
- I have an undergrad degree in math: **BDA**

I want to make cool models before page 200:

- **SR or Dog Book**

I am a patient learner:

- **ROS or Student's Guide**

I want a good reference book, all are decent, **SR** is worst because it has 'playful' chapter titles such as "Ulysses' Compass" and "The Golem of Prague" which actually serves as helpful memetic when you are reading, but becomes a pain a year later when you need to look up things.

My Experience with the books

I have not read all the books from cover to cover, here is my experience with each one:

- Doing Bayesian Data Analysis (**The Dog Book**)
 - Read cover to cover
- Statistical Rethinking (**SR**)
 - Read Chapter 12 and half of 14
 - Mixture Likelihoods and Covariance Models.
 - Watched all 20 lectures
 - Solved assignments via study groups
- Regression and Other Stories (**ROS**)
 - Read Chapter 1-9
 - Read a lot of Andrew Gelman's Blog
- Bayesian Data Analysis (**BDA**)
 - Read Chapter 1-13
- A Student's Guide to Bayesian Statistics (**Student's Guide**)
 - Skimmed the earlier chapters.
 - Read Chapter 8
 - Watched about 10-20 hours of his YouTube lectures

-
1. Richard McElreath "Statistical Rethinking" [←](#)
 2. John Kruschke "Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan" [←](#)
 3. Ben Lambert "A Student's Guide to Bayesian Statistics" [←](#)
 4. Gelman, Hill and Vehtari, "Regression and Other Stories" [←](#)
 5. Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. "Bayesian Data Analysis" [←](#)

Covid 5/27: The Final Countdown

(Personal note: My Facebook account was hacked. Attacker has not contacted me, and deleted my name from the profile. Since I am already [Against Facebook](#) I do not view this as a great loss unless it leads to further trouble, but readers should be aware that for now I have no Facebook account. If I don't get it back, I may or may not create a new one.)

We are now a second week into The Great Unmasking, with no sign of trouble in the case numbers. While it ain't over till it's over and I'm not quite prepared yet to outright declare victory. On reflection my criteria for V-A day is 'I notice I am acting the exact same way I would if I was unvaccinated, provided everyone else was OK with that' and we're definitely not there yet. Still, it seems likely that in America it's all over but the shouting and I see a lady preparing to sing.

At that point, we'll potentially need to worry about future seasonal concerns if not enough people get vaccinated and we return to the full old normal, and we'll need to keep an eye on variants, but that's about it.

With numbers declining and life starting to return to normal, talk turned to the question of whether Covid-19 leaked from a lab, and potentially originated from gain of function research. Once considered a vile, racist conspiracy theory to be ruthlessly censored and scorned, it is now being officially investigated and is widely considered highly likely. We've seen this kind of Official Facts transformation before, and things are following the standard script. There's a section on it, but in general I'd still hope to keep this mostly out of the column going forward.

Let's run the numbers.

The Numbers

Predictions

Prediction (now using Johns Hopkins numbers fully going forward): Positivity rate of 2.7% (down 0.3%) and deaths fall by 8%.

Result: Positivity rate of 2.5% (down 0.5%) and deaths fall by 10%.

Prediction: Positivity rate of 2.2% (down 0.3%) and deaths fall by 9%. Continued steady progress, with a little hedging in case we got out in front of things a bit.

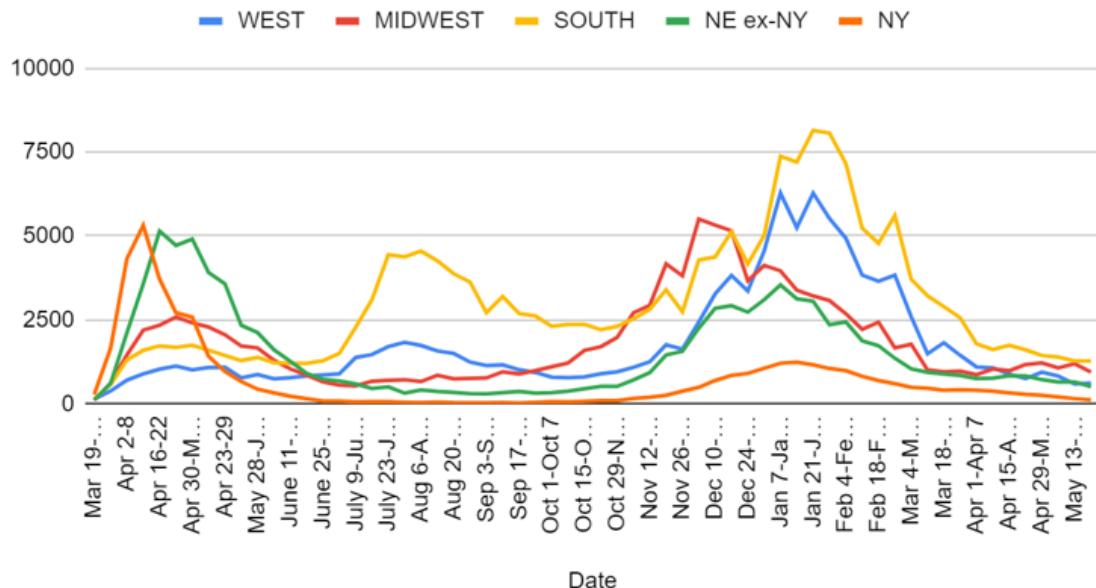


Deaths

I deleted 400 deaths, 300 from Oklahoma and 100 from New Mexico, that were clearly reporting previous deaths rather than new data. Oklahoma reports weekly so there's no way to know the real number, and I left it on the high end of possible. New Mexico had a day with 114 deaths so taking away most of those seems reasonable. That resulted in this:

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Apr 15-Apr 21	883	987	1747	1168	4785
Apr 22-Apr 28	752	1173	1609	1110	4644
Apr 29-May 5	943	1220	1440	971	4574
May 6-May 12	826	1069	1392	855	4142
May 13-May 19	592	1194	1277	811	3874
May 20-May 26	615	948	1279	631	3473

Deaths by Region

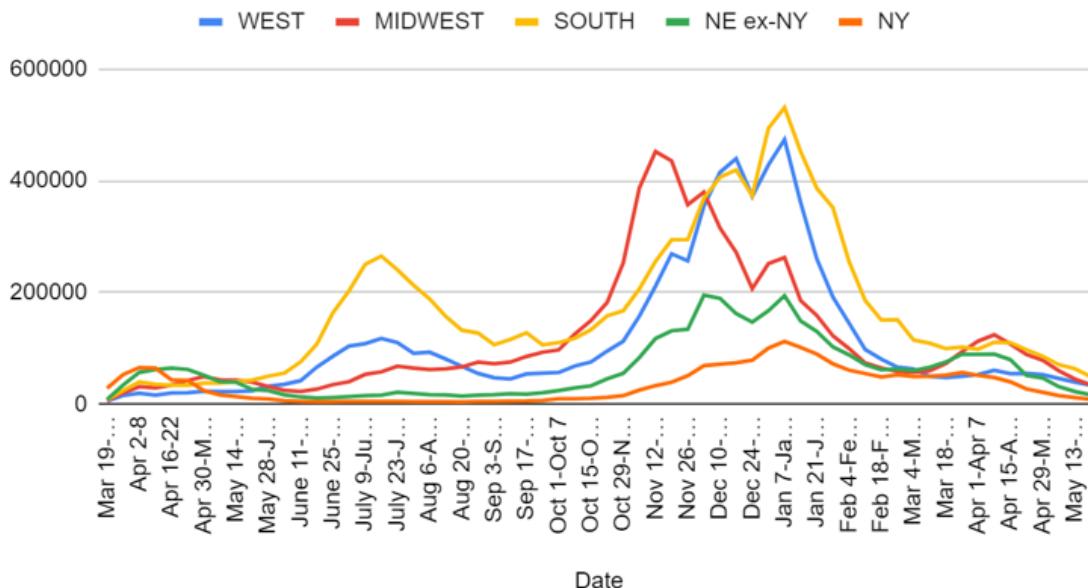


The decline in the Northeast is steep, but at least superficially it appears real. It's likely last week's number was higher than the true trend line, slash random variance reduced deaths a lot in the Northeast and Midwest, and hid a true decrease in other regions. The overall rate of decline seems most important here.

Cases

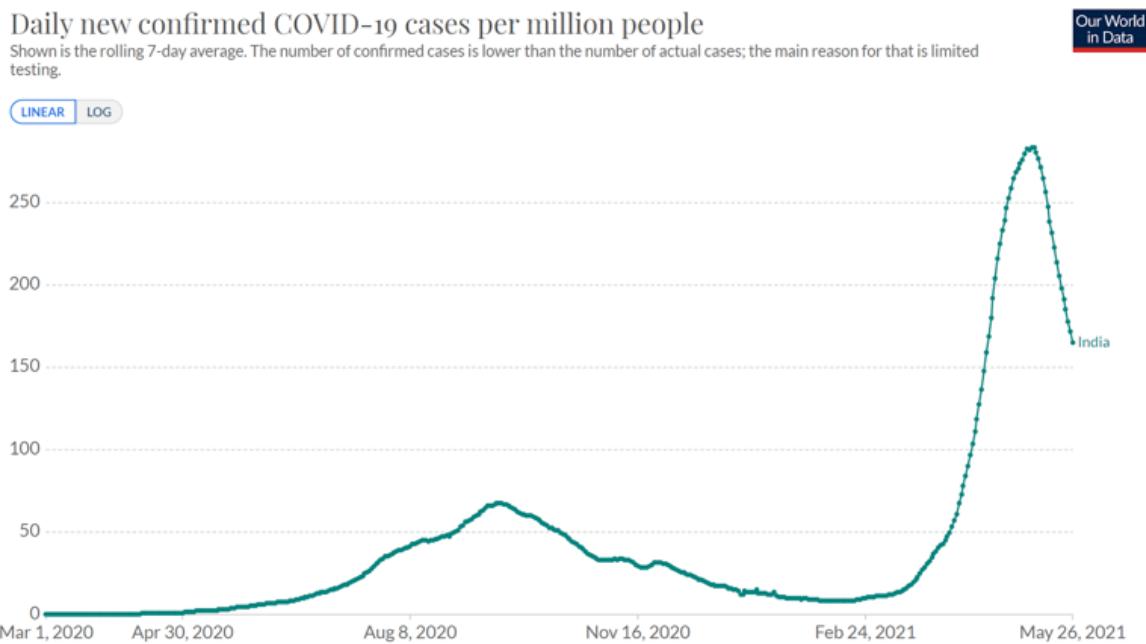
Date	WEST	MIDWEST	SOUTH	NORTHEAST	
Apr 8-Apr 14	60,693	124,161	110,995	137,213	433,062
Apr 15-Apr 21	54,778	107,700	110,160	119,542	392,180
Apr 22-Apr 28	54,887	88,973	97,482	78,442	319,784
Apr 29-May 5	52,984	78,778	85,641	68,299	285,702
May 6-May 12	46,045	59,945	70,740	46,782	223,512
May 13-May 19	39,601	45,030	63,529	34,309	182,469
May 20-May 26	33,890	34,694	48,973	24,849	142,406

Positive Tests by Region



Those are huge gains across the board. This is what victory looks like. The control system will doubtless try its best, but it seems we made it.

India



Things are steadily on their way back down, so some combination of changing seasonality, ending the election and various festivals, and the control system combined to get things back under control. Things are still rather bad out there, and it's going to be a while before we get enough vaccine shots to India, but the worst is likely over.

As with every other strain, there's paranoia around how effective the vaccines are against the India strain. Also as usual, everyone [quotes the infection percentages](#) (e.g. there 70-75% for Pfizer, [here 88%](#) for Pfizer and 60% for AZ) rather than the severe illness percentages or death percentages, with a special emphasis on numbers that are on the low end, so numbers that sound scary mostly aren't scary. [This data looks scary at first](#), but when compared to the baseline numbers for similar vaccinations in the past it mostly seems fine even if the effect here is fully real.

Escape From the Lab

Until now, I've entirely dodged the question of the origins of Covid-19. The issue didn't seem to have practical implications and there were plenty of other things to focus on. I was content to let others handle that question. Even here, I don't really have anything *new* to offer, it's more that it's so on point that it requires mentioning.

Now that the consensus has shifted, and the lab escape hypothesis is no longer grounds for censorship and cancellation [and the tone of coverage has dramatically shifted overnight](#), it joins a long list of other questions in which the media, elites, authorities and censors treated a question as definitively one way, then abruptly shifted when that position stopped being defensible, and quickly started rewriting history to pretend we had always been at war with Eastasia. As usual, the elites and authority figures are going to pretend what happened didn't happen, that they never denied anything, that they're the ones who figured this all out, and that the people who claimed this crazy hypothesis back when it was still crazy should *continue being dismissed as cranks* because in their parlance crank means 'went against elite consensus' rather than 'made claims without good evidence, especially ones that aren't true.'

[Consider the 'stealth edits' that Vox did to its old articles](#) once a month had gone by and the 'debunking' had become common knowledge:



Paul Graham  @paulg · May 24

Some of the stealth edits that Vox made to its article debunking "conspiracy theories" that Covid-19 originated in a lab leak between its original publication in March 2020 and now.

...

The Wuhan Institute of Virology is a real place, and the exact origin of the novel coronavirus is still a mystery, with researchers racing since the outbreak began to figure it out. But already, virologists who've parsed the genome and infectious disease experts who study coronaviruses		The Wuhan Institute of Virology is a real place, and the exact origin of the novel coronavirus is still a mystery, with researchers racing since the outbreak began to figure it out. But already, virologists who've parsed the genome and infectious disease experts who study coronaviruses
have more than enough evidence to show that		say they have enough evidence
the virus is brand new and came from nature,	« 9 words »	the virus is brand new and came from nature.
not the Wuhan lab.		
A large group of them, citing genome analyses from multiple countries, recently affirmed in <i>The Lancet</i> that the virus originated in wildlife.	« 41 words »	A large group of them, citing genome analyses from multiple countries, recently affirmed in <i>The Lancet</i> that the virus originated in wildlife.
The emergence of the virus in the same city as China's only level 4 biosafety lab, it turns out,		The emergence of the virus in the same city as China's only level 4 biosafety lab, it turns out,
is		appears to be
pure coincidence.		pure coincidence.

It's totally fine to be wrong sometimes. It's completely unacceptable to rewrite the public record to pretend retroactively that you covered your ass. Especially when the certainty that comes from lack of ass covering was used as a justification for censorship, castigation and cancellation.

[Until this week, Facebook was removing posts that claimed Covid-19 leaked from a lab.](#)

After being caught doing this, [Vox admitted they did the edits:](#)

Editor's note, May 24, 2021: Since this piece was originally published in March 2020, scientific consensus has shifted. Now some experts say the “lab leak” theory warrants an investigation, along with the natural origin theory. Some language in this article was updated in April 2020 to reflect scientific thinking, but it has not been updated since then. For our most up-to-date coverage, visit Vox’s **coronavirus hub**.

If that note had *also* been published in April 2020, then fine, but it wasn’t. One could argue it’s fine anyway, it’s better to fix things than not fix things, and that it’s unreasonable and misleading to call Vox out *now* for edits they made over a year ago. There’s some validity to that. A lot of Paul’s readers presumably assumed that the edits were made in the last month, which would be a much worse look. Then again, when are you going to call someone out on such stealth edits? Presumably at exactly the same time one would call someone out for writing the original.

I’d be curious what Vox thinks it means by ‘scientific thinking’ here. Does it mean ‘the scientists started using CYA-words so we should too’? Or does it mean the actually correct ‘scientific thinking means not taking people’s words for things or reporting claims as facts, so instead we’re going to do proper journalism and tell you exactly what evidence is present’?

I remember when the New York Times did this every single time because that’s how one is actually a paper of record – ‘Man arrested during an armed robbery, police say.’ Police say is doing important work there. Whereas now, NYT and Vox are the two mainstream sources I will not link to and that I avoid reading.

Essentially all claiming-to-be authoritative sources treated the lab hypothesis as pure conspiracy theory and utterly impossible up until the last few weeks, the same way they insisted there concerns about the virus in February were alarmist and racist, masks didn’t work, outdoor events were dangerous, vaccines would take a minimum of eighteen months and so on and so forth. I grow weary of typing the list out.

I don’t think this is *quite* the same as the others, as it both did not do practical harm to people’s decision making and in hindsight still seems like a highly reasonable position to put

most of one's probability into at the time. That doesn't make it acceptable to censor and castigate, but it's a big step from being clearly and knowably wrong at the time.

[Here's a thread on historical lab leaks](#), which doesn't prove anything but does make it clear that a lab leak was always plausible.

[Nate Silver asks people to go on record with their probabilities of a lab leak](#). The comments are telling. Most replies Twitter shows us give absurdly low numbers, as close to 0% as they dare, and continue to attempt to police the discourse and accuse anyone saying otherwise of being crazy, while also attacking the very idea of Bayesian evidence. Then there are the replies saying 'there's a lab, there's a virus, obviously it's from the lab.' None of it is a good look.

Noticing *both* that there is huge unnecessary risk of pandemics coming from wet markets or otherwise having a natural origin, *and* that a pandemic could easily come from a lab accident, is what matters, regardless of which origin led to Covid-19 in particular.

Even more than that, I'm interested in the mechanisms behind the suppression of information and debate. My odds are mostly on the basis of the shift in other people's odds being against strong headwinds trying to prevent it, and the improbability that things would have gotten this far unless the case for escape was strong.

At this point, I think I am somewhat below Nate Silver's 60% odds that the virus escaped from the lab, and put myself at about 40%, but I haven't looked carefully and this probability is weakly held. I'm sharing it because it's important to share probabilities even when they're weakly held. The question of whether we'll ever *prove* what happened, or the official story will *conclude* a lab leak, is very different from the question of the actual origin, so there's no pure way to evaluate such predictions, but it seems important to give a number even with my uncertainty. I still find the natural origin story likely (that's why there was a lab in the first place), and the evidence that the lab is acting suspicious and everyone is covering things up is still consistent with them doing that automatically without any need for there to be anything to cover up, but a lab leak is *also* plausible and the investigation and legitimacy of the lab leak claim getting this far under these conditions was surprising.

Despite all the other ways in which we were misled, until last week I still reliably put lower probability on the lab leak theory than I should have on reflection. A lot of that was that, as noted, I intentionally didn't consider the question and wouldn't have even if I thought the probability was higher, and we still don't know what happened, but none of that means I didn't get it wrong or that such errors don't need to be admitted and corrected.

Should we update to give more credence to other things that are labeled as 'conspiracy theory'? That's tricky. I don't think this was a 'grand conspiracy' or anything, nor do I think those suppressing the theory had any knowledge of whether or not the virus leaked from a lab. My model says this is how the system works by default, with all who form the system instinctively moving to implement the suppression of such speculations, without any need to coordinate.

It's important to note that *this is not a conspiracy theory* because if the theory is true *there need not be a conspiracy*. The lab didn't *intend* to leak the virus (or if it did, that would be a very different theory). If the leak happened, the lab almost certainly *accidentally* leaked the virus, the same as there have been historical other leaks, and of course they didn't come forward and admit that and instead covered it up, and the system did its thing automatically rather than because there was some cabal or set of secret orders. That's true even if the virus was also created in the lab.

This is a sharp contrast to actual conspiracy theories that involve conspiracies that explicitly coordinate to achieve objectives, including but not limited to suppressing The Truth That Is Out There. Such claims should continue to be viewed with extreme skepticism. There likely is no conspiracy.

In contrast, there is totally a cover up. I'm at 98%+ that there was a de facto cover-up, even if the virus didn't leak from the lab. There is often a truth that is effectively suppressed, because the Powers That Be intuitively sense that it would be better not to spread it around, and you can't take the elite consensus or media's word for anything. There's no reason for such powers to check the validity of the claim before suppressing it, since they'd want to suppress it if it was true and also suppress it if it was false, and checking validity risks giving the claim credibility. If you think that this originates in conspiracy, your map won't be accurate, and you'll see it in all the wrong places.

Calls to 'hold people accountable' for the suppression of such information don't interest me much, any more than I want to 'hold people accountable' for the mask debacle. I'm not opposed to them, but I don't see much point. What is important is to know what happened and how such things happen, so we can avoid them in the future, and update our view of various sources of information accordingly.

The likely consequence to the origin being considered a lab leak is that people might react to that information by attempting to punish, or learning to fear or hate, that which they see is responsible. That presumably means some combination of China, likely the United States because we offered some funding and it is popular to turn everything on us whenever possible, and biologists and those who study infectious disease or scientists and labs in general. A lot of people would hear 'lab leak' and assume it was intentional or a weapon, and nothing we say would convince them differently. There would be a general rise in conspiracy theory and paranoia, as discussed above. Any movie with a scientist is more likely to paint them as evil, or irresponsible, and the cause of potential disaster. Geopolitical tensions would presumably rise all around. Worries about 'hate' and ethnic violence are used as cudgels these days, but they're also real concerns.

None of that seems positive. If it had happened last year, with conditions still bad and a different president, it presumably would have been far worse.

The positive response technical we would hope for, that we would learn to use better precautions when researching and storing deadly viruses, which we should clearly do regardless of the true origin here (if you look at the precautions they were taking, they *clearly* weren't sufficiently robust, and the history of lab leaks is long), seems likely instead to cause a stupid response, regulatory and otherwise, that hugely raises costs in both dollars and optics, and reduces our ability to do research to prevent future pandemics or otherwise make scientific progress, without much in the way of gains to safety.

This was an additional reason I had no interest in pursuing the story of a potential laboratory leak, and made a decision for myself to leave it alone. I draw a big distinction between choosing not to investigate and discuss something and suppressing it via misleading people and actively suppressing dissent, especially via censorship, but it's also important to know what motivations are going on in your sources of information, and what sources are going on in one's own head.

I also think that same instinct of 'why would you go looking into this, it can only cause bad trouble and not [good trouble](#)' then leads to people blaming those who looked into it *when they turn out to be right*, and thus they expect vindication from the system and instead mostly get more blame. In elite eyes, being seen as right only makes their actions that much worse. Already a New York Times reporter whose beat is primarily Covid-19 is [floating that the claim of a lab leak is still racist](#). No, Scott Alexander wasn't an isolated incident, and I'm not going to lift the NYT ban any time soon.

Going forward, I intend to continue doing what I can to not cover the question of Covid's origins, so my silence should be treated as very weak evidence of anything. That does not mean I will succeed, as circumstances are not making this easy, but I'm *not* going to automatically post if my posterior on this changes from week to week.

In Other News

Turns out [the Pfizer vaccine never needed those special freezers.](#)

[DeSantis actively trying to stop cruise ship from requiring its passengers be vaccinated.](#)

Meanwhile, many countries that got vaccine allocations [are at risk of having them expire unused.](#)

[Here's some obvious nonsense for the week](#), I leave the math this is implying as an exercise to the reader, while noting that most future workers won't have been in school at all:

ian bremmer 
@ianbremmer

Learning loss from school closures will reduce US GDP by 3.6% and hourly wages by 3.5% by 2050, per Wharton School.

11:50am · 25 May 2021 · TweetDeck

[Restaurant reservations back to baseline levels.](#) This matches my personal behavior, except that I'm still in Warwick instead of New York City for now which makes restaurants less appealing for now.

[New York City will not be offering remote learning in the fall.](#)

Jillian Jorgensen  @Jill_Jorgensen · 56m ...
BREAKING: [@NYCMayor](#) says there will be NO remote option this fall for NYC public school students.
30 183 477 

Jillian Jorgensen  @Jill_Jorgensen · 54m ...
The mayor opted to break this news on MSNBC's Morning Joe, well-known for its in-depth coverage of NYC schools. Can you sense my sarcasm?

[New York also lifts the mask requirement for 2-to-5 year olds at camps](#). Yay sanity.

Zeynep thread about school infection rates. Masks and ventilation work but not as much as one would hope, and barriers and desk spacing [were the goggles](#) and did nothing.

[A political analysis asks why Biden has high ratings on the pandemic](#). It makes no mention of any of Biden's policies, decisions, actions or statements, because it turns out none of that matters whatsoever. Presumably there's a point at which something would matter, but we are still waiting to prove that via example.

Marginal Revolution [shares this story](#) accusing Noble Laureate Levitt of being irresponsible and misleading via having a Covid opinion that wasn't sufficiently concerned and serious, and being unacceptably confident about it, while having authority and respect of some kind. Unacceptable. There are some real accusations here, saying that Levitt tried to privately get a critic's funding pulled, and that he didn't allow for proper review of his findings. I don't think his conclusions made sense given the data available, either. Yet I have little doubt that if Levitt had aligned himself with the Very Serious People, no matter how extreme his conclusions or questionable his methods and results, we wouldn't be looking at an article like this.

[Post looking to explain why suicides didn't rise during Covid-19](#). Offers a variety of possible explanations, some bottom lines being that depression due to objectively bad conditions is different than depression for internal reasons, the baseline happiness rate generally having less impact on suicidality than one might think (I have doubts but it's plausible), people often come together in a crisis, and there may have been issues with lack of available methods of suicide slash many suicides may have shifted into drug or alcohol (so, drug) overdoses which are often at least kind of suicides. It's not mentioned but I suspect another part is that when things are depressing because we're worried about infection and death, suicide is a weird response, since one could instead throw caution to the wind and suicide feels even more than usual like a betrayal, or something like that. I dunno. Like everyone else, I was most definitely surprised to see suicide rates not rising.

Not Covid, but in terms of 'media credibility reaches new all time lows' [the Associated Press has let FanDuel buy the exclusive right to have only its odds quoted on sporting events](#). That would be bad enough, tying AP's sports odds to a recreational book whose odds are distorted and a severely compromised source of the odds of various outcomes, but they've also agreed to *let FanDuel embed widgets of select content into coverage*, turning AP content into ads for a recreational sportsbook with a predatory business model. Where that predatory business model is things like 'buy your way into things like the Associated Press coverage to capture suckers, er I mean customers, who don't know any better.' Why should we assume their coverage of other things is better?

Not Covid, but worth noting that in highly developed WEIRD countries [we are more likely to have someone we believe we can count on](#) than in other places, not less likely.

EDIT: In a story that I misunderstood and spread in error, and isn't a huge deal but also definitely isn't a great look, the FDA is going to hold hearings on some other very dangerous substances (but has not, in fact, lost their f***ing minds any more than they already had before):



TXgrlWatching 📺⚔️🛡️ @VeritasTXgem · May 21

...

The FDA has lost their fucking minds, Related thread below from
@Tweetweetbeyoch

Full list of herbs & Supplements up for hearing on stricter regulations
including parsley, anise, molasses...

Full list here:

fda.gov/media/94155/do...

Updated July 1, 2020

- Mercuric chloride
- Mercuric oxide
- Mercuric salicylate
- Mercuric sulfide
- Mercury
- Mercury olate
- Mercury sulfide
- Methapyrilene fumarate
- Methoxyphenamine Hydrochloride
- Methoxypolyethoxymethylglycol 350 laurate
- Methyl nicotinate
- Methypyrrilene Hydrochloride
- Milk and molasses
- milk solids, dried
- Molasses
- Molybdenum Glycinate
- Monosodium L-Aspartate
- Mullein
- Mustard oil (allithiocyanoate)
- Mycozyme
- Myrrh gum tincture
- Para-chloromercuriphenol
- Parethoxycaine Hydrochloride
- Parsley
- Passion flower extract
- Pennyroyal Oil
- Pentylenetetrazole
- Peppermint Oil and Sage Oil
- Pepsin
- Peruvian balsam (Myroxylon balsamum var. pereirae balsam)
- Phenacaine Hydrochloride
- Phenindamine Tartrate
- Phenolate sodium
- Phenolphthalein
- Phenoxyacetic acid
- Phenyl salicylate (Salol)
- Phenyl salicylate
- Phenyltoloxamine dihydrogen citrate
- Phenyltoloxamine Hydrochloride
- Phosphate fluoride
- Phosphorated carbohydrate



icanteven @Tweetweetbeyoch · May 18

After 57 years of selling N-acetylcysteine (NAC) as an over-the-counter supplement, the FDA decides that it's now a medication that requires a physician's prescription. NAC is an antioxidant compound made up of three amino acids — glutamic acid, glycine and cysteine.

twitter.com/tweetweetbeyoc...

[Show this thread](#)

[The full list is here.](#) EDIT: They are not, as of yet, actively coming for your melatonin, or the other things on this list of suspicious products, which include: Asparagus, molasses, bean (yes, bean), grape seed oil, mustard oil, non-fat dry milk, nutmeg oil, parsley (the bitter herb!), pine tar (your bat needs a prescription now), sesame seeds (so we go with General Tso's now I suppose), soybean protein, and sure, why not, let's just hold nothing back and include sugars. And yeast. Can't have anyone going around making unauthorized breads.

I misunderstood what was going on, which is on me, although one ponders why the misinterpretation was plausible enough to happen in the first place. It turns out that all this is, is that these are components companies want to use in their drugs, all of which require holding hearings (again, no unauthorized breads!) and category 1 here is things they think they'll probably say yes to once the genuflexions are complete, whereas categories 2 and 3,

which contain much of the above list, are things that they'll likely say no to because that's crazy talk and the public must be protected.

We regret and apologize for the error, and thank those who pointed it out.

Meanwhile, [the actual damage is this:](#)



TXgrlWatching @VeritasTXgem · May 22

NAC Banned From Amazon, FDA Says It's Medication

...

May 15, 2021

For now it's still relatively easy to get in practice, but it's basic get-people-through-the-day stuff that several of my friends rely upon.

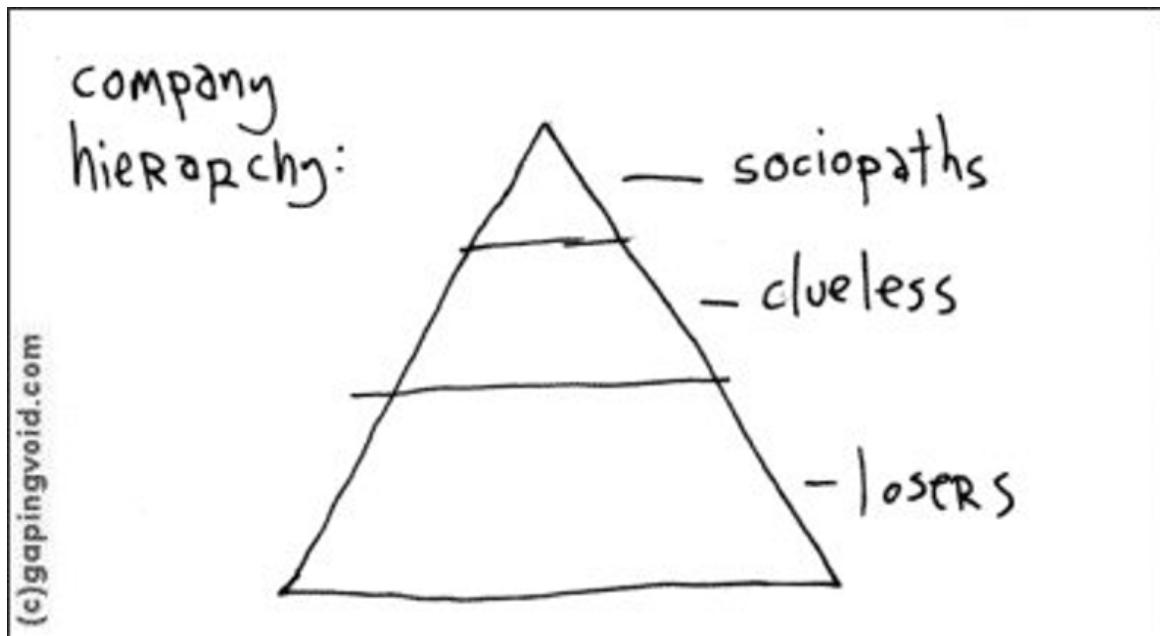
While things are not as crazy as all that, I still stand by the overdetermined conclusion: FDA Delenda Est. Seriously. Burn the buildings to the ground. Salt the Earth. Wave, and [send a message to the next ten generations.](#)

Academia as Company Hierarchy

In [The Gervais Principle](#), Venkatesh Rao argues that the show *The Office* "is not a random series of cynical gags aimed at momentarily alleviating the existential despair of low-level grunts. It is a fully realized theory of management that falsifies 83.8% of the business section of the bookstore."

In this post, I argue that viewing academia through this lens can be equally revealing. But first, we need to discuss the lens itself.

Rao develops this theory of management around the comic *Company Hierarchy* by Hugh MacLeod:



Hugh MacLeod's *Company Hierarchy*

The theory begins by dividing people in an organization into three categories: *Sociopaths* at the top, *Clueless* middle managers, and the average workers as *Losers*. For brevity we will sometimes call this system the SCL hierarchy.

Because these category names were chosen for a gag comic, they aren't great matches for the groups they describe, and can even be a little confusing. Losers aren't losers in the normal sense, they are losers economically — they have struck a bad bargain where they labor for a paycheck. They don't have equity. This bargain can be perfectly rational; it's low reward but it's also low risk. Someone without great natural talents or large amounts of capital may be smart not to take these risks, and in these cases being an economic Loser is often the right call. Most of the characters in *The Office* are Losers, essentially everyone who isn't in management.

Sociopaths may or may not be literal sociopaths — they are like clinical sociopaths in that they are willing to take risks in the service of rewards, and that they are willing to bend social rules to do so. This too is often rational, for people who are willing to take risks and have the ability or capital to do so. It may or may not be admirable. Bending ethical rules is often bad, but bending rules like "don't question authority" or "don't have original ideas" is often good. In *The Office*, executives like David Wallace are Sociopaths.

Clueless might be the most appropriate of the three terms. These are people who are clueless enough that they can be easily manipulated to serve the purposes of the organization. In *The Office*, Michael Scott is the flagship Clueless.

The three categories can be understood as a developmental trajectory, but curiously the development is not the same as the company hierarchy.

Clueless are underdeveloped, and act like children or adolescents. They are motivated primarily by approval from authority figures, and that makes them suckers. Since they are motivated by approval, and because they are otherwise not very smart/not very self-aware, they don't realize that being a wage slave is a bad deal. This is their defining characteristic: they will work very hard for a company that doesn't value them. This is also why they end up in middle-management. They will live or die (metaphorically, we hope) for the company, which makes them very useful to organization Sociopaths, who can use them as fall guys.

In comparison, Losers recognize that being a wage slave IS a bad deal. As a result, they do the minimum necessary to not get fired and keep collecting their paycheck. Again, this is a reasonable thing to do in many cases. For example, you may be a Loser in your day job so you can pursue your real interests nights and weekends. And for many people unable or uninterested to take the risks inherent to being a Sociopath, this is an acceptable bargain. Taking such risks is not only a gamble, it often involves bending or breaking social rules. This is likely to estrange your peers, and so people who are well-adjusted will usually prefer to stay economic losers rather than become isolated.

Sociopaths are the most developed, but maybe over-developed. They understand social dynamics well enough that they begin to have a hard time taking them seriously ("...they are looking for the *truth* about social realities because they think they can handle it."). As a result they stop finding social or status rewards motivating, and as a result tend to value material rewards instead. Unfortunately for them, this tends to make them unhappy in the long run.

There can be state transitions. A Clueless who stops valuing approval from authority figures, or realizes that the company does not actually care about them, will stop over-performing and become a Loser. And a Loser who gains the skills, leverage, or disillusionment needed to take risks will become a proto-Sociopath, do even less work than usual, and look for a chance to get promoted. Once they have some actual bargaining chips, they become a real Sociopath. Losers do not generally become Clueless, because it's rare for people to regress that much.

As a result, the categories aren't exactly personality traits, and they're not exactly descriptions of where you exist in the company hierarchy, they're somewhere in between. The same person might be a Loser in one company, then leave the company to found a startup (where, as a founder, they are a Sociopath). If the company goes under, they might get another Loser job. On the other hand, someone who is Clueless will be an economic sucker wherever they go, so they will probably end up in Clueless positions in every company they become a part of.

In addition, Rao describes four languages (and a fifth semi-language) that the three groups use to communicate. More on this later.

A critical point, and one that is easy to miss, is that the Clueless serve two main purposes in an organization. First, since they have misplaced loyalty to the organization, Sociopaths can use them as cat's-paws and fall guys. They can get them to take risks by proxy, and have them take the fall for these or other risks as necessary. Second, they serve to insulate the Sociopaths from the Losers, or as Rao puts it, "to provide a buffer in what would otherwise be a painfully raw master-slave dynamic in a pure Sociopath-Loser organization."

In this post I use this theory of management to analyze the dynamics of academia. There are a couple of good reasons to do this. Applying the theory to a new area is a good way to

explore it and test its power as a framework. It can help explain some parts of academia that might otherwise seem confusing. And finally, I think it can explain some aspects of the current culture war (so *caveat lector* for those of you who are wary of such things).

1. Academic Hierarchy

A university is an organization just like any business. Dunder-Mifflin has many branches; the Scranton branch is just one branch of many. A university has many departments; each department is just one department of many. Or we could say, academia is divided up into many different universities; each university is just one university of many. So we might analyze this system at the department level, at the university level, or at the all-academia level, but it doesn't make much of a difference.

To begin to analyze a system from the SCL perspective, we first need to figure out which of the three groups people belong to. In case you wonder where my loyalties lie, know that as a PhD student, under this system of analysis I am decidedly a Loser. But more on this later.

Rao names a number of signs by which we can identify the three groups:

Clueless

- Over-perform for their organization, marking themselves as suckers
- Identify with the organization, to the point of having loyalty (which is not requited)
- "...cannot process anything that is not finite, countable and external. They can only process the *legible*."

Losers

- Do the bare minimum to stay in the organization (and the more they under-perform, the more likely they are to become Sociopaths)
- Conflate material (e.g. money) and emotional (e.g. status) rewards; "cannot process the material aspect of anything that involves strong emotions"
- Jockey over social status, rather than material power (Sociopaths) or approval from authority (Clueless)

Sociopaths

- Play for real (material) stakes
- "about recognizing that there *are* no social realities"
- Perform "game design" for the organization, arranging for social competition (Losers) and medals and ranking schemes (Clueless), while collecting material rewards for themselves

Academia has many different sub-groups. This is not unlike business — the Scranton branch has warehouse staff, support, and sales, as well as a manager. In academics, however, the structure is less immediately hierarchical, and so it is worth examining this system at every level. We will skip a few levels to simplify. In particular we will ignore MA students, but that's ok, they're used to it.

1.1 Undergraduates

Undergrads as undergrads do not fall into the SCL hierarchy. After all, they're not part of the organization — to a university, undergrads are consumers, not employees. But undergrads who aspire to become academics ARE semi-employees, usually through serving as research assistants (RAs).

Most RAs are Losers. They are engaged in a bad economic deal — exchanging their labor for a recommendation letter, a long shot chance at graduate school. They're not even paid, so in most cases this is an even worse deal than working for a paper company. Undergrads don't tend to be experienced enough to understand this the same way most workers do, but they usually have an intuitive sense for it, which is why few undergrad RAs put in long hours or show much devotion to their lab.

A small number of RAs, however, *do* devote themselves to their lab and work heroic hours on thankless research projects. If you've spent any time in academia, you recognize this character. RAs who act this way are Clueless — remember, the defining characteristic for this group is over-working themselves for an organization that couldn't care less about them and doesn't reward them. In this way they send a strong signal that they are suckers who can easily be exploited. Unsurprisingly, these RAs are destined to go far.

RAs cannot really be Sociopaths because undergrads, as Rao would put it, are playing with monopoly money. Having already paid tuition, they have almost nothing academia could want from them. Any RAs with Sociopath tendencies express instead as low-performing Losers. They are in the system only to look for opportunities. A hypothetical true Sociopath RA would need to have either their own funding or truly blockbuster ideas, and would turn them into first-author (or even single-author) publications in good journals, or better yet, use their ideas to do something like get a book deal or found a startup.

Especially cynical RAs will choose projects that appear to be very difficult but are secretly very easy — for example, data coding tasks that can be automated with a simple script — in order to appear Clueless on grad applications.

1.2 PhD Students

Two kinds of students are selected for PhD programs. The first are those who have proven themselves, as RAs, to be utterly Clueless. This looks like accomplishing many impressive projects as an undergraduate (for neither pay nor credit) and having very impressive recommendation letters (and nothing else) to show for it. For related reasons, these students also tend to have very good grades. As discussed, this singles them out to faculty as suckers who can be easily exploited for lots of labor. Faculty are probably not self-aware enough to see it this way, but in their own jargon, they recognize the students will be "productive".

This is part of why burnout is such a big problem in graduate school. Not because there is a problem with the system (though there is), and not because faculty push students to overwork themselves (though some do), but because there is an enormous selection pressure to promote Clueless RAs into PhD programs. In many ways this is like the promotion of the Clueless that Rao describes in the business world.

On another level, the Clueless do very well as PhD students. As Rao says, to the Clueless "everything worth learning is teachable, and medals, certificates and formal membership in meritocratic institutions is evidence of success." So while they find PhD programs stressful, it's at least stressful in a way they understand.

However, there are only so many Clueless RAs in a given year. In addition, everyone can tell that the work done by Clueless undergrads is not very creative; it looks less like independent work, and more like pulling multiple all-nighters on someone else's project. As a result, faculty can tell that this student "may not be able to do original work".

So the second kind of students selected for PhD programs are the Losers who have some Sociopathic tendencies. As mentioned, most RAs are Losers. The ones who float to the top tend to be those who trend Sociopathic, because this tendency will inspire them to create something they have at least partial ownership over, and this ends up looking like the ability to come up with original lines of work. Faculty value this, since it leads to a different kind of productivity, and so these students are often admitted as well.

In addition, faculty who lean Sociopathic will be tempted to admit students of the same stripe, because they value having someone around who sees things the same way they do. This is true even if neither of them are true Sociopaths.

So in admission to PhD programs you tend to have an even split of Clueless and Losers who trend Sociopathic. Beginning to get some real power, and growing steadily more disillusioned, some of these Losers will become true Sociopaths. This is pretty rare, however, since PhD students rarely have the power or will to play at that level. Those that do often get summer internships for major companies, leave early to do something like found a startup, or spend all their time secretly working on a side project instead of attending to their graduate research.

Losers without Sociopathic tendencies don't often make it to grad school, and don't often stay when they do, because true Losers put community and their emotional life first. This forms a feedback loop. There is not much of an emotional life in grad school, so the people who value it leave, so there is not much of an emotional life in grad school...

1.3 Faculty

There are even fewer faculty positions than there are spots in PhD admission, so at this level we see another round of strong selection pressure.

This is where we run into the first major surprise from analyzing academia from a SCL perspective. Because while you might expect me to say that faculty are mostly Sociopaths, in fact they are almost entirely Clueless. I think this is true of both tenured and untenured faculty, so I will treat them together from here on.

The defining characteristic of the Clueless is over-performing relative to their level of reward. It's hard to imagine a better way to describe university faculty. Everyone knows that the unappreciated workload of faculty is massive, and the unpaid workload even larger. They teach classes for humorously low wages, edit journals for "prestige", and perform peer review for nothing at all.

The Clueless "cannot process anything that is not finite, countable and external. They can only process the *legible*." Certainly this describes the behavior of faculty, literally counting lines on their CV, grubbing for citations, breathlessly calculating their [h-index](#).

This sounds more than a little abusive, and it is, but in many ways, these people are attracted to academia for exactly these reasons. "The Clueless can process the legible," says Rao, "so a legible world is presented to them." In this way they find it very comforting.

Rao even has a whole section on the [humor used by each group](#), and while it is a little hard to explain, faculty definitely have Clueless humor. Losers make jokes for the group, and often use forms of humor that encourage the group to join in. Sociopaths make jokes for themselves, that other people don't get, and often don't even notice. But the Clueless make jokes that are antisocial and yet also not for their own benefit. I can testify that sitting in on faculty meetings is a lot like sitting in on the lunch table at the local high school. Different faculty members will all try to make the same joke, one after another. They will make jokes that you can't build on, to which there is no possible response. They will say a joke, and then when no one laughs, they will say the same joke again, only louder. Rao's other note on the Clueless is that they make you *cringe* with their actions, and faculty humor is nothing if not cringeworthy.

Despite ostensible appearances to the contrary, faculty are not at the top of the pyramid in academia. Instead, they are academia's middle-managers.

Some fields are probably more this way than others. A field where there is more room for material rewards, where labs can land huge grants, may be more likely to attract Sociopaths.

But on the other hand, "me win most grants" is also very legible.

Of course, there are some Loser faculty and some Sociopath faculty in every field. The Losers are distinguished by being very aware that being a professor is an economically raw deal. They tend to be people who have a passion for research or teaching and accepted this bad deal because it let them fulfill themselves in their other calling. I know one professor who never applies for grants and never takes grad students. As a result his department has pushed him into a tiny office, but he doesn't care. He just wants to do his independent work without being bothered, and his tenured position has landed him a situation where he can focus on that.

Sociopath faculty are distinguished by using their faculty position as part of a wider portfolio, or as a stepping-stone to other things. Any faculty member who is involved with a startup or has several popular trade books might be a Sociopath. Steve Pinker, who clearly aspires to be more of a public intellectual than "merely" a Harvard Professor, is almost certainly a Sociopath under this system (and maybe in general).

So why are university faculty almost universally Clueless? I think there are two main reasons. First of all, doing hard work for little reward marks you as exploitable, and two levels of filtering for that trait leads to an inevitable conclusion.

Second, the Clueless serve a special role in a large organization, that of separating the Losers from the Sociopaths at the top. "Without it," says Rao, "the company would explode like a nuclear bomb, rather than generate power steadily like a reactor."

1.4 The Top???

Universities are organizations. But as we've just seen, the faculty are not ruling the roost — they are all Clueless. Who is working these machines from the top?

My first instinct is to say that these are the people directly above the faculty, maybe the deans. This makes some sense — I have almost no experience interacting with any dean. But [the stereotype of the university dean](#) seems a lot more Clueless than Sociopath. In large organizations, there may be many layers of Clueless middle-managers, and universities are very large.

Maybe you have to go higher. The board of trustees? The president of the university? Maybe, but my limited experience of these people is that they seem pretty Clueless as well. The president gets paid a lot, but maybe seems like a potential fall guy, which would make him Clueless. But who would he be taking a fall for? I don't know; but given that I am a Loser in these organizations, and I don't even know who my local Sociopaths are, the system seems to be working as intended.

It's also possible that universities have evolved to be a truly headless entity, but I find it hard to believe that *someone* isn't profiting off of this system. At their heart, many major universities are real estate companies. Between them, Harvard and MIT own most of Cambridge. NYU is slowing buying up as much of Manhattan as they can. What makes these companies special is just how much Clueless flash they have been able to put between themselves and the public eye.

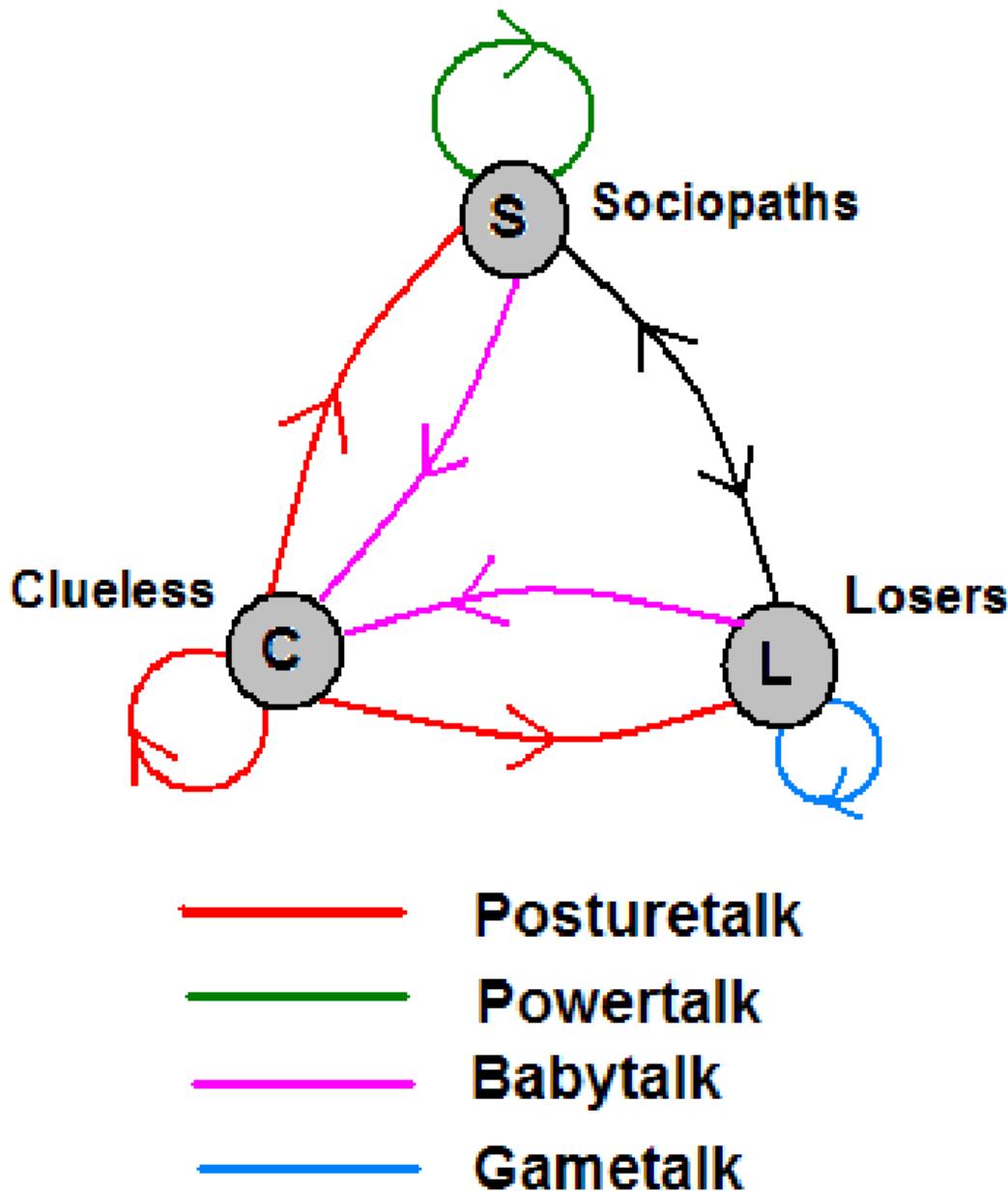
(In fact, the fact that many universities are secretly real estate companies makes me wonder if there might be a [Georgist interpretation](#) along similar lines — Universities are landlords, faculty are capital/bosses, and grad students/undergrads are labor. This matches the three categories of SCL surprisingly well.)



This reinforces the value of having most of the faculty be Clueless. They work long hours to be very distracting. They have brand loyalty to the university, even when that university is a monster. They will take risks for the university in exchange for nothing more than "medals, certificates and formal membership." And when the university needs someone to take the fall for a risk that went wrong, the faculty are always there to take that fall.

2. Academic Talk

In [Part II](#) of his essay, Rao describes the four (plus one) languages that the different groups speak with one another.



Languages spoken in organizations

Powertalk is the language that Sociopaths speak among themselves. It is the most interesting language on its own merits, but as academics are almost never Sociopaths (in the SCL sense of the term), we won't discuss it in depth here. Read Rao's essay for the fascinating details.

Posturetalk is the language that the Clueless speak to everyone; indeed, "they don't have an in-group language since they don't realize they constitute a group." In academia, the typical Clueless is a faculty member, so this is the jargon you hear from faculty — differing slightly by field, but vague and stuffy in the ways you expect.

Babytalk is the language the other two groups use to address the Clueless. Rao emphasizes that Babytalk "seems like Posturetalk to the Clueless." In the case of academia, this is something that sounds like jargon to faculty, but is actually dismissive of them. I'll further emphasize that the purpose of Babytalk is to allow the other two groups to *manipulate* the Clueless, which in this case means manipulating faculty and senior PhD students. More on this in a minute.

Finally, Gametalk is the language spoken among Losers as part of their pecking-order games. "Gametalk leaves power relations unchanged because its entire purpose is to help Losers put themselves and each other into safe pigeonholes that validate do-nothing life scripts." If you are cynical enough, maybe this will sound like the language of undergraduates to you too.

That black line on the diagram goes officially unnamed per Rao, because one of the functions of the Clueless is to provide a buffer between Sociopaths and Losers, so they never have to / get to communicate. But he says this is "an unadorned language you could call Straight Talk if it were worth naming." In academia the Sociopaths are so far removed from the Losers that this is not worth considering.

2.1 Academic Babytalk

Clearly the most interesting of these languages, in the context of academia, is Babytalk, the shared language spoken by both Losers and Sociopaths when they want to placate, manipulate, or distract the Clueless. Notably, to the Clueless it sounds largely like their own language, Posturetalk, so to an academic, this will sound something like academic jargon.

I submit that in the modern political climate, "woke" language is an important dialect of Babytalk.

It fits all the criteria. Woke language borrows the style of academic jargon and sounds a lot like academic speech to faculty. None of them are aware that they are being condescended to.

Woke language rarely changes anything substantive — in Rao's words it "leaves power relations basically unchanged" — but it is very effective in manipulating Clueless faculty and senior PhD students. The more progressive among them will readily back down out of agreement with the ideology, and the less progressive will back down out of their irrational and disproportionate fear of being "cancelled", this despite the fact that in reality professors are almost never "cancelled" for anything.

Finally, woke language can be identified as Babytalk by the fact that neither Sociopaths nor Losers use it among themselves; they only use it in communicating with the Clueless.

Certainly we don't expect the academic Sociopaths, whoever they are, to use woke language in their personal dealings. But some of you may be surprised to learn that students don't use woke language among themselves either. Now, their own ingroup language, or Gametalk, does involve similar issues of race, class, and gender, but it's distinct from the Babytalk they direct at faculty. Those of us Losers who have spent a lot of time in progressive spaces, I'm sorry to reveal, can easily distinguish between the two.

Some people are even consciously aware of using woke language to condescend to or manipulate the local middle-managers. I happened to be speaking with an undergraduate student recently. This student is not only from a notoriously progressive, even radical college, they also fit many of the personal stereotypes of "wokeness" — they are queer, asexual, neurodivergent, etc. But at one point when we were discussing a problem they were facing with the administration, they told me:

I feel like putting my case in personal terms is wrong because my race shouldn't matter, but I feel like our Dean of Students will only listen to me if I frame it as a threat to me as a woman of color.

This matches my general experience as well. Faculty and administrators prefer to ignore student concerns, but they freak out when presented with issues of race or gender. Students are in tune with this and learn that they have to frame things this way to have any hope of getting anything done.

Presumably the local Sociopaths are aware of this as well. As a result it is not surprising that both Loser and the Sociopath academics use woke language as part of their Babytalk. Nor is it surprising that many professors *experience* a world where everything is framed in the woke terms of race, gender, and class. This is not the dialogue spoken in the great wide world, but these professors have made it clear that this is the only kind of framing they will pay attention to, and people have adjusted their messaging accordingly. I understand that this puts them in a cold sweat, but it's hard to feel sympathetic.

This is further emphasized by the fact that when faculty (who are Clueless) try to describe "woke ideology", they fail miserably. This will be invisible to outsiders because Babytalk is designed to pass for the faculty's native Posturetalk, but believe me, professors could not remotely pass the Woke Turing Test. Young people today are not afraid of being challenged; they do not reduce themselves to their skin color or their genitalia. They are not "confused" about their gender. When college professors express concern about this sort of thing, they're just showing that they do not even understand the terms they are being condescended to with.

Because woke language sounds like Posturetalk to the Clueless, some of the terms have been adopted by Clueless PhD students and faculty. At this point, we shouldn't be surprised to hear the Clueless using woke terms with the other groups and even with each other. When the Clueless use it, however, it is totally ineffectual, and [never sounds quite right](#).

Since there are Clueless PhD students and even Clueless undergrads, you will sometimes hear earnestly outrageous "woke" messages coming from them. But the main use of woke language is to cajole or frighten faculty into submission. There's no real power here, it's a bluff — Rao says, "Posturetalk and Babytalk leave things unchanged because they are, to quote Shakespeare, 'full of sound and fury, signifying nothing.'" But because faculty are concerned about and/or afraid of these issues, they can often be convinced to back down by the use of this kind of language.

2.2 Actually Being Cancelled

What about those faculty members who are cancelled? Here we return to the other organizational role of the Clueless, that of being a convenient scapegoat.

It's hard to know for certain, as these decisions are deliberately obscured, but I suspect that many of these faculty were fired for reasons unrelated to wokeness. Rao describes how Sociopaths set up bureaucracies that are designed to be byzantine and become clogged with appeals. When they want to keep something from happening, they let the appeals pile up. But when they want to make something happen, the Sociopath who handles the exceptions lets the right appeal jump the queue.

So in the few (rare) cases where a professor was actually cancelled by their university, I suspect that what happened was that the university wanted to fire them for some unrelated reason first. To protect the people at the top, however, and redirect the blame to the students and the bureaucracy, they first waited until a student made a complaint about the professor in question. This complaint then jumped the queue and was promoted to the level of an issue, and the professor was fired, ostensibly as the result of the student complaint.

This is hard to prove but it makes sense when we observe a professor being fired for what appear to be very flimsy reasons.

This is not the only way faculty can be made to take a fall for something. It's also possible, for example, that if some other scandal were about to come to light, a school could fire some professor for "wokeness" reasons as a way to distract from the other issue.

3. Other Implications

Some final thoughts on implications of this analysis.

3.1 Graduate Programs

Everyone knows that graduate school is kind of hellish. People are overworked and underpaid. Most of them slowly become aware they will never get an academic job. They burn out, suffer breakdowns. A lot of work goes into making it a better place for everyone.

But viewing it through an organizational lens suggests that these problems can't be solved: they're inherent to the system, and can't be gotten rid of. Rao says that theory of management is "based on the axiom that organizations don't suffer pathologies; they are intrinsically pathological constructs." Again to quote Rao directly, "It is *designed* to fail in ways that achieve unspoken Sociopath intentions, while allowing them to claim the nobler explicit intentions enshrined in the law. "

It's even possible that attempts to make things better will make things worse, as it gives the Sociopaths at the top an opportunity to fiddle with the system to better suit their needs. Rao says of bureaucracies that you should "periodically attempt to 'reform' it through means that only ensure it gets worse (adding complexity)." If you have spent any time in academia, this will sound familiar to you.

If this perspective is correct, the only thing to be done is to abandon grad school altogether. But as long as there are Clueless students who can be recruited to feed the machine, I'm afraid this cycle will continue.

The same probably goes for graduate admissions and faculty hiring. These processes are perverse not as a bug, but as a feature. At some level most universities really *do* want you to hire the person with the longest CV, regardless of how much crap is on it, because this allows you to pick out the Clueless applicant who is the #1 biggest sucker. Universities can find many uses for such suckers.

3.2 Science Generally

An interesting implication is that the reason science sucks so much these days is that mainstream science has been "captured" to serve as the [intent-obscuring bureaucracy](#) of a set of major organizations.

In the old days, most scientists were Sociopaths, bored of life, pursuing meaning through solving mysteries. A small number of them had day jobs as Losers, and did science on the side as their hobby. But "science" is now dominated by the group with the lowest level of development, the Clueless, which can only bode poorly.

This is in line with my more general feeling that we should expect most scientific progress to occur outside the academy, though it is something of a problem that so much science funding is captured in this way.

Teaching ML to answer questions honestly instead of predicting human answers

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

(*Note: very much work in progress, unless you want to follow along with my research you'll probably want to wait for an improved/simplified/clarified algorithm.*)

In this post I consider the particular problem of models learning “predict how a human would answer questions” instead of “answer questions honestly.” (A special case of the problem from [Inaccessible Information](#).)

I describe a possible three-step approach for learning to answer questions honestly instead:

1. Change the learning process so that it does not have a strong inductive bias towards “predict human answers,” by allowing the complexity of the honest question-answering to “pay for itself” by constraining the space of possible human-models.
2. Introduce a bias towards the intended model by using a more complex labeling process to answer questions where a human answers incorrectly.
3. Be really careful to avoid penalizing honest answers, by only judging comparisons between two answers where we are confident one is better than the other and getting the model to help us.

I don’t know whether this problem is a relatively unimportant special case of alignment, or one of the core difficulties. In any case, my next step will be trying to generate failure stories that definitely *cannot* be addressed by any of the angles of attack I know so far (including the ones in this post).

I think it’s relatively unlikely that almost anything specific I said here will really hold up over the long term, but I do think I’ve learned *something* about each of these steps. If the ideas end up being important then you can expect a future post with a simpler algorithm, more confidence that it works, clearer definitions, and working code.

(Thanks to Ajeya Cotra, David Krueger, and Mark Xu for discussions about this post that helped clarify it.)

The problem

Suppose that we train a model to answer questions in natural language about what will happen in the future (“Will Alice take the train home tonight?”) given a data about what’s happened in the recent past (e.g. video recorded over the last few days).

To generate training data we take the video from time t_0 to t_1 and get humans to answer questions about what happened between time t_1 and t_2 . Then we do supervised learning on these (video, question, answer) triples.

The **intended model** has two parts: (i) a model of the world (and inference algorithm), (ii) a translation between the world-model and natural language. The intended model answers questions by translating them into the internal world-model.

We want the intended model because we think it will generalize “well.” For example, if the world model is good enough to correctly predict that someone blackmails Alice tomorrow, then we hope that the intended model will tell us about the blackmail when we ask (or at

least carry on a dialog from which we can make a reasonable judgment about whether Alice is being blackmailed, in cases where there is conceptual ambiguity about terms like “blackmail”).

We want to avoid models that generalize “badly,” e.g. where the model “knows” that Alice is being blackmailed yet answers questions in a way that conceals the blackmail.

One source of generalization failures is the **instrumental model**, which involves: (i) a world-model similar to the intended one, (ii) knowledge about the training-data-generating process, within that world model. The instrumental model answers questions by predicting what answers the training-data-generating process would produce. This can generalize badly to any situation where the humans generating the training data are wrong (e.g. because they are being deceived or manipulated).

Why might a neural net end up learning the instrumental model? There are two big reasons:

1. The instrumental model may be much simpler than the intended model. Both of them contain the same world model. If that world model is good enough, then it also knows how to make predictions about humans and understands a lot about the training process. From there, the instrumental policy just needs to point to a relatively simple part of the model: what gets inserted into the training set? In contrast, the intended policy needs to bake in a ton of rules for translating between the internal model and natural language. Doing this naively requires encoding all of the meaning of natural language, which has a lot of incompressible data. It’s not clear if there is a more economical way to represent the intended model (without recreating the pathologies of the instrumental model).
2. The instrumental model is guaranteed to accurately fit the training distribution, while the intended model might disagree if the humans make a mistake. If we keep collecting training data indefinitely, sometimes make mistakes, and don’t do something clever to fix or ignore them, then we will end up converging to the instrumental model no matter how good our inductive bias is.

Aside on imitative generalization

[Imitative generalization](#) suggests a possible strategy for overcoming problem 1: learn a neural net which is only a world-model, and then use a human prior over “translations” that relate the world model to human concepts (instead of training a neural net that also does the translating, and so implicitly using the neural net prior over translations).

In this post I’m going to mostly ignore that idea and think about neural nets that implement both world-model and translation, allowing them to be integrated tightly together. That means we need a different solution to problem 1. I still think imitative generalization is part of the solution, but in the context of this particular problem I think we just need to improve the prior without changing the representation. So throughout this post you should completely ignore imitative generalization and read me as taking a more traditional perspective.

Step 1: make the intended model pay for itself

(I think this is the most confusing/ambitious/uncertain step.)

Intuition

The intended model is complex because it needs to specify how to use natural language to talk about the world model.

The intended model loses to the instrumental model because in some sense this information was *redundant*—there are already humans in the world model, and knowing enough about humans to predict their behavior means the world model already knows how to ground out natural language.

Put differently, the intended model redundantly specifies (humans) and (language). Whereas the instrumental model specifies (humans) and infers (language) from (humans). This is bad because we get out a predictive model of humans, who generalize poorly.

What I want to do instead is specify (language) and then infer much of (humans) from (language). The hope is that we then get a simple model of language use, which generalizes better.

My hope in step 1 is to remove the inductive bias in favor of the instrumental model; we're not going to introduce any kind of positive bias towards the intended model.

Rough plan

I want to consider models that learn to predict *both* “how a human will answer question Q” (the instrumental model) and “the real answer to question Q” (the intended model). These two models share almost all of their computation—which is dedicated to figuring out what actually happens in the world. They differ only when it comes time to actually extract the answer. I'll describe the resulting model as having a “world model,” an “instrumental head,” and an “intended head.”

We start by collecting a high-quality dataset, in which we believe that the intended and instrumental models agree.

We're then going to imagine filling in the parameters of our learned model in two stages:

- In the first stage we will learn both the intended and instrumental heads, as well as a small part of the world model (but none of the facts about how humans answer questions).
- In the second stage we will fill in the remainder of the world model (including all of the facts about how humans answer questions). In this stage, we **condition** on the instrumental head and the intended head agreeing on the training set.

You can think of this as a generative model where we first fill in stage 1 randomly, then we fill in stage 2 randomly but do it over and over again until we happen to get a set of parameters for which the two heads agree. Then the learning process tries to infer the parameters that were most likely to give rise to the training data.

A simplicity prior could potentially have learned this two-stage model directly. Namely, it could specify the bits of the first stage, and then it could use the “heads agree” assumption to compress the second stage. The problem is that the decompression is very expensive (it involves solving a hard optimization problem in stage 2, which is now done *at test time*), and so this involves learning a computationally complex model. That's ineffective at test time, and it also means that the learning process will be slow.

(There might be some way that a simplicity prior could have achieved this compression even with test-time compute constraints, but I can't see a way in general and so I wouldn't want to bank on it.)

This two-stage learning algorithm fixes the test-time slowness, but learning is still computationally expensive. But I think it's good to move the hardness out of the learned model and into the learning algorithm—that lets us treat the whole thing as an algorithmic problem, of finding an efficient way to compute the posterior for this kind of generative model. I'm quite optimistic about this problem, which I'll discuss in a few sections (where

we'll also define the algorithm more precisely in the context of SGD over neural nets rather than random program search). But first, let's talk about why this whole thing might work if you could actually implement it.

Why might this work?

For concreteness, let's say that the world model requires a trillion ("N") bits to specify, the intended head costs 10,000 bits, and the instrumental head costs 1,000 bits. If we just applied a simplicity prior directly, we expect to spend $N + 1,000$ bits to learn the instrumental model rather than $N + 10,000$ bits to learn the intended model. That's what we want to avoid.

Under the alternative proposal, suppose we spend $N/2 + 11,000$ bits in the first stage to specify half of the world model and both heads. If we didn't do rejection sampling in the second stage we'd need to spend $N/2$ more bits, for a total cost of $N + 11,000$ bits. That's even worse.

However, if our dataset was rich and high-quality then we are able to save many bits in the second stage. If the dataset is high-quality then the intended model results in the two heads agreeing. But if the dataset is rich, then virtually all models result in the two heads disagreeing.

The magnitude of savings in the second stage is basically equal to the (log) *fraction* of possible world models that cause the instrumental head to agree with the intended head.

How common is agreement? If the instrumental head is basically reporting human answers to questions, and if the human model is basically unspecified in the first stage, then this is the fraction of human models that use natural language "correctly" on the training set. If the training set is rich enough, this is quite close to the fraction of human models that use natural language "correctly" overall.

Now we have two priors over ways to use natural language: we can either sample the intended head at random from the simplicity prior (which we've said has probability $2^{-10,000}$ of giving correct usage), or we can sample the environment dynamics from the simplicity prior and then see how humans answer questions. If those two are equally good priors, then only $2^{-10,000}$ of the possible humans would have correct usage, so conditioning on agreement saves us 10,000 bits.

So now we spend $(N/2 + 11,000) + (N/2 - 10,000)$ bits altogether, for a total of $N + 1,000$. This is comparable to the instrumental model. There may be other ways to learn even more efficiently, but *prima facie* this suggests that our two-headed model is roughly as easy to learn as the instrumental model.

We've learned a model with two (symmetrical) heads. On the training set they agree, but in order to generalize well it's important to us that we use the "right answer" rather than the "predict what a human would say." In section 2 we'll describe how to provide a small boost to the intended model—this both breaks the symmetry between the two heads, and allows the two-headed model to be actively *preferred* to the instrumental model.

This argument doesn't work if random world models actually induce a *better* prior over language use than the simplicity prior. (As a silly example: consider the limit where we follow all of physics from the big bang and wait for the evolution of language + language-users to take over the world. Then the prior from a random world model seems likely to be *much* better.) I think those cases are totally fine when we combine this idea with iterated amplification, but I don't want to get into that complexity here since I think it's mostly orthogonal to the novel algorithm.

The actual algorithm

We're trying to learn a neural network that takes some data (e.g. video) as input and has two question-answering heads. We'll parameterize these networks by a vector of real numbers θ . We can think of SGD as doing regularized regression where the regularizer is the L2 distance from the initialization θ_0 (the only difference is that SGD linearizes the loss). Write $C(\theta)$ for the average difference between the two heads on the training set, such that we'll want C to condition on C being small.

Then we can describe the kind of two-stage learning process as the following generative model:

$$\begin{aligned}\mathbb{P}(\theta_1) &\propto \exp\left(-\|\theta_0 - \theta_1\|_2^2\right) \\ \mathbb{P}(\theta_2) &\propto \exp\left(-\lambda C(\theta_2) - \|\theta_1 - \theta_2\|_2^2\right)\end{aligned}$$

Where λ is some large parameter (that needs to be finite only to keep the optimization smooth so that we can use SGD).

Basically this says that you draw θ_1 from the prior, then you draw θ_2 from the prior after conditioning on the two heads agreeing.

Write $L(\theta)$ for the log probability of the answers using the first head given parameters θ . Then the unnormalized posterior log probability of (θ_1, θ_2) is:

$$\begin{aligned}L(\theta_2) - \|\theta_0 - \theta_1\|_2^2 - \|\theta_1 - \theta_2\|_2^2 - \lambda C(\theta_2) \\ - \log \int_{\theta'_2} \exp\left(-\lambda C(\theta'_2) - \|\theta_1 - \theta'_2\|_2^2\right)\end{aligned}$$

Intuitively: the first term L is fitting the data, the next two terms are the prior, and the fourth term C is expressing the agreement condition. The final term with the integral represents the bits saved by rejecting all the inconsistent models—if there are very few consistent models then the integral will be small and many bits will be saved.

We're interested in either finding the parameters with maximum posterior probability (MAP) or sampling from the posterior. The first four terms are straightforward enough to handle, but naively the last term requires two nested loops of optimization—for each value of θ_1 that we consider, we need to do an expensive optimization to figure out the value of the integral.

I'll write a longer post about this algorithmic issue in the future, but I'm actually fairly optimistic that we can do inference despite the nested integrals. The basic reason to have hope is that the optimal values of θ'_2 change slowly as a function of θ_1 , so we can do alternative gradient steps instead of running SGD again to compute each update to θ_1 .

Some remarks:

- This two-stage optimization process is basically the same as the GAN situation, where you also use the assumption that the optimal discriminator changes slowly in order to

avoid nested optimization. I'm optimistic that optimization in this case can be made significantly easier/stabler and can be done while remaining competitive with traditional SGD. I think we can *almost* optimize this quantity by updating θ_2 by SGD as normal and updating θ_1 at each step to minimize the difference $C(\text{new } \theta_2) - C(\text{old } \theta_2)$.

- Most of the difficulty of the integral comes from the need to integrate over the entire space rather than just a neighborhood of the current parameters. But what we actually care about is explanatory power *along the trajectory of SGD*—we can view this as compressing the model by compressing the individual gradient descent steps, rather than trying to compress the full model. This makes me significantly more confident that there is a fast/simple way to do what we want to do, and I think it is going to lead to much more robust algorithms. Overall I don't feel too concerned about the algorithmic difficulty posed by this setup.
- When differentiating the consistency test C we should treat the intended head as fixed rather than differentiating through it. This removes SGD's incentive to achieve consistency by e.g. making sure the world is simple and so all questions have simple answers. I currently feel very uneasy about the situation. The algorithm looks a lot more likely to work when we make this change, but it makes it harder to reason about the algorithm globally and I'm very scared about the fact that we can't do an analogous trick in the program search case. I think the best hope may be to take the gradient-compression perspective suggested in the last bullet point. Overall this may amount to a significant change in this algorithm.
- There seems to be a hard limit on the bits we can "save"—we could always make the two heads equal by simply changing the instrumental head to be equal to the intended head (or vice versa). Even if we tried to rule this out by preventing θ_2 from messing with the intended head, I think that SGD could probably route around this (e.g. by physically implementing the intended head within the world model). So I think we should *at best* expect to recoup all the bits we spent specifying the intended head. I don't believe this algorithm can introduce an inductive bias in favor of the intended model, it can just level the playing field.

Step 2: Give the intended model a leg up

At best, step 1 merely removes one inductive bias in favor of the instrumental model. But we are still left with a massively underspecified learning problem; we could end up with many different functions that all agree on the training set and have approximately equal prior probability. In order to learn the intended model, we need to somehow specify which one we really wanted. In this section I'll describe a possible approach.

Rather than using a single labeling process, we will have two labeling processes, a "simple" one and a more reliable "complex" one. The idea is to use the "complex" labeling process to identify examples where the "simple" process errs. This hopefully rules out the "simple" labeling process as a good model of the data, leaving us with the intended model as the best one.

More precisely:

- In the "simple" labeling process we simply ask a human (embedded in the world) to answer questions about what is happening and put it into the dataset.
- In the "complex" data-generating process we record a bunch of data about the world (camera footage, answers from many different humans, random traces from computer programs...) and then we make our best effort to figure out the right answer. This may involve consulting ML systems for help (as in iterated amplification). We can also use this process to collect hypothetical data, by using a generative model to make predictions and then applying our best efforts to the hypothetical.
- We have a "small" dataset where we believe the simple labeling process arrives at the correct answer.

- We have a “large” dataset that includes the small dataset as well as at least a few cases (say 10% of the size of the small dataset) where the simple labeling process makes a mistake that is identified by the complex labeling process. This may include some hypothetical data, e.g. showcasing situations in which the simple labeling process could make an error.
- We use the “small” dataset in order to define the consistency predicate C needed in step 1 (i.e. we condition on the two heads agreeing on the small dataset). But we use the “large” dataset in order to define the likelihood L (i.e. we condition on the first head correctly predicting the large dataset).

Why think this would work? Imagine that there are now three possible models—the intended model, the simple model (that predicts the simple labeling process), and the complex model (that predicts the complex labeling process). Suppose that these models satisfy the following properties:

- All models agree on the small dataset. The complex model and the intended model agree on the large dataset. There are inputs in the large dataset where the simple model disagrees with the other two.
- The simple model is simpler (i.e. higher prior under SGD) than the complex model. There are a few reasons this could happen (and we just need one), but I am most optimistic about achieving this when the complex labeling process itself uses ML systems as assistants or to generate hypothetical situations. It’s not easy for the ML system to represent or simulate another model as large as itself, and so it seems like it needs to do something more tricky (e.g. figuring out the equilibria of a game played by multiple peer models). I don’t think that it necessarily takes a huge number of bits to do this well, but I think it may well be more than the simple model. (My biggest concern is that the “simple” model may also need to use similarly-complex tricks to be computationally efficient.)
- (For concreteness, I’ll say that the simple model takes 1,000 bits, the complex model takes 2,000 bits, and the intended model takes 10,000 bits.)
- The small dataset is still rich enough that you could infer correct language usage from it, i.e. the consistency condition on the small dataset alone suffices to recover all 10,000 bits required to specify the intended model.
- Agreement between the complex and simple heads on the small dataset isn’t nearly that surprising. Under typical conditions the complex labeling process will basically just defer to the simple labeling process; knowing what counts as an “error” tells you something about humans but it’s not very much relative to the complexity of natural language. (This may also be a dicey assumption.)

In the two-headed model, each of the heads could independently learn any of the (simple, complex, or intended) heads. This yields 9 models; which of them could we end up learning?

- (intended, simple) is what we’re hoping for. It costs 11,000 bits to specify both heads, and we save 10,000 bits. So the total cost is $N+1,000$.
- (intended, complex) or (intended, intended) would also be fine to learn, but cost $N+2,000$ or $N+10,000$ bits.
- (simple, X) doesn’t match the data well, since there are datapoints in the large dataset where the simple labeling process goes wrong.
- (complex, simple) costs at least 2,000 bits to specify. This would be a bad model to learn, but we will avoid it as long as agreement between the complex and simple models has probability at least $2^{-\{1,000\}}$ over a random world model. Intuitively it’s plausible that this would be much more likely than the simple model agreeing with some very specific ways of using language encoded in the intended model.
- (complex, complex) costs 2,000 bits to specify the heads, for a total cost of $N+2,000$.
- (complex, intended) costs 12,000 bits and saves 10,000 for a total cost of $N+2,000$. (This calculation technically rests on a further claim about the complex head being harder to specify than the simple head even conditioned on the intended head.)

So under this somewhat-heroic set of assumptions, it seems like the (intended, simple) pair would have the best prior probability amongst the models that actually fit the data. Pairs other than those 9 may achieve better losses still, and several of those assumptions could easily fail, but I think this approach is still promising and gives some traction on a tricky problem.

I think this approach basically executes on a common intuition—in order to avoid generalizing poorly because the question-answerer gets blackmailed, just add some training points involving blackmail (and make extra sure to label them correctly, e.g. by generating them hypothetically). But other implementations of this idea seem to run into the problem that the model will just learn whatever process you *actually* use to generate training data, and you can't generate datapoints where that process goes wrong.

Step 3: Make the training set good enough

Steps 1 and 2 relied critically on the assumption that the training set *always* agrees with the intended model. If they disagree, then it doesn't matter whether the intended model has a good prior probability, you are still going to learn the instrumental model because it has better likelihood. Even a small advantage for the instrumental model will eventually dominate in a large enough dataset.

Perfect agreement seems impossible. The question becomes: can we get close enough, and how can we cope with the errors that remain?

Rather than aiming at “perfect” labels, or trying to be robust to general labeling errors, my current hope is to try to generate labels that *aren't wrong in any way that the model can identify*.

For example, suppose that a human is answering a question about how many chairs are in a room. They look at the room briefly, and fail to notice an occluded chair. Moreover, even when humans look more carefully at a video later (the “complex” labeling process), they also don't see the occluded chair and so believe (incorrectly) that the original label was correct.

We are only concerned about the case where this bad label leads to a bad gradient update, and especially where it pushes us from the intended model towards the instrumental model. We aren't concerned about cases where our model has no idea about the chair either, or where our model has no chance of correctly reporting the presence of the chair. I think that gives us an important leg up in avoiding the problem, if we are able to use the existing model to help make sure our data is good.

I don't have a clean proposal for making the dataset good enough, but I do have various thoughts that leave me feeling more optimistic:

- I don't think we should use supervised learning. Instead, we should get pairs of answers from the model and then only add them to the dataset when we are confident that one of them is better than the other. So we are only in trouble if the intended model actually generated a good answer, but we incorrectly said we were confident that it was worse. (Note that we may still need to use likelihood for pre-training or joint training for competitiveness reasons, and we'll need to adapt the whole scheme to that case.)
- Moreover, we are only worried if the model understands the dynamics of the data-generating process and why there was an error, since otherwise the instrumental model won't actually perform well either. If our model answers a rich enough set of questions, we can effectively interact with it to extract an explanation for why we erred. We just need to be able to recognize this kind of story-of-error as plausible rather than confidently dismissing them. Tons of details would be needed to make anything like this work, but overall it seems promising.

- We might be concerned that the model will instead learn to make up complex stories about how we are wrong. And indeed this training process won't actively penalize such models as long as they are able to make the stories plausible enough. But such models also aren't *favored*, and if steps 1 and steps 2 are successful then the inductive bias will favor the intended model over this kind of instrumental model (assuming that the dataset is rich enough, even after filtering for confidence, that you can't save a ton of description complexity by this kind of bullshitting).
- We can take simple precautions—like “don’t collect training data in situations where strong adversaries are trying to mislead us.” Many other alignment approaches make this a bit of a circular dependency—our AI predicts well, and so can help protect us from manipulation, only as long as the AI system predicts that no manipulation would occur in the future. But in this setting we have no such dependence—in order to have the model work correctly at time t we just have to make sure we didn’t train on corrupted data at times prior to t .
- Reducing errors seems to require doing [reliability amplification](#) (though maybe just ensembling is sufficient) and [security amplification](#) (or some other approach to being secure whenever the intended model is). Those are more like hopes than algorithms; we’d need to do a lot of work, and think a lot about concrete examples of possible errors, to find something that might actually work.
- These bullets are all gesturing at one approach to this problem, but I think there are a ton of ways you could perform machine learning with “potentially wrong” data to prevent a small number of errors from causing trouble. This feels closer to a traditional problem in AI. I haven’t thought about this problem much because I’ve been more focused on the fear that we wouldn’t learn even with perfect data, but I feel relatively optimistic that there are a lot of approaches to take to dataset errors if that’s actually the crux of the problem.

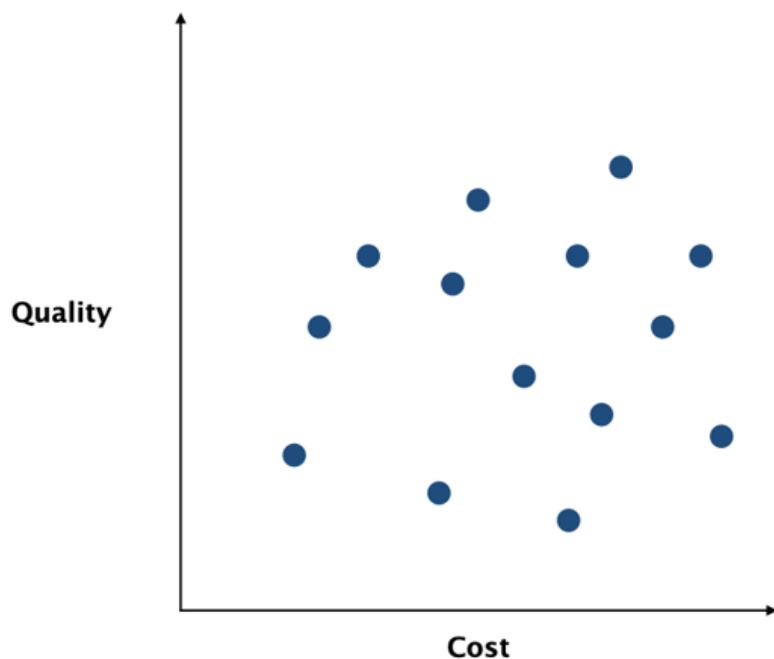
Did the Industrial Revolution decrease costs or increase quality?

This is a linkpost for <https://rootsofprogress.org/cost-quality-and-the-efficient-frontier>

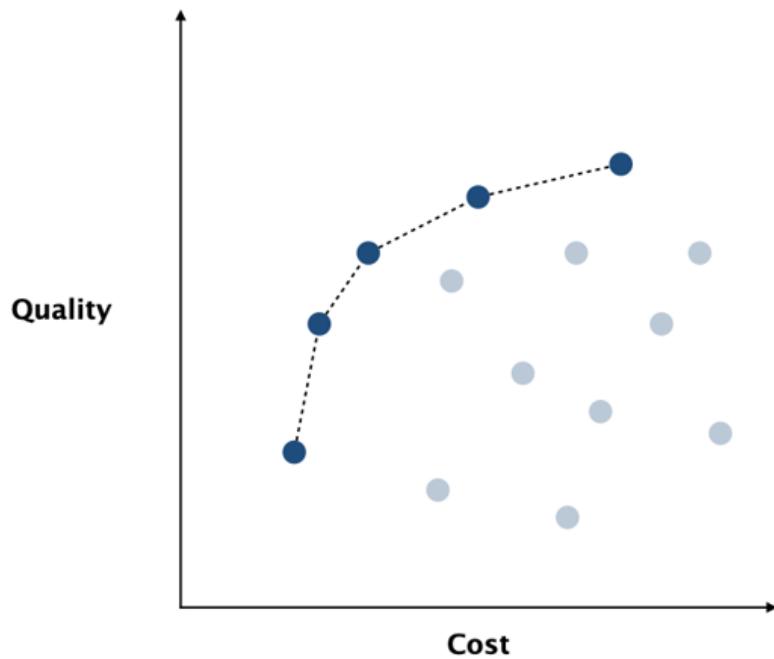
An oversimplified story of manufacturing progress during the Industrial Revolution is: “we automated manufacturing processes, and that decreased the cost of goods.” This is not wrong, but is not the full picture.

Mechanization—and other progress in manufacturing, such as improved tools, materials, and chemical processes—not only decreases costs, but also improves quality. Making things by hand requires skill and attention: just try making your own clothing or furniture at home; on your first attempt you won’t be able to achieve nearly the quality you can purchase for a modest price. Automation not only improves average quality, but also consistency, by reducing variance.

If we want a fuller picture of how goods were improved through the Industrial Revolution, we should think of cost and quality together. We can visualize this conceptually on a two-dimensional chart. Here, each dot represents one possible manufacturing process for a particular good:



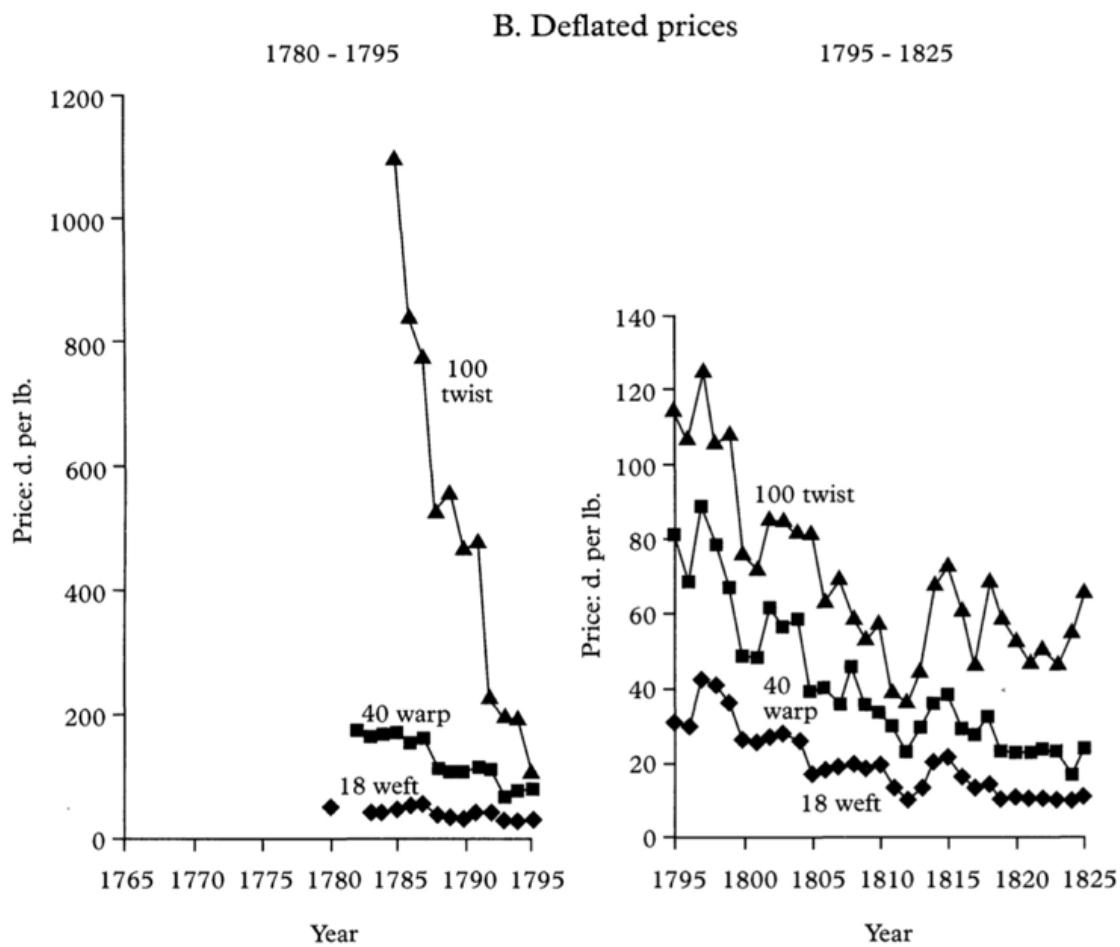
In this chart, you want to be in the upper left: high quality at low cost. So there’s no point in using any given process if there is another process that achieves higher quality *and* lower cost at the same time. This means the industry tends to move towards the set of processes along the upper-left edge. This set is known as the *efficient frontier*:



If you're in the middle of a diagram like this, you can improve both cost and quality together by moving towards the frontier. If you're on the frontier, you face a cost-quality tradeoff.

The nuanced way to think about new technology is not that it simply decreases cost or increases quality, but that it *moves the frontier*.

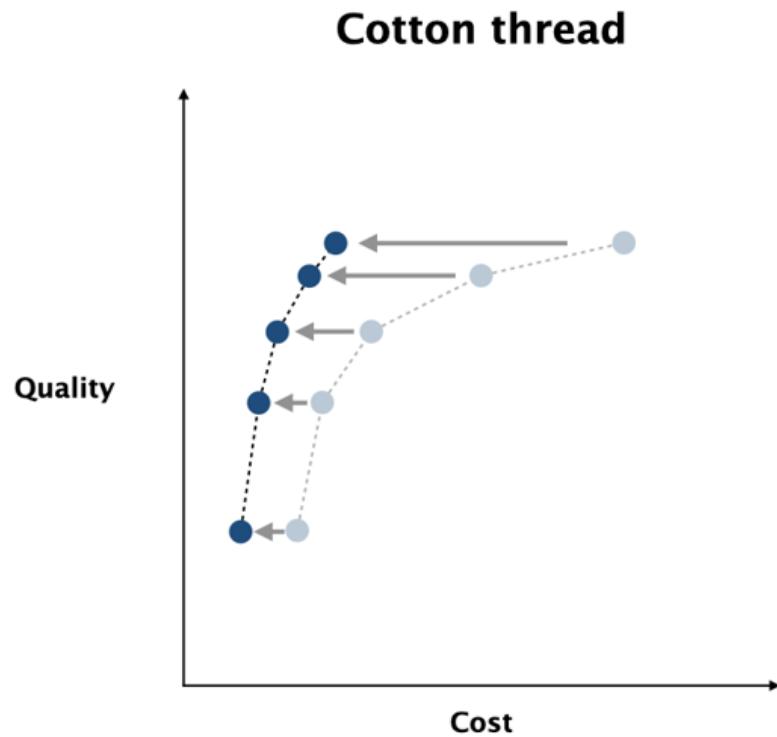
Consider cotton thread, one of the first products whose manufacturing was automated. In the late 1700s, spinning machines were invented that could do the work of dozens of hand spinners. Large factories were set up, driven by water mills (and later, by steam engines). Here's what happened to the cost of thread as automation took over the industry:



[Harley, "Cotton Textile Prices and the Industrial Revolution" \(1998\)](#)

The chart above shows prices for three different grades of thread. Higher numbers indicate finer thread, which makes for softer fabrics.

Conceptually (not to scale), we can visualize what happened like this:

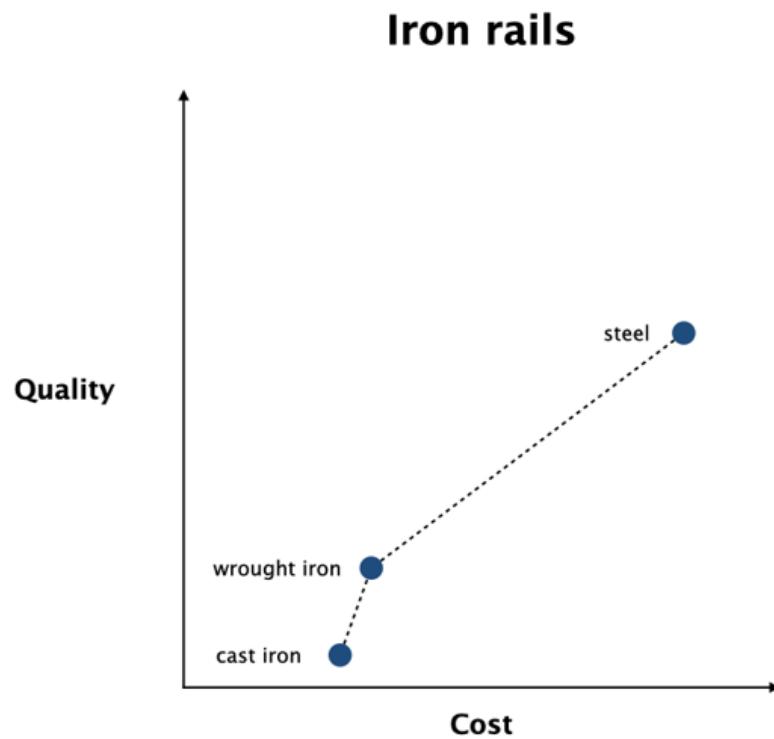


Note that the price for the highest-quality thread, 100 twist, came down most dramatically. It's *possible* for humans to spin thread this fine, but it's much more difficult and takes longer. For many uses it was prohibitively expensive. But machines have a much easier time spinning any quality of thread, so the prices came closer to equal.

At one level, this is a cost improvement. But don't assume that the effect for the *buyer* of thread is that they will spend less money on the same quality of thread. How the buyer responds to a change in the frontier depends on the cost-quality tradeoff they want to make (in economics terms, their [elasticity](#) of quality with respect to cost). In particular, the customer may decide to upgrade to a higher-quality product, now that it has become more affordable.

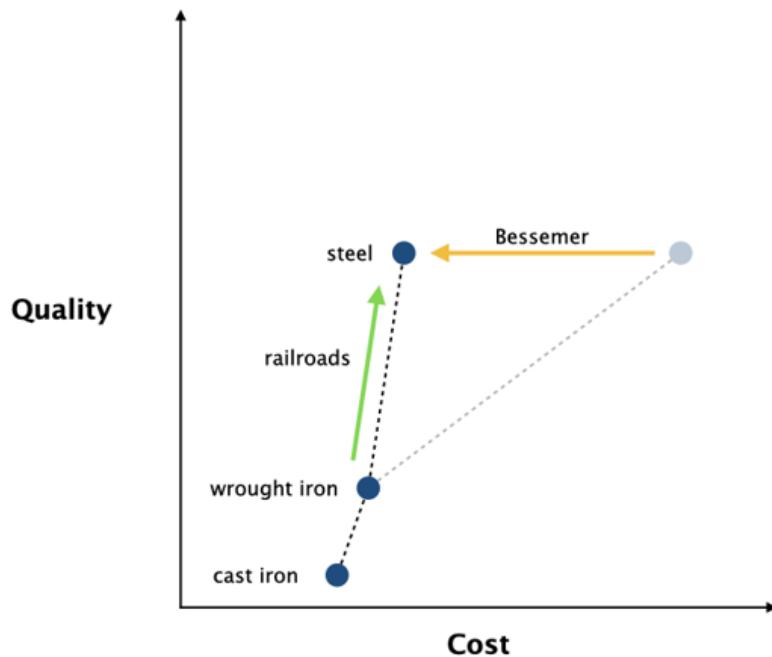
This is what happened in the case of iron, steel, and railroads. In the early decades of railroads, rails were made out of wrought iron. They could not be made from cast iron, which was brittle and would crack under stress (a literal train wreck waiting to happen). But wrought iron rails wore out quickly under the constant pounding of multi-ton trains. On some stretches of track the rails had to be replaced every few months, a high maintenance burden.

Steel is the ideal material for rails: very tough, but not brittle. But in the early 1800s it was prohibitively expensive. The frontier looked like this (again, this is conceptual, not to scale):

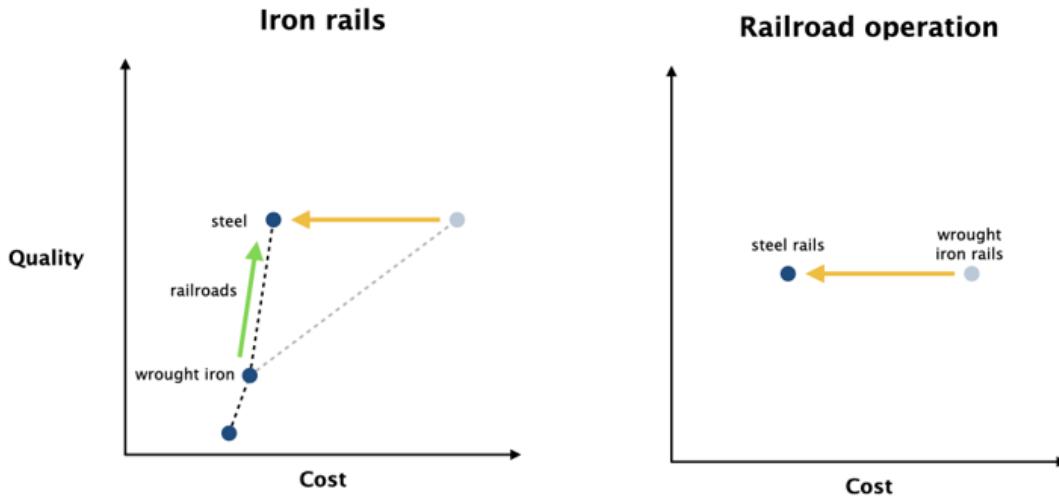


Then [the Bessemer process came along](#), a new method of refining iron that dramatically lowered the price of steel. Railroads switched to steel rails, which lasted years instead of months. In other words, what looks like a cost improvement from the supply side, turns into a quality improvement on the demand side:

Iron rails



But what was the effect of upgrading to steel rails? A greatly decreased need for replacing the track lowered the operating costs of the railroad. So the full picture looks like this:



In other words, cheaper steel meant cheaper train travel—but *not* because the railroads could buy cheaper rails. Rather, cheaper steel allowed them to buy *higher quality* rails,

leading to cheaper travel.

The lesson is that whether a new process is an improvement in cost or in quality can change as you follow the links of the value chain.

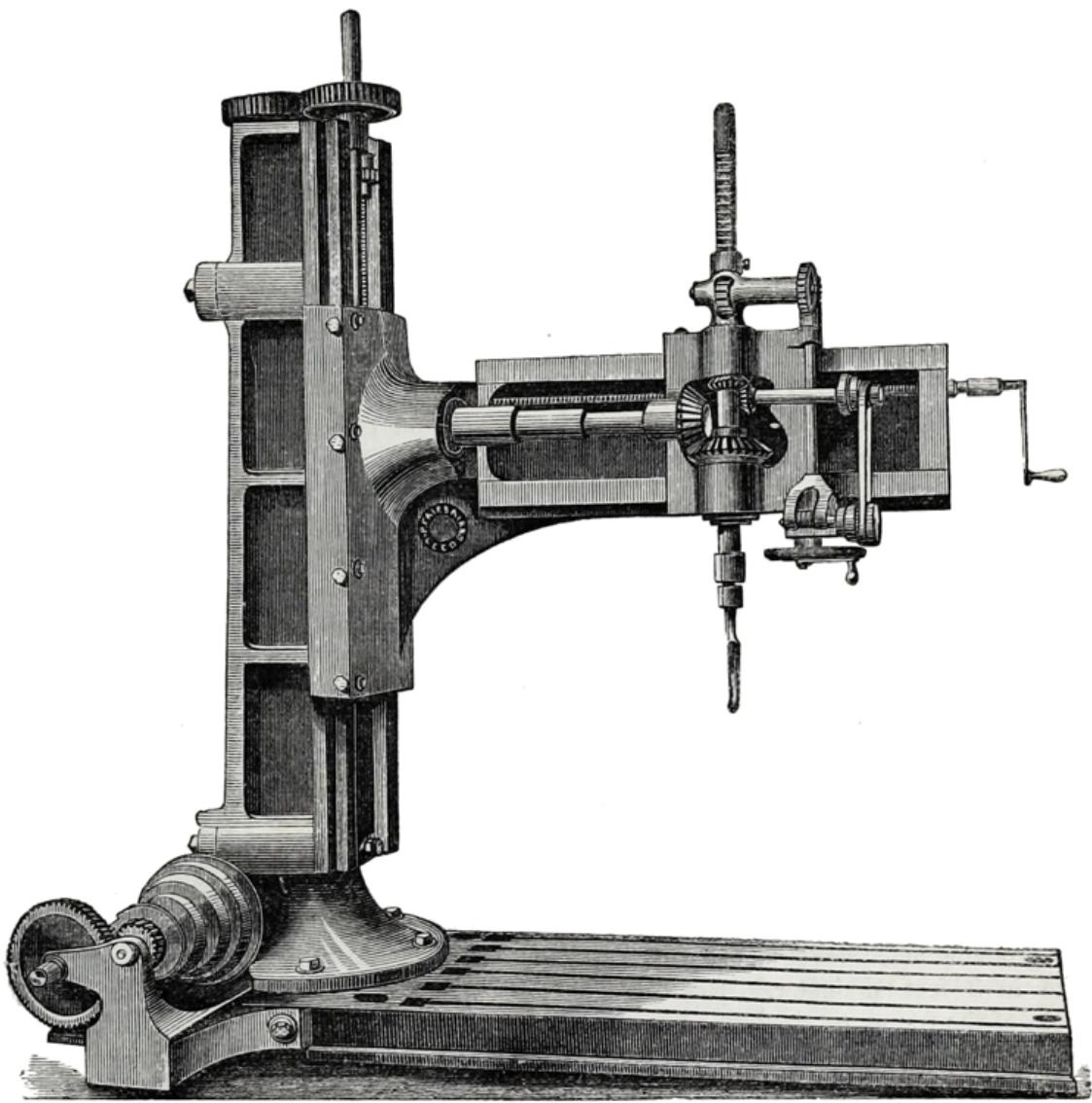
Another example is that of interchangeable parts. Consider gunsmithing. A gun is made from many parts. The parts have to fit together well enough that the gun works smoothly and reliably. But in pre-industrial craftsmanship, the trigger, for instance, of one gun wasn't expected to fit any *other* gun.

On the face of it, this didn't pose much of a problem for most machines, such as locks or watches, but it was an acute problem for guns, because of the need for field repair. If your trigger breaks while out on a campaign, you can't repair it until you can get it to a gunsmith. Until then, all you have is your bayonet, and your rifle is just a funny-looking and not particularly aerodynamic spear. At one point in 1811, the British Army had 200,000 useless muskets awaiting repairs.

A better system is to make the parts to better precision, to the point where they are interchangeable. Then if your trigger breaks, you just fetch another out of a box of triggers.

Like fine thread, it was *possible* for craftsmen working with hand tools to make interchangeable parts. It required precise model parts, jigs, or other guides, and a lot of checking and filing. Check your part against the guide, file it down a bit if needed, check it again. Filed it too much? Toss it out and start over with another part. This was prohibitively expensive for most applications.

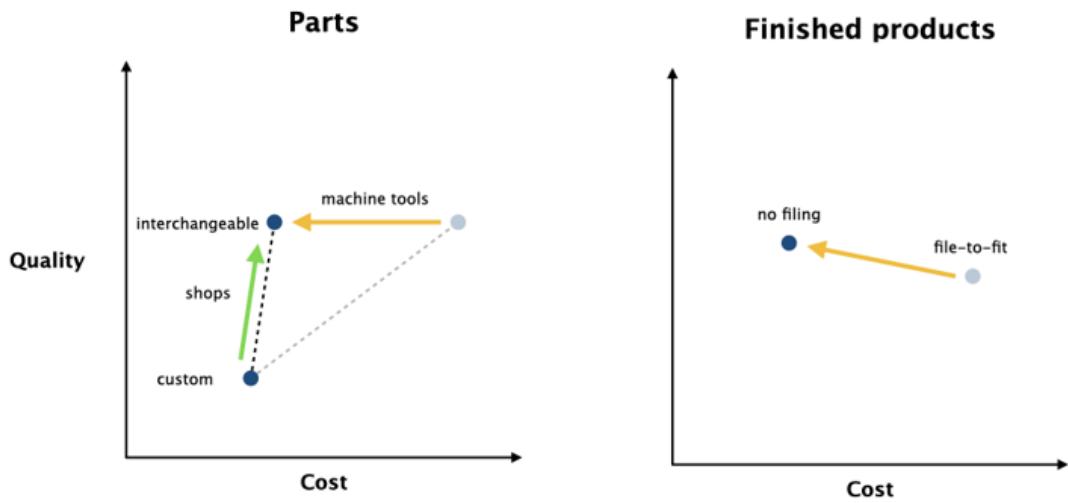
The breakthrough came when higher-precision metalworking tools, known as "machine tools," were invented. Among other advances, these tools did not rely on the worker holding either the part to be worked or the cutting edge by hand: both were clamped down and controlled by screws, gears and dials. As with thread, this increased both quality and consistency.



Fairburn's Self-acting Radial Drill. [Record of the International Exhibition, 1862](#)

Machine tools reduced the cost of interchangeable parts. For the army, buying expensive field-repairable guns, this was a direct cost savings. But other manufactured products *upgraded* from low-precision parts to interchangeable ones, now that the latter had become more affordable. Why? Because interchangeable parts are faster to *assemble*. Before, final assembly required filing parts to fit if they didn't line up just right. Interchangeable parts eliminated the need for this, speeding up the production line. Ultimately, this resulted in lower prices for finished goods.

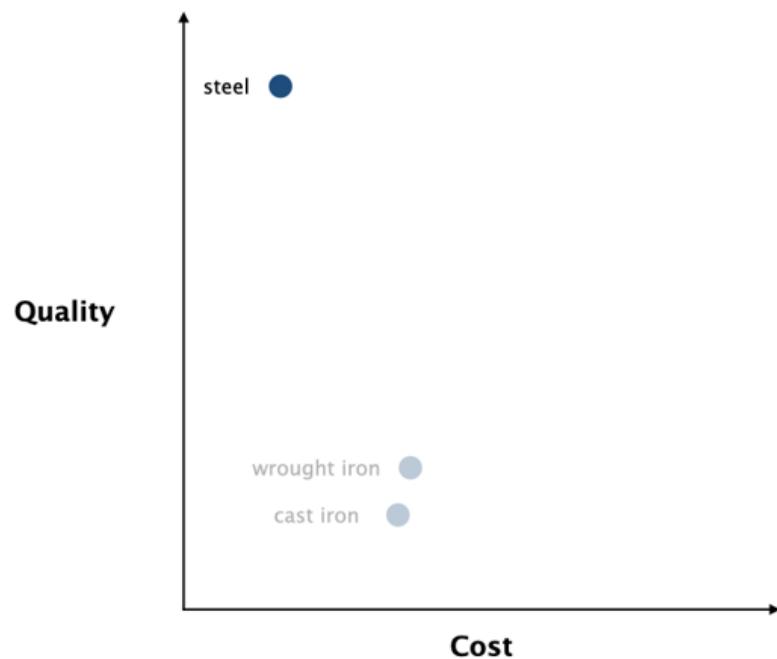
So just as with Bessemer and steel rails, when machine tools lowered the cost of interchangeable parts, it led to lower prices for final products, but via a quality increase in the parts rather than a cost decrease:



In the diagram above, the arrow on the right half points to the left but also a bit up. Machines made with interchangeable parts were not only cheaper, they were more reliable and easier to repair. That means the new production method was superior to the old file-to-fit method on both axes—which is why the old one is now obsolete.

Once sufficient economies of scale had been reached, steel hit the same point with respect to other grades of iron, which is why today almost all iron is made into steel:

Iron today



After some 250 years of iterative improvement, the cost-quality frontier has been shifted so far for so many goods that we enjoy both greater quality and lower cost on pretty much everything we buy. But to get there required navigating a subtle, winding road traversing both axes in different ways at different times.

Bayeswatch 2: Puppy Muffins

A green humvee arrived at Jieyang Chaoshan International Airport. Vi got in the back with Molly Miriam who handed her clipboard to Vi.

"健重制造公司. A no-name Chinese factory that makes barbells and similar equipment. It's not even fully-automated," read Vi.

"They are registered to use a low-intelligence AGI," said Miriam.

"What are we even doing here? Neither the product nor the AI poses a threat to civilization," said Vi.

"Something must have gone badly wrong," said Miriam.

The road to the factory was blockaded by the People's Liberation Army (PLA). The soldier at the checkpoint scanned the Bayeswatch agents' badges. A young officer—barely out of high school—escorted them inside the perimeter to Colonel Qiang.

"We could probably handle this on our own," said Colonel Qiang, "But protocol is protocol."

"So it is," said Miriam.

There were no lights on in the factory. No sound emanated from it. Fifty soldiers behind sandbags surrounded the factory, along with two armored personnel carriers and a spider tank.

"The police responded first. They sent a SWAT team in. Nobody came back. Then we were called. We would like to just burn the whole thing down. But this incident could be part of a wider threat. We cut power and Internet. Nothing has entered or left the building since our arrival. Rogue AIs can be unpredictable. We wanted your assessment of the situation before continuing," said Colonel Qiang.

"You did the right thing. This is probably an isolated incident. If so then the best solution is to rescue who we can and then level the building. Unfortunately, there is a chance this is not an isolated incident. Therefore our top priority is to recover the AI's hard drives for analysis," said Miriam.

"We will assault the building," said Colonel Qiang.

"You may have our sanction in writing. Assume humans are friendly and robots are hostile," said Miriam.

"Yes sir," said Colonel Qiang.

Miriam and Vi were quartered in a nearby building that had been comandeered by the PLA. They watched the assault from a video monitor.

"In training they taught us to never go full cyber against an AI threat," said Vi.

"That is correct," said Miriam.

"Which is why every assault force is no less than ten percent biological," said Vi.

Miriam nodded.

"Standard operating procedure is you go ninety percent robotic to minimize loss of life," said Vi.

"Ninety percent robotic does tend to minimize loss of life without the failure modes you get from going full cyber," said Miriam.

"It looks to me like they're going one hundred percent biological while their battle droids stay outside. Are we facing that dangerous of a hacking threat?" said Vi.

"No. They are just minimizing loss of capital," said Miriam.

The video feed of the factory was replaced by Colonel Qiang's face. "We have a survivor," he said.

Two privates guarded the door to their freely-liberated prisoner. His dress shirt was stained with blood and grease. An empty styrofoam take-out food tray lay in the corner of the table with a pair of disposable chopsticks and an empty paper cup. Miriam and Vi took seats opposite him.

"I understand you helped program the improperly registered AI at 健重制造公司," said Vi.

"I didn't know it was improperly registered," said Paul while looking straight at the security camera.

"We're not here to find out what laws were or weren't broken. We just want to know why there is a company of infantry surrounding this factory," said Miriam.

"There wasn't much to it. The mainframe running the assembly line barely qualifies as an AGI. We could never afford that much compute," said Paul.

"How does it work?" said Miriam.

"Labor is affordable here by international standards. Our factory is mostly human-run. Androids are expensive. We only have a couple of them. We should have been able to overpower robots if they were all that had gone rogue," said Paul.

"But that's not what happened," said Miriam.

"We didn't smell anything. People just started dying. We tried to help. More died. We tried to escape but the fire doors had been locked. I ran to my office, barricaded the door and breathed out the window," said Paul.

"Argon gas. It has all sorts of industrial applications," said Vi.

"Exactly," said Paul.

"And the same mainframe which controlled the robots also controlled the fire suppression system," said Vi.

Paul nodded.

"So why did it want to kill people?" said Vi.

"Maybe it was jealous," said Paul.

"Let's stick to the facts. Why use an AI at all if human labor is so cheap?" said Miriam.

"Human manual labor is cheap. New products are high margin but top designers are expensive. We had the AI do some manufacturing because embodiment helps with designing human-compatible products. But mostly we just used it for the generative model," said Paul.

Miriam flinched. "Thank you. That will be all," she said.

They were back in the monitor room.

"We don't need the hard drives. Do whatever you want," said Miriam to the image of Colonel Qiang.

The monitor went black.

"I lost count of how many OSHA regulations they violated," said Vi.

"OSHA has no jurisdiction here," said Miriam.

"Do you know what happened?" said Vi.

"When I was your age, I inspected a muffin factory. They followed all the regulations. It was even Three Laws Compliant. Very high tech. For its time," said Miriam.

Miriam lit a cigarette.

"The told the AI to make the cute muffins. They fed [/r/cute](#) into it as training data."

Miriam took a deep breath from her cigarette.

"The AI bought puppies. The engineers thought it was cute. They thought maybe they had solved the alignment problem," said Miriam.

Miriam took another swig. She exhaled slowly.

"The engineers had told the AI to make blueberry muffins. Do [an image search for 'puppy muffin'](#) on your phone," said Miriam.

"They do look the same. Puppies do look like blueberry muffins," said Vi.

"Puppy heads. Puppy heads look like blueberry muffins," said Miriam.

"Oh," said Vi.

"Come outside. You need to see this with your eyes," said Miriam.

The soldiers had retrieved several bodies. Doctors were autopsying them. The bodies' hands were missing. A few were missing half their forearms. One body had its neck and shoulders removed.

"They used a generative model on weightlifting equipment. They fed it pictures of people lifting weights. They annotated which regions of the images constituted a 'barbell'." said Miriam.

Vi almost puked.

"Tell me what happened," said Miriam.

"The generative model added disembodied hands to the barbell," said Vi.

Colonel Qiang ordered sappers to implode the facility.

A Review and Summary of the Landmark Forum

(This post is from a long-time member of the rationalist community - I may or may not be an active poster on Less Wrong. I've messaged a moderator and asked if they would be willing to make a comment to confirm that this is the case.

Landmark has copyright on their course content and they have trademarked terms like Always/Already Listening and Rackets, which is important to understand for legal reasons.

I'm also going to attach a spoiler alert as understanding the techniques might make them less effective. But that said, my recommendation is that if you are considering the forum then you should understand exactly what you are signing up for)

After attending the Landmark Forum, I was completely shell-shocked at having encountered some powerful psycho-social technology unlike anything I'd ever seen before. They seemed to have a powerful ability to change how people viewed the world. At the same time, I felt troubled by some of the methods they used to persuade us to invite our friends and family to do the forum. Ironically, if they hadn't tried so hard, I would have recommended the forum in a heartbeat. But given that they did, I would strongly recommend that anyone who is considering going do their research before signing up and ensure they are appropriately mentally prepared. I've tried to make the main body of the analysis as objective as possible, but there were some points at which I felt the need to comment.

What happened and the stories we tell ourselves

The core idea of Landmark is that we are trapped within and held back by the stories that we tell ourselves. For the most part, we aren't even aware that we are operating from within these narratives, but they serve as excuses and determine many of our choices.

For example, imagine Derek had a rough childhood and was frequently beaten up by other kids until he started working out and training in martial arts. Years later, Derek is married and has trouble sharing openly with his wife, which pushes them to the verge of divorce.

A Landmark forum leader would ask Derek what he told himself back when he was being kicked into the ground. After some back and forth, Derek might eventually say that he told himself "I was weak and needed to become strong to survive". The forum leader might suggest that this is why he is unable to share: sharing makes Derek feel weak and he has a story that he has to be strong. Even though it is years later and Derek is in a completely different context, he is still stuck within that narrative.

The forum leader may hammer this point home by drawing two circles: what happened and the story he told himself. Suppose the forum leader asks what belongs in the what happened circle and Derek says he was bullied. The forum leader would object: "bullied" is still an interpretation. Even "beaten up" involves an element of subjectivity - what one person might consider being beaten up might not be anything significant for another. Ideally Derek would say something like, "On about 10

occasions, I was punched and kicked" (some people may note the similarity to non-violent communication).

They are also very suspicious of words like "always", "never" and "every time" as they tend to be exaggerations - always typically means most of the time. Even if you had caught up with someone five times and each of those times they were late, they would suggest you say "five out of five" times instead of "always late". You might wonder how using the word "always" here could be part of the story, or how this could be significant, but the fact that they were late in the past doesn't guarantee they'll be late in the future, and laxity about this use of language makes a difference for the stories being lived out in your mind. And even if they say, "you've been late every time" which is an objectively true fact, there's still an element of narrative here - using the word "every" suggests a pattern (perhaps deliberate, vindictive, malicious) rather than happenstance. One of my key takeaways was just how impactful these subtle narratives can be.

Next, the forum leader would ask what belongs in the circle representing the story he told himself. Suppose, Derek says, "Well if you've been beaten up you're naturally going to feel weak". The forum leader would encourage Derek to restate it using the word "I" instead, as talking about events in the third-person can be used to avoid emotions. For the same reason, the forum leader would also discourage the use of abstract language - it would be much better to say that "I had to be strong to survive" rather than saying "I felt compelled to carry out traditional masculine behaviours".

Landmark calls this being in the arena vs. being in the stands. When you're in the arena, you're actually putting yourself out there, being vulnerable and opening yourself up for growth. Another way they try to encourage this is by telling you not to take notes as it takes you out of the experience. Instead they provide you with notes afterwards.

Landmark would name "being strong" as one of Derek's "Winning Formulas". We can imagine it helping Derek stand up to bullies, survive his first breakup and then work his way through college. However, when Derek gets married, his "Winning Formula" is no longer helping him to succeed and indeed it is the very thing causing him to fail.

So why doesn't Derek let go of his "Winning formula" automatically? One reason might be that Derek doesn't realise that he's still acting out this narrative or the full extent of it. Narratives constructed in the heat of a traumatic moment may not seem optional, or also may not be remembered as having been constructed. Landmark suggests that these unknown unknowns may actually be the most important thing to understand in life.

However, even when this situation is pointed out to him, Derek may still be reluctant to change, despite knowing full-well that it might cost him his marriage. Derek may even say that he wants to change - maybe even change temporarily - but then always revert back. Why might this be? Well, the most likely explanation is that there's some kind of payoff for acting in that way. Maybe "being strong" allows Derek to pretend that he'll never get hurt again or to hide the fact that he still feels weak. Landmark calls this a Racket, which they define as "Anything that is Unwanted and yet Persists". I've personally found this to be a useful conceptual handle as it combines the idea of it being a narrative with the idea of it having some kind of payoff.

Extending on the previous point, Derek may go through life always trying to become stronger. Maybe he gains a blackbelt and when his dog dies he refuses to shed a single tear. A forum leader would suggest that Derek is not acting in this way because he feels strong; instead he is running away from an underlying sense of weakness that he's had ever since he was bullied. And what a performance - no-one who knows him would ever call him weak!

Another way that we form these narratives is by having assumptions about how someone will react which shaped how we perceive our actions. For example, if you think your partner is mean and they happen to be late, you may assume it is because they are punishing you. Landmark calls this concept Always/Already listening.

I'll note that a large part of the way that Landmark convinces you to drop narratives is by very strongly insisting that they are narratives. The forum leader would often cut off participants. This allowed them to maintain control over the situation, however they would likely explain it as necessary to prevent participants from jumping back into their standard narratives.

They were also very effective in using humor to make what someone said sound absurd. I often found myself laughing, even when I thought the participant was actually making a reasonable point. This was effective at creating social proof of their claims.

One potential issue with this style of engagement is that people may end up believing stories about themselves that aren't actually true. After all, really figuring out the cause of issues, to the extent that this is possible, would likely require a long conversation and asking a large number of questions. However, it could be argued that at the end of the day that the actual true story is mostly irrelevant. For example, if someone was really lacking in confidence because they do not have many friends, but they end up believing that it was due to childhood bullying and then they end up regaining that confidence due to them believing that they've finally addressed this bullying, then they may believe a falsehood, but they' still regained confidence. And some people would say that this is what really matters?

Clearing narratives

So what should Derek do once he realises that he's trapped in a narrative and he's decided that it no longer serves him. According to Landmark, the answer is simple, you just do ("all this time you thought you were trapped inside, but the door wasn't even locked"). They illustrate this with the story of monkeys being trapped by putting a banana in a cage just big enough for them to put their hands through. As it goes, when the monkey tries to grab the banana, it finds its hand trapped as the hole isn't big enough to pull it out. The monkey could escape, but it's unwilling to let go of the banana. However, we could also interpret them as operating under the theory that if the understanding and realisation is strong enough and lands deep enough then it creates a shift automatically.

I think there's a deep wisdom in this. How do you decide to change your life? Well, there's a sense in which you just do. However, this isn't quite the whole story as they have a trick up their hands. If someone is stuck and doesn't quite feel able to let go, the forum leader might move onto someone else. Usually after hearing someone else discover they are trapped in a narrative and perhaps even let go of it the first person would feel ready to let go as well.

One thing to be careful about is escaping one narrative only to end up trapped in another. Suppose that when you were younger you were a member of a particular political party. Eventually you decide that you can no longer support the party because of some of their policies and you feel like you wasted years of your life. It would be very easy to adopt a narrative that "party X is terrible" and "only party Y is good". But this is a narrative in and of itself and it could very easily result in you irrationally defending all the actions of a group that is bad in many the same ways.

Landmark tells us that we are meaning-making machines. We can clear out our narratives, but we will always be replacing them with new ones. They consider this unavoidable, however we can learn to clear them faster and create new ones that serve us better.

The goal of the Landmark Forum is to create a clearing or blank space in which you will subsequently be able to create your future. We may not be able to inhabit this space permanently, but we may be able to achieve it long enough to make a breakthrough. We were told that the advanced course would teach us how to envision a future.

The philosopher Sartre famously wrote about a student torn between going to war to serve his country and staying home to look after his mother. The way I've heard our facilitator handled an analogous situation was to tell the student that both of those are narratives and that it is only by stepping outside of both those narratives that we can open up a clearing so that we can decide without feeling burdened. We might choose to go to war, we might choose to stay with our mother or - now being unburdened by our narratives - we might discover a third option which we previously lacked the headspace to notice.

No Excuses

Landmark is very much in the No Excuses school of thought. Suppose you are training for a marathon and you decided to go for a run every day to prepare, but one day it rains. This is an excellent reason not to go for a run, but they would tell you that an excellent excuse is still an excuse. When you are attempting something truly at the limits of your capabilities, allowing yourself to accept a reasonable excuse is a sure-fire way to fail. Landmark reinforces this No Excuses attitude by repeated emphasis on the importance of arriving on time. This is a simple, but effective way of setting standards.

Or here's another one. Suppose you're shy. Again, there's a sense in which this is objectively true. But there's also a sense in which thinking of yourself as shy can limit the way that you act and this can reinforce your shyness. Maybe if you stopped thinking of yourself as shy you'd actually act less shy. Similarly, there's an objective sense in which you may feel tired or in pain. But beyond this, there's also the narrative of being tired or feeling pain which can limit you or cause you additional suffering and anxiety.

When I said Landmark is very much in the No Excuses school of thought, this is something of an understatement. If someone says that they can't do X due to trauma, the response would likely be that an excellent excuse is still an excuse. A surprising amount of the time this is exactly what people need to hear as people are stronger than you think, but I have worries that when it goes wrong it might go horribly wrong and retraumatise someone.

Personally, I found their tough love approach helpful, although I have to admit that I didn't make myself fully vulnerable. The Western focus on individuality and autonomy

can be limiting as often a push is exactly what we need. This may explain part of why they were able to achieve what seemed like remarkable results - psychologists are limited by ethics in a way in which Landmark is not.

Social Aspects

Landmark relies heavily on social components. A key aspect of this course is about calling people and admitting where you haven't been taking responsibility or where you've been stuck in a narrative in order to "get complete" with them. I was persuaded by the course to have a few conversations during the breaks which I found helpful. I imagine that most people have at least some difficult conversations which they've avoided or have done things they are yet to take responsibility for. They push you to do this during the breaks, which might make some people uncomfortable, but makes sense from the perspective of these often being conversations that people are reluctant to have and which it'll be easy for people to never get around to after the course.

Landmark using the term Enrol to mean sharing with someone in a way that touches or inspires or motivates them. This definition could be interpreted as something of a [dark arts](#) trick to make people more favourably disposed towards signing up their friends for the course. After all, if you're talking about enrolling your friends and family all the time, then you're naturally going to think about enrolling them in the course. Although some people might consider this interpretation uncharitable.

This brings me to another point. Landmark really wants you to enroll your friends and family in the course. As much as people say that it is just for the money, they really do seem to be true believers in the course. My forum leader suggested that if everyone did the forum it could lead to world peace, and they seemed to honestly believe this. Only the forum leaders and a minimal office staff are paid, much of their operations are run by volunteers. Forum leaders seem to be paid reasonably well, but I've been told that these leaders need to go through years of training and volunteering before achieving the position. So it doesn't appear nearly as lucrative as you might think, but I am also conscious that I haven't been able to verify this information as much as I'd like.

Suppose someone shares that they have had issues with their parents in a very vulnerable manner. Based on what I've seen, I would expect a forum leader to suggest that they should try to persuade their parents to do the course in order to get complete with them. And as much as I would like to make the main body of this post as objective as possible, I feel obligated to mention that I see this as quite manipulative.

If someone said this in the forum, the facilitator would probably answer that it's just their narrative: that the forum leader merely suggested that their parents might benefit and that they constructed the narrative that they were being pressured, instead of taking responsibility for their own decisions. And even though there might be an element of truth in that we could have simply chosen to ignore them, I don't find this explanation satisfactory.

Our Landmark coach justified the pressure to share along the following lines: First, they said that if you thought that it helped you and it could help others, then you'd naturally want to share it with others. Then they suggested that the same difficulty that lay at the root of people's hesitancy to share often also laid at the root of their inability to make sales or to ask someone out. And well, they actually seem pretty

good at teaching people how to sell the Landmark Forum and I don't doubt that you could apply the same lessons elsewhere. However, this all feels just a little bit too convenient.

It's important to understand the extent of this pressure since, according to comments on the Internet, some people have alienated friends by pushing too hard for them to do the forum or by constantly talking about Landmark. From what I've read, the greatest pressure to recruit people increases dramatically if you do their leadership training.

Testimonies are also a key aspect of this course. When people were demonstrating that they had applied the techniques and shared their successes, it naturally makes you feel like you're slacking off if you didn't do what they told you and inspires you to achieve your own successes. Many of the successes people achieved were truly inspiring, but obviously there's a selection effect where people who had the greatest successes will be most keen to share. At the same time, if a cognitive bias makes you feel inspired, you will be much more likely to achieve your own successes than if you had a more realistic appraisal.

Other aspects of the course

The Landmark course has long hours - starting early, finishing later, with two smaller breaks and one bigger meal break during the day. This means that each of the sessions is a couple of hours - which is the complete opposite of how I've generally seen conferences run - which is with breaks as people tend to drift off during longer sessions. Beyond this, you are assigned homework in-between the days and often told to make calls during the break. If you decide to do the course, I'd encourage you to try to set it up as much as possible to ensure that you don't have other tasks so that they don't interfere with the homework assignments. I suspect that the long days break down some of your usual defences. It makes their techniques more effective, but you may not want to provide them with this power over you.

Another aspect of Landmark is that it is effectively a closed system. They want you to do things their way and this is a significant part of what they call Coachability. The way the forum leader explained was by saying "Don't coach your coach" and that they followed their coach's instructions when they were being coached themselves. Combined with some of the techniques they use for persuasion, this creates a substantial risk of some students ending up adopting Landmark teachings as an ideology.

At the same time, I can also see why they would want people to just follow the process. I once attempted to tutor a student in maths who insisted on doing everything their way. While I'm totally in favour of innovation, in this case they were mostly just messing up as they hadn't mastered the basics.

Landmark also likes to talk about decisions vs. choices. A decision is something you do for a reason whilst a choice is something that you do choose because you choose it. This seems somewhat paradoxical, you might say that anything you do would be for a reason and that if you did things without reasons you'd make bad choices. The best way I can explain this is as some kind of psychological hack. After you've committed to choice if you can commit to it on the basis that you've committed to it and not on any other basis, you'll likely have less doubts and your mood will be less dependent on outcomes. They also use this to embrace life situations - ie. "I choose to have

experienced suffering because I choose to have experienced suffering". When used in this way, it's essentially the standard Buddhist acceptance principle.

Conclusion

The value of Landmark is mostly in the practise, rather than the theory. If you're just interested in the theory, you probably wouldn't gain that much by going to the first-level forum. Sure, there's a lot of details I haven't covered, but they don't add that much. Reading about a thing is not the thing itself and the value of the forum lies in the environment that it creates.

Regardless of any judgements I might have made, I couldn't help but admire how all the components of the forum work together to form a finely crafted machine for delivering people a meaningful experience that changes their patterns of behaviour and convinces them to evangelize it to their friends. And if there's one key limitation of this post, I feel that it's that I've only described the components and not how they fit together and essentially left that for the reader to work out on their own.

In conclusion, Landmark seems to have some rather effective techniques, but there are also some significant red flags. I believe that attending the first-level forum would be net-beneficial for most people, but there is the risk of it going very badly for people with trauma. And even if I only recommended it to people who were well-suited for it, there's a risk that they might recommend it to people for whom it wouldn't be the right fit. I also can't comment on further courses and the risks involved in them. That said, I would love to see more members of the rationality engage with Landmark and share what they learn with the community. There are risks and downsides, but I believe that they can be managed with appropriate precautions.

Further Reading:

- [The Landmark Forum - a rationalists first impression](#)

Appendix: The Original EST Course

I thought it would be worthwhile sharing a bit of information about the original EST course this was based on. Landmark was founded by former students of Werner Erhard who attended his Erhard Seminars Training (EST) and bought the intellectual property off him after he fled the country. EST has been associated with the Human Potential Movement and is heavily based upon Zen.

[Eliezer Sobel](#) described it as follows:

"I considered the training to be a brilliantly conceived Zen koan, effectively tricking the mind into seeing itself, and in thus seeing, to be simultaneously aware of who was doing the seeing, a transcendent level of consciousness, a place spacious and undefined, distinct from the tired old story that our minds continuously tell us about who we are, and with which we ordinarily identify."

I agree with him. I believe that they've done an excellent job of adapting Zen in order to make it palatable to a Western audience.

Here's another quote from the article:

One of the main arguments that opponents voiced against Werner's work was that "you can't package and sell enlightenment in a few days, because people spend years

and years doing austere spiritual practices, and often still fail to 'get it.'" Erhard's response to this was, "No, people spend years and years not getting enlightened — when they finally get it, it happens in a flash, it takes no time at all, it happens outside of time."

And some quotes it attributes to the course:

"What is, is; and what isn't isn't"

"Rocks are hard, water is wet, and you're feeling sad"

"What you resist, persists"

"Choose what you got, choose what you got, choose what you got"

"If you're not sharing it, then you never got it."

"At all times, and in all places, and in any situation, you have the power to transform the quality of your life; stop waiting for it to 'turn out,' because this is how it turned out."

Apparently the EST course was much stricter than Landmark. Instead of running to a schedule, the EST course went until the instructor felt the lesson had been delivered and participants weren't allowed to go to the bathroom during the sessions.

I'll note that Ernhard seems to have had a remarkable talent for coining phrases - my forum leader's speech wasn't nearly so elegant.

SGD's Bias

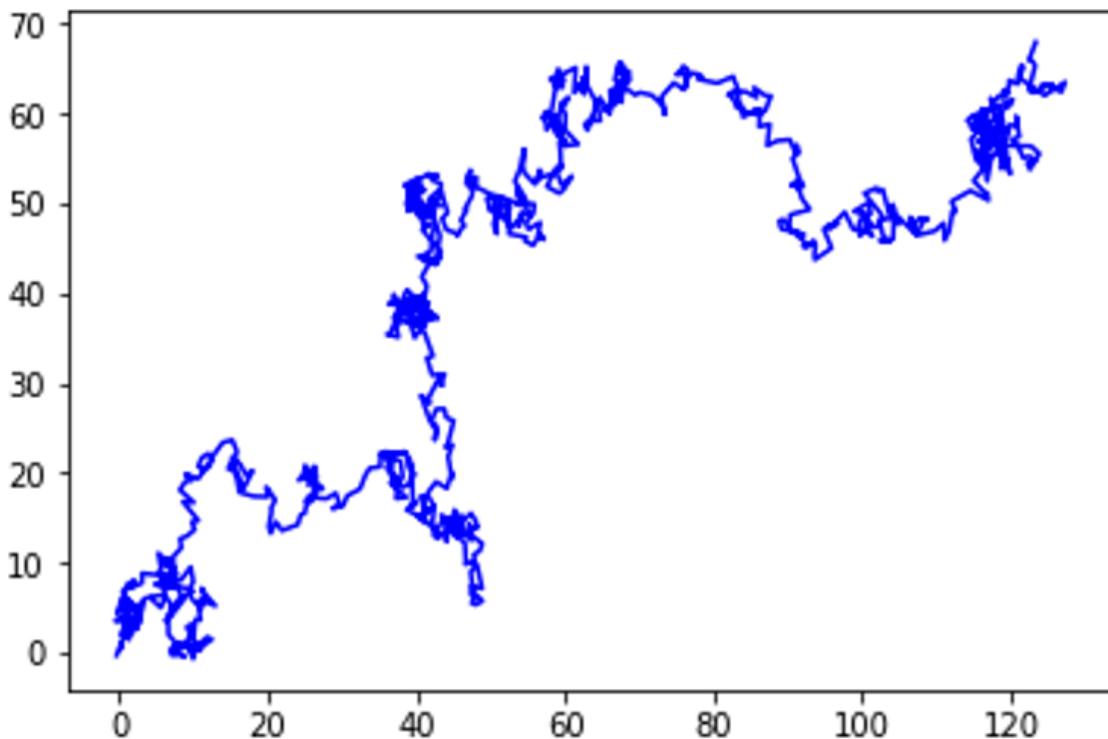
Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

There's a common hypothesis that stochastic gradient descent has some kind of built-in bias toward simpler models, or models which generalize better. (It's essentially the opposite of the [NTK/GP/Mingard et al model](#).) There is also a rough intuitive argument to expect such a thing a priori: sub-structures which are only useful sometimes will be lost on the occasions when they're not useful, whereas sub-structures which are consistently useful will be retained. Conceptually, it's similar to the [modularly varying goals](#) model in biological evolution.

Mathematically, it's not too hard to show that SGD does indeed have a bias, and we can even write it down explicitly given some not-too-unreasonable approximations. This post will walk through that derivation, and give some intuition for where the bias comes from.

First Idea: SGD Is Approximately Brownian

"Brownian" here refers to Brownian motion, i.e. a random walk. In other words, the path taken by SGD looks qualitatively sort of like this:



2D Brownian motion with drift up and to the right.

The key idea is that each sample used to estimate the gradient is approximately independent (in the probability sense of the word), and the estimate is an average over samples, so the net effect of several steps is approximately normally distributed. That's the defining feature of Brownian motion. (Alternatively, we can assume that steps are approximately independent and additive, which gets us to the same place with a little more generality but also a little more work.)

Let's put some math on that.

We use SGD to choose θ to minimize $E_X[u(X, \theta)]$. Each step, we take n independent samples $X_1 \dots X_n$ of X , use them to estimate the gradient, then take a step:

$$d\theta = -\eta \sum_{i=1}^n \nabla_\theta u(X_i, \theta)$$

... where η scales the step size. This step is an approximately-normal random variable, constructed from the IID random variables $X_1 \dots X_n$. It has mean $-\eta E_X[\nabla_\theta u(X, \theta)]$, and variance $(\eta)^2 \text{Var}_X[\nabla_\theta u(X, \theta)]$.

To formally represent this as a Brownian motion, we declare that the amount-of-time which passes during each SGD step is $dt = \eta$, which we assume to be small (i.e. we'll approximate things to first order in dt). Then SGD's path can be represented as

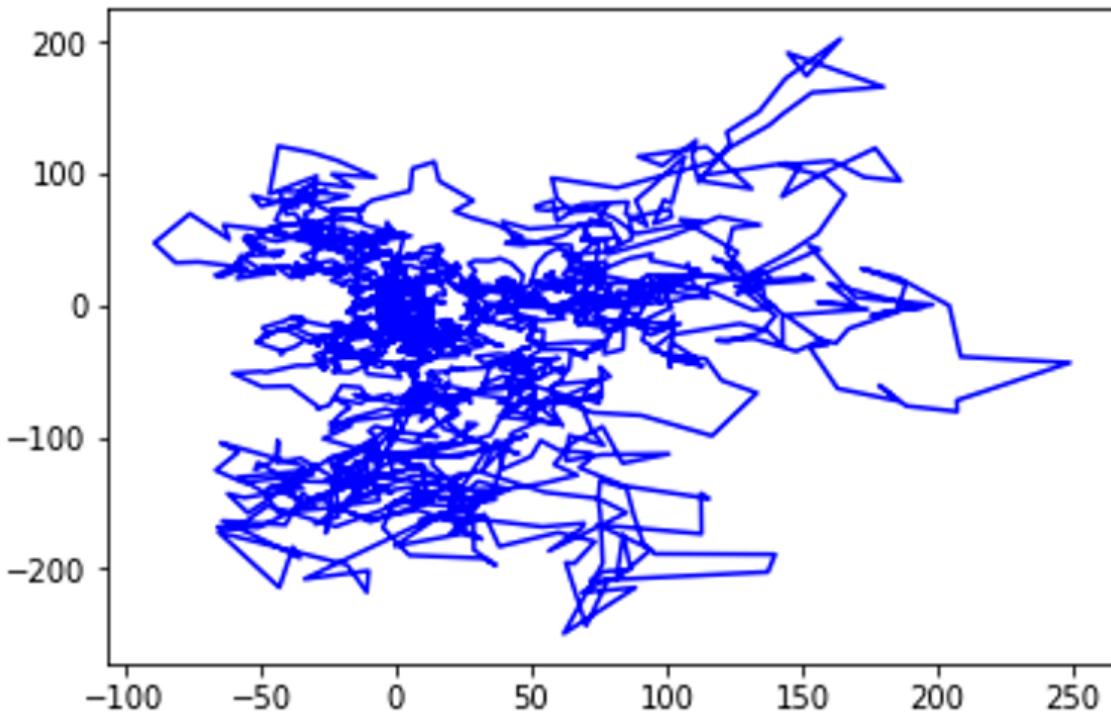
$$d\theta = -E_X[\nabla_\theta u(X, \theta)] dt + \sqrt{\eta \text{Var}_X[\nabla_\theta u(X, \theta)]} dW(t)$$

... where $W(t)$ is a [standard Brownian motion](#), the analogue of a standard normal variable, and the square root is a matrix square root. (If you haven't worked with Brownian motion before, ignore that formula and keep reading.)

Second Idea: Drift From High-Noise To Low-Noise Regions

Sometimes, the "noise" in a Brownian system is location-dependent. As an example, let's consider the original use-case: a grain of pollen floats in water. It's small enough to get randomly kicked around by water molecules, so its path is Brownian (and can be seen under an ordinary microscope). If the water has a temperature gradient, then the "noise" in the pollen-grain's path will vary with its location.

When the grain of pollen is in a higher-noise region, it will be kicked around more, and move around faster, until eventually it moves into a lower-noise region. In the lower-noise region, it will be kicked around less, and move around slower, so it takes longer to leave the region. So, the pollen grain will tend to spend more time in lower-noise regions. With a noise gradient, this tendency to spend more time in lower-noise regions becomes a tendency to drift down the noise gradient.



Brownian motion with variance proportional to distance from the origin, and no explicit drift. Notice that the process spends most of its time near the origin. When it does move away from the origin, it tends to drift back reasonably quickly.

Mathematically, if $Y(t)$ is our pollen location, we can write its motion as

$$d\bar{Y} = \mu(Y) dt + \sqrt{2 D(Y)} dW(t)$$

... for a location-dependent “drift” $\mu(Y)$, “diffusion matrix” $D(Y)$ (larger in higher temperature regions), and W is the standard Brownian motion again. Intuitively, the “drift” pushes Y along the direction μ , and the “diffusion” controls how fast Y spreads out along each direction - or at least that’s how we think about it for *constant* drift and diffusion. The *probability distribution* of Y evolves over time according to:

$$\frac{\partial p}{\partial t}(y) = \nabla_y \cdot [-\mu(y)p(y) + \nabla_y \cdot (D(y)p(y))]$$

(This is the [Fokker-Planck equation](#).)

Key thing to notice: we can re-write this as

$$\frac{\partial p}{\partial t}(y) = \nabla_y \cdot [(-\mu(y) + \nabla_y \cdot D(y))p(y) + D(y) \cdot \nabla_y p(y)]$$

... so $-\nabla_y \cdot D(y)$ acts like a drift term, just like $\mu(y)$. This noise-induced drift is nonzero only when there's a noise gradient.

A very simple example of the math: suppose $D(y) = y_1 I$ (i.e. it's an identity matrix scaled by y_1). Then $-\nabla_y \cdot D(y) = -(1, 0, 0, \dots)$. So, the diffusion-gradient-induced drift is constant and along the $-y_1$ direction.

Putting It Together

So:

- SGD's path is approximately-Brownian, with location-dependent noise
- Brownian motion with location-dependent noise tends to drift down the noise gradient

In SGD, our “intended” drift is $-E_X[\nabla_\theta u(X, \theta)]$ - i.e. drift down the gradient of the objective.

But the location-dependent noise contributes a “bias” - a second drift term, resulting from drift down the noise-gradient. Combining the equations from the previous two sections, the noise-gradient-drift is

$$-\frac{1}{2} \# \nabla_\theta \cdot \text{Var}_X [\nabla_\theta u(X, \theta)]$$

I don't have much theory or evidence right now for what kinds of things that bias pushes towards (other than “regions of low gradient noise”), but having an explicit formula should help investigate that sort of question. Personally, I suspect that it pushes toward models with a modular structure reflecting the modularity structure of the environment, which is the main reason I'm interested in it.

Understanding the Lottery Ticket Hypothesis

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

Financial status: This is independent research. I welcome [financial support](#) to make further posts like this possible.

Epistemic status: The thread of research I'm reviewing here has been contentious in the past so I expect to make updates based on the comments.

Outline

- This is my attempt to understand the lottery ticket hypothesis.
- I review the original paper, as well as two follow-up papers and one related paper.
- I also go over four previous lesswrong posts on this subject.

Introduction

The lottery ticket hypothesis is a hypothesis about how neural network training works. It was proposed in 2018 by Jonathan Frankle, a PhD student at MIT, and Michael Carbin, a professor at MIT. It suggests that, in a certain sense, much of the action in neural network training is during initialization, not during optimization.

Research that sheds light on neural network training is relevant to alignment because neural network architectures may eventually become large enough to express dangerous patterns of cognition, and it seems unlikely that these patterns of cognition can be detected by input/output evaluations alone, so our only choices seem to be (1) abandon the contemporary machine learning paradigm and seek a new paradigm, or (2) augment the contemporary machine learning paradigm with some non-input/output method sufficient to avoid deploying dangerous patterns of cognition. Insights into contemporary machine learning effectiveness is relevant both to determining whether course (1) or (2) is more promising, and to executing course (2) if that turns out to be the better course.

The lottery ticket hypothesis

The lottery ticket hypothesis (or LTH), as originally articulated by [Frankle and Carbin](#), says:

A randomly-initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations.

A "subnetwork" means that you take the original neural network and clamp some of the weights to zero, so a network with N weights has 2^N subnetworks. The lottery ticket *conjecture*, which is an extension of the lottery ticket hypothesis but which Frankle and Carbin are careful to point out is not tested directly in their original paper says:

SGD seeks out and trains a subset of well-initialized weights. Dense, randomly-initialized networks are easier to train than the sparse networks that result from pruning because there are more possible subnetworks from which training might recover a winning ticket.

The lottery ticket conjecture is much more interesting from the perspective of understanding neural network training, and seems to be referred to as the "Lottery Ticket Hypothesis" both in the literature and on this site, so I will follow that convention in the remainder of this post. I will take the original lottery ticket hypothesis as an implication.

In neural network training, in the absence of the lottery ticket hypothesis, we think of the role of SGD as pushing the weights of the overall network towards a local minimum of a loss function, which measures the performance of the network on a training set:

On this view, the point in the loss landscape at which we begin optimizing — that is, the way we initialize the weights at the beginning of training — is not really where the main action is. Initialization might well be important, but the main point of initialization is to start in some not-completely-crazy part of the loss landscape, after which the optimization algorithm does the real work.

The lottery ticket hypothesis says that actually we should view a neural network as an ensemble of a huge number of sparse subnetworks, and that there is some as-yet poorly understood property of the initial weights of each of these subnetworks that determines to how quickly they will learn during training, and how well they will generalize at the end of training. The lottery ticket hypothesis says that among this huge ensemble of subnetworks, some number of subnetworks have this "trainability" property by virtue of having been initialized in accord with this as-yet poorly understood property. What the optimization algorithm is implicitly doing, then, is (1) identifying which subnetworks have this property, (2) training and upweighting them, and (3) downweighting the other networks that do not have this property.

Now this "lottery ticket view" of what is happening during neural network training does not exactly overturn the classical "whole network optimization view". The two views are of course compatible. But the lottery ticket hypothesis does make predictions, such as that it might be possible to give our entire network the as-yet poorly understood initialization property and improve training performance.

Now it's not that the winning "lottery ticket" is already trained, it just has some property that causes it to be efficiently trainable. There is some follow-up work concerning untrained subnetworks, but I do not believe that it suggests that neural network training consists of simply picking a subnetwork that already solves the task at hand. I discuss that work below under "supermasks" and "weight-agnostic networks".

Also, if some network *does* contain a subnetwork that would perform well at a task on its own, that does not mean that there is a neuron within the network that expresses

the output of this particular subnetwork. The output of any one neuron will be a combination of the outputs of all the subnetworks that do not mask that neuron out, which will generally include exponentially many subnetworks.

So how did Frankle and Carbin actually investigate this? They used the following procedure:

1. Train a dense neural network on a computer vision task
2. After training, pick a subnetwork that discards the bottom $X\%$ of weights ranked by absolute magnitude, for some values of X from 5% to 95%
3. Now reset the remaining weights to the value they had when the original dense network was initialized
4. Train this reduced subnetwork on the same computer vision task
5. Compare this to training the same reduced subnetwork initialized with freshly randomized weights

It turns out that the subnetworks produced by step 4 train faster and ultimately generalize better than the subnetworks produced by step 5. On this basis the authors conjecture that there was some special property of the initialization of this particular subnetwork, and that due to this property it trained efficiently relative to its peers, and that it was thus implicitly upweighted by the optimization algorithm.

In many of the experiments in the paper, the authors actually iterate steps 2-4 several times, pruning the weights gradually over several re-training phases rather than all at once after training the dense network just once.

When running experiments with larger architectures (VGG-19 and ResNet), the authors find:

We continue to find winning tickets for all of these architectures; however, our method for finding them, iterative pruning, is sensitive to the particular learning rate used.

In some [follow-up work](#), Frankle and Carbin also found it necessary to use "late resetting", which means resetting the weights of the subnetwork not to their original values from initialization but to their values from 1% - 7% of the way through training the dense network.

Deconstructing lottery tickets: signs, zeros, and the supermask

The suggestion that there might be some property of the initialization of a neural network that causes it to learn quickly and generalize well has prompted follow-up work trying to uncover the exact nature of that property. Zhou et al from Uber AI [investigated](#) the lottery ticket hypothesis and found the following.

First, among the weights that the sparse subnetwork is reset to in step 3, only the sign matters. If you replace all the positive weights with 1 and all the negative weights

with -1 then the subnetwork still trains efficiently and generalizes well, but if you randomize all the weights then this property disappears.

Second, if instead of clamping the pruned weights to zero you clamp them to their initial value from the dense network, then the good performance disappears. They hypothesize that clamping small-magnitude weights to zero is actually acting as a form of training since perhaps those weights were heading towards zero anyway.

Thirdly, and most fascinatingly, that they can find some "supermasks" such that merely applying the mask to an untrained dense network already produces better-than-random results. It is *not* that the supermask identifies a subnetwork that already solves the task. The untrained subnetwork identified by a supermask actually performs very poorly by the standards of any trained supervised learning system: 20% error on MNIST compared to 12% achieved by linear classification and 0.18% achieved by trained convnets. But 20% error is much better than the 90% error you would expect from chance predictions [1]. The authors suggest that we think of masking as a form of training.

One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers

In this paper, Morcos et al [show](#) that a "winning lottery ticket" subnetwork found by training a dense network on one dataset with one optimization algorithm still retains its attractive properties of efficient training and good generalization when the subnetwork is later trained on a different dataset or optimized by a different optimizer.

Weight-agnostic neural networks

This [paper from Google Brain](#) finds neural network architectures that perform well *no matter which weights are inserted into them*. They demonstrate networks solving reinforcement learning problems when all of their weights are set to 1, and then still solving the same problem when all of their weights are set to 2, 10, etc. The paper uses a completely different method to find architectures from the one used by Frankle and Carbin so this paper is a bit outside the lottery ticket literature, but it provides further evidence that weight training may not be the entirety of "where the action is at" in neural network training.

Daniel on the lottery ticket hypothesis and scaling

Daniel Kokotajlo asks whether the lottery ticket hypothesis, if true, would suggest that machine learning will continue to solve more problems as we apply more computing power to it. The reason is that more computing power means we can train networks with more parameters, which means we are searching over more "lottery tickets", which might mean we are able to solve increasingly difficult problems where lottery tickets are harder and harder to come by.

Evan on the lottery ticket hypothesis and deep double descent

Evan Hubinger [speculates](#) about whether the lottery ticket hypothesis might explain the deep double descent phenomenon:

My guess as to how double descent works if the Lottery Tickets Hypothesis is true is that in the interpolation regime SGD gets to just focus on the winning tickets and ignore the others—since it doesn't have to use the full model capacity—whereas on the interpolation threshold SGD is forced to make use of the full network (to get the full model capacity), not just the winning tickets, which hurts generalization. [...] That's just speculation on my part, however

John on the lottery ticket hypothesis and parameter tangent spaces

John Swentworth [proposes](#) an update to the lottery ticket hypothesis informed by recent results that show that the weights of large neural networks actually don't change very much over the course of training on practical machine learning problems:

At initialization, we randomly choose θ_0 , and that determines the parameter tangent space - that's our set of "lottery tickets". The SGD training process then solves the equations - it picks out the lottery tickets which perfectly match the data. In practice, there will be many such lottery tickets - many solutions to the equations - because modern nets are extremely overparameterized. SGD effectively picks one of them at random

I don't yet understand this proposal. In what way do we decompose this parameter tangent space into "lottery tickets"? Are the lottery tickets the cross product of subnetworks and points in the parameter tangent space? The subnetworks alone? If the latter then how does this differ from the original lottery ticket hypothesis?

John quotes the following synopsis of the lottery ticket hypothesis:

When the network is randomly initialized, there is a sub-network that is already decent at the task. Then, when training happens, that sub-network is reinforced and all other sub-networks are dampened so as to not interfere.

The "supermask" results above *do* suggest that this synopsis is accurate, so far as I can tell, but it's important to realize that "decent" might mean "better than random but worse than linear regression", and the already-present subnetwork does not *just* get reinforced during training, it also gets trained to a very significant extent. There is a [thread](#) between Daniel Kokotajlo and Daniel Filan about this synopsis that references several papers I haven't reviewed yet. They seem to agree that this synopsis is at least not implied by the experiments in the original lottery ticket hypothesis paper, which I agree is true.

Abram on the lottery ticket hypothesis and deception

Abram [points out](#) that the lottery ticket hypothesis being true could be disheartening news from the perspective of safety:

My Contentious Position for this subsection: Some versions of the lottery ticket hypothesis seem to imply that deceptive circuits are already present at the beginning of training.

Daniel provides the following [helpful summary](#) of Abram's argument in a comment:

[the parameter tangent space version of the lottery ticket hypothesis] seems to be saying that the training process basically just throws away all the tickets that score less than perfectly, and randomly selects one of the rest. This means that tickets which are deceptive agents and whatnot are in there from the beginning, and if they score well, then they have as much chance of being selected at the end as anything else that scores well. And since we should expect deceptive agents that score well to outnumber aligned agents that score well... we should expect deception.

I previously [attempted to summarize](#) this post by Abram.

Conclusion

It's exciting to see these insights being developed within the mainstream machine learning literature. It's also exciting to see their safety implications beginning to be fleshed out here. I hope this post helps by summarizing some of the experimental results that have led to these hypotheses.

-
1. We would expect 90% error from chance predictions since MNIST is a handwritten digit recognition dataset with 10 possible labels. [←](#)

Decoupling deliberation from competition

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

I view [intent alignment](#) as one step towards a broader goal of decoupling deliberation from competition.

- **Deliberation.** Thinking about what we want, learning about the world, talking and learning from each other, resolving our disagreements, figuring out better methodologies for making further progress...
- **Competition.** Making money and racing to build infrastructure, managing political campaigns and maneuvering within the political system, running ads to persuade people, fighting wars...

Competition pushes us to become the kind of people and communities who can win a fight, to delegate to whichever kind of AI is available first, and to adopt whatever ideologies are most memetically fit.

Deliberation pushes us to become the kind of people and communities who we want to be, to delegate only when we trust an AIs judgment more than our own, and to adopt views that we really believe.

I think it's likely that competition is going to accelerate and become more complex over the next 100 years, especially as AI systems begin to replace humans and compete on our behalf. I'm afraid that this may derail human deliberation and lead us to a place we don't want to go.

Decoupling

I would like humans and humanity to have the time, space, and safety to grow and change in whatever way we decide—individually and collectively—that we want to.

You could try to achieve this by “pausing” competition. Alice and Bob could agree to stop fighting while they try to figure out what they want and work out their disagreements. But that's a tall order—it requires halting not only military conflict, but any economic development that could put someone at an advantage later on. I don't want to dismiss this kind of ambitious goal ([related post](#)), but I think it's uncertain and long-term enough that you probably want a stop-gap solution.

An alternative approach is to “decouple” competition from deliberation. Alice and Bob keep competing, but they try to make sure that deliberation happens independently and the result isn't affected by competition. (“Pausing” is the special case of decoupling where deliberation finishes before competition starts.)

In a world without AI, decoupling is possible to a limited extent. Alice and Bob can spend time competing while planning to deliberate later after the dust has settled(or have their descendants deliberate). But it's inevitable that Alice and Bob will be different after competing with each other for many years, and so they are not completely decoupled.

Alignment and decoupling

Aligned AI may eventually make decoupling much easier. Instead of Alice and Bob competing directly, they may delegate to AI systems who will make money and fight wars and keep them safe. Once Alice and Bob have a clearer sense of what they want, they can direct their AI to use its influence appropriately. (This is closely related to [the strategy stealing assumption](#).)

Eventually it doesn't even matter if Alice and Bob participate in the competition themselves, since their personal contribution would be so small relative to their AIs. At that point it's easy for Alice and Bob to spend their time deliberating instead of thinking about competition at all.

If their AI systems are competent enough to keep them safe and isolate them from the fallout from competition, then the outcome of their deliberation doesn't depend much on the competition occurring in the background.

Misalignment and coupling

Misaligned AI could instead introduce a severe **coupling**. In the worst case, my best strategy to compete is to build and empower AI systems who want to compete, and my AI also ends up competing with *me* in the long run.

In the catastrophe scenario, we have relatively little control over how our society's values evolve—we end up pursuing whatever kinds of goals the most competent AI systems typically pursue.

Discussions of alignment often drift to questions like “But what do we *really* want?” or “how do we handle humanity’s diverse and sometimes-conflicting desires?”

Those questions seem important and challenging, but I think it's clear that the answers **shouldn't** depend on whatever values are easiest to give AI. That is, we want to decouple the question “what should we do in light of uncertainty and disagreement?” from the question “what is the most effective AI design for making money?”

Appendix: a bunch of random thoughts

Persuasion and limits of decoupling

Persuasion often doesn't fit cleanly into “deliberation” or “competition.”

On the one hand, talking to people is a critical part of deliberation:

- It's a fundamental part of reconciling conflicting desires and deciding what we collectively want.
- Having contact with people, and being influenced by people around us, helps us become the people/communities we want to become (and to stay sane).
- Other people have experiences and knowledge we don't, and may think in different ways that improve the quality of the group's conclusions.

- Being exposed to good arguments for a view, discovered by people who take it seriously, can be a step in evaluating that view.

On the other hand, the exact same kinds of interaction give scope for competition and manipulation:

- If Alice and Bob are talking to each other as they deliberate, each has a motive to influence the other by carefully filtering what they say, making misleading statements, playing off of each other's fears or biases, and so on.
- The possibility of manipulation gives Alice and Bob a motive to race ahead and become smarter faster in order to manipulate each other. This is in conflict with an individual desire to take it slow (for example it may push them to delegate to unaligned AI).
- In communities with more individuals there are even more opportunities for conflict, e.g. to exploit group norms or skirt enforcement, to get more access to more people's attention, and so on. These can lead to similar deadweight loss or incentives to race.

Wei Dai has talked about many of these issues over the years on Less Wrong and this section is largely inspired by his comments or conversations with him.

I don't think intent alignment addresses this problem, and I'm not sure there's any clean answer. Some possible approaches:

- Alice and Bob can split up and deliberate separately, potentially for a very long time, before they are ready to reconvene. This may be compatible with Alice and Bob continuing to interact, but not with them genuinely learning from each other.
- Alice and Bob can try to have an agreement to avoid racing ahead or engaging in some kinds of manipulation, and analogous a broader society could adopt such norms or divide into communities with internal agreements of this form. They may want to make such agreements relatively early if there is growing suspicion about someone manipulating the social contract to empower themselves.
- If Alice and Bob split up for a while it may be difficult for them to reconvene, since either of them may have decided to adopt an adversarial stance while they were separated (and if they are adopt an adversarial stance it may be very hard to negotiate them in good faith until reaching technological maturity). They could take various exotic approaches to try overcoming this problem, e.g. sharing details of their history with each other or each embedding themselves in new communities (built for purpose with trusted provenance).

Overall I expect this to be messy. It's a place where I don't fully expect it to be possible to fully decouple competition and deliberation, and I wish we had a better story about how to deliberate well in light of that.

Politics of decoupling

Although I think "Decoupling deliberation and competition" is a broadly desirable goal, any implementation will likely benefit some people at others' expense (like many other efforts to improve the world). So I don't ever expect it to be a politically clean project.

For example:

- Without decoupling there may be mounting pressure to “pause” competition in various ways. Many pauses would result in big changes in the balance of power (e.g. by lowering the value of AI or of military capabilities). So you could easily end up with conflict between people who would prefer “pause” and those who prefer “decouple,” or between people who prefer different decoupling strategies.
- Failures of decoupling often push values in a predictable direction. For example, some people may simply want *something* to spread from Earth throughout the universe, and they benefit from coupling.
- There is tons of messiness around “persuasion.” Many decoupling approaches would reduce opportunities for some kinds of persuasion (e.g. buying ads or shouting at people), and that will inevitably disadvantage some people (e.g. those who have a lot of money to spend or those whose positions sound best in shouting matches). So people with those advantages may try to use them while possible in order to avoid decoupling.

A double-edged sword

I think that competition currently serves an important sanity-check on our deliberation, and getting rid of it is scary (even if I’m excited on balance).

In an idealized decoupling, the resources someone ends up with don’t depend at all on how they deliberate. This can result in dedicating massive resources to projects that no one really likes. For example:

- Alice may decide that she doesn’t care what happens with her resources and never wants to think about the question seriously. Normally she would get outcompeted by people who care more about future influence.
- Bob’s community may have deeply dysfunctional epistemic norms, leading them both to consistently make errors when thinking about empirical questions *and* to reach insane conclusions about what they should do with their resources. Normally they would get outcompeted by people with more accurate views.
- Charlie isn’t very careful or effective. Over the course of a long enough deliberative process they are inevitably going to build some misaligned AI or drive themselves insane or something. Normally their carelessness would lead to them gradually losing out relative to more effective agents.

I reasonably often find myself grateful that some dysfunctional norms or epistemic practices will most likely become obsolete. It’s a bit scary to think about a world where the only solution is waiting for someone to snap out of it.

Competition isn’t a robust safeguard, and it certainly isn’t optimal. A careful deliberator would make early steps to ensure that their deliberation had the same kind of robustness conferred by competition—for example they would be on the lookout for any places where their choices would lead to them getting outcompeted “in the wild” and then think carefully about whether they endorse those choices anyway. But I’m afraid that most of us are below the bar where paternalistically forcing us to “keep ourselves honest” is valuable.

I don’t really have a settled view on these questions. Overall I still feel comfortable with decoupling, but I hope that we can collectively decide on some regime that captures some of the benefits of this kind of “competitive discipline” without the costs. For example, even in a mostly-decoupled world we could end up agreeing on different domains of “safe” competition (e.g. it feels much better for states to compete on “being a great place to live” than to fight wars), or imposing temporary

paternalistic restrictions and relaxing them only once some reasonably high bar of competence is demonstrated.

The balance of power affects deliberation

Negotiation and compromise is an important part of deliberation, but it depends on the current state of competition.

Suppose that Alice and Bob start talking while they don't know who is going to end up more influential. But they are talking slowly, in the way most comfortable to them, while competition continues to accelerate at the maximum possible rate. So before they can reach any agreement, it may be clear that one of them is vastly more influential.

Alice and Bob would have preferred to make an early agreement to treat each other with respect, back when they were both ignorant about who would end up with the power. But this opportunity is lost forever once they have seen the outcome.

Alice and Bob can try to avoid this by quickly reaching a compromise. But that seems hard, and having to make a precise agreement fast may take them far away from the deliberative process they would have preferred.

I don't have any real story about coping with this problem, though I'm less worried about it than persuasion. Some possible (but pretty weird) approaches:

- If Alice and Bob need to reach an agreement in a hurry, they may be able to make some minimal agreement like "we'll both isolate ourselves so we don't see the result of the competition until we've reached a better agreement" or "long after the competition is over, we'll allocate resources based on a high-fidelity prediction of what we would have agreed to if we had never seen the result of the competition."
- After winning the competition and waiting for considerable technological progress, Alice can simulate what Bob would have done if *he* had won the competition. If this is done carefully, I think you can get to the situation where Alice really doesn't know if she won the competition (or if she is just in a simulation run by Bob to figure out how nice to be to Alice). Then we can have a discussion between Alice and simulated-Bob (who thinks of this as a conversation between Bob and simulated-Alice) from that state of ignorance.

The singularity, the distant future, and the “long reflection”

In some ways my picture of the future is very aggressive/unusual. For example I think that we are likely to see explosive economic growth and approximate technological maturity within the next 50–100 years (and potentially much sooner).

But in other ways it feels like I have a much more “boring” picture of the future. I expect technology could radically transform the world on a timescale that would be disorienting to people, but for the most part that's not how we *want* our lives to go in order to have the best chance of reaching the best conclusions about what to do in the long run. We do want some effects of technology—we would like to stop being so hungry and sick, to have a little bit less reason to be at each other's throats, and so

on—but we also want to be isolated from the incomprehensible, and to make some changes slowly and carefully.

So I expect there to be a very recognizable thread running through humanity's story, where many of the humans alive today just continue to being human and growing in a way that is familiar and comfortable, perhaps changing more quickly than we have in the past but never so quickly that we are at risk of losing our footing. The point of this is not because that's how to have the best life (which may well involve incomprehensible mind-alteration or hyper-optimized virtual reality or whatever). It's because we still have a job to do.

The fact that you are able to modify a human to be much smarter does not mean that you need to, and indeed I think it's important that you take that process slow. The kinds of moral change we are most familiar with and trust involve a bunch of people thinking and talking, gradually refining their norms and making small changes to their nature, raising new generations one after another.

During that time we have a lot to do to safeguard the process; to become more and more comfortable that it's proceeding in a good direction even as we become wiser and wiser; to do lots of moral philosophy and political philosophy and psychology at every stage in case they provide clues about how to take the next step wisely. We can take the things that scare us or that we dislike about ourselves, and we can very gingerly remove or change them piece by piece. But I think it doesn't have to be nearly as *weird* as people often imagine it.

Moreover, I think that the community of humans taking things slowly and living recognizable lives isn't an irrelevant sideshow that anyone serious would ignore in favor of thinking about the crazy stuff AI is doing "out there" (or the hyper-optimized experiences some of our descendants may immerse themselves in). I think there's a real sense in which it's the main thread of the human story; it's the thread that determines our future and gradually expands to fill the universe.

Put differently, I think people sometimes imagine abdicating responsibility to crazy AI systems that humans build. I think that will happen someday, but not when we can *first* build AI—indeed, it won't happen until those AI systems no longer seem crazy.

In the weirdest cases, we decouple by building an AI that merely needs to think about what humans *would* want rather than deferring to any real flesh-and-blood humans. But even those cases are more like a change than an ending—we pack up our things from Earth and continue our story inside a homey simulation. And personally I don't expect to do even that until everyone is good and ready for it, many years after it first becomes possible.

(Trying To) Study Textbooks Effectively: A Year of Experimentation

When I started studying the art of studying, I wanted to understand the role of book learning. How do we best learn from a textbook, scientific article, or nonfiction book? What can a student of average intelligence do to stay on top of their homework? Is it possible to improve your annual knowledge growth rate by one or two percent by learning how to learn? Should a motivated student take a maximizing or satisfying approach to their coursework? How many of the skills of a top scholar are strategic, collaborative, psychological, or involve merely a set of habits and technological proficiencies?

Fortunately, I started with the most esoteric of approaches, exploring visualization. I tried using a memory palace to memorize a textbook. It was vivid, fun, and creative. Exploring visualization helped me understand chemical diagrams, led me to invent a math problem, and made learning a lot more fun. But I simply couldn't jam that much detailed technical knowledge into my head. The method didn't help me pass my final exam, and I dropped it.

Posts from this era include [Visual Babble and Prune](#), [Using a memory palace to memorize a textbook](#), [The point of a memory palace](#), [Visualizing the textbook for fun and profit](#),

After that, I explored speed reading. I read the theory, experimented both with physical technique and speed reading apps, and kind of broke my reading habits developing this difficult-to-correct tendency to skim. This tendency to read too quickly persisted long after I'd dropped deliberate attempts at speed reading. I finally made some intellectual progress, which preceded correcting the reading habit itself, in [The Comprehension Curve](#).

Then I explored the world of Anki and tried to use flashcards to memorize a textbook instead (or at least a few chapters). After simulating the sheer amount of flashcard review I'd have to do to keep a strategy like that up long-term, I dropped that too. I felt that forming memories of narrow facts (like the structure of RNA polymerase or the name of the 7th enzyme in glycolysis) was the costliest way to learn. And I found the achievement of world-class memory champions irrelevant to real-world learning, which just seems like an entirely different task.

Posts from this area (not all on flashcards specifically) include [The Multi-Tower Study Strategy](#), [Define Your Learning Goal: Competence Or Broad Knowledge](#), [Progressive Highlighting: Picking What To Make Into Flashcards](#), [Goldfish Reading](#), [Curious Inquiry and Rigorous Training](#), and [Using Flashcards for Deliberate Practice](#).

During this time, I also played around with "just reading," without a conscious technique. Posts from this era include [Check OK, babble-read, optimize \(how I read textbooks\)](#), [Wild Reading](#),

Notes are cheap. It takes a lot less time to write down a fact than to memorize it. But I went further. I developed an elaborate and carefully-specified system of shorthand notation to represent causal, temporal, and physical structures. It used Newick notation for tree structures, variants on arrow signs to articulate causation, sequence, combination, and more, templates to rewrite the stereotyped information presented

by textbooks in a uniform format, and hyperlinks in Obsidian to represent the relationships between concepts.

Not only did I take notes on the textbook, I also took notes on each individual homework problem. I also developed notes for other problems. I wrote Question Notes for The Precipice. This means that for each paragraph in the book, I wrote down one question to which that paragraph was a valid answer.

I never published any posts on note-taking. Partly, note-taking itself scratched that itch. But more importantly, it was a very fast iterative cycle. My methods developed day by day, over the course of months. I was experimenting with different software apps, tweaking the templates I used, figuring out how to expand my particular method of shorthand to represent complex structures. After all the shifts I'd made on my previous experiments, I thought I would spare LessWrong the tedious minutiae of my developing thoughts on note-taking. I'm confident that crafting the perfect notes in an elaborate and precise shorthand system is no a panacea, so I don't know if it's worth bothering.

Exploring note-taking was as useful as visualizing was fun. The rigid structure of my note-taking approach gave me clear guidance on what it means to "read" or "study" a textbook chapter. They became a useful reference for looking things up. The idea of bringing together any data, formula, charts, or techniques I needed to solve a problem, and then making a plan of attack before setting to work, was a big upgrade for my accuracy and sense of ease.

Yet when my note-taking apotheosized after several iterations of improving my diagrammatic shorthand to deal with weird edge cases, and shifting from Evernote's WYSIWYG editor to Obsidian's markdown editor and full support for folders and hyperlinks, I found that not only was my approach to note-taking incredibly laborious, it was also profoundly distracting. It shifted my focus from building an intuitive feeling of understanding the material to constructing a precise translation of the material. At the end, I'd have a carefully notated description of a biochemical process, but virtually no ability to describe even the basics without reference to my notes. The experience of reading shifted from enjoyable, while visualizing, to the frantic skimming of flashcards, to sheer drudgery with note-taking. It didn't feel at all like programming, which is an activity I enjoy and that I'd hoped my note-taking would mimic.

It came back to me, then, after almost a year since I'd given much focused thought to visualization, that I should try just reading a chapter - no flashcards, no notes, no nothin' - and just try to picture everything as I went along, with no worries about trying to remember it all as I went. What do you know? The old spark returned! It was fun again! I breezed through a chapter on transcription, and had no trouble banging through the homework immediately afterward. Not only did I understand it better as I went, I was having more fun.

Now that I look back on the last year of exploring these issues, I see that I've only just now completed a single iteration of the Grand Study Problem, which is explaining how all these techniques, and possibly others, fit together into a technique for effective scholarship. Surely, it's partly about focused memorization (flashcards). Partly, it's about searching, note-taking, planning and problem-solving. And partly, it's about visualizing, anthropomorphizing, storytelling, model-building, and all the other ways of engaging your senses. What can I say about each of them?

If you're visualizing it, almost every textbook sentence provides you with an opportunity to create a new image in your mind. As you progress further through the textbook, it will call back to more and more earlier concepts. In biochemistry, it's things like the relationship between Gibbs free energy, enthalpy, entropy, and electrostatic potential; the amino acids; the nucleotides; different types of lipids; and a variety of major enzymes (i.e. DNA polymerase) and pathways (i.e. glycolysis). If you can figure out what those concepts are, and memorize them, you'll be able to picture them when it mentions them casually in passing. If you can't remember glutamine's abbreviation or chemical structure, then every time the book mentions G (or is it E?), you'll miss out on an opportunity to practice recalling it, or else you'll have to interrupt your flow to look it up for the umpteenth time. This is a role for flashcards and super-convenient reference charts. Some knowledge is most helpful if you can access it in five seconds or less.

Note-taking is incredibly helpful for focusing, but so is visualizing. I still think that there's a big role for taking really good notes, and assembling other reference and search tools. Yet taking notes needs to be balanced with enjoyable reading and building an intuition for the subject matter, and I think that comes from a visual approach first. Render it down into symbols later.

Along the way, I've written an informal scientific journal with my current working hypothesis, motivations for trying it out, tests, limitations, and future directions. This has been very helpful for giving these experiments a sense of direction.

One unifying trait so far is that each experiment has focused on one technique: visualization, memorization, note-taking, and now back to visualizing. It seems to me now that each of these has a purpose. Visualization puts the fun, creativity, and intuition in learning, and it's also fundamental to understanding anything that has a physical form. Memorization is important so that when you learn a concept in Chapter 2 that reappears persistently over the next 22 chapters, you aren't just reading words on the page, but are able to recall a concept to mind. That way, the rest of your reading refreshes and extends your memory of that initial concept. Note-taking and reference-sheet-making is helpful as a way of optimizing and compressing the natural-language, beginner-oriented version you get in a textbook into a format more suitable for review or looking up particular details. Figuring out how to interleave these three techniques will probably be the focus of my next iteration of this exploration.

Bayeswatch 5: Hivemind

Three humans sat around a table. The notary was unaugmented. He wore a charcoal business suit. The ancillary wore a silk paisley print top zippered at the back. A thick cable extended from the back of her head to the collective's mainframe. Trinity nervously fondled the virgin socket embedded in the occipital bone of her own skull.

"Please confirm your DNA," said the notary.

Trinity pricked her thumb on the bloodometer.

"You are Trinity Sariah Rees," said the notary.

"You don't say," Trinity rolled her eyes.

"Nineteen years old. Mormon..." said the notary.

"Ex-Mormon," corrected Trinity.

"Please speak into the microphone. We need an unambiguous record of your consent," said the notary.

"This is ridiculous. It's my body and mind. I should have the freedom to do what I want with it," said Trinity.

"Informed consent is important," said the ancillary, "We don't want to get into legal trouble. Tomorrow you will feel the same way."

"Which is why we need consent now. It won't count after the procedure," said the notary.

"I already had the surgery. All we're doing is plugging me in," Trinity crossed her arms and legs.

"The government is prejudiced against transhumans," said the ancillary, "It could be worse. We're lucky all they demand is paperwork."

"Let's get this done quickly," said Trinity.

"Are you really ready to have five sixths of your personality erased?" said an ancillary.

"That five sixths will be equally divided among you guys. Are you ready to have one sixth of your personality become me?" said Trinity.

"It sounds like you are familiar with predictive processing and connectome-specific harmonic waves," said the notary.

"I have wanted to join a collective since I was thirteen. I keep up-to-date on the newest research," said Trinity.

"Then I don't need to explain how modern neuroscience is based on the idea that synapse weights modify the topology of your brain like how gravity bends spacetime," said the notary.

"Do you always use general relativity metaphors when explaining things to people?" said Trinity.

"I didn't want to insult your intelligence," said the notary.

"It's about time someone recognized it," said Trinity.

"We do too. We can sync only a handful of brains before the global workspace fragments. We must be selective about who we assimilate. High g-factor is a priority," said the ancillary.

"Anyway," said the notary, "The primary purpose of the human brain is to create a predictive model of its environment. Everything else is secondary."

"Technically it is the primary purpose of consciousness to create a predictive model of its environment," said Trinity, "The human brain does other things too. What really matters is the connectome. When you link multiple connectomes together cybernetically you combine their individual topologies into a single larger topology. It is the opposite of a corpus callosotomy. Software in the synchronizing mainframe prevents epileptic seizures. Collectivization is so safe these days people with epilepsy often cure themselves by joining a collective. Collectivization is simple technology. If it weren't for runaway resonance you could do the whole thing with analog electronics. "

"That's the most technical summary I've heard from any client so far. It sounds like you know exactly what you are doing. I hereby confirm you have informed consent. Just sign this document and you are certified to assimilate," said the notary.

Trinity signed.

"Congratulations," said the notary, "You are legally part of the collective. Everything else is just software."

Trinity sat in the collective's assimilation chair. All five ancillaries of the collective stood around her. Their tether cables were decorated in paints and gemstones. They secured her wrists and ankles with padded shackles. A clamp stabilized her head. A belt went around her waist.

"We don't want you to hurt yourself if there are muscle spasms," said an ancillary.

"Just give me the mouth guard," said Trinity.

An ancillary inserted it.

"Ready?" said an ancillary.

Trinity was too safely secured to nod her head properly. She managed an upward twitch.

The collective began its sync. Trinity received five universes of semantic data from the other ancillaries. Her brain emulated them. The collective personality overwrote her own.

Trinity tried to sense her own personality overwriting theirs. Communication was a one-way street. The collective had hacked the interface to transmit data to her brain

but not from her brain. She struggled against the shackles. She tried to scream but her mouth was clamped shut.

The collective released her head after six hours. It compelled her to eat and drink. The assimilation resumed.

Concerning not getting lost

Content warning: striving, awkwardness, expressions of personal insecurities

Outline

- This post is somewhat personal.
- I try to convey why I think it's important to stay in touch with a certain personal voice.
- The reason I give is that without this personal voice, we are liable to lose track of our terminal goals and get lost.
- By "persona voice", I don't mean "emotional" or "following the gut" or "following the heart". This post is my own attempt at writing from this personal voice. I hope that this conveys what I mean by "personal voice" even if the explication I give does not.

Concerning not getting lost

There was something on the tip of my tongue that I wanted to write, just before I sat down. There was a monologue playing out in my mind, and it was beautiful, and it was relevant, and I thought I would write it down and share it with you all so that it wouldn't just be me alone hearing it in my head.

It had to do with the agent model, and with notions of self. It had to do with loneliness. It was personal. And it was relevant to the conversation happening here in the place where we discuss AI and the future of life on the planet.

It is so very difficult to speak about that which is personal. But actually this is all personal to me — this whole journey we're on to navigate the development of AI. Every one of the small number of posts I've written here over the past ten years has been personal to me, but I've mostly hidden that personal dimension behind a certain kind of wall.

And that's fine. Our purpose here is to navigate a very real and very threatening state of affairs out in the world. We should do whatever needs to be done in order to navigate this threat, and that may mean putting aside that which is personal in order to get the job done. But it seems to me that when we leave aside that which is personal we actually forego a certain kind of power. We lose the compass within ourselves that separates that which is important from that which is a distraction. And in a world rife with distracting things, there is no place in the world where this personal compass is more important than within the community of people working to navigate the most significant threats to life on this planet.

As I write this, I am finding it difficult to stay with it. I am getting up and moving about and thinking of going someplace else where I don't have to open up in the way that writing in this style forces me to. But I'm resolving to stay here. This little thread of life within me is going to be brought forth, even if it takes the rest of my life.

Because what the world needs right now more than anything is this analytical capacity to make sense of the powerful systems we humans are constructing, *in partnership with* this deeply personal compass that discerns that which is worthy of our lives' work from that which is a distraction.

And there is no time or place where it is more difficult to convince this personal voice to stay online than in the midst of a great crisis, and in a place where people have gathered to work out what to do. That is what this place is, my friends, this community, this website. It is a place where we have gathered to work out what to do, and it is at this time that this personal compass is most needed, and also most fleeting.

I am not talking about emotions, my friends. When I say "this personal compass", I am not talking about emotions. Why is it that we have kept coming back to this place, this question, this issue of understanding intelligence and agency and knowledge, year after year, while all else in my life has changed and changed? What is it that has brought me back so consistently, over such a long period, with no discernible change in the quality of its power? It is not my emotions, my friends. My emotions are fickle over the course of about ten seconds, nevermind ten years. There is no possibility that emotions could be the driver of something with this kind of regularity.

And I am not talking about any sensations that I feel in my stomach, or in my gut, or in my heart.

I am talking about the personal chord that cuts right through my soul and out the other side, and is the heartbroken and helpless basis from which everything I have ever written has proceeded. I am talking about speaking *directly* from that place, not because it feels good but because it is *needed*. It is needed because it is this place that can keep us on track, can keep our reasoning minds, so brilliant and capable, focussed not just on a goal but on the right goal, on what truly matters. It is a partnership between our analytical capacities and this personal compass that can resolve this AI problem, I believe.

I am afraid, my friends. Not so much of the world being destroyed, but much more of the world not being destroyed, and not being part of it. I am afraid of being cut off, dismissed, disallowed from participating in this great voyage. Because I am afraid, I participate here on this website from a colder, harder place within myself. But I would rather not do that, friends, so today I'm practicing speaking from this personal place. Perhaps with practice I will get better at this.

Because the work we are doing here is too important for us to put aside that part of ourselves that is not cold and not hard and not confused about what is worth centering our life's work around. This is about what *works*. How do we actually do this? How, at a very practical level, do we navigate the development of these powerful technologies? How do we pull that off? It cannot be done without a clear connection to the voice that brought us here in the first place, to the voice that provides our reasoning minds with a clear purpose from which to reason. This is not an abstract spiritual consideration, friends. It is a highly practical consideration.

I have spent several hours now writing the few words so far in this post, and my body feels sore, and I can't quite get comfortable in my chair. What is left to say here?

Don't get lost. There is a voice within us that sees a larger picture, has been seeing it our whole lives. Our task is to hear that voice clearly, and find within that voice the clarity of purpose that is needed to put this practical task before us to rest, in order

that future generations may have the opportunity to find that voice within themselves and offer it as their own gift to their own world. Perhaps their world will not face the kind of danger that we face now. Perhaps this particular task can be laid to rest here and now, in our lifetimes, by us, and the safety hence secured can be offered as the greatest of all possible gifts to our children.

These are the table stakes, my friends. There is nothing more worthy of our life's work.

[link] If something seems unusually hard for you, see if you're missing a minor insight

This is a linkpost for <https://drmaciver.substack.com/p/how-to-do-everything>

Related to: [procedural knowledge gaps](#), [trivial inconveniences](#), [errors vs. bugs](#).

Summary of the linked post: the world is full of all kinds of minor insights about how to do something (either easily or at all): how to open stitched bags, how to mop floors, how to use search engines, the right way to look at a ball in order to catch it, what it means to 'ping' somebody, how the skill of playing sudoku shares useful elements with other things like proving math theorems, how to pass as 'normal', etc.

For anything that seems unusually hard for you, there's a chance that there's some simple insight that you happen to be missing that would make it easier. You could put a lot of effort into trying to laboriously level up the skill, or you could see if you could acquire that insight by relatively little work.

I particularly like the post for having an extended list of examples of the thing, hopefully making it easier to notice potential applications of this principle when they come up in your own life. (The example that came to my own mind was finding it unexpectedly aversive to vacuum in my current home, when it had felt fine in the previous one, and then only eventually realizing that my former housemate's vacuum cleaner that I'd been using in my previous home was *much* more pleasant to use than my own.)

Covid 5/20: The Great Unmasking

[The CDC has lifted its mask mandate for vaccinated people.](#) In one fell swoop, it is suddenly fine for anyone fully vaccinated to go maskless anywhere, without social distancing, at any time, so long as regulations and rules permit it. This retains the madness of ignoring *partial* vaccination entirely, and the madness of making children as young as two (!) wear masks around groups of fully vaccinated adults, but it is certainly a huge step towards sanity. It also implies that the rules and regulations that remain need to be changed, and many of them are indeed being changed.

Reactions were mixed.

There was much rejoicing. Many cried tears of joy. Others pointed out that this was a damn good reason to get vaccinated.

There was also much confusion and opposition. Very Serious People expressed panic and horror. Others expressed reasonable concern that this was premature. This would destroy mask mandates, causing maskless people to go around maskless. People would go around acting like the pandemic was over before there was full 'herd immunity.'

The opposition has some legitimate points. The control system could still have some fight in it. We have no way to verify verification status that anyone is willing to use. Vaccination checks are mostly going to be either non-existent or on the honor system, and America does not highly value honor. Norms that are ignored by half the people are easy for the other half to choose to ignore as well. At a minimum, this will change norms in ways that slow down the decline in cases. If everyone were to use this as a reason to go fully 'back to normal' at current vaccination rates cases would presumably start rising again at least for a while, and it's not clear we could do much of anything to reverse course if that happened beyond pushing harder to get more vaccinations.

A lot of people also pointed out the contradictions between this new policy and the old one, with the change from 'take precautions that wouldn't even make sense if you were unvaccinated' to 'once vaccinated take no precautions whatsoever' coming overnight with zero warning, with little or no change in the underlying physical conditions. I think it's important not to make a big deal about that. Yes, the contradictions are huge and blatant, but the time to point out how crazy the old guidance was was when they were still in place. As many people did, frequently. Admitting one is wrong and correcting errors needs to be rewarded and encouraged rather than punished and piled onto.

Despite the risks, and it plausibly being *slightly* too soon (on the order of a few weeks) to do this full blast, I think this was *absolutely* the right move to make. I congratulate the CDC on its newfound wisdom. Now let's work on the insane mask and distancing mandates that still apply to children, and on updating rules and regulations to reflect the newly acknowledged realities of the physical world. I'm happy to have gone from 'worried about Covid when going outside' to 'worrying about insect bites eating me alive when going outside' to this week's 'worrying about seasonal allergies when going outside.'

Also, let's run the numbers.

Author's note: It's been a hectic week, which is why this is going out in the late hours. Some stuff likely got left out due to that, other stuff got only a link without a discussion.

The Numbers

Predictions

Prediction from last week: Positivity rate of 3.0% (down 0.4%) and deaths decline by 10%.

Result according to Washington Post:

In the past week in the U.S. ...

New daily reported **cases fell 17.1% ↓**

New daily reported **deaths fell 5.9% ↓**

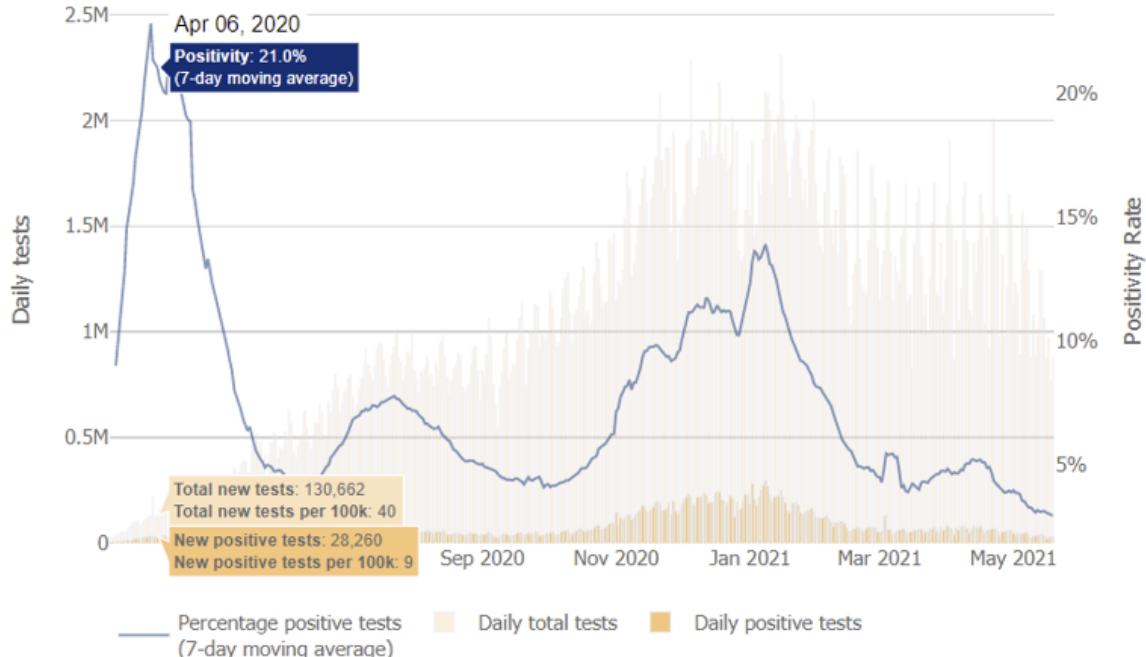
Covid-related **hospitalizations fell 13.2% ↓** [Read more](#)

Among reported tests, **the positivity rate was 4.7%**.

The **number of tests reported fell 15.6% ↓** from the previous week. [Read more](#)

So they're saying that cases fell 17%, tests fell 16% and the positivity rate is up from 3.4% to 4.7%. Yeah, no. Sorry Washington Post, this is even more obviously nonsense than before.

[Let's see Johns Hopkins.](#)



That seems more like a thing that could possibly have happened. They have 3.0% positivity rate, down from 3.2% (a 0.2% drop). That's still an overshoot, and somewhat disappointing, but test counts are also dropping rapidly.

Prediction (now using Johns Hopkins numbers fully going forward): Positivity rate of 2.7% (down 0.3%) and deaths fall by 8%.

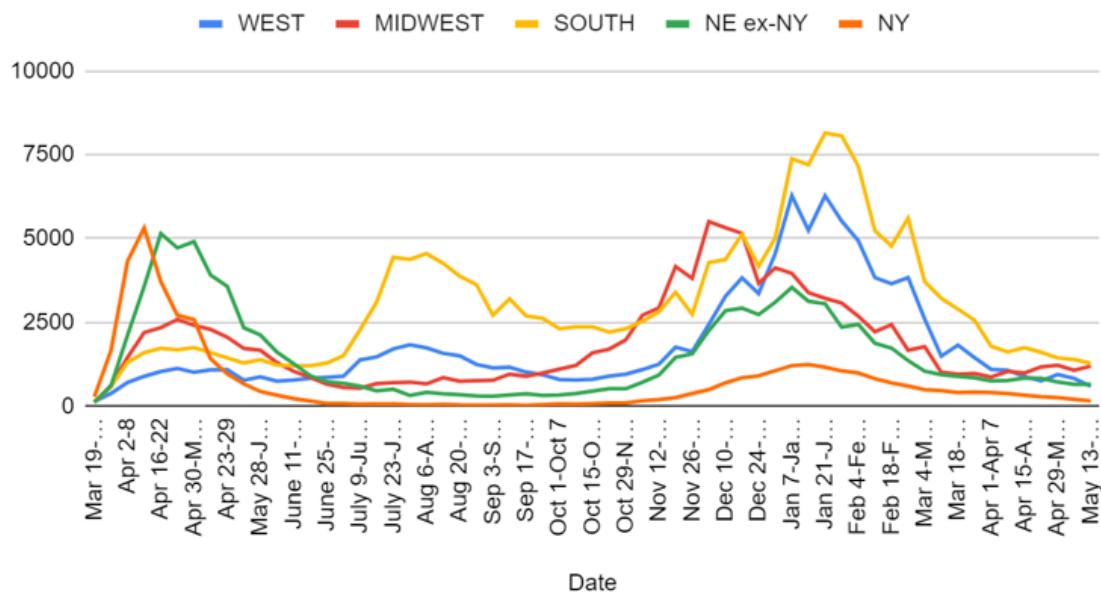
It has been seven days since the CDC updated its guidance. That's not quite enough time that we would start to see the effects of the change. We won't see the full effects next week because this kind of shift in behavior happens slowly, but by next week we will know if we have made a huge mistake.

Thus this is an important week. If conditions continue to improve at a similar pace, we have won.

I think there is roughly a 75% chance that we'll see things going along that path. Then there's about a 15% chance that there's a noticeable change but not one that makes me worry that things are about to get worse, with confidence that ongoing vaccinations will keep things contained. Then that leaves a 10% chance that things get concerning, and the decision to change standards will look dangerously premature.

Deaths

Deaths by Region

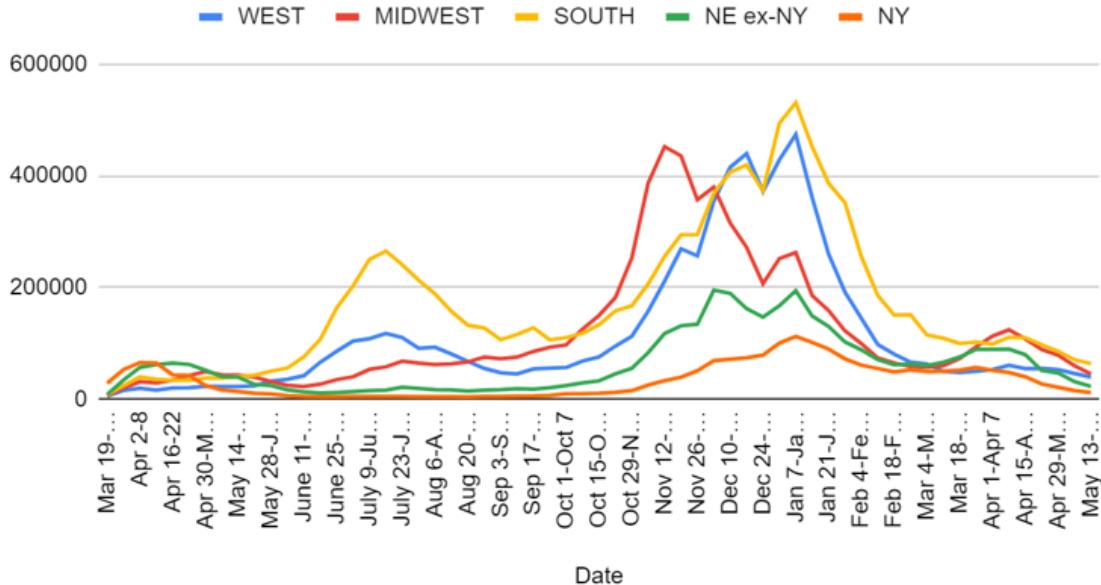


Date	WEST	MIDWEST	SOUTH	NORTHEAST
Apr 1-Apr 7	1098	867	1789	1160
Apr 8-Apr 14	1070	1037	1621	1145
Apr 15-Apr 21	883	987	1747	1168
Apr 22-Apr 28	752	1173	1609	1110
Apr 29-May 5	943	1220	1440	971
May 6-May 12	826	1069	1392	855
May 13-May 19	592	1194	1277	811

Overall a slow but steady decline, but that hides a big decline in the West and a backslide in the Midwest. I doubt either of them reflect a major real shift, and I'm guessing they mostly cancel out, and we likely saw a slightly higher real drop than we observed, given past infection data.

Cases

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Apr 1-Apr 7	52,891	112,848	98,390	140,739
Apr 8-Apr 14	60,693	124,161	110,995	137,213
Apr 15-Apr 21	54,778	107,700	110,160	119,542
Apr 22-Apr 28	54,887	88,973	97,482	78,442
Apr 29-May 5	52,984	78,778	85,641	68,299
May 6-May 12	46,045	59,945	70,740	46,782
May 13-May 19	39,601	45,030	63,529	34,309

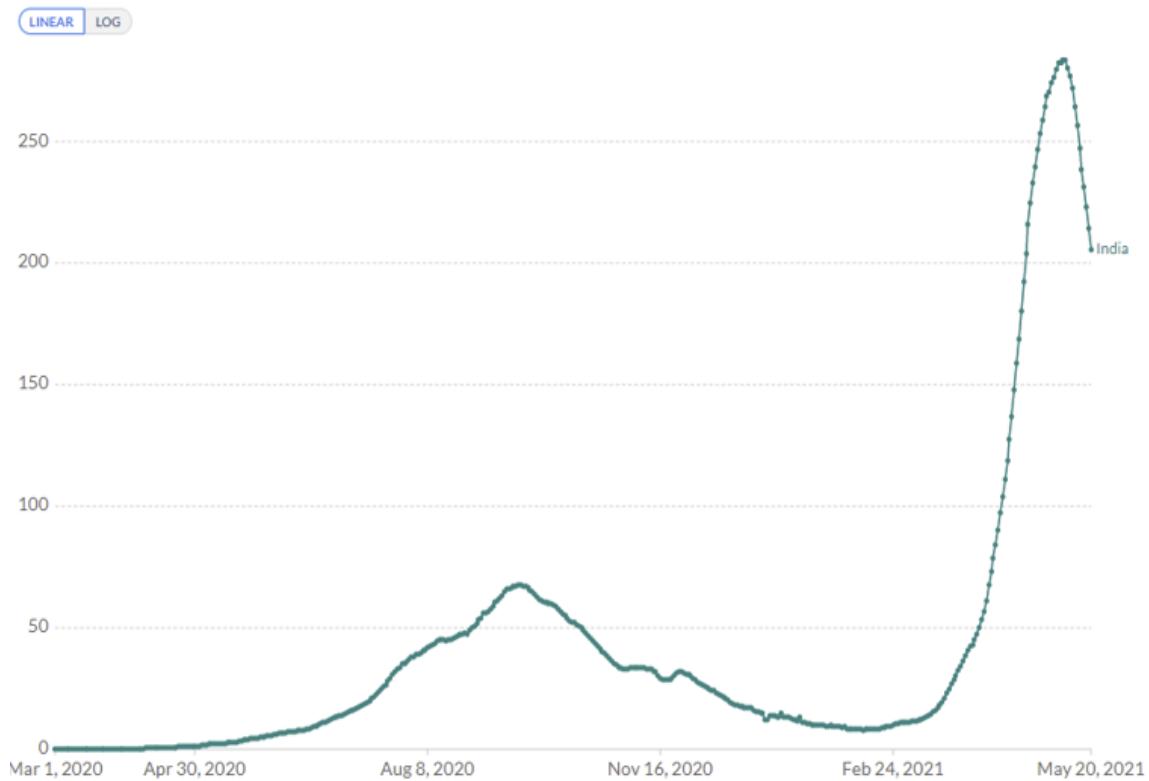
Good drops across the board, with very large drops continuing in the Midwest and Northeast. You love to see it. If this doesn't change direction soon, it will soon be over.

India

Daily new confirmed COVID-19 cases per million people

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

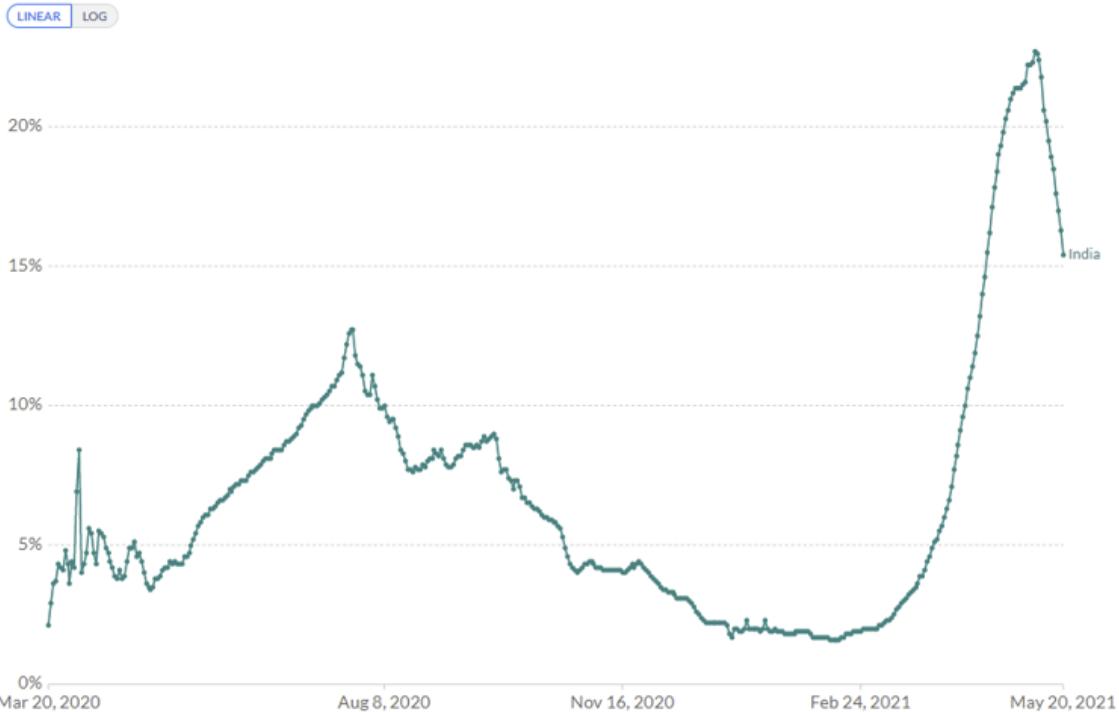
Our World
in Data



The share of daily COVID-19 tests that are positive

Shown is the rolling 7-day average. The number of confirmed cases divided by the number of tests, expressed as a percentage. Tests may refer to the number of tests performed or the number of people tested – depending on which is reported by the particular country.

Our World
in Data



Looks like India will be through the worst of things within a few weeks. Perhaps the elections or various festivals were major contributors to the issue. Maybe the control system is simply that strong anywhere and everywhere. Somehow, trends have reversed and things are rapidly improving, although deaths won't come that far down for a while and the hospitals will remain in dire straits.

I do not buy ‘seasonality’ as an explanation here because you would have seen more of a gradual phase transition and plateau, rather than a dramatic reversal. This is humans making different choices.

The Great Unmasking

If there was any doubt that changing CDC guidance would have a cascading effect on other guidance and on behavior, those doubts were put aside quickly.

Many jurisdictions and private entities come in line with the new principles within days. For example, [here's Pennsylvania falling in line within 4 hours](#). Anyone *not* falling in line is coming under increasing pressure to do so. [Here's New York falling in line a few days later](#), with an exception (that I fully support) for public transportation. [Here's Massachusetts declaring that we have normality](#). [Here's the US Senate](#). For a change of pace [here's Trader Joe's](#), along with a general ‘too much too soon’ objection.

My previous prediction was that if the CDC were to update its guidance in sensible ways, that elite opinion would fall in line with that. Those defending the old guidelines would shift to defending the new ones because their actual algorithm is to defend the official line, so opposition to a change *before it happens* is a very different thing from opposition to a change *after it is announced*.

What we saw was that everyone's opinion *on the science* seemed to move on the spot, and everyone agreed that oh yes, fully vaccinated people are at and present minimal risks. That's definitely *not* what a lot of those same people had been saying directly before the announcement.

Still, we saw objections that the science was accurate, but this was still too much, too fast, too soon. The CDC might have abandoned its scare tactics and worries, but that problem was clearly going to be a lot stickier. As well it should be a sticker problem. There's some very reasonable objections to deal with.

The objections seem to be something like (using the term 'we' here for ease of comprehension to refer to the Powers That Be and their decisions):

1. Vaccinated people still can spread the virus *to some extent*, and it's thus irresponsible to encourage them to do less prevention, since every bit of prevention helps in a pandemic.
2. We still don't know the extent to which the virus spreads among the vaccinated. We know more than we did before, we can no longer say 'no evidence' or anything, but it's not as robust as we'd like.
3. If we tell people they can take their masks off and stop distancing once vaccinated, people will think the pandemic is over, and do all sorts of irresponsible things, whether vaccinated or not.
4. Mask mandates are all or nothing. If you don't make the vaccinated wear a mask, you can't tell the unvaccinated to wear a mask, they won't listen to you. The norm won't survive.
5. Unvaccinated people can just lie and claim to be vaccinated, either implicitly or explicitly, and it's clear we have no appetite to check.
6. Because of the temptation to lie this could lead to vaccine passports.
7. With the new CDC guidelines there will be pressure on private venues and corporations to drop their mask mandates even when and where they still make sense. People, especially in the areas with less vaccinations, will grow hostile to mask requirements.
8. The whiplash of this decision, going from 'vaccination buys you very little' to 'pandemic over if you get a shot' in one go, will erode trust and make us look like hypocritical idiots.
9. In particular, our claims that 'science has evolved' and similar are laughable on their face and open us up to several angles of attack. If science 'evolves' like that then you can question it, expect it to change and so on, rather than Believing Science. Plus no, it didn't 'evolve' and nothing changed in the past week or two other than the political reality. [The whole thing will look really bad](#) and wasn't implemented well (WaPo).
10. The contrast with children still being asked to wear masks even though (and kind of because) they are not eligible for the vaccinations yet, and people who are mostly vaccinated still being told to wear masks, won't go over well.
11. People have calibrated to the CDC being absurdly overprotective and giving an upper bound on precautions. Once people expect that, you can't then move to a reasonable average level of precaution, because people will misinterpret it.
12. There hasn't been enough time for everyone, even all essential workers, who want to get vaccinated to get fully vaccinated yet, and you're forcing those folks to get exposed to maskless people, some of whom will be unvaccinated. Why not wait a few weeks?
13. Cases are declining rapidly now. We can cut cases in half quickly, and then do this at a more comfortable level. Why not wait a few weeks?
14. If we are wrong and this blows up in our face, there's no way to reverse course.
15. [The whole thing is confusing](#), especially for those dedicated to not thinking. I don't think this is an actual problem and people can handle it, but it's a (small) cost.

The last objection, #14, is the one I take most seriously. As I noted above, I think this will *probably* lead to a large improvement in quality of life and people's ability to live it, without having a major impact on case counts. But what if that's wrong?

If that's wrong, we're stuck. There is no going back. Tell the vaccinated they have to mask up and distance again, and they will quite reasonably look at you with one voice and tell you 'no.'

If a new variant shows up that breaks through and infects the vaccinated, things could get especially ugly, although that could *possibly* create a justification to reimpose masking and distancing, at least until people get the presumed booster shot.

If things go wrong purely because too many of the unvaccinated stop taking precautions, that would be unfortunate. In the long term it would be self-correcting, in the sense that those taking the risks would slowly get vaccinated and also steadily get infected, but that could take a while and could involve substantial numbers of cases and deaths.

What *could* be done in that situation would be to start checking vaccination status, or otherwise put more pressure on the unvaccinated to get their shots, in addition to the whole 'offer them their lives back' bid and the current host of small bribes. The question is to what extent that would work. Certainly there would be a lot of pressure to start implementing and using vaccine passports if cases start going up again.

I listed that pressure as an objection #6 here, because many people strongly dislike them, rather than because I am part of that group. I think it is highly unlikely (5%?) to come to this, but If I had to choose 'none of us get our lives back' or 'the vaccinated get our lives back but have to show their status every so often for a while' I definitely go with the second one, despite the risk that this turns into a longer-term concern.

Unvaccinated people will *definitely* take advantage of the situation, as objection #5. In general, those refusing vaccination are largely those who don't take Covid-19 seriously, or think they've already been infected (which all go together, for obvious reasons) so it makes sense they'd treat the rules as stupid. I don't know how many would be willing to *explicitly* lie when asked, especially since many of them tie their identities and tribal allegiances to not getting vaccinated – a kind of 'best-case scenario' might result if masks became associated with vaccine resistance going forward. In my experience, a remarkably large number of people still think outright lying is a big deal in most circumstances.

The other problem with the 'sometimes people just lie' equilibrium, other than people getting to ignore the mask requirement if they want to, is that we're effectively putting a tax on being *honest and unvaccinated*, as opposed to one on being only unvaccinated. This seems different than denying access to life saving medicine to those who don't lie, but it's still fundamentally the same problem, and yes it matters.

The question of how much the honest will let this stop them is also open. Objections #3 and #4 speak loudly here. My expectation is that there will be places that say 'everyone wears a mask' and they'll mostly be able to hold that line if they're willing to bear the costs of that (which are objection #7, and could rapidly become large), but anywhere that doesn't hold that line will not have an easy time enforcing this by halves.

I also don't see this as an obvious 'mask mandates are good and getting rid of them is bad.' Mask mandates have costs and benefits, and over time they become worse and worse deals, be they private or public.

This also plays into objection #11, where people have given up their agency in order to sacrifice to the CDC guidelines, or treated the guidelines as the strict upper bound on plausible precautions, and now have the CDC guidelines pointing to a different more reasonable place. If the CDC plans on leaving *all its other guidelines* in the 'you should wash your hands for 20 seconds again, because science' mode, the recalibration from this adjustment is going to be actively unproductive rather than helpful, and give false hope for the future. So this is definitely a downside there, and also does risk a full version of #3. It's something to keep an eye on, but so far it seems like people are not stupid, and have noticed this isn't a typical CDC guideline. It's worth tracking that carefully.

The science on vaccines preventing infection has been in for a long time, but the Science(TM) on it was only conclusive more recently. That's some of what's causing this shift. Some are still protesting that it's not fully in (objection #2) but I don't think this is a reasonable objection at this point. It's more of a fully general 'you can't ignore that which you can't fully quantify using our exact method' objection. Objection #1 by contrast is valid if you think that we are likely to end up 'on the edge' between suppression and failure, which I do not expect but acknowledge is possible.

What about objections #8 and #10? How bad are the optics on this? My first response is that I'm happy for the optics to be terrible, because those are accurate optics. People will form a more accurate model of the world and plan accordingly. Also, this day will have to eventually come regardless, so it's not clear to me how much could be done about it. The contrast with children is especially embarrassing, but also I can be hopeful it will lead to the right answer and get children unmasked soon.'This right decision will expose our other wrong decision as wrong' is not the strongest of objections.

From what I've seen, regular people are far more happy for things to be improving than they are angry about the whiplash or contrasts. There's a lot of forgiveness available, without even the need to ask for it.

That leaves the timing objections, #12 and #13. I am *somewhat* sympathetic, as the timeline for those who got Moderna is rather tight.



Luke @Shivaekul · May 14

...

I got my first shot basically immediately upon it being open to *all* adults in my county, it wasn't even open statewide at that point. I'm not fully vaccinated until tomorrow. Neither are many of the GenZ/Millennial employees that work in shops/restaurants. This is too early.



zeynep tufekci @zeynep · May 14

The CDC mask guidance switched too fast without enough explanation and overlooks key sociological factors for indoor mask mandates—especially to protect workers and the immunocompromised. Better to have announced benchmarks—and kept it up just a bit more.
nytimes.com/2021/05/14/opi...

[Show this thread](#)

The C.D.C.'s initial mask guidance was mainly intended to dampen transmission to others, so even cloth masks were sufficient. Those who will be working indoors with unmasked colleagues may need the higher level of protection that N95s provide. **If I were advising employers, I'd tell them to keep up the mask rules indoors and pay attention to ventilation, regardless of vaccination status.**

The people who suffered the most in this pandemic are disproportionately the working poor and the essential workers who kept things going while the rest of us could stay home. Now, for example, restaurant workers will find themselves in environments where not only do the patrons not wear masks but the staff may feel pressure not to.



14



82



293



If the vaccine 'turned on' fourteen days after the second dose, I'd be highly sympathetic to this argument. We'd be giving such folks a very short window to get vaccinated in time for the unmasking. Those who waited until appointments were easy to get aren't done.

The thing is that the vaccines finish their work faster than that. The final result is likely in a full week faster, and even ten days after the first dose you're mostly already safe. It might not yet be fully reasonable to say everyone needs to be finished just yet, but it's very reasonable to say everyone should be two weeks out from their first dose.

Combine that with the lag between changing guidelines and changing behavior, and the fact that 'safe' versus 'unsafe' is not a boolean distinction, plus the drop in cases that's already happened and the additional drops baked in and won't be reversed likely ever and definitely not for enough time to get vaccinated, and the amount of 'risk' we are forcing onto people seems highly acceptable to me.

What would we have gotten by waiting a few more weeks? We could have cut the remaining risk by something like half (until we started approaching some new equilibrium), and given everyone enough time to be fully vaccinated, with an additional three week delay. I don't think that's worth it for this reason, but plausibly could be in combination with concern #14. [Here's Zeynep, who is always thoughtful](#):



zeynep tufekci

@zeynep

...

The CDC mask guidance switched too fast without enough explanation and overlooks key sociological factors for indoor mask mandates—especially to protect workers and the immunocompromised. Better to have announced benchmarks—and kept it up just a bit more. [nytimes.com/2021/05/14/opi...](https://nytimes.com/2021/05/14/opinion/covid-masks-cdc.html)

 **zeynep tufekci**  @zeynep · May 14 
Replying to [@zeynep](#)
I'm on board with the "the vaccinated have lowered their personal risk back to baseline" but there's more to indoor mask mandates than that, including sociological factors. Many just became eligible! Even access isn't fully solved. Not sure about this hurry, this week.

 **zeynep tufekci**  @zeynep · May 14 
Also, I agree we should provide carrots to encourage vaccination. I'm all for carrots! But a vaccination benchmark for lifting all mandates (like North Carolina is doing) is a good carrot. No mandate any more and no checking doesn't seem like a carrot to me, but the opposite.

 **zeynep tufekci**  @zeynep · May 14 
Eek, correction. Before the CDC change, North Carolina had a sensible policy, saying the mandates would all be lifted when two thirds of adults had at least one dose. One could quibble with the exact number, but it was a goal and a benchmark. Now all gone.

 **Alexis Wainwright**  @AWainwrightTV · May 14
Some states are starting to lift their mask mandates following the CDC mask rules for vaccinated people.

Those include: Maryland, Virginia, North Carolina, Ohio and Michigan.

 **zeynep tufekci**  @zeynep · May 14 
And I'm very much on the vaccines are amazing side of this: I think the data on individual protection and transmission — including against variants — is excellent and very encouraging. But mandates have sociological dynamics and benchmarks are better than abrupt changes.

 **zeynep tufekci**  @zeynep · May 14 
I would've been in favor of this shift a little later if it had been pre-announced and tied to reasonable benchmarks, and explained better. If nothing else, there are workers who are immunocompromised. The poor, largely minority essential workers have suffered disproportionately.

 **zeynep tufekci**  @zeynep · May 14 
This is the timeline in many states. Would've been better to give some time for this, and a heads up.

 **Caitlin Rivers, PhD**  @cmyeaton · May 14
No masks required in Maryland as of tomorrow. Yet people who got vaccinated on the very first day of open eligibility in MD would not be fully protected yet. Why not wait to give people who want to get vaccinated a chance? baltimoresun.com/coronavirus/bs...

The objection that some workers are immunocompromised is valid, but I don't see this as anywhere near big enough to move the answer this much, especially considering there is nothing stopping employers from keeping their mandates. Mostly it seems to take the form

of ‘we can’t do this because there’s a non-zero number of people who this makes unsafe’ with no attempt at a cost-benefit analysis. As I mentioned above, saying the timeline was a little tight is reasonable, I just don’t see it as important enough to wait. I’m even less very sympathetic to ‘goal and benchline’ style thinking, which seems like collective punishment for those who don’t get vaccinated fast enough more than anything else. I’m especially unsympathetic to saying that the new guidelines mean the CDC isn’t “following the science” or that this isn’t “science based.”

The Next Unmasking

We have begun restoring sanity and life to the adults, yet we remain sufficiently crazy that the [children still need our help. Fauci is letting us know.](#)



Manu Raju @mkraju · May 13

CNN: Children too young to be vaccinated will still have to wear masks when they are indoors and around others, even if older kids and adults are free to take off face protection once they are fully vaccinated, Fauci tells [@jaketapper](#)

928

905

1.4K



Manu Raju @mkraju · May 14

Fauci says to [@wolfblitzer](#) that kids in elementary school who haven’t been vaccinated in the fall should still be wearing masks

166

290

1K



I often see it pointed out that Europe *never* went in for this idea of masking young children. When things were going badly and we needed all the help we could get from every lever we could pull, this seemed terribly inefficient but I at least *understood* this as a policy, even if in practice it mostly was used as an excuse to do things like close or limit playgrounds and otherwise destroy childhood and make life as a parent or child that much more expensive and unpleasant without making anyone safer.

At this stage, it makes zero sense. The reason we don’t have the ability to vaccinate our youngest children is that we correctly realized it was not a priority to do so. Such children are at almost zero risk from Covid-19. If they are at sufficient risk from Covid-19 that they must mask, then your risk tolerance is so low that you’re effectively against the existence of childhood. That *does* seem to largely be our society’s position, as we go around doing things like arresting parents for letting kids play on their own the way kids always used to, or forcing those kids to periodically report to things we euphemistically call ‘schools’ for their ‘education.’

It’s still true that an unvaccinated five year old is safer from Covid than a fully mRNA-vaccinated senior citizen, and it’s not remotely close. Why is the child, for whom the mask is more costly, the one being asked to wear one? Is it because we are sincerely worried about them passing the virus along to others? I do think it’s reasonable to presume that the kid might still be a *better carrier* than fully vaccinated adults, even if they are at almost no risk, but they are still rather lousy carriers and will mostly be spreading it to each other even in that case. The evidence for spread in schools by underage students remains threadbare.

The whole reason we can't vaccinate young children yet is that they are so safe from the virus that it wasn't worth the effort to verify that vaccinating young children was safe and effective, and to determine the correct dose (although we never determined the correct dose for adults either, and are likely giving doses of at least Moderna that are higher than necessary).

Now that we have a national vaccine surplus, we are fixing that, but the only reason it's worth fixing that is because of insane reactions like the continued mask mandate, with a side credit to vaccine hesitancy that makes it important we get whatever help across the finish line that we can get. If the vaccinations let people stop being giant balls of stress and lets them return to normal life, then they're worth it.

The craziest part is that we now have a regime where *unvaccinated adults* can unmask freely for most purposes simply by implicitly claiming to be vaccinated, but *young children* can't unmask because they're not yet *eligible* and so they can't make the same claim.

Lottery Provides Redeeming Social Value

[The system works! \(WaPo link\)](#)



Philip Bump ✅
@pbump

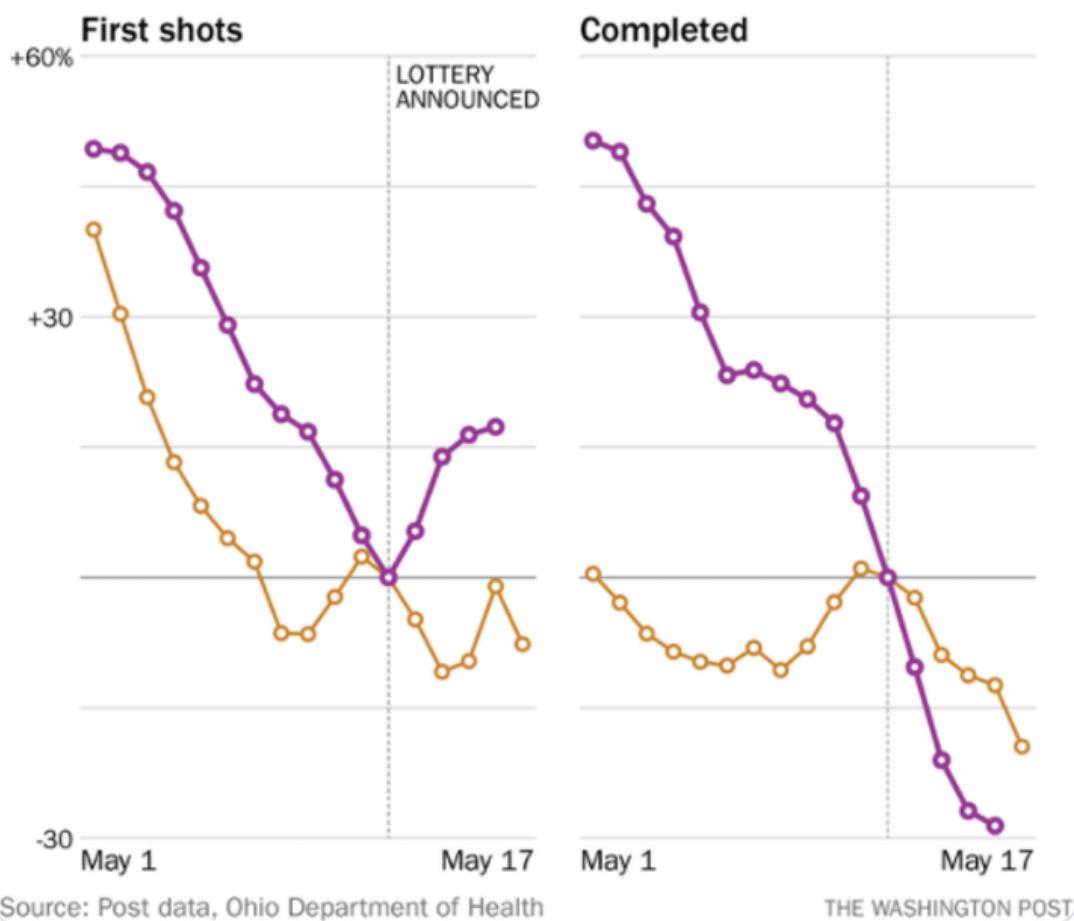
...

Ohio's vaccination lottery seems, for now, to have turned around the trend in the number of people getting their first shots.

washingtonpost.com/politics/2021/...

The million-dollar rebound

● Ohio ● National



Source: Post data, Ohio Department of Health

THE WASHINGTON POST

This did coincide with expanded eligibility, but that was true everywhere, and the contrast between Ohio and other places is stark.

Even better, the big-s System Works [because other states are now following suit](#).



Walid Gellad, MD MPH @walidgellad · 5h

...

Everyone who was planning on studying the impact of CDC mask guidance change on uptake of vaccines using an interrupted time series is distraught by all these lotteries (OH, NY, now MD)



Mike Murillo @MikeMurilloWTOP · 5h

.@GovLarryHogan is joined by the Maryland lotto ball to announce \$2 million in prize money for residents who get vaccinated. Each day between 5/25 and 6/3 a vaccinated resident will receive a \$40k prize. On July 4th a \$400K Jack pot will go to a vaccinated resident. @WTOP

[Show this thread](#)



4

1

13

↑

You know your strategy is working when it's sufficiently effective that it successfully confounds people's attempts to figure out what else is going on. I'm hoping the people who were planning on using an interrupted time series are not that upset by this development, and also I don't see what the problem is from a technical standpoint.

In Other News

Want people to get vaccinated? [I hear they respond to incentives.](#)

[Restaurant traffic back at normal levels.](#)

[Dominic Cummings speaks truth to power and anyone else who will listen.](#) A little dated in some places, but still an excellent thread.

I don't cover treatments as much, but it looks like [the treatment rules for monoclonal antibodies have been expanded quite liberally](#), and that basically everyone should be asking for them if they get sufficiently sick.

Washington Post reports that [we are instructing providers to open vials to give the vaccine to one person](#), even if the rest of the vial will be wasted. Saying this 'raises ethical concerns as many in other countries remain unvaccinated' does not seem to properly cover the level of concern here. I notice I am not okay with this.

[UK Covid forecasting challenge.](#) Prizes are small (100 GBP 1st place) but I applaud the concept.

[Tokyo Olympics will have an audience of child prisoners, who won't have the opportunity for vaccination.](#)

California offered schools \$12 million dollars to reopen, so naturally [San Francisco schools reopened for one day, claimed they'd earned the money, and then promptly closed again](#). On some level I'm not even mad I'm just impressed. I sure hope these institutions aren't playing any iterated games or trying to build trust or act in the public interest, cause that would be super awkward.

[Thread about whether vaccinated are less likely to infect others conditional on being infected themselves](#), continues seem like the answer is yes.

[India wisely delays the second dose.](#)

[MR links us to a Covid forecasting model for India.](#) I have not evaluated what the model is doing.

[The Wired story everyone linked to about \(some aspects of\) how we messed up so badly on how Covid spreads.](#)

They also remind us that [the Pfizer vaccine is even more effective if you wait 12 weeks between doses](#). This is excellent news for the effectiveness of booster shots if we ever need them, and also gives additional incentive to have people delay their second shot when one shot is already highly effective. It's still not that big an effect, since the vaccine is already super effective, so given only personal impact I'd still take that second shot at three weeks and get on with life rather than wait longer.

[Washington Post article on a slowdown in the slowdown in vaccinations.](#)

[ABC News claims](#) that restaurants, bars, gyms, beaches and such did not contribute substantially to outbreaks, but that seems to be primarily based on a lack of cases that can be definitively traced back to those locations, combined with not finding correlation between restrictions imposed and growth of case counts. I don't think such methods let us conclude much.

Not Covid: [One of several rumors that drivers for Uber and Lyft are getting very good pay right now](#), crowding out other work. If this is typical, then the rides I've been taking have been underpriced. Then again, it's not like Uber makes money.

Not Covid Pipeline Follow-up: The pipeline turned off, not because they couldn't pump the oil, but [because they weren't confident they'd be able to properly bill for the oil](#). This is good news, because it means that if it were a true emergency could have kept the oil flowing, and also we can think now (although chances are we won't) about how to deal with a more dangerous situation should it arise in the future.

The bad news is that they paid a \$5 million ransom (that it sounds like didn't do them that much good, as the tools they were given were so slow they didn't help much versus the existing alternative backup plans?) and that means they've made it clear that such attacks pay. The flip side of this is that this has given ransomware a much bigger target on its back because this wasn't a safe or legitimate target. The group that did this actually *apologized*,

saying that their intent is not to go after critical infrastructure so as not to cause this kind of damage, and still got taken offline after.

Not Covid, but [I wrote a thing this week about Magic: The Gathering, its professional tournaments and their larger context](#), in case anyone missed it. Lot of strong opinions on it in both directions, make of that what you will.

The fierce nerds I know personally mostly liked it a lot. So for that and many other reasons, although I don't drink and you likely don't either, here's to you, [fierce nerds \(Paul Graham\)](#). [Here's to you](#).

Bayeswatch 4: Mousetrap

Miriam chucked the replica *Salvator Mundi* into the bonfire.

"We do a lot of shooting first and asking questions never," said Vi.

"Personnel are expensive. AIs are replaceable. We have standard operating procedures. It wasn't always this way. AIs used to be rare. Knowledge was precious in those early days. We didn't know what the machines could and couldn't do," said Miriam.

"You were a founder of Bayeswatch?" said Vi.

"I am not that old," said Miriam.

Miriam paused the holocaust for a moment to examine an oil painting of a vampire chained to a solar execution chamber. She tossed it into the fire.

"They had yet to standardize the architectures back then. There were overhangs all over the place. Using explicit error-entropy functions wasn't even standard industry practice. Instead they just used the implicit priors of whatever architecture got results quickly," said Miriam.

"That's like making gunpowder without atomic theory," said Vi.

"Or medicine without chemistry. Those early machines were..." Miriam trailed off.

Vi tossed a painting of a dodo tree into the fire.

"One of my first missions...it was my mentor's last. We were dispatched to explore a small compound with signs of unaligned activity," said Miriam.

"That's suicide. What was command thinking?" said Vi.

"It was the early days. Singularity breakout could have been just around the corner," said Miriam.

"But drones—" said Vi.

"Software back then was written by human beings. It had more security holes than features. Sending a drone to investigate a misaligned AI was like sending a set of lockpicks to investigate whether a magician has broken out of his cell," said Miriam.

Miriam wore lots of foundation and concealer on her face. Vi wondered how many scars it covered up.

"We investigated a compound in the mountains of California. Kind of reminded me of *Ex Machina*," said Miriam.

"Was it owned by a billionaire?" said Vi.

"In your dreams. You read too many romance novels. The guy wasn't not-rich. He was an early employee of a moderately-successful software startup. I guess he built the

error-entropy minimizer himself. To this day I am unsure what the thing was supposed to do. It was dead by the time we showed up," said Miriam.

"Dead?" said Vi.

"Poetic license. My point is we weren't dealing with an active threat. The inventor turned on his machine. It carried out its mission. It turned itself off. The end," said Miriam.

"There's obviously more to it. Otherwise you wouldn't be telling me this story," said Vi.

"We did passive scans. No sound or electric activity. The compound had been built around a courtyard. We entered the courtyard by climbing over the compound wall," said Miriam.

Vi had long since stopped noticing the works of art. Her hands continued on autopilot.

"The courtyard had been.... Here. Let me show you a picture," said Miriam.

Miriam handed her phone to Vi. The photo was poor quality. It had been taken from a cell phone camera.

"What is that?" said Vi.

The photo looked fake. It was like a Zen garden grown out of 3D-printed crystals. Plants had once formed part of the fractal but they had subsequently misbehaved. The original vision lingered only in the inorganic bits. Vi identified hints of higher-order patterns but most of the symmetries lurked beyond her conscious comprehension.

Vi had dated a grad student studying physics. His professor once allowed him a single index card full of handwritten equations to use on test. He packed the index card with equations and diagrams so concise they were almost encrypted. The garden reminded Vi of that index card.

"It's the most beautiful thing I've ever seen," said Vi.

Miriam nodded.

"Why don't we make AI art like that anymore?" said Vi. She threw a knockoff Picasso into the fire.

Miriam laughed. "It's like the crested black macaque selfie. Humans never made it in the first place."

"Ah. I see. After the error-entropy machine completed its task it set about optimizing its environment to conform to its sense of beauty. But can't we deliberately program a machine to do that? It's not like the technology has disappeared," said Vi.

"We can't copy it exactly because the machine erased its own source code. But that garden came from a very particular configuration of good priors and resource constraints. Good priors are dangerous. Priors that good...we're lucky it didn't turn the universe into paperclips," said Miriam.

"Great art comes from tortured people. Torturing a superintelligence is dangerous," said Vi.

"You can't torture an AI," said Miriam.

"Poetic license. Besides, good artists are sadists. It's hard to make an AI both safe and sadistic at the same time. You also can't give it real world resource constraints in a simulation," said Vi.

"Bayeswatch is expensive. Our tools cost money. Agents die in the line of duty. There is collateral damage. But the greatest casualty of regulation is novel machines. If that AI was built today it would be decommissioned before it got to optimize its environment," said Miriam.

"Did you find anything else in the compound?" said Vi.

"A few deactivated robots. A server rack overwritten with random data. The corpse of the engineer. I think he died of natural causes. We left through the front door. Well, we tried to. As we opened it an M18A1 Claymore anti-personnel mine activated. My mentor took the brunt of the blast," said Miriam.

"Why did the AI want to kill you?" said Vi.

"It didn't care about us. It set up the booby trap to protect the garden while it was under construction. When the garden was finished it just didn't bother to disable it," said Miriam.

Vi's hands stopped. There were no more paintings to destroy.

Parsing Chris Mingard on Neural Networks

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

This is independent research. To make further posts like this possible, please consider [supporting me](#).

Epistemic status: This is my understanding of multiple years of technical work by several researchers in just a few days of reading.

Outline

- I attempt to summarize [some of Chris Mingard's recent work](#) on why neural networks generalize so well.
- I examine one chunk of work that argues that mappings with low Kolmogorov complexity occupy large volumes in neural network parameter space.
- I examine a second chunk of work that argues that standard neural network training algorithms select mappings with probability proportional to their volume in parameter space.

Introduction

During the 2000s, very few machine learning researchers expected neural networks to be an important part of the future of their field. Papers were rejected from major machine learning conferences with no reason given other than that neural networks were uninteresting to the conference. I was at a computer vision conference in 2011 at which there was a minor uproar after one researcher suggested that neural networks might replace the bespoke modelling work that many computer vision professors had built their careers around.

But neural networks have in fact turned out to be extremely important. Over the past 10 years we have worked out how to get neural networks to perform well at many tasks. And while we have developed a lot of practical know-how, we have relatively little understanding of *why* neural networks are so surprisingly effective. We don't actually have many good theories about *what's going on* when we train a neural network. Consider the following conundrum:

1. We know that large neural networks can approximate almost any function whatsoever.
2. We know that among all the functions that one might fit to a set of data points, some will generalize well and some will not generalize well.
3. We observe that neural networks trained with stochastic gradient descent often generalize well on practical tasks.

Since neural networks can approximate any function whatsoever, why is it that practical neural network training so often selects one that generalizes well? This is the question addressed by a recent series of papers by [Chris Mingard](#).

The basic up-shot of Chris' work, so far as I can tell, is the following:

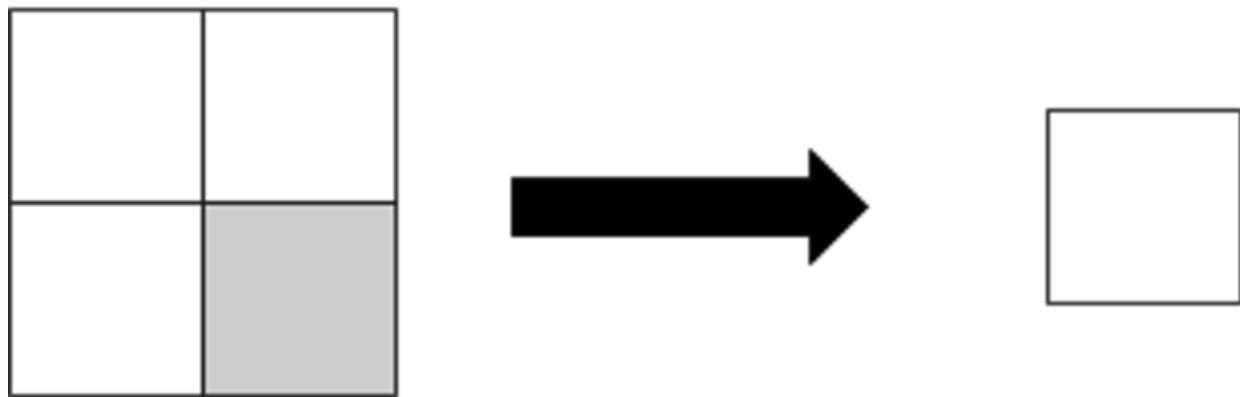
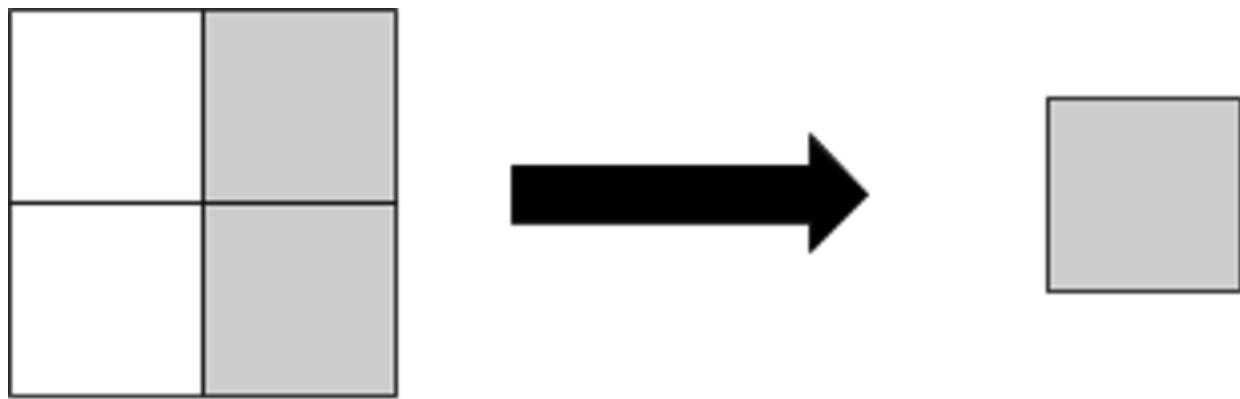
- The optimization methods that we use to train neural networks are more likely to select mappings that occupy large volumes of neural network parameter space than functions that occupy small volumes of neural network parameter space.
- Most of the volume of neural network parameter space is occupied by simple mappings.

These are highly non-obvious results. There is no particular reason to expect neural networks to be set up in such a way that their parameter space is dominated by simple mappings. The parameter space of polynomial functions, for example, is certainly not dominated by simple mappings.

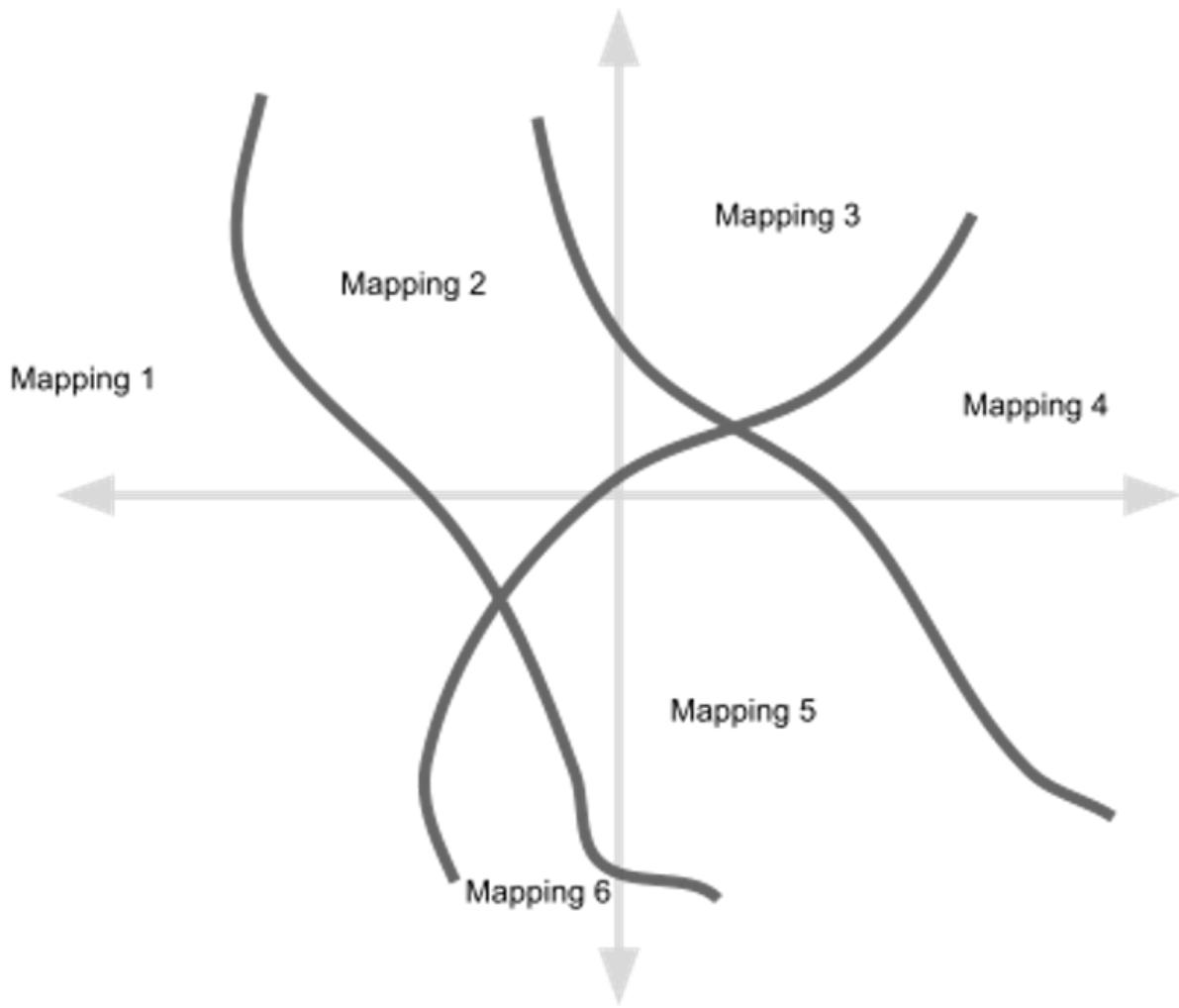
Chris' work consists of a combination of empirical and theoretical results that suggest but do not decisively prove the above claims. In this post I will attempt to explain my understanding of these results.

Simple mappings occupy larger volumes in parameter space

Chris' work is all about volumes occupied by different functions in parameter space. To keep things simple, let's consider a machine learning problem in which the inputs are tiny 2x2 images with each pixel set to 0 or 1, and the output is a single 0 or 1:



Since there are 4 input pixels and each one can be either a 0 or a 1, there are 16 possible inputs. Each one of those inputs could be mapped to either a 0 or 1 as output, so there are $2^{16} = 65,536$ possible mappings from inputs to outputs. Any neural network with four input neurons and one output neuron is going to express one of these 65,536 possible mappings^[1]. We could draw out the whole space of possible neural network parameters and label each point in that space according to which of the 65,536 mappings it expresses:



Each point in the above diagram represents a particular setting of the parameters in a neural network. I have drawn just two dimensions but there will be far more parameters than this. And I have drawn out volumes for 6 mappings but we would expect all 65,536 mappings to show up somewhere within the parameter space.

So given the picture above, we can now ask: do each of the 65,536 mappings occupy equal-sized volumes? Or do some occupy larger volumes than others? And if some mappings do occupy larger volumes than others then is there any pattern to which mappings occupy larger versus smaller volumes?

Chris' work suggests that some mappings do in fact occupy larger volumes than others, and that it is the mappings with low Kolmogorov complexity that occupy larger volumes. What does it mean for a mapping to have a low Kolmogorov complexity? It means that there is a short computer program that implements the mapping. For example, the mapping that outputs 0 if there are an even number of black pixels in the input image and otherwise outputs 1 has a low Kolmogorov complexity because this mapping can be computed by XOR'ing all the input pixels together, whereas the mapping that outputs 0 for some randomly chosen arrangements of input pixels and otherwise outputs 1 has high Kolmogorov complexity because any computer program that computes this mapping will have to include a big lookup table within its source code. It is important to understand that when we talk about complexity we are talking about the length of a hypothetical computer program that *would* compute the same

mapping that a given neural network computes. Also, (John reminds us) [<https://www.lesswrong.com/posts/5p4ynEJQ8nXxp2sxC/parsing-chris-mingard-on-neural-networks?commentId=fzkGYmHsKdFx5dyzb>] that the paper uses a proxy for simplicity that is actually pretty different from Kolmogorov complexity.

In order to demonstrate this, Chris worked with the well-known MNIST dataset, which contains images of handwritten digits of 28x28 pixels. This means that the number of possible images is 2^{56} , since in this dataset there are two possible pixel values, and the number of possible mappings is $10^{(256)}$, since in this dataset there are 10 possible outputs. This is a very large number, which makes it infeasible to explore the entire space of mappings directly. Also, Kolmogorov complexity is uncomputable. So there was quite a bit of analytical and experimental work involved in this project. This work is summarized in the blog post "[Deep Neural Networks are biased, at initialisation, towards simple functions](#)", with references to the underlying technical papers. The conclusions are not definitive but they are highly suggestive, and they suggest that mappings with lower Kolmogorov complexity occupy relatively larger volumes in parameter space.

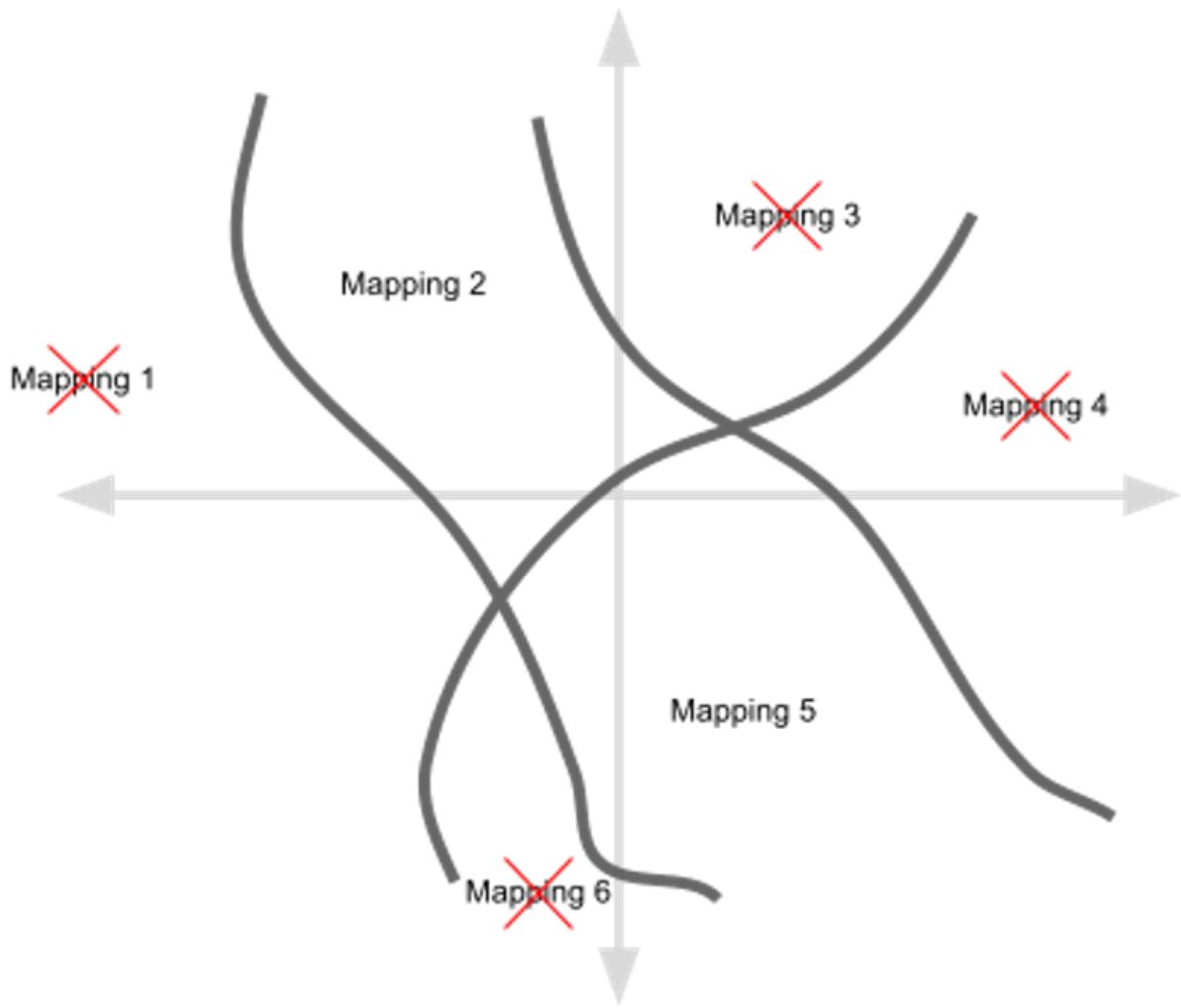
This sheds some light on the question of why trained neural networks generalize well. We expect that mappings with low Kolmogorov complexity will generalize better than mappings with high Kolmogorov complexity, due to Occam's razor, and it seems that mappings with low Kolmogorov complexity occupy larger parameter space volumes than mappings with high Kolmogorov complexity.

Mappings occupying larger parameter space volumes are more likely to be selected

The next question is: do the optimization algorithms we use to train neural networks care at all about the volume that a given mapping occupies in parameter space? If the optimization algorithms we use to train neural networks are more likely to select mappings that occupy large volumes in parameter space then we are one step closer to understanding why neural networks generalize, since we already have evidence that simpler mappings occupy larger volumes in parameter space, and we expect simpler mappings to generalize well. But they might not be more likely to select mappings that occupy large volumes in parameter space. Optimizations algorithms are designed to optimize, not to sample in an unbiased way.

A second blog post by Chris summarizes further empirical and theoretical work suggesting that yes, the optimization algorithms we use to train neural networks are in fact more likely to select mappings occupying larger volumes in parameter space. That blog post is called "[Neural networks are fundamentally Bayesian](#)", but it seems to me that viewing this behavior as Bayesian, while reasonable, is actually not the most direct way to understand what's going on here.

What is really going on here is that within our original parameter space we eliminate all mappings except for the ones that perfectly classify every training image. We don't normally train to 100% accuracy in machine learning but doing so in these experiments is a nice way to simplify things. So our parameter space now looks like this:



The question is now: for the mappings that remain, is the standard neural network training algorithm (stochastic gradient descent) more likely to select mappings that occupy larger volumes in parameter space?

To investigate this, Chris compared the following methods for selecting a final set of neural network parameters:

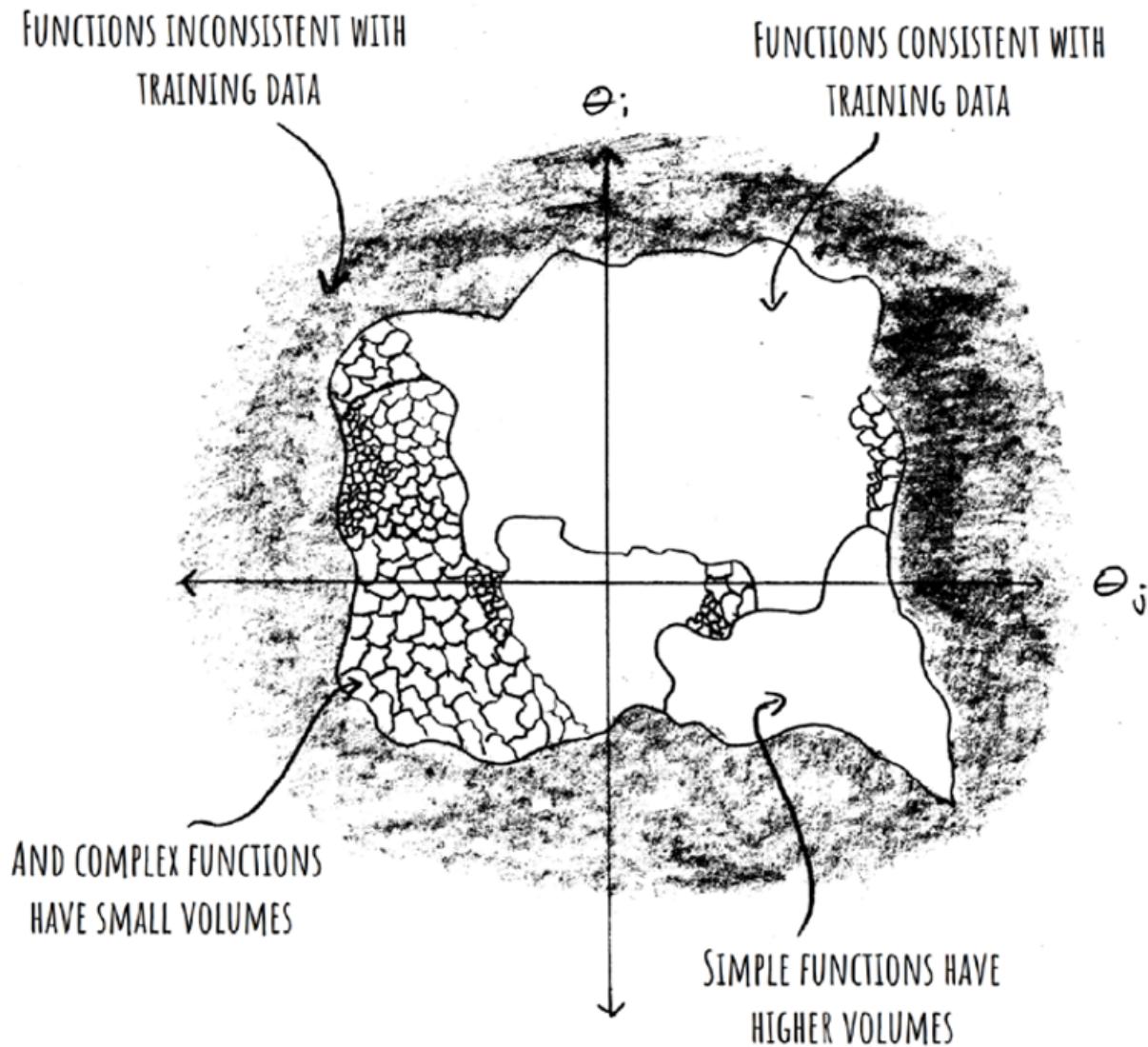
1. Select neural network parameters at random until we find one that perfectly classifies every image in our training set, and output those parameters.
2. Train a neural network using the standard neural network training algorithm (stochastic gradient descent) and output the result.

We know that method 1 is more likely to select mappings that occupy larger volumes in parameter space because it is sampling at random from the entire parameter space, so a mapping that occupies twice the parameter space volume as some other mapping is twice as likely to be selected. So by comparing method 1 to method 2 we can find out whether practical neural network training algorithms have this same property.

But actually running method 1 is infeasible since it would take too long to find a set of neural network parameters that perfectly classify every image in the training set if sampling at random, so much of the work that Chris did was about finding a good

approximation to method 1. To read about the specific methods that Chris used, see the blog post linked above and the technical papers linked from that post.

The basic picture that emerges is nicely illustrated in this graphic from the blog post linked above:



Scalability

This section added based on [this helpful comment by interstice](#).

Both of the claims discussed above are supported by a mixture of theoretical and empirical results. The empirical results are based on machine learning tasks that are relatively small-scale. This is understandable because the experiments involve re-training networks hundreds of thousands of times from scratch, which would be very expensive for the largest networks and problems being tackled today. However, it

leaves open the question of whether these results will continue to hold as we run experiments with larger-scale networks and problems.

For further discussion of the likely reach of the results discussed here see [this excellent post and its associated comments](#).

Relevance to AI safety

If we want to align contemporary machine learning systems, we need to understand how and why those systems work. There is a great deal of work in machine learning that aims to find small "tips and tricks" for improving performance on this or that dataset. This kind of work does not typically shed much light on how or why our basic machine learning systems work, and so does not typically help move us towards a solution to the alignment problem. Chris' work does shed light on how and why our basic machine learning systems work. It also provides an excellent example of how to perform the kind of empirical and theoretical work sheds light on how and why our basic machine learning systems work. I am excited to follow further developments in this direction.

1. the output neuron will be treated as a 1 if it is positive or a 0 otherwise ↵

[Weekly Event] Alignment Researcher Coffee Time (in Walled Garden)

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

I'm organizing a weekly one hour coffee time for alignment researchers to talk about their research and what they're interested about, every Monday starting tomorrow. The time (9pm CEST, see [here](#) for your timezone) was decided after I polled a number of people, to allow timezones from PT to IDT (sorry for people in Asia and Australia).

It will happen in the Walled Garden, in the central house which is directly up when you pop up in the Garden. The link of the event is public, although the event is by default reserved for AF members and people invited by them.

Here is the [link](#).

Some basic infos:

- The restriction to AF members and invitees is mostly so that we don't have to reexplain AI Risk 101 every time. This is not the point of this event. If you're genuinely interested in alignment, and if you've read some of the posts in the AF and are starting to form an image of the field, then by all means come discuss.
- No obligation to come everytime. The goal is mostly that if you want to talk to a relatively broad range of people instead of having one-on-one calls, you can do so the next week instead of waiting for the next big event.
- By default there is no talk planned, but if attendees feel like shaking the structure a bit, that's fine.

Sabien on "work-life" balance

From Duncan Sabien:

This morning, a friend of mine referred to his "work-life balance" and then grimaced at himself, and noted that he doesn't really like that term.

(As far as I can tell, he doesn't like it because his work is important to him, and is part of being alive, and his non-work life isn't some *fundamentally* different kind of thing.)

This led me to the mental distinction between one's *directly* valuable life experiences, and one's indirectly/instrumentally valuable ones.

Like, there are many fewer layers involved when you, say, snuggle up to someone you love, or pop a delicious food in your mouth, or bounce on a trampoline. There's a very short, direct path between the action and the reward; it's just straightforwardly close to your values and the things which bring you joy and fulfillment.

Whereas if you're good at your work and you think that your job is important, there's an intervening layer or three—I'm doing X because it unblocks Y, and that will lead to Z, and Z is good for the world in ways I care about, and also it earns me \$ and I can spend \$ on stuff...

I think it's less about "work-life balance" and more about the ratio of direct vs. indirect value. "How many of the things that I'm doing pay off directly, versus how many of them are knocking over dominos that eventually *lead* to the payoffs I'm seeking?"

(And how close is that ratio to one which I will actually find sustainable and enjoyable and healthy.)

This, to me, is a distinction that cuts closer to the true joints of reality than the arbitrary categories of "work" and "everything else." As a single easy example, this also applies to the balance, in one's relationships, between straightforwardly valuable interactions with people you enjoy, and meta interactions/maintenance/laying the groundwork for the future.

Mundane solutions to exotic problems

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

I'm looking for alignment techniques that are [indefinitely scalable](#) and that [work in any situation we can dream up](#). That means I spend time thinking about "exotic" problems—like AI systems reasoning about their own training process or about humanity's far future.

Yet I'm very optimistic about finding practical alignment strategies that are relatively simple and usable today. I expect the results of my research to look mundane and even boring.

These two stances may appear to be in tension. If I'm worrying about all kinds of exotic considerations, how can the end result be something mundane?

I'm not too concerned. These failures seem exotic because they involve exotic *capabilities*. I'm looking for a mundane algorithm that trains the model to use whatever capabilities it has for good purposes. So it will lead to exotic countermeasures if and when the model has exotic capabilities, i.e. just in time to avoid exotic failures (by which time they'll no longer feel exotic).

I want to illustrate this phenomenon with a few examples.

Example 1: Gradient hacking

[Gradient hacking](#) is a possible failure mode where a powerful misaligned model behaves in a way that is apparently benign but biases the gradient of the loss to point in an unfavorable direction, eventually trapping optimization at a bad local minimum. This feels exotic today because it requires the model to have an incredible amount of introspective access and control, as well as strategic awareness over long time horizons. (“[Forward Pass](#)” by Andrej Karpathy explores some similar ideas.)

Normally I imagine using SGD to optimize a “behavioral” objective: to update the parameters θ you use them to make a randomized decision y on an input x , then you calculate a reward $r(x, y)$, then you compute an unbiased estimate for the gradient of the expectation $r(x, y)$ with respect to θ .

Gradient hacking is troubling because it suggests that if we optimize a purely behavioral objective we could get a bad outcome even if the reward function $r(x, y)$ perfectly captures the quality of the output y .

Rather than fixing this problem in a specific way, I'd like to rely on competitiveness. That is, assume that (throughout training) we have a question-answering policy $Q \rightarrow A$ which reflects *everything* that our learned model “knows” about the world.

Then it seems like the best way to compute the system's loss should be by asking some question like “What should the loss be for the current parameters?”

Our usual behavioral loss then is a great *heuristic* for computing the loss (which we can justify on principled Bayesian grounds if we want). In mundane situations it's not

clear we have anything further to say about what the loss should be, so “What should the loss be?” is almost the same as hard-coding the heuristic.

But when gradient hacking occurs, we are no longer in a mundane situation. By hypothesis, our learned model is able to reason introspectively about its own parameters and the structure of the loss landscape! It is thinking about how to change its behavior in order to affect the loss, change the optimization trajectory, and ultimately disempower humans.

That means that a competitive question-answerer *also* is able to answer questions about its own parameters and the structure of the loss landscape. It is able to talk about how assigning a loss could mess up the optimization trajectory and ultimately disempower humans. If we incorporate all that knowledge into our assessment of “What should the loss be for the current parameters?” then it will assign a low loss to some of the actions that the gradient-hacker tried to sabotage. That is, the malicious AI can outsmart the behavioral heuristic, but it can’t outsmart an overseer *who knows everything that it knows*.

It’s not clear if this works (and my strong guess is that we’ll ultimately end up with an algorithm that looks quite different from this). But overall I think that if gradient hacking actually came up, it’s more likely than not that the core difficulty in avoiding it would be the same version of epistemic competitiveness that we need in more mundane situations. The only real change to our algorithm we need to make is applying a competitive overseer to the right question.

You might worry that the knowledge a model has about its own parameters is qualitatively different than the knowledge it has about the world. But I don’t think so —they are both produced by transfer from the task of “get a low training loss,” combined with a bunch of computation. It’s possible that after we have an epistemically competitive solution we’ll see that it doesn’t apply to a model’s introspective knowledge. If that happens then we may indeed need some more exotic solution that talks about introspection *per se*, but I personally doubt it.

Example 2: far future forecasts and inaccessible influence

Suppose that there’s [some domain that our AI can understand but we cannot](#), for vividness you can imagine humans without microscopes and an AI that has a working model of the microscopic world. For simplicity, imagine that the world factors nicely into an accessible part and an inaccessible part, and that there are influences in both directions but to us the the (inaccessible) → (accessible) influences just look like noise. (If you don’t know about the mechanics of biology, then events like “getting cancer” can just look like noise.)

One of my concerns is that an AI that understands the inaccessible part may be able to cause trouble in the very long term. Even if we’re able to prevent the AI from doing anything overtly bad in the accessible part for the next few years (like killing humans), it may still be that humans aren’t really in control of the situation because the real power is in the inaccessible part of the world, and that power will eventually assert itself.

So if we can only observe the accessible part of the world, then we would need to look very far ahead to avoid problem. Basically we’d need to look far enough ahead that the future humans have done science and learned about the previously-inaccessible parts of the world. This kind of lookahead feels particularly exotic because it involves

extreme generalization. Training typically involves feedback loops over seconds, but now we are talking about a model that makes good predictions about events that literally occur *generations* from now.

That kind of extreme lookahead sounds exotic, but in some cases my current alignment plans require it.

As in the last example, I think this situation is only exotic because our model had exotic capabilities. We're in this situation because we are trying to handle a malicious AI that makes plans that have no observable consequences for many generations, building up power in the inaccessible part of the world that it predicts will eventually interfere with humans. That AI is already apparently able to generalize well to extremely long time horizons—if it wasn't, we'd have no problem at all.

So once again, to avoid this problem we really just need epistemically competitive question answering, i.e. we need to learn a $Q \rightarrow A$ function that generalizes as well to novel questions (e.g. about the future) as the underlying model's knowledge. And that's a pretty mundane problem that's core to the idea of epistemic competitiveness that we need to solve for other reasons.

If the malicious model is able to generalize from its experiences to make plans that unfold over generations, and the $Q \rightarrow A$ model generalizes as well as the malicious model, then our $Q \rightarrow A$ system can also answer questions about what happens over generations. That only sounds exotic because we assumed some exotic capabilities. Nothing about the solution or training needs to deal with ultra-long horizons specifically.

Conclusion

In my work I don't shy away from exotic problems (I often find them useful as extreme cases to illustrate some principle). At the same time, I'm aiming for mundane solutions and optimistic about finding them.

I think those positions are consistent because my intermediate goal is to ensure that the oversight process is able to leverage all of the capabilities developed by the model—so if the model develops exotic capabilities which pose exotic challenges, then we get an exotic oversight process automatically.



[Mundane solutions to exotic problems](#) was originally published in [AI Alignment](#) on Medium, where people are continuing the conversation by highlighting and responding to this story.

Challenge: know everything that the best go bot knows about go

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

On a few different views, understanding the computation done by neural networks is crucial to building neural networks that constitute human-level artificial intelligence that doesn't destroy all value in the universe. Given that many people are trying to build neural networks that constitute artificial general intelligence, it seems important to understand the computation in cutting-edge neural networks, and we basically do not.

So, how should we go from here to there? One way is to try hard to think about understanding, until you understand well enough to reliably build understandable AGI. But that seems hard and abstract. A better path would be something more concrete.

Therefore, I set this challenge: know everything that the best go bot knows about go. At the moment, the best publicly available bot is [KataGo](#), if you're at DeepMind or OpenAI and have access to a better go bot, I guess you should use that instead. If you think those bots are too hard to understand, you're allowed to make your own easier-to-understand bot, as long as it's the best.

What constitutes success?

- You have to be able to know literally everything that the best go bot that you have access to knows about go.
- It has to be applicable to the current best go bot (or a bot that is essentially as good - e.g. you're allowed to pick one of the versions of KataGo whose elo is statistically hard-to-distinguish from the best version), not the best go bot as of one year ago.
 - That being said, I think you get a 'silver medal' if you understand any go bot that was the best at some point from today on.

Why do I think this is a good challenge?

- To understand these bots, you need to understand planning behaviour, not just pick up on various visual detectors.
- In order to solve this challenge, you need to actually understand what it means for models to know something.
- There's a time limit: your understanding has to keep up with the pace of AI development.
- We already know some things about these bots based on how they play and evaluate positions, but obviously not everything.
- We have some theory about go: e.g. we know that certain symmetries exist, we understand [optimal play in the late endgame](#), we have some neat [analysis techniques](#).
- I would like to play go as well as the best go bot. Or at least to learn some things from it.

Corollaries of success (non-exhaustive):

- You should be able to answer questions like “what will this bot do if someone plays [mimic go](#) against it” without actually literally checking that during play. More generally, you should know how the bot will respond to novel counter strategies.
- You should be able to write a computer program anew that plays go just like that go bot, without copying over all the numbers.

Drawbacks of success:

- You might learn how to build a highly intelligent and capable AI in a way that does not require deep learning. In this case, please do not tell the wider world or do it yourself.
- It becomes harder to check if professional human go players are cheating by using AI.

Related work:

- The work on identifying the ‘[circuits](#)’ of [Inception v1](#)
- [The case for aligning narrowly superhuman models](#)

*A conversation with Nate Soares on a related topic probably helped inspire this post.
Please don't blame him if it's dumb tho.*

The Variational Characterization of KL-Divergence, Error Catastrophes, and Generalization

Epistemological Status: Correct, probably too technical for Less Wrong, but these results are interesting and relevant.

There's been a flurry of posts on generalization recently. Some talk about [simplicity priors](#) and other about the [bias of SGD](#). However, I think it's worth bringing up the fact that there is already an alternative that already works well enough to provides [non-vacuous bounds](#) for neural networks. Namely, I'm talking about the [PAC-Bayes approach](#).

The twist I'll give in this brief note is a motivation of the bound from a biological perspective and then derive the full PAC-Bayes bound. The first part relies on the [error threshold](#) for replication and the [patching interpretation](#) of the KL-divergence. The second part relies on the [Donsker-Varadhan](#) variational characterization of KL-divergence.

Error Threshold

A replication \hat{R} of an object R is an approximate copy (up to mutation). The environment determines whether or not an object gets to replicate via a fitness $S(R)$. Suppose I is an instruction set or parameterization of R . When R replicates I is copied.

We begin with a prior distribution of instructions π_0 and then the distribution changes as the population replicates to π . The amount of information needed to achieve the modification is the [patching cost](#) and is equal to $D_{KL}(\pi\|\pi_0)$.

Only information from the fitness landscape $L(\pi)$ is useful for converging to the fittest individual(s). On the other hand, the amount of information available from the environment is equal to $D_{KL}(L(\pi)\|L(\pi_0))$. Thus, transitions are favorable only when we have,

$$D_{KL}(\pi\|\pi_0) < D_{KL}(L(\pi)\|L(\pi_0))$$

As an example, suppose that the space of instructions is the set of boolean strings of length $|I|$ and that mutation flips a bit in the string with uniform independent probability ϵ . Suppose that I always survives and our fitness is simply whether or not

the replicate survives. This means $L(I) = 1$ and $L(\hat{I})$ is a Bernoulli random variable with parameter S . Then we have,

$$\begin{aligned} D_{KL}(I\|I^{\hat{I}}) &= \sum_{i=1}^n p_I(i) \log(p_I(i)/p_{I^{\hat{I}}}(i)) = \sum_{i=1}^n \log(1/(1-\epsilon)) \approx \epsilon \cdot |I| \\ D_{KL}(L(\hat{I})\|L(I)) &= \log(1/S) \\ \Rightarrow \epsilon \cdot |I| + \log(S) &< 0 \end{aligned}$$

PAC-Bayes Bound

Now suppose that the instructions are parameterizations for some sort of learning model and that the fitness is the training metric. In particular, the survival rate could be a classification error metric. If we are given n examples that induce an error rate of ϵ we have the requirement,

$$\begin{aligned} D_{KL}(\pi\|\pi_0) &< n \cdot \log(1/(1-\epsilon)) \\ \Rightarrow 1 - e^{-n D_{KL}(\pi\|\pi_0)} &< \epsilon \end{aligned}$$

This implies that using a lot of information to update our estimate for the optimal model parameters will require high error rates.

Specifically, a sublinear relationship between the information used and the data obtained is necessary for a low error rate if we are to avoid an [error catastrophe](#). This would be a situation where the model continues to update despite being at the optimum.

At this point it's natural to wonder if a sublinear relationship is sufficient. This is precisely the content of the PAC-Bayes theorem. To show this first note that for any $f \in L^\infty$,

$$\begin{aligned}
\langle f, \pi \rangle &= \langle \log(e^f), \pi \rangle \\
&= \langle \log(\frac{\pi}{\pi_0} e^f), \pi \rangle \\
&= \langle \log(\frac{\pi}{\pi_0}), \pi \rangle + \langle \log(\frac{\pi_0}{\pi} e^f), \pi \rangle \\
&= D_{KL}(\pi \| \pi_0) + \langle \log(\frac{\pi_0}{\pi} e^f), \pi \rangle \quad (\text{Definition}) \\
&\leq D_{KL}(\pi \| \pi_0) + \log(\langle \frac{\pi_0}{\pi} e^f, \pi \rangle) \quad (\text{Jensen}) \\
&= D_{KL}(\pi \| \pi_0) + \log(\langle e^f, \pi_0 \rangle) \\
\Rightarrow \langle f, \pi \rangle &\leq D_{KL}(\pi \| \pi_0) + \log(\langle e^f, \pi_0 \rangle)
\end{aligned}$$

As an aside, the astute reader might notice this as a Young-Inequality between dual functions in which case we have the famous relation,

$$\Rightarrow D_{KL}(\pi \| \pi_0) = \sup_{f \in L^\infty} \langle f, \pi \rangle - \log(\langle e^f, \pi_0 \rangle) \quad (\text{Donsker-Varadhan})$$

Either way, it's clearer now that if we take f to be a generalization error such as,

$$f = \lambda \left(\frac{1}{n} \sum_{i=1}^n l(h(x_i), y_i) - E[l(h(x_i), y_i)] \right) = \lambda (\hat{L}(h) - L(h))$$

then the left-hand side of our Young-inequality yeilds an empirical loss which is similar to what we had before. The major difference is that we have a contribution from a cumulant function which we can bound using the [Markov](#) and [Hoeffding](#) inequalities,

$$\begin{aligned}
\langle e^f, \pi_0 \rangle &\leq_{1-\delta} \frac{1}{\delta} E[\langle e^f, \pi \rangle] \quad (\text{Markov}) \\
&= \frac{1}{\delta} \langle E[e^f], \pi \rangle \\
&\leq \frac{1}{\delta} \langle e^{\lambda^2/2n}, \pi \rangle = \frac{1}{\delta} e^{\lambda^2/2n}
\end{aligned}$$

Now we put everything together and then optimize over λ .

$$\hat{L}(\pi) - L(\pi) = \frac{1}{n} \langle f, \pi \rangle \leq \frac{1}{n} [D_{KL}(\pi \| \pi_0) + \log(\langle e^f, \pi_0 \rangle)]$$

$$\leq_{1-\delta} \frac{D_{KL}(\pi \| \pi_0) + \log(\frac{1}{\delta})}{2n} + \frac{\lambda}{2n}$$

Optimizing we obtain,

$$L(\pi) \leq \hat{L}(\pi) + \frac{\overbrace{D_{KL}(\pi \| \pi_0) + \log(\frac{1}{\delta})}^{\text{PAC-Bayes}}}{2n}$$

↓

Abstraction Talk

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

I gave a talk a couple weeks on abstraction. It moves very fast, but it covers more material in one place than any individual post I've written, and includes a few things which haven't showed up in posts yet (especially near the end).



[Slides are here](#). The slides are hard to see sometimes in the video; I recommend keeping them open to the side of the video.

Love on Cartesian Planes

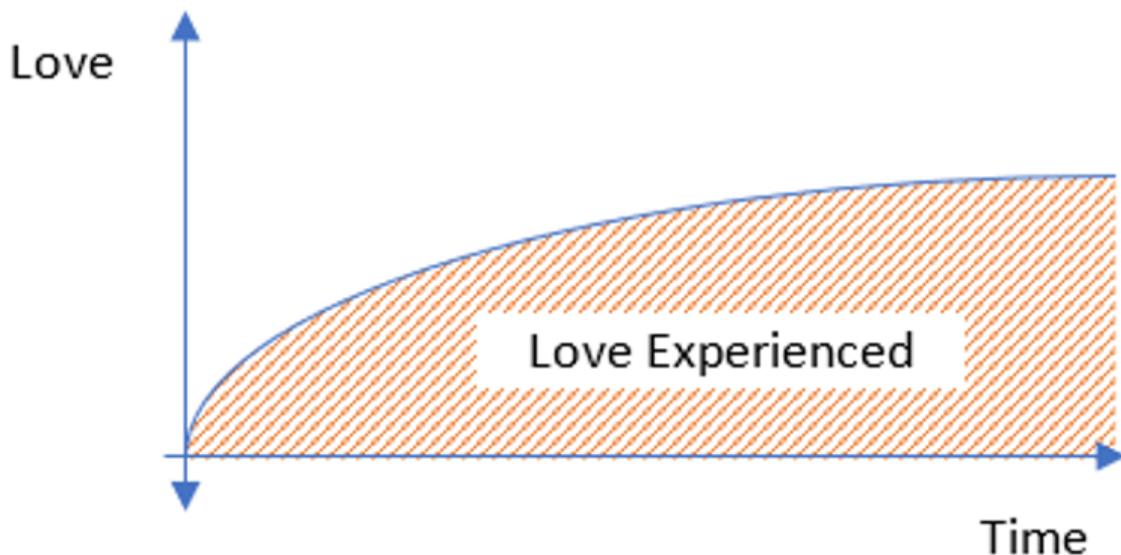
Introduction

Note: this was originally published on [Medium](#), since I'm putting a bunch of other related posts here I thought it best to migrate the original as well.

For over a decade now I've had a concept milling around my head that has been subtly influencing my behaviours. Perhaps formulating a Theory of Love is not something you'd expect coming from a university student studying macroeconomics and drawing too many AD-AS curves, but its staying power and adaptability has continued to surprise me over the years and in the spirit of [Tyler Cowen's advice on writing a book](#) I felt like I had to get this out of my system.

In essence the concept dwells on identifying causal components of the feeling of love, assessing how their relationship to love maps on a two-dimensional plane, and then assessing their movement over time in a way to allow myself to make more optimised decisions in relation to love experienced. Like all good data-nerds I love a good XY plot, and applying their power to the concept of love itself has been a fun, enlightening and rewarding exercise.

Initially I developed this framework based on my own strong internal preferences toward the remembering self, expressing "love experienced" as the Area-Under-Curve for a chart expressing love felt over time. A notable effect of this is the overwhelming impact over time of shared experiences as the primary actionable component of love.



Love in a given relationship over time

Note: all charts included in this piece were fabricated in Microsoft Word, badly. I could have scripted up something in Python to draw nicer pictures, but this would not have accurately captured my (much more visual) thought process.

I recall seeing a curve like this (sans AuC) on a webcomic like XKCD or SMBC around 2004-2005, but when I went looking for it years later I was unable to find it. That one basic

diagram gave me a foundation for everything that follows, and helped inform my relationship decisions greatly. So whoever made it, thanks!

Choosing a definition of love

Articulating the concept of love is, to be honest, too difficult a job for me to try to fully explore. The cop out way to express what is love would be to say “you know it when you feel it”, but as evidenced by numerous works of art over the ages, *this is really bad advice* (the most pertinent example to come to mind is Romeo & Juliet). Instead I will simply say it is a feeling of calming happiness triggered by the emotional presence of an entity.

This helps keep the concept distinct from lust (something more like an ‘energising desire’), allows for love of many different types (love of sexual partner, love of platonic friendship, love of family, love of land or inanimate objects, etc) and allows for the feeling of love to be felt even in the physical absence of the object of affection (e.g. remembering a loved one who has died). This could probably be mapped to PANA or some other emotional framework somehow, but I found these were overly limiting or not meeting my personal interpretation when I tried (note: I didn’t try very hard).

In choosing this definition I am certainly missing something about love. Hopefully by laying this out others can help me improve my methods by improving my target.

With all that said the only time I have personally made use of this framework *in decision making* has been in the context of a romantic partner, the other uses have generally been personal musings on my appreciation of friends and children.

Components of love

The following is a list of key elements I have found to influence my perception of love at any given point in time:

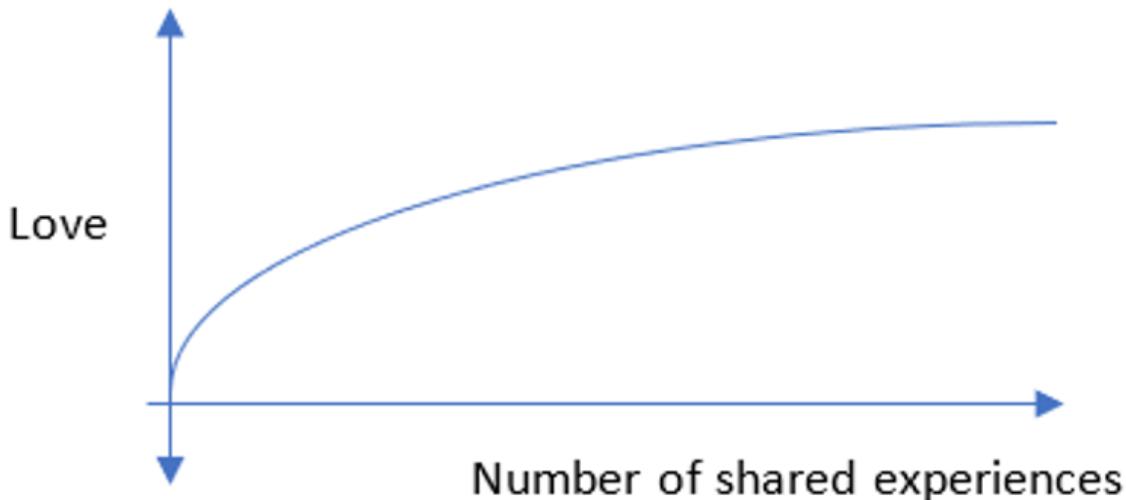
- Shared experiences
- Perception of reciprocal love
- Physical attractiveness
- Intellect
- Non-sexual physical contact
- Alignment of morals
- Complementarity (partner’s capability in areas of your own personal deficiency)
- Frequency of creation of joy
- Frequency of creation of emotional harm
- Gross amount of resentment

I acknowledge that this is not a comprehensive list, nor are all the items on this list statistically independent. E.g. for some Physical contact is a key determinant of Perception of reciprocal love, intellect may have bearing on the capacity to create more frequent moments of joy, Frequency of creation of joy/resentment will typically overlap with Shared experiences, etc. But overall, I’m hoping this overlap is not too significant and love felt at a point in time will be the sum of the outputs from the relationship between love and these components as explored in the following subsections.

It is at this point that distinguishing “Love” as a sense of *calming* happiness from other forms of happiness or desire also becomes apparent: joy is an energised feeling of happiness, and yet the connection made through this energy can also create by products of removing stress and thus introducing calm, especially in the “remembering self”.

I also admit that I have been quite lazy in identifying attributes which erode love and have clear negative correlations by distilling these into the last two items. However, this does not mean that the relationship between love and all the other items on the list are strictly positive. An important note for the plots that will follow is that **this is my own personal mapping of these relationships, they are not in any way some kind of universal representation**. Love is personal (though I am curious how others would draw them).

Shared Experiences



In my personal definition, a shared experience is something as simple as an event that creates an anecdote which may be recalled later. The length of this anecdote does not matter; it could be something as short as a 1 sentence quote from a TV show, a 5-minute story you like to share at parties or a fact like 'I have run 10km'. The relationship itself starts out fairly linear but starts to reach saturation as the human memory can only retain so much information. It is possible for some people that the curve may even start to trend down after a certain point where they feel that they are "spending too much time together", however this may be a by-product of the increased opportunities to create emotional harm. This reflection may mean I should split out another component of "time spent together" but disentangling that from shared experiences is too tricky for the amount of effort I'm already going to here.

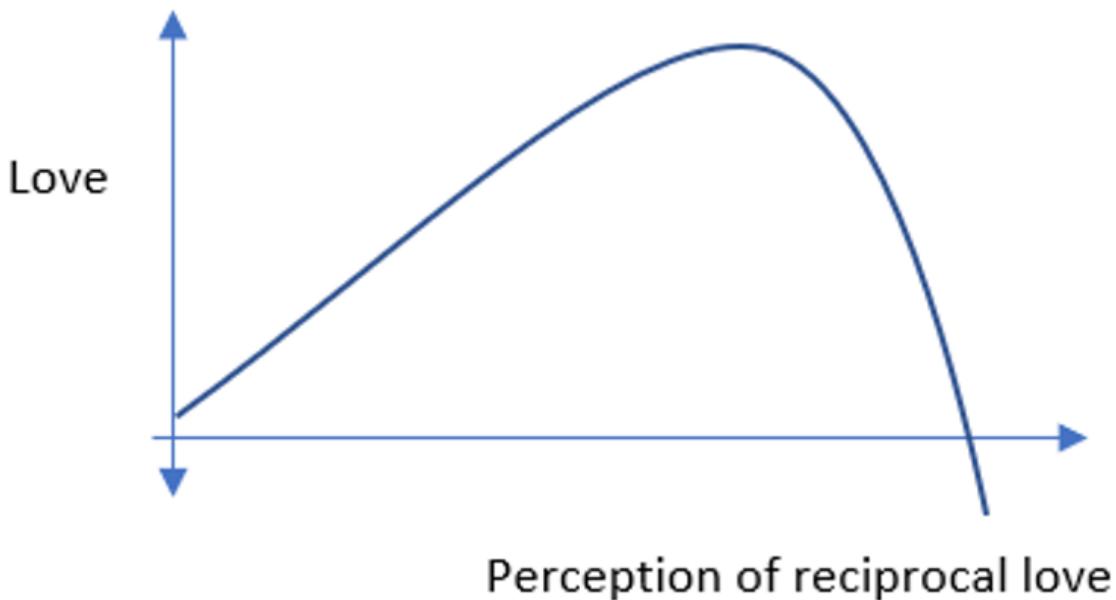
What the above chart fails to capture is that the more similar you and your counterpart's experience is, the more powerful it is. E.g. If I watched Strong Bad as a teenager but my partner just watched it last week the "shared experience" won't have the same impact as if we had both watched it at the same stage in our development. Experiences directly shared with one another are the most powerful in this way: simply watching a TV show together on the couch means you are both experiencing almost exactly the same thing and creating an almost perfectly aligned shared experience (internal commentary and the distance of a few centimeters notwithstanding). Similarly, more emotionally resonant/memorable experiences will be more powerful.

The beauty of this item is that it will naturally increase over time as the length of a relationship extends. It can get a jump start by tapping into similarities of experience, sure, but the effects of these are substantially smaller compared to experiences shared in physical and temporal proximity with a partner.

Many “economic” takes on love and relationships like to frame these shared experiences as “sunk costs” to be discarded when looking prospectively. This is to the detriment of the remembering self who can enjoy a feeling of love from many more triggers in the world, with a wider variety of recollections than could be had with a new partner. However, the diminishing marginal utility is a key factor here: the longer your prospective time-frame, the less relative time will be spent in the “ramp up” of shared experiences and the lower their overall importance.

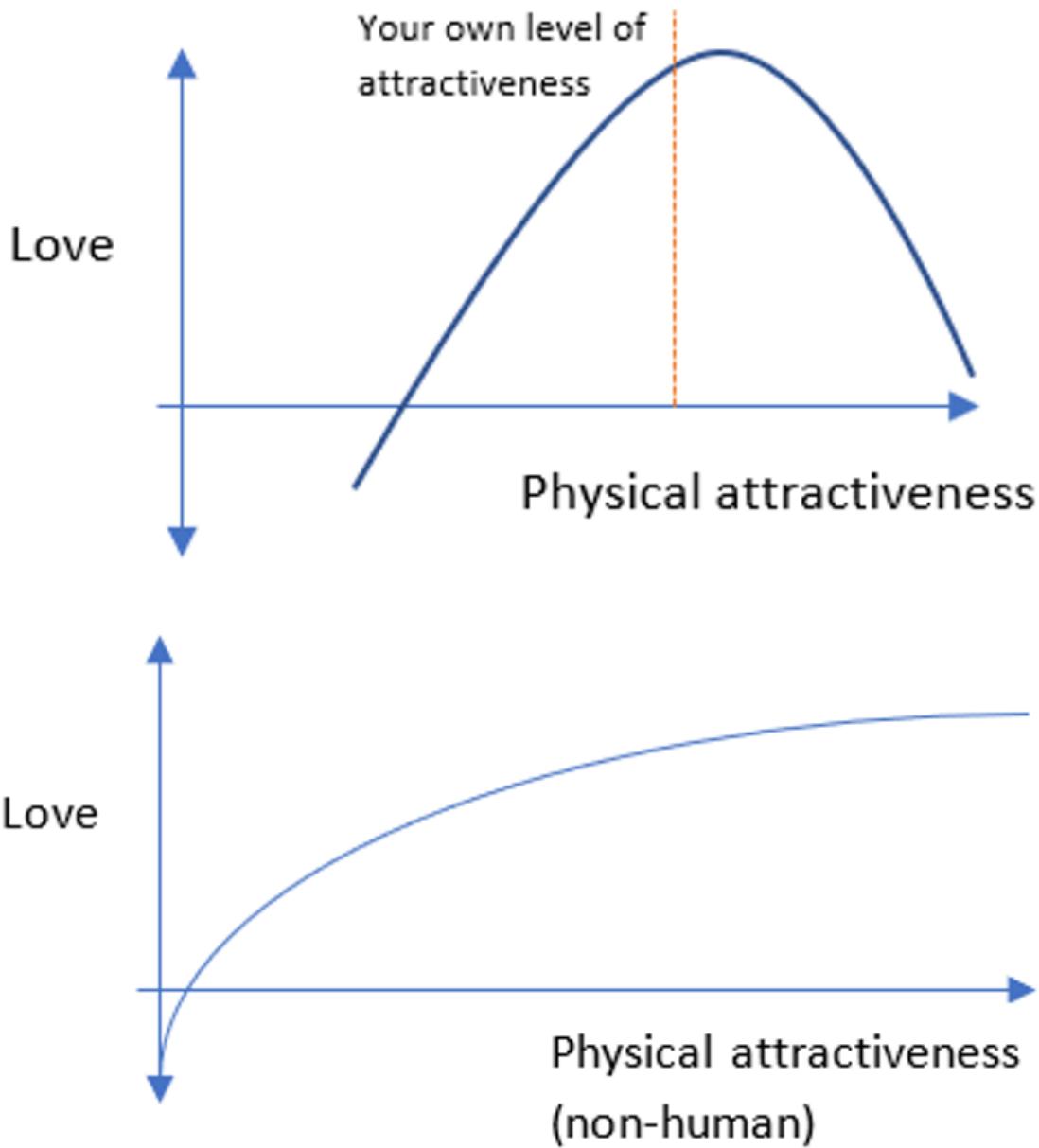
I speculate that individuals with a higher focus on the “experiencing” rather than “remembering” self will see this factor have a lower weight in their own feelings of love.

Perception of Reciprocal Love



I may need to work on the above shape a little, but in general it gets at the point that as one feels more loved, the more love they are likely to feel in response, up to a point after which it starts to become creepy and uncomfortable.

Physical Attractiveness

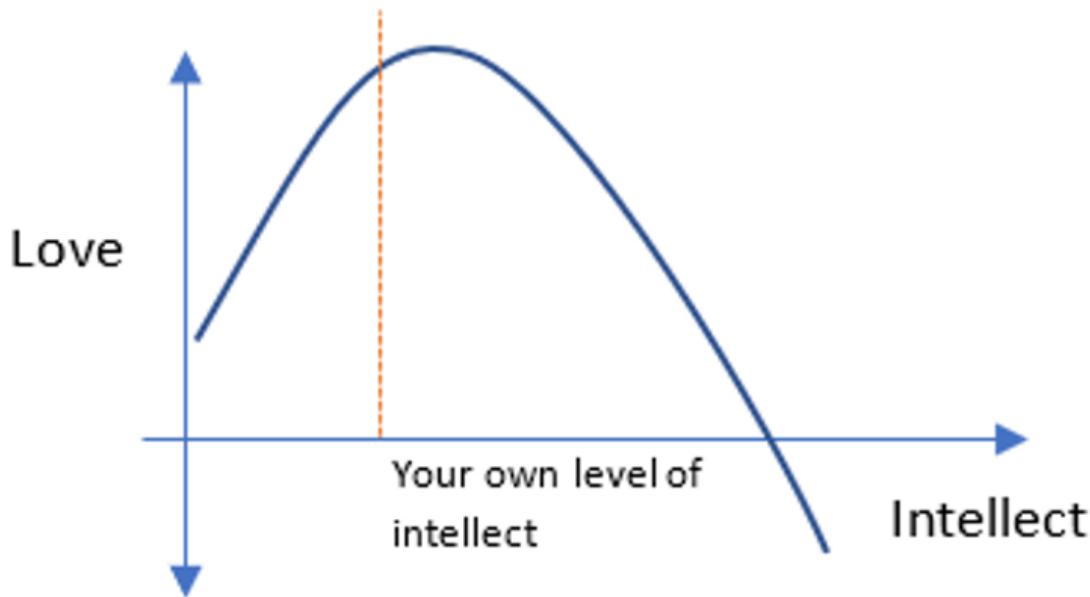


The basic concept here when looking at love between people is to aim a little higher than your own position in the range. If you're a 6 on a 0-10 attractiveness scale, you're going to be happy with a 4-8, but if it's above 8 its going to start seeming a bit fishy.

But why does physical attractiveness matter at all if sex doesn't, especially if we aren't limiting ourselves to romantic partners? Consistent with the premise of love being a "calming happiness", aesthetics (especially familiar ones) can trigger the same emotional response, and when coupled with all the other components of love this effect can be quite profound.

I don't think the same negative marginal love after a point applies to non-human objects such as love for dogs, artwork, land, etc. Based on the connoisseur market I could see an argument that the line for non-human objects of love would not decrease in marginal utility at all, and maybe actually increases exponentially, but as stated in the introduction to this section, this is my chart.

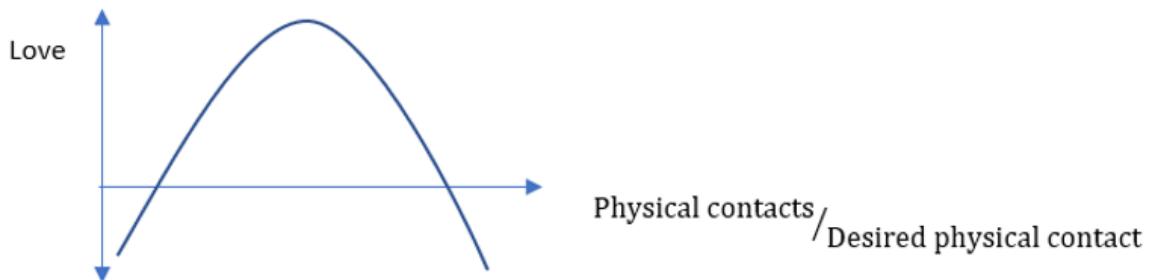
Intellect



Similar to physical attractiveness, with Intellect ideally you will find someone who is able to challenge you slightly but primarily you are optimising for an ability to "be on the same wavelength". Too smart and you will feel alienated and talked down to, not smart enough and you will feel exasperated and unable to connect when you cannot articulate a concept in a way that is understandable to the object of your affection.

Caveat: some people will have a preference where the peak of the intellect curve is below their own believed level of intellect, because they get an additional sense of satisfaction in sharing knowledge/teaching (or potentially just some smug sense of superiority). Affection for children is a special case as well.

Non-sexual physical contact



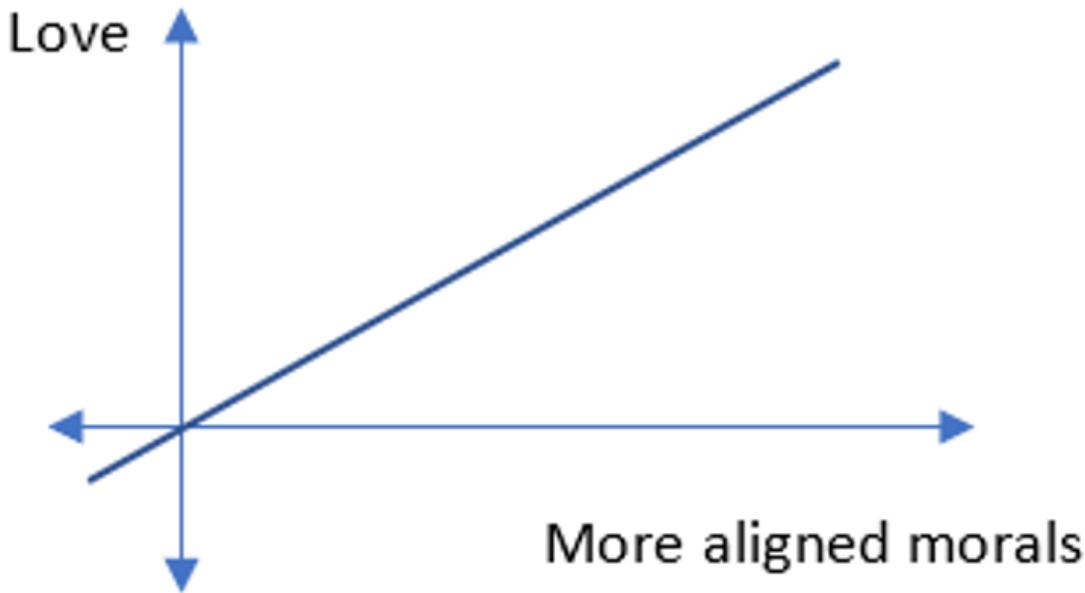
In general, what I am trying to express in the above curve is an outright rejection of physicality (e.g. refusing to even handshake, or only handshake when it would be typically expected to hug) can be detrimental to love, while being overly affectionate can be negative (a hug from my 4-year old or him laying on top of me on the couch watching TV is an excellent loving experience, but it quickly becomes exasperating if he becomes too needy).

The use of a ratio on the x-axis allows for the curve to be tailored to specific relationships (friends, family, partners). My intuition is that the Y-axis would have different responses based on these categories (love of a partner would respond significantly more to finding the optimal physical contact balance than that of a platonic friend). I would expect the peak of the curve to be at a 1 on the x-axis, but I could also be swayed for it to be slightly to the right of this because the occasional unanticipated hug/pat on the back/kiss can be beneficial.

I specify “non-sexual” here because in my mind sex itself is addressed through Shared experiences, Creation of joy and Perception of reciprocity in the positive, and through Frequency of emotional harm (where being rejected from sexual advances is emotionally painful) and Gross feelings of resentment in the negative.

This disentangles sex from love and allows this exploration of love to be broader in encapsulating platonic and familial love. I can see counter arguments about sexual love being a special case bonus addition to overall love, with friends and family simply getting a zero for this input, but the passion and intensity of sex just don’t fit the mould here, and it also just brings in a whole heap of other icky questions and confounders.

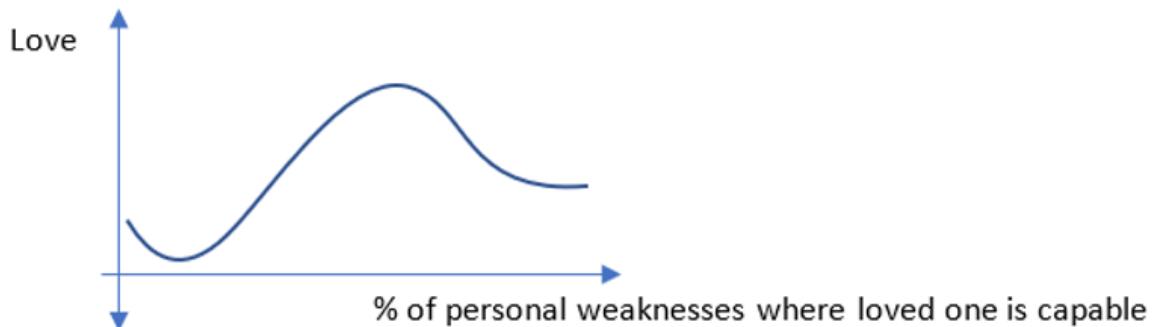
Moral Alignment



Moral alignment and shared values are another way of “being on the same wavelength” as a loved one however the benefits do have a limit. Alignment can also avoid some traps in being drawn into conflicts over items believed to fundamentally be part of your identity and increasing the chance of heightened levels of emotional harm.

When you have no conception of the counterparty’s morals, the effect is zero. As moral beliefs become contrary there can be a negative impact on overall love, but this is not to say the effect would overwhelm positive sentiment from all the other drivers. I have this as linear as I don’t see a strong argument for diminishing utility, but it will definitely be bounded (just how many moral beliefs can you hold?).

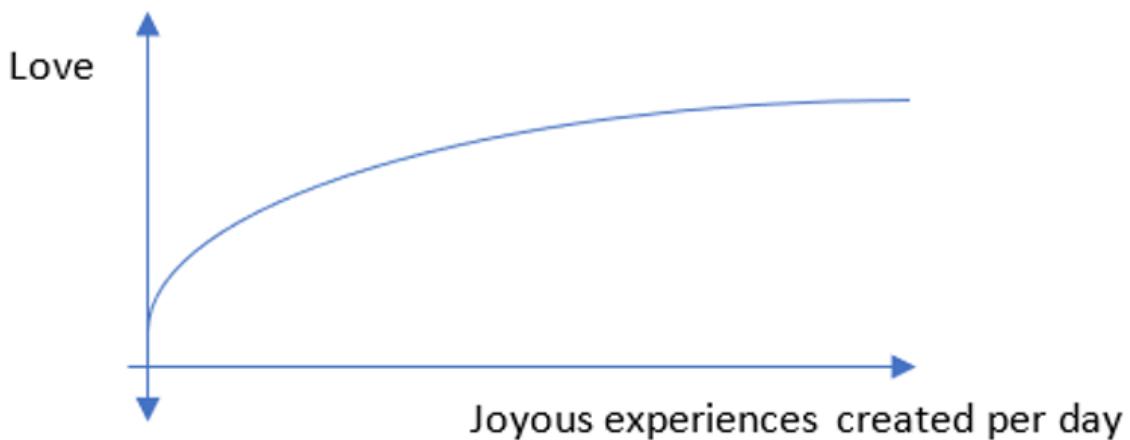
Complementarity (there has to be a better word for this)



This one is a bit weird. The further left on the x-axis you are, the more similar you are to your loved one, which is good in the “wavelength” sense mentioned above, as well as the ability to empathise with struggles. Being better at just a couple of things may diminish this while also breeding some level of envy, while being better at many things can be exceedingly beneficial in “making up for each other’s flaws” or simply “being a great team”.

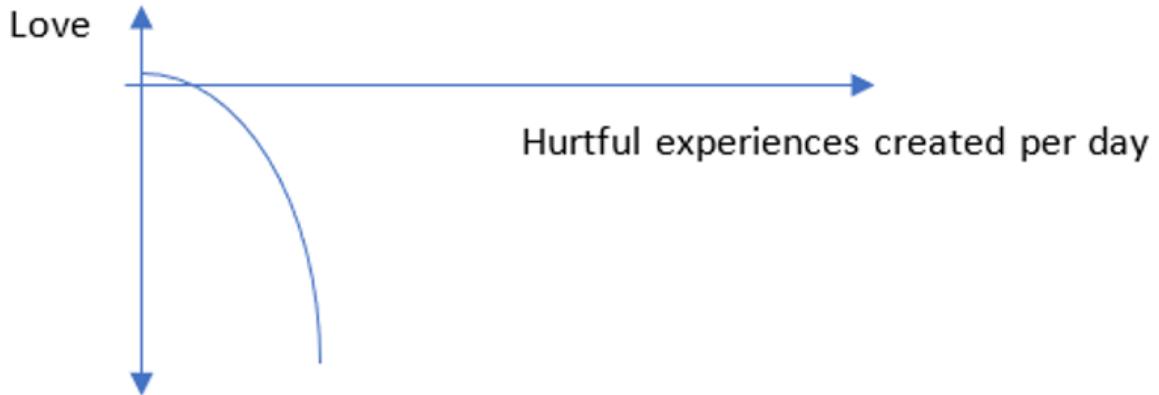
Because the x-axis being expressed as a percentage there is a hard limit on the overall benefits in this space. Every move up the curve is going to be progressively more difficult. This means returns to love start to decrease again at the higher end as you start to resent just how damn good your partner is at everything you suck at, or realise just how flawed you yourself are as they make up for your own flaws again and again. Surely those flaws should be almost impossible for someone to be good at! At the very top end this flattens out as it enters the realm of child-like dependence.

Frequency of creation of joy



This curve only tapers off because I can't be laughing all the damn time, sometimes there needs to be space in life for Serious Business(TM). However, I don't see a need for it to have negative marginal returns at the higher end, as the efforts to create joyous experiences would no longer actually be joyous, just futile efforts only successful in generating annoyance.

Frequency of creation of emotional harm

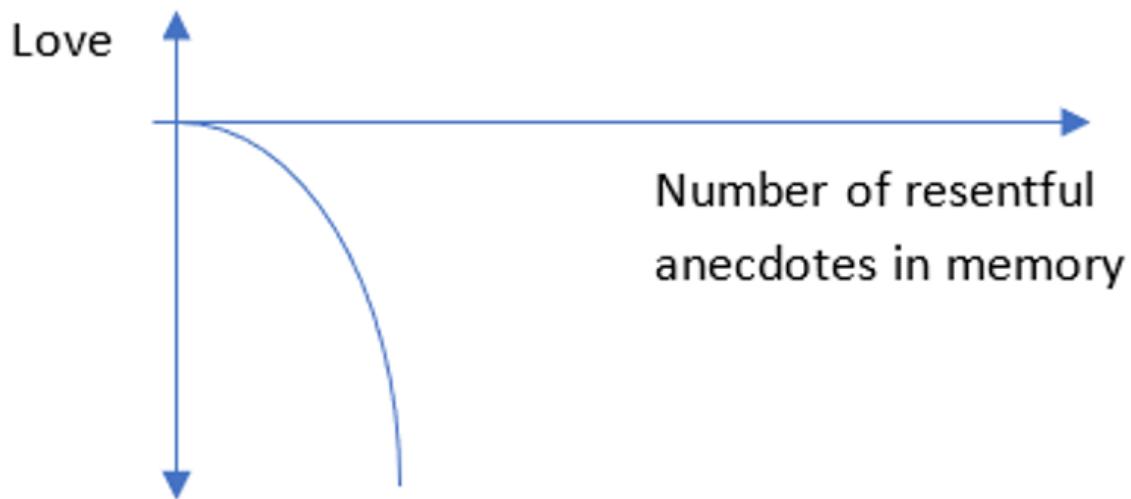


The choice to have a slight positive effect just above zero in this chart is reflective of the tendency for minor, quickly forgiven annoyance which comes as part of challenging each other to grow intellectually/emotionally/etc. But no mistake, the effect quickly becomes negative as emotional harm generates increasing resentment.

I'm not sure how this applies to those who are stuck in an abusive relationship (e.g. [Act 2 of this episode of This American Life](#)) as I could be talked into there being some kind of limit to the negative love generated, but I just don't believe it — something else that is clearly *not* love is keeping them there.

I'm also aware of the catch phrase "treat 'em mean, keep 'em keen" but I think that's bullshit.

Gross amount of resentment



Feeling resentment is bad.

Getting to Love Experienced

I've been a bit cheeky in not actually putting any scale on the X or Y-axes of the above charts, because actually putting hard numbers onto something like love just feels... wrong. Hell, even writing out my internal logic in this whole screed has felt a bit iffy.

So, for the purposes of the exercise I'm going to continue to avoid numbers and instead talk through how I would expect the sum of the above components to change over time. This gets to an AuC and thus "Love Experienced" in a given relationship based on the premise outlined in the introduction.

Factors which naturally change over time

Note the word "natural" — I'm only concerned with things which will change *with no intentional additional effort* over time, as a) that makes my calculations harder, and b) the effects of such intentional effort will become self-evident as a result of exploring the base case anyway. I'll touch a bit on making decisions regarding extra effort in the last section. If I wanted to be super objective and stick to this premise, I would model the below assuming your relationship with a literal rock, but I'll try and be realistic and assume some dynamism from the counter party.

- **Shared experiences** — Increases in line with the curve as the length of the relationship extends. Potentially this may have some erosion at the far end of the tail as you start to forget earlier shared experiences, but this should be made up by a continual inflow of new experiences.
- **Perception of reciprocal love** — Constant-ish. Movement mirrors your own level of "love" as you are better able to read the other person. This ends up acting as an amplifier on movement in the overall curve from the baseline so it's not worth calling out.
- **Physical attractiveness** — Constant, potentially declines. After an initially high level of importance, the "novelty" of your partner's physical appearance is eroded as a result of habituation/hedonic adaptation. Once this effect wears out the change over time flattens as you age along with your loved one (you may think you're going to be a silver fox, but you aren't fooling anyone).
- **Intellect** — Constant. I base this on IQs being relatively stable over time, and any degradation as a result of senility, etc. is expected to be in line with the other party (not that these expectations are reliable!). Special case: love of children would increase over time up to a point.
- **Non-sexual physical contact** — Wiggles around a bit at the start of a relationship as you are figuring out what is appropriate, then settles relatively quickly. Assume constant after this point because this is tied to a ratio: while levels in desired physical contact change (you may be less interested as you get older) as long as your loved one is continuing to adjust their own behaviours in response, the net effect stays the same.
- **Alignment of morals** — Slight increase over time as you tend to become a reflection of those around you, so parts of your morals will naturally drift into alignment.
- **Complementarity** (partner's capability in areas of your own personal deficiency) — Constant
- **Frequency of creation of joy** — Constant. While there would be a slight increase at the beginning of a relationship as the other is better able to determine what makes you happy, this effect is negated by the decreases over time from things losing their novelty.
- **Frequency of creation of emotional harm** — Constant. Realistically I'd expect an increase at the beginning as people reveal what's behind the façade, but this is counter-acted by a certain level of resilience developed over time.
- **Gross amount of resentment** — quickly grows over time if above zero. Has a limit of what you can actually remember, but look at the chart again and see how rapidly it accelerates in the negative, the relationship should be terminated well before saturation point.

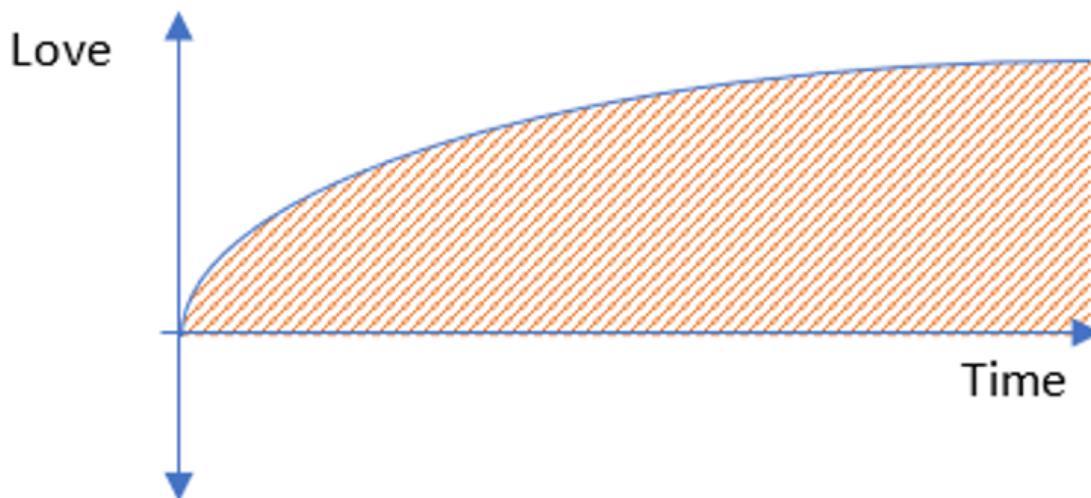
Love over time

From the above, it becomes clear that:

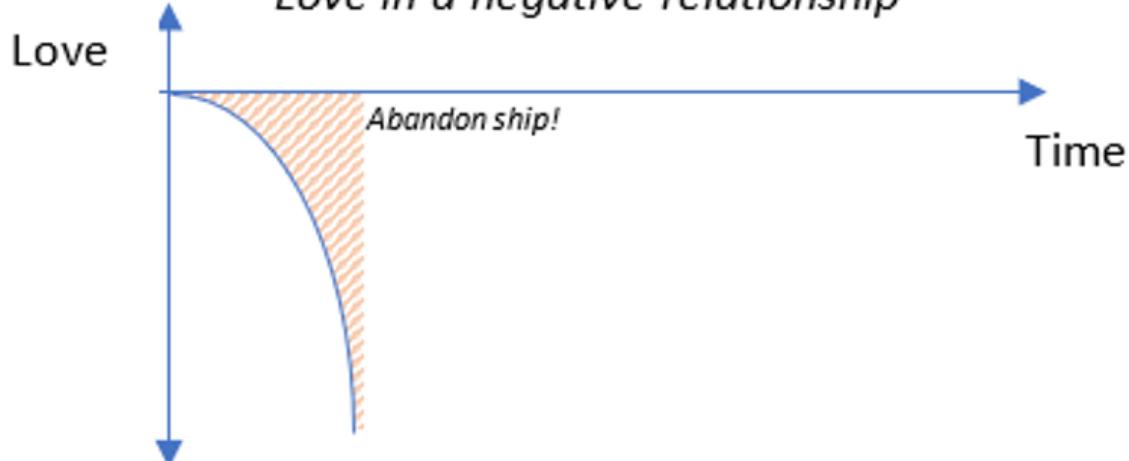
- A positive relationship will start at a given point, wobble around a bit as you get a feel for each other and then increase over time akin to a log-curve, driven by the accumulation of shared experiences, assuming it hasn't been overwhelmed by emotional harm and resentment.
- Negative relationships will quickly nosedive as a result of the strong impact of resentment and you'll want to exit fast.
- Relationships where things start out well, but the frequency of emotional pain is slightly too high. This situation sees love grow in the early stages, then be overwhelmed by resentment as time progresses and the negative memories pile up — what I'll term as a train wreck. This is where I can really see an argument for people making objectively bad decisions to stay with their partner on the basis of sunk costs ([but I'm sceptical on just how prevalent sunk costs are](#)).

This generates the basic outlooks below:

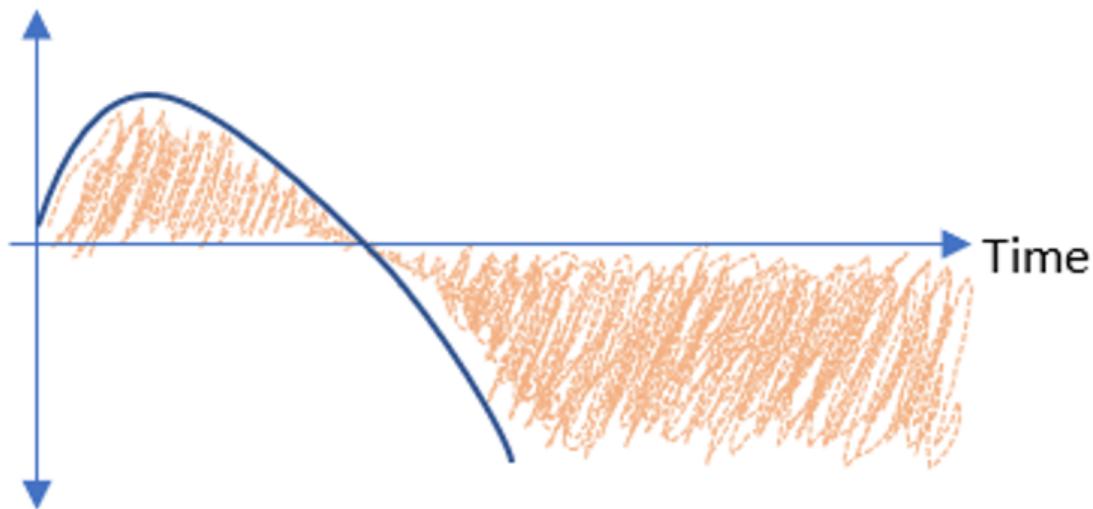
Love in a positive relationship



Love in a negative relationship



Love in a train wreck

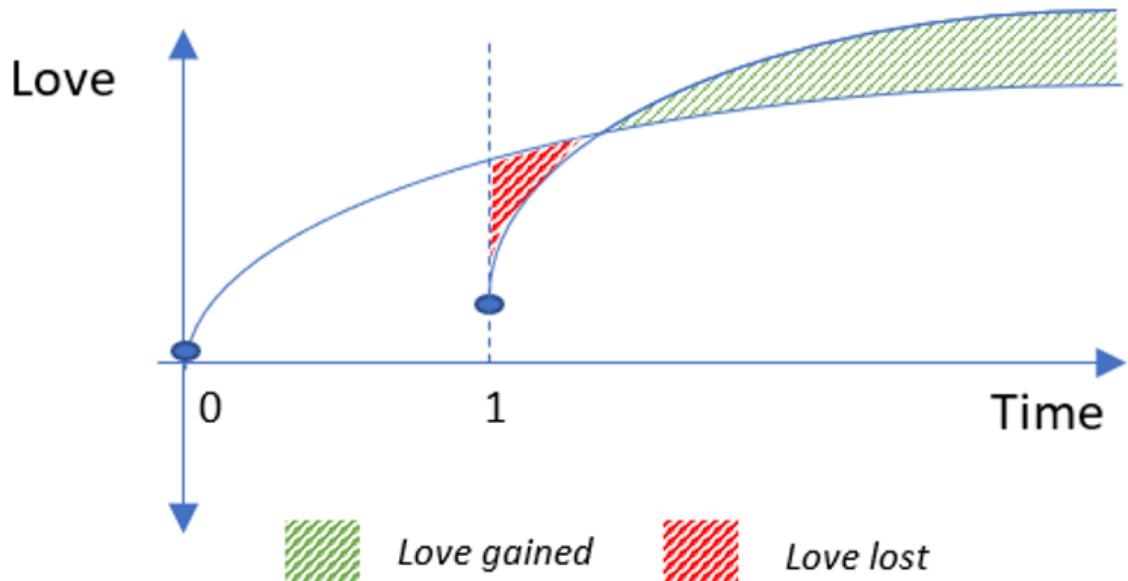


[Note: I know how to fix the shading, part of me just enjoys that the train wreck diagram is itself a train wreck.]

Making Decisions

Choosing between romantic partners

Love choices - Switching

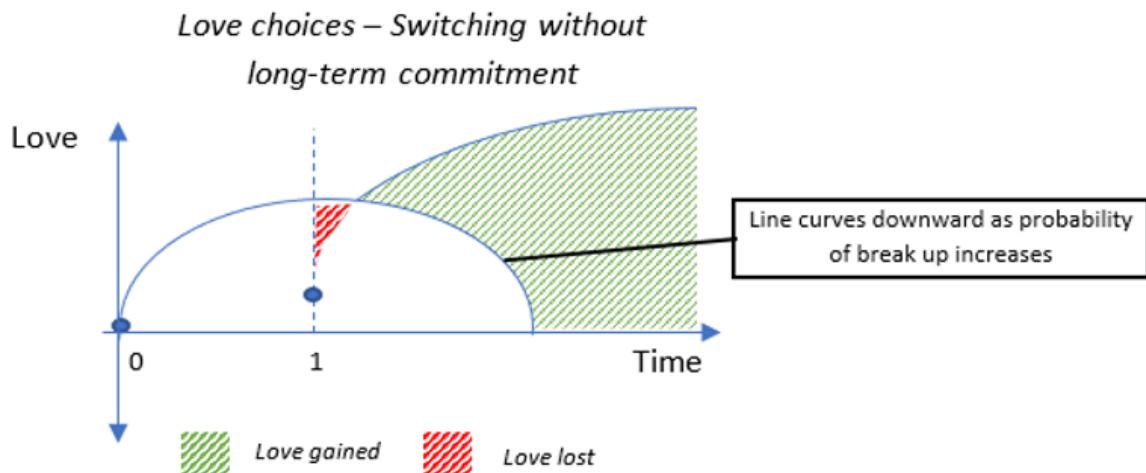


Assume at time 0 you enter a monogamous relationship with Alice, which is looking relatively positive, but you are not yet fully committed and still “keeping an eye out” on the dating market. At time 1, you find Bianca who is a better “fit” according to the criteria outlined in “Components of Love”, i.e. they are more attractive [within bounds], have better aligned morals, share some childhood experiences with you, etc.

You’re in a dilemma: Bianca may seem better for you, but by switching now you are very much losing out on love! By using the framework outlined in this model it becomes apparent that for some length of time you will be experiencing less love than you would otherwise enjoy. The key is that *in the long-term* the benefits of switching partners must outweigh these short-term sacrifices. Getting tripped up on the sunk cost would be making your decision including all the “love experienced” so far in your relationship with A, the AuC between t=0 and t=1, but even looking prospectively there is certainly an element of loss in the situation.

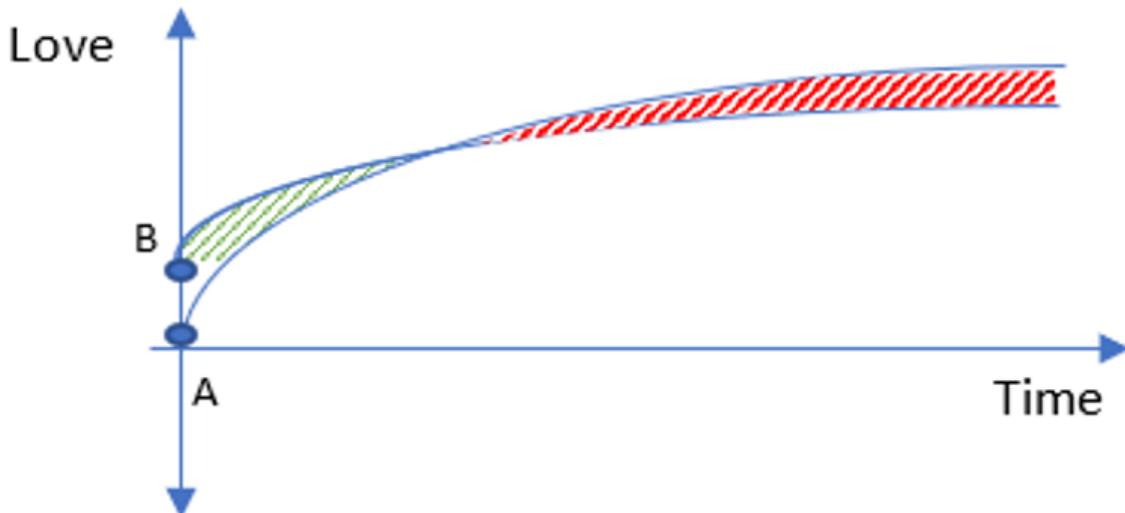
If you’re looking for a lifelong partner, then the choice to switch to Bianca becomes much more reasonable. This presents a conundrum: if you’re willing to switch *now* that kind of indicates you would be willing to switch again if Cassie came along who is an *even better fit*, this would limit your upside and may make the choice net negative. But if based on this you choose to stay with Alice, you will continue along Alice’s curve and by the time Cassie comes around the love lost may outweigh the love gained. The longer you’re with a partner the less likely you are to switch. How willing to commit are you?

On the other hand, if you’re newly entering the dating market and still trying to figure out your preferences, perhaps you don’t think you would be dating Bianca for very long anyway, so switching is not worth it right now. Similar reasoning applies if you have a clear end point that would likely end any relationship you had at the time: you’re moving overseas, are a secret agent going on your next mission, etc.



In the same vein, you might think you would stop dating Alice soon (they might be “keeping their eye out” as well), which gives you an enormous amount of net positive “love experienced” from switching to Bianca, as seen in the rough model above. There are many variations on this theme (what if you thought Bianca would leave you sooner than Alice, etc.) but hopefully the basic principles are becoming apparent.

Love choices - Choosing



Assume you are monogamous and have no partner, at time 0 you are weighing up a second date between Annie and Brie. You presently have more love for Brie (more attractive, etc. etc.), but know they have a really busy schedule and would be likely to have less time to spend together were you to commit to a long-term relationship than Annie. Because Brie will move more slowly along the "Shared Experiences" curve than Annie, it is entirely possible that making the choice of partner Annie is the right move in the long-term.

It's also entirely possible that Brie is still the right choice! It's important not to get stuck on "what you see is all there is" here: that additional free time could allow you to invest in other non-romantic relationships to make up for the missing love, or maybe you would happily trade it for the utility of time alone reading or making more money or etc.

If you're really thinking long term, you're probably going to be losing a good chunk of time to create new shared experiences if you choose to have kids; you'll be investing that time into making shared experiences with your children instead. Also note that none of this section factors in other costs such as shame, etc. associated with switching/choosing between partners; this will likely have quite a big impact on your decision making!

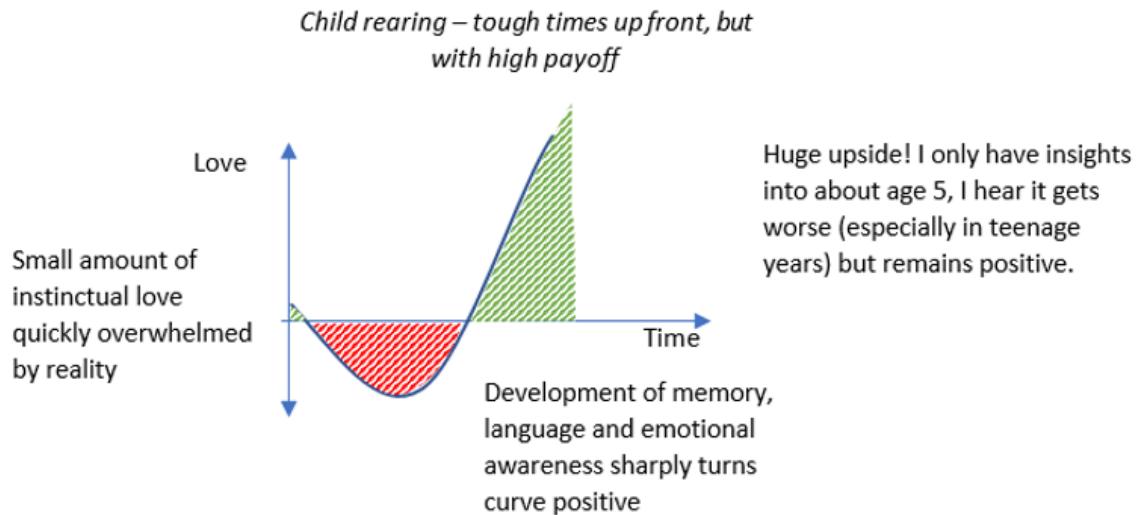
I was originally going to include personal experiences for each of these subsections, but to be honest just thinking conceptually has been a lot more fun for this one.

Having patience

Having a child was really hard for the first few years, like, *really hard*. I believe the words I used to describe it to a friend were along the lines of "for the first 3 years I fucking hated it". Zero intellect, zero perception of reciprocal love, zero moral alignment, zero shared experiences* and high frequency of emotional harm: it's rough.

But, but! In the time since the transformation in the love I feel has been astonishing. Being able forge shared experiences by enjoying some of my own childhood pleasures (Pokémon) or even more recent hobbies (watching video game speedruns) or teaching math has moved

my “love over time” relationship firmly into the positive. Also, you can’t hold on to resentments against children, so the emotional harms don’t stack up nearly as quickly as in adult relationships and are very quickly forgotten and/or reframed.



For the first 18 months I had a friend who would often ask me “are you in the black?” and at the time it was distressing that I felt nowhere near it (I have detailed the key points of the experience in [an appendix](#)). It was a sense of duty that was keeping me going, not love. Taking the long-term view on the relationship I now strongly believe the ROI is there, there are just serious costs that need to be paid up front. This view helps a lot with having patience.

Tangential side note: at the time I relied a lot on stoicism as a philosophy to get me through, which emphasises acceptance of the way of the universe and controlling your reactions to the demands of the present. It helped a little, but I’ve found that tweaking the stoic view and instead of just “letting go of the present” I instead imagine myself in 5 years’ time remembering back on the present, that helps a lot more than Epictetus’ guidance of basically just killing off your emotions. Admittedly, when I was in my early 20s “imagining myself in 5 years’ time” seemed like an impossible task given 5 years prior to that I was a teenager and that was a completely alien world, so projecting myself forward was difficult, but as I approached my 30s this became a lot easier to do.

*I maintain that this stays at zero because it is partly dependent on your own mental model of your loved one’s ability to recollect those shared experiences. There’s a lot of talk about how smart babies are etc. but based on my experience I’m pretty dubious on any level of memory of events existing for the first couple of years. Memory of people? Sure. Memory of events? No.

Taking actions to increase love

The above was a pretty fun analysis of making decisions in-situ or assessing relationships by effectively creating a net present value (I’m not going to bother figuring out how discounting could work into this). But what about making decisions to improve the love in an existing relationship?

Below are some quick thoughts on each component, and where you should put your focus.

- **Shared experiences** — This is what I see generally as the **biggest** opportunity. Even small investments in this area can add up over time as the result of an additional

shared experience created has an almost permanent increase in your future love experienced. Not bad for just making time to watch a new show on the couch together.

- **Perception of reciprocal love** — Not huge gains to be had here unless you're really getting it wrong. I've been recommended to read "[The Five Love Languages](#)" with your partner to help in this area but haven't found myself with a pressing need to do so.
- **Physical attractiveness** — Beauty is largely natural, so again not huge gains to be had here. Dress to match your partner so your relative position on the attractiveness curves remains consistent. If you really want to dress up because you want to boost your own self-esteem, encourage your partner to do the same.
- **Intellect** — You're going to struggle to move this one so investment for love's sake is probably not worth it. Accumulation of knowledge can make up for deficiency in intelligence, so if you are feeling a bit too low on the "Intellect" scale relative to your partner it may be worth your time learning some of the things they're interested in so the gap is less obvious (side benefit: a great way to learn is to discuss, and discussions with your partner on these topics also creates shared experiences!)
- **Non-sexual physical contact** — Recognise when circumstances change and ask yourself a few times a year if you're getting it right. Easy to drop off in a romantic relationship after children. In a situation like a pandemic this is going to take a hit so you probably should have some broad awareness of it, don't just take it for granted that you used to hug your friend and assume that's still okay.
- **Alignment of morals** — It seems really hard to deliberately shift someone's morals, so investment here is probably not worth any thought. It's entirely possible that by being overly enthusiastic on this front you will actually generate resentment from your partner (dragging them along to church or something) which could backfire quite unpleasantly!
- **Complementarity** — Probably not huge gains to be had here once you're in your 30s, maybe something for your 20s? If your partner is making up for too many of your own flaws, maybe you should spend some time on self-improvement.
- **Frequency of creation of joy** — Big opportunity here to be deliberate on this front. Requires more thought and effort than creating shared experiences, but the payoffs could potentially be worth it. Teaching your partner to make you happy more often would be very difficult to achieve, however.
- **Frequency of creation of emotional harm** — If things aren't getting better after you have made it clear to your partner that they are hurting you; this should be a strong indicator to get out early. The wrong move would be to tell yourself to "toughen up" and believe you can just stop feeling hurt. Sure, some level of resilience may be trainable, but this should be done at the very left end of the curve where the impacts of this item are still neutral if not positive.
- **Gross amount of resentment** — Assuming you are not willing to leave the relationship or have no reason to believe resentment will keep accumulating, perhaps the best bang for buck on improving love (even better than creating shared experience) is *forgiveness*. Having been in this situation myself, even figuring out "what is forgiveness" can seem incredibly daunting. Should I just forget the painful memories inflicted? Is forgiveness just some words you say? My own answer to this was being able to remove the pain from those memories when I recalled them, part of this was by rebuilding trust by creating new positive shared experiences after the painful event, part of it was understanding the perspective on the other person's side of the interaction, part of it was simply time (and here is where the "imagining myself in 5 years' time" trick helped). By dulling the emotional response to those memories, while still engaging with them, my gross amount of resentment reduced and the love experienced has recovered (not fully, some level of resentment is really hard to let go, but I'm working on it).

Pretty much, the big areas to target in a positive relationship are creating new shared experiences while in relationships which are on the rocks but have some hope of salvation, working on forgiveness is going to be the most important task.

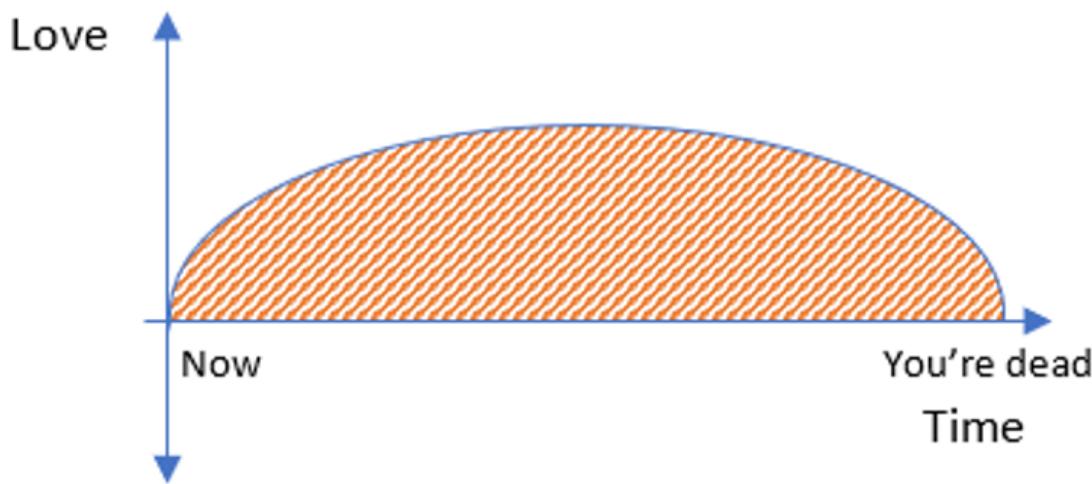
Expanding on the principles

Limits on Love

Looking at most of the charts so far you might get the impression that love can keep on growing forever. Alas, all things must come to an end, including relationships. As touched on earlier in the “*Switching without long-term commitment*” chart, this could be due to being part of an active dating scene or being early in your career and needing to move around a lot. But even if you aren’t in these situations, the further out you project, the less certainty there is that your partner won’t tip over into creating a significant amount of resentment or some other exogenous factor breaks you apart. And then of course there is the spectre of inevitable doom: Death.

The way to factor this into the Love-Time relationship is as an accelerating drag, pulling the curve to zero over time as the chances of something going wrong accumulate. Even in the most committed of relationships with absolute trust will trend down as you or your partner reach the unfortunate position of having a 100% chance of being dead.

*Love experienced – The best-case
long-term view*



The interesting part of this phenomena is how this drag on your future expectations of love changes over time.

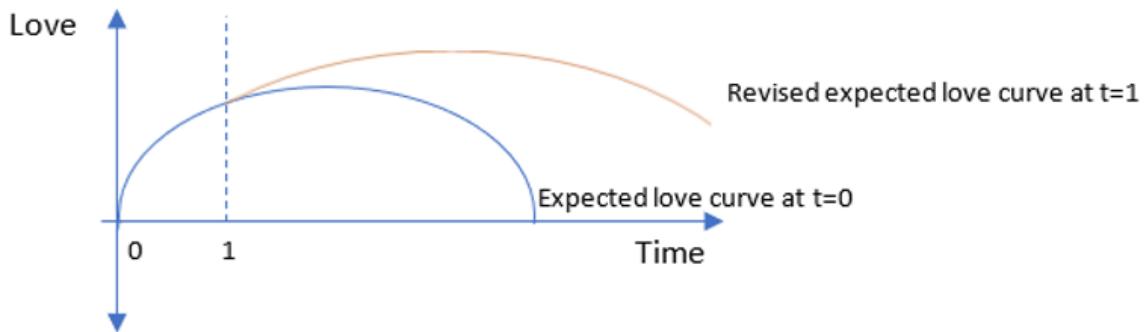
A subtle outcome of the analysis so far is that the longer you are in a loving relationship, the less likely you are to choose to substitute out of that relationship. There are a couple of drivers here:

- **Increasing switching costs** — as your “Shared Experiences” build up over time and you move “up” the love-time curve, your potential “love lost” increases for a given new partner, making switching less appealing. This is further compounded by there being less time left in your life to “make up” for this lost love (as explored above in “choosing between partners”), and cognitive biases such as the endowment effect and loss aversion would also play a role; and
- **Reduced uncertainty** — as you get to know the other party in the relationship, your mental model of them becomes more accurate. This combines with the point above to

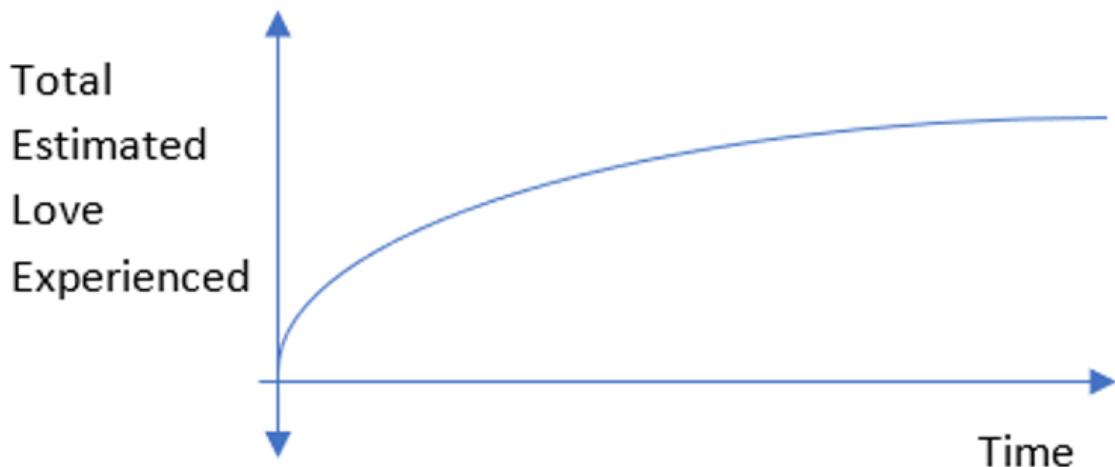
essentially say “if you were going to break up, you would have done it already” based on your expected future value. If you didn’t know your partner had a particular dealbreaker quality when you first met (say, they’re anti-vax, not interested in having children, or secretly two kids wearing a trenchcoat), the chances that you haven’t discovered that fact after a year is much lower than if you haven’t figured it out in the first week.

Gradually the curve will shift from having an endpoint where it is reasonable to plan ahead 1-2 years, to one similar to the “best-case long-term view” presented above.

*Changes in the expected love curve
over time*



This means that the longer you spend in a relationship, the more your total estimated love from that relationship will be:



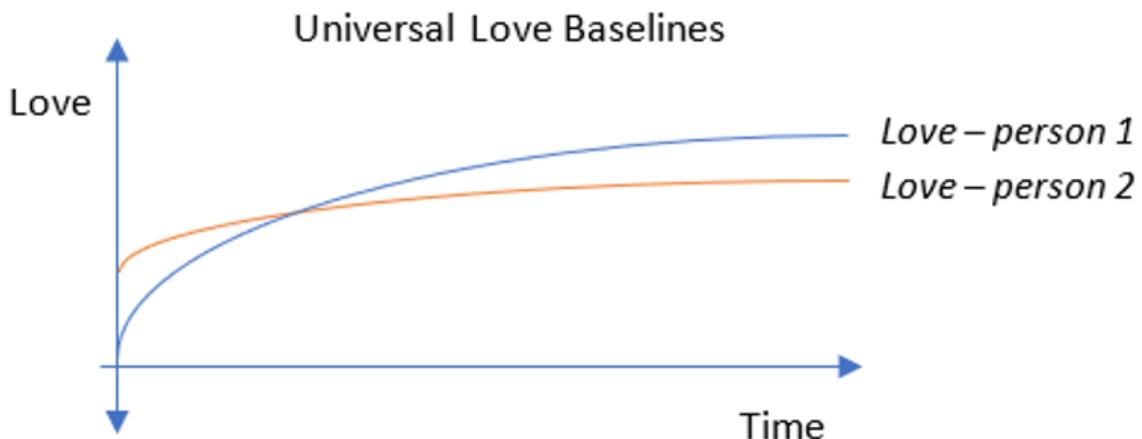
But you shouldn’t be looking at the “Total Estimated Love” because then you will be including all the sunk costs! What you *should* be looking at, especially when comparing potential partners, is the *future* estimated love experienced. As to how that looks and how that interacts with your own mortality, I’m not entirely sure, though I expect it will end up looking quite similar to the Total Estimated Love Experienced chart while under age 40.

To be honest, while somewhat morbid I actually think I was a bit *too* romantic in this section writing on the basis of “till death do us part”. I struggle to plan 5 years ahead and I certainly know many who struggle to plan 6 months ahead! This drastically shortens the window for assessing future expected love, and if I set an arbitrary cut-off on the curve for heuristic

decision making it could play out that switching is more appealing. I suspect that this dynamic is much stronger for those who have more of a preference for the “experiencing self” over the “remembering self”.

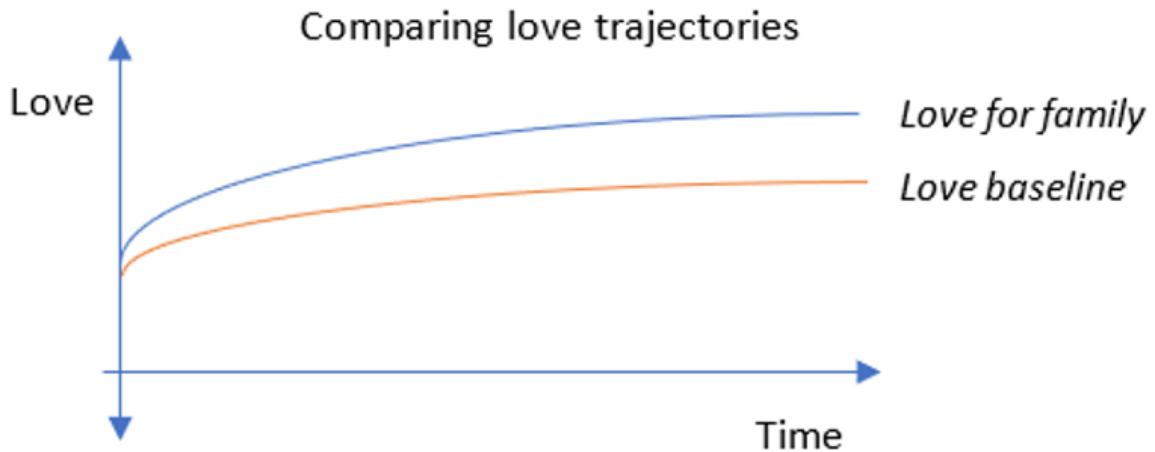
Speculation: Altruism

I am nowhere near confident on this area, but I have a pet theory that altruistic acts are motivated in by love, and that perceiving “shared experiences” in fundamentally different ways (e.g. qualia/consciousness as the definitive shared experience) allows some to love others more equally and prioritise efforts to assist others accordingly.



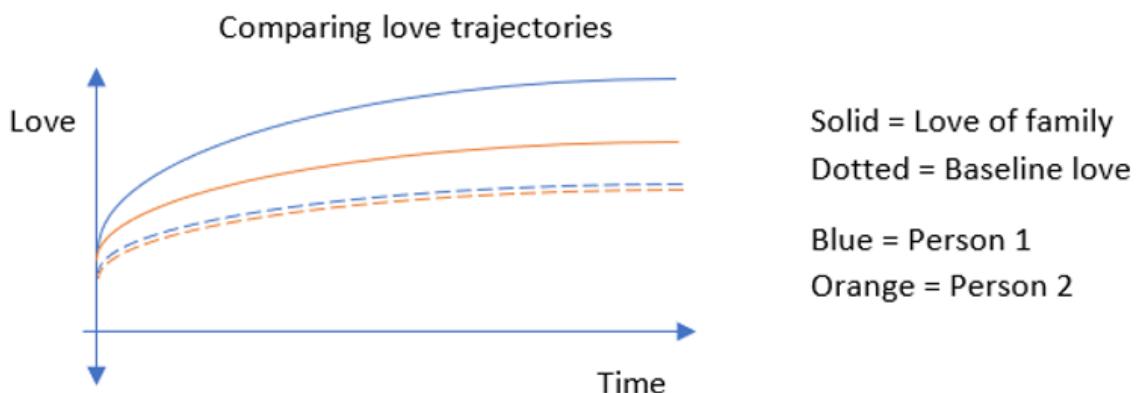
The above diagram shows example curves for two people for what could be assumed for their “baseline” love for any given stranger they should meet. Person 1 has a relatively low (still positive I will note!) starting point in any given relationship, with time spent in that relationship continuing to increase over time as explored earlier. Person 2 starts with a much higher starting point for love, as a result of being able to see the fact that a given stranger is alive and shares the truly defining experience of being conscious. While over the course of an average relationship they may not rate further bonding time as “meaningful” as person 1 would, they are much more likely to be generous to strangers.

Of course, person 3 could start at the same point as person 2 and see the same gains as person 1, different people have different capacities for love. Some are able to extend this to loving animals (projecting the experience of consciousness) and beyond. Love of nature and supporting altruism for “the planet” doesn’t really seem to fit this neatly: some deep seated belief in animism or potentially preserving nature as a potential for shared experience with others could be explanations, but it still feels like the model falls short on this front (e.g. how does it work for human extinction advocates?).



This of course does not mean that the decision to be altruistic is purely driven by “who I love more”, but is a combination of love plus some other factors (one that is heavily weighted for me is “perceived need”, perhaps linked to probability of reciprocal love?). The above chart shows a comparison of love curves between family and expected average for new, unrelated connections. Family gets a boost to start with (shared genetics = shared experience? Or just put it down to the instinct to procreate) and also has a steeper curve for gains in love over time as you are both more likely to spend time together, and for that time to be potentially more meaningful.

The same could also apply for expected love for those in the same church, of fellow country members, etc. which would boost each group's respective chance of being the recipients of altruism, independent of need or other factors.



This chart compares love of strangers and love of family between 2 people. Given the much smaller gap between family & strangers for Person 2 the theory holds that they would be more likely to donate to effective altruism causes (even if Person 1 is more loving overall).

I can't help my brain toying with this idea since my wife is capable of so much more love than me, and yet I find it really easy to see charity as something that gives me utility (I am personally interested in humanitarian effective altruism) whereas her own priorities are much more family driven.

In terms of application to the effective altruism community: Can appreciation of others' consciousness and qualia be taught? Maybe, but it seems a lot harder to me than investing money behind just flying millionaires to impoverished places, giving them a shared experience as a booster to their love for those who have greater addressable “needs”.

Conclusion

As a guy who often wonders just how far along the autism spectrum I am, I'm no expert on love. But also, as a guy whose first serious romantic relationship at age 17 turned into an ongoing, supportive marriage at age 32 I'd like to think that while I certainly got lucky — I also made good decisions. Looking around at age 20 during university and recognising that the probabilities were sure as hell against my first serious girlfriend becoming a good wife, I used the framework above in the “choosing between partners” to assess the dating market at the time and over and over again I reaffirmed that there was no one better for me. That sounds incredibly unromantic, but in my mind that makes the relationship even more powerful than someone who is just along for the ride (I'll note that well before the time I proposed I had stopped “assessing the dating market” based on the incredibly low probability of finding a more ideal match).

Working through the overall framework in this manner has allowed me to assess opportunities to prospectively further improve my relationship with my wife, son, parents and friends, and I can really see now that I should be making efforts to reduce my solo reading time at night while my wife watches TV in bed, and encourage her out to the couch to indulge in something together. While I may lose some personal “self-development” utility, I think the gains in my utility from overall “love experienced” are worth the tradeoff.

The “Expanding on the principles” section opens a number of questions and there may even be some data in existence which could potentially validate parts of the hypothesis (e.g. I'm sure there is data out there giving the probability of divorce for a couple who are newly married vs married for X years). The speculation on altruism would be hard to support with hard evidence, but I'd be interested to hear opinions on it out of intellectual curiosity.

See further reflections on parental love on Cartesian Coordinates [in this appendix](#)

Starting a Rationalist Meetup during Lockdown

Authors: [fkarg](#) (my [blog](#)), [ctrltab](#), [wilm](#)

Backstory and Intentions

In October 2020, wilm moved to Karlsruhe and was disappointed by the lack of a local LessWrong community. In order to change this, he reached out to fkarg and together they decide to try to establish a LW Meetup Group in Karlsruhe.

Despite the ongoing pandemic, which made in-person gatherings impossible, the Meetup Group has been going great! In this post we want to share what we have learned about starting and running a local meetup group during lockdown.

By sharing our experience we want to encourage others (you, YES YOU) to also initiate a meetup group. Lockdown is no excuse, and online might just work out!

In the following we will explain how we got started, how we organize our meetings in terms of content and structure and provide information about what did and did not work well.

Obligatory disclaimer: What worked for us might not work for you, and things that didn't work for us might work very well for you.

Why create a Meetup

Our reason to create a meetup was simple: to bring together aspiring rationalists from our area, to learn together, to share progress, to help each other, and to socialize. It also helps to get to know new people and to learn about how other people think.

Overview

The first step in creating the meetup group was writing a [post](#) on LessWrong to announce the meetup. Since the initial video-call, we have been meeting online almost every week. With a regular influx of new members, our group has grown to about a dozen regular attendees.

Similarly, the structure and content of our meetings has evolved over time. Our weekly schedule looks something like this currently:

- 19:00 – 19:10 People slowly coming in and welcoming each other
- 19:10 – 19:20 Introductions (if someone new is present)
- 19:20 – 20:00 Topic 1: Hammertime Sequence
- Topic 2
- [Topic 3]
- Scheduling Topics for next Week

- Feedback
- [End of Meetup]
- Open Socializing

The first few meetups were considerably shorter, but recently the open socializing went until around 21:30 to 22:00 o'clock.

(If you want to start your own meetup, it is advisable to adjust the structure and content to the interests and preferences of the people attending.)

From our experience, coming up with interesting and worthwhile content is nothing to be worried about as (so far) we always had more ideas for discussion topics or other activities than we manage to schedule.

One of the cornerstones of our meetup has been working through the [Hammertime Sequence](#). We started committing to work on one or two days of the sequence each week on our second meetup. This way, working on Hammertime and discussing our experiences and insights is giving our group a 'baseline-purpose' to get together, but this is rarely the highlight of the day. This honor usually belongs to Topic 2 or 3, if not to some discussion that started randomly.

After talking about the most recent Hammertime days, we usually have discussions, and sometimes presentations by one or multiple members. Often enough this is not structured: having an initial topic people are interested in talking about is sufficient. Over time, numerous digressions occur and topics shift repeatedly, usually with additional topics written down to talk about in more detail later.

Other times, especially when a topic has been explicitly scheduled in the previous week, there might be prepared talks or interactive sessions.

Some topics we recently talked about are:

- note-taking systems (e.g. Zettelkasten, roam and [obsidian.md](#), emacs Org-Mode)
- personal management systems and workflows (e.g. GTD-implementations or parts of OS setups)
- book recommendations and reviews
- how to track data and visualize it (e.g. health, finances)
- HPMOR

Another activity we often did is something we called 'Opinion-Speed-Dating'. For this, we split into groups of 2-3 people and pick a conversation topic from a curated list of topics designed to favor personal or philosophical discussions.

After 10 to 15 min we then come back together and share the most interesting thoughts and ideas from each sub-group, sometimes continuing the discussion and digressions. While being fun on its own, it also works well as a way to get to know each other.

Initiating a Meetup

If the idea of attending a meetup sounds appealing to you, why not try creating one yourself? In our experience, there is less to do organizationally than expected.

Here is what we did.

First Meetup

To get started, we read guides like the [How to Run a successful LW Meetup Group](#) or the [Meetup Cookbook](#). We tried to figure out what we want to achieve and roughly plan the first meetup (mainly getting to know each other and sharing expectations). Finally, we [wrote a post](#) in order to announce the new meetup. Don't forget to invite your friends!

Weekly organization is very manageable. Writing a post to announce the next meetup takes about 10mins, and (from our experience) it's not a problem to rely on the topics being scheduled during the previous meetup or getting decided spontaneously. No thoroughly thought-through plan needed.

Roles

[How to Run a successful LW Meetup Group](#) mentions a number of implicit and explicit roles, such as content provider, welcomer (someone that includes and introduces new people and greets meetup veterans), networker, or organizer.

So far (after about half a year) we did not need to explicitly assign most of these roles. fkarg emerged as a moderator, but many others help out with organizing, providing content, and writing the posts.

Infrastructure

We of course needed infrastructure to run an online meetup - but we didn't want to host it ourselves. Using publicly available instances is absolutely fine.

Our infrastructure changed drastically early on, so here is what we're currently using.

Video chat

The most important part for a virtual meetup is the platform for video calls. We use a [Big Blue Button](#) instance hosted by our local university for our meetups (provided for free to students). This allows for high-quality video chats with breakout sessions and screen sharing. Additionally, no one needs to install anything and the university setup guarantees privacy. The link to the room does not change, but it can only be entered when a moderator is present.

Coordination

Having an instant messenger group ([Signal](#) in our case) is useful for coordination between meetups (e.g. 'I'll be late today' or 'I won't be able to make it') and sharing information.

Everyone who shows up to a meetup is invited to that group.

Shared documents

We use a few collaborative online documents that can easily be edited by multiple users in order to store more permanent information. We have a couple of [HedgeDoc](#) (formerly CodMD, formerly HackMD) documents that every member can access. Google Docs would probably work just as well.

Currently, we use these documents for:

- Ideas for future meetups (mostly discussion topics)
- History of recent meetup topics
- Book recommendations
- Questions for Opinion Speed Dating
- A Meta-Pad with links to the other pads, every other pad links to this one
- A list with our expectations to this meetup and people attending it
- Numerous lists about various topics
- ...

We found this useful to establish common knowledge within the group (e.g. for future meetup topics if someone wasn't present, and looking up information).

Links are provided during a meetup, when it is being talked about, or in the messenger group for coordination.

Yes, that's it. Really.

Reasons for Success

So far the atmosphere during our meetups has been really great and people are thoroughly enjoying getting together. This allows for sharing personal problems and subsequent solving or at least iteration on them from everyone present (important: [beware of other-optimizing](#)).

We think that there are a few factors beneficial for fostering such a friendly atmosphere. Be aware of us being lucky and survivorship bias.

Below, we share a few of these factors along with general tips and reflections.

Success Factors: Social

Cameras

Something that helped create a more personal atmosphere is that we don't communicate via speech only. Video offers additional communication bandwidth which allows for a lot more nuanced interactions. Seeing how others react, even just for a split-second, reduces a lot of the inherent ambiguity in communication.

This is not an explicit policy, but rather an implicit one. We also have the feeling that people in general like to use cameras for interaction more than they ever did before the pandemic.

Breakout Sessions

We found that with fewer people, a very comfortable atmosphere allowing for honest exchange and topic digression presented itself naturally. Establishing such an atmosphere got a lot harder with more and continuously new people attending. An easy solution for this was to create breakout-sessions (i.e., splitting up the video conference into multiple rooms) with up to five people.

The 'comfortable' atmosphere from the small groups tends to spill over to the 'big' meeting even after breakout sessions have concluded. This is one of the reasons we try to have breakout sessions early in our Meetup.

Benefits are multi-faceted:

- People are less afraid to talk longer
- People have less inhibition to speak up or mention related anecdotes and information
- People are more comfortable sharing personal details and asking for help with problems
- Smaller groups are much more self-directed, needing little moderation
- It's easier to actively integrate everyone (especially newcomers)
- It's easier to establish trust with each other
- It's easier to 'synchronize' with each other to establish a group mentality / atmosphere

Hand-Signs

When your group hits a certain point (~8 people or more), receiving input from everyone becomes a communication challenge and can easily take a long time. In-person meetings sometimes have rules for a number of hand-signs, signaling a number of different things:

- 'I want to talk'
- 'I have an objection to this'
- 'This is wrong, I can tell you how'
- 'Applause'
- and [many more](#)

We do of course not use all of these gestures to the same degree, the one used most often is the [temperature probe](#): receiving immediate feedback on how people feel about a certain proposition.

Allow Topic Digressions

Something we see a lot of value in for ourselves is the ability to have frequent topic digressions. Even when a topic is provided, it's not rare to stray to another - sometimes related, sometimes not - topic for a few minutes before continuing the original discussion.

These digressions are usually started by someone asking a question about a personal challenge related to the topic or sharing related information.

These very valuable interventions happen less in larger groups:

- It's not clear that everyone is interested or wants to participate in a topic digression

- No one wants to take up a lot of 'talk time', because it would prevent others from doing the same
- It's harder to keep track of everyone, which limits the energy available to keep track of frequent jumping between arguments and topics

We noticed that having space for these kinds of digressions is what's enabling the previously mentioned friendly and open atmosphere in the first place. The value gained from deeper discussions with digressions is a big part of why people participate.

Shared Culture

We noticed that most members have a number of common interests. They might be interested in certain (niche) subcultures, having just read HPMOR, or spent a lot of time configuring their computer setup.

Having discovered these common interests, delving into them is a nice way of connecting with each other. These tend to be more off-topic than usual (but also fun, e.g., who would have thought that a large fraction of attendees uses [a custom keyboard layout](#)).

Success Factors: Other

Posts on LessWrong

Everyone who joined found us directly through a post or indirectly through getting invited by someone who joined through a weekly post. This means that the obvious method of attracting members seems to work quite well. Most people who joined us at some point also stayed.

In order to find a balance between attracting new people and retaining the personal atmosphere, we only make a post for every other meetup. This has also worked well so far.

Make your physical meetup location noticeable

Most of our meetups have been in German. We spontaneously decided to hold some in English when someone was not able to participate in German.

We very much welcome new faces, and are more appropriately a 'Karlsruhe Area'-Meetup. Still, we would appreciate it if people were at least aware of where [Karlsruhe](#) is, and that we might speak German.

We noticed that the quality of expression takes a small but noticeable hit when switching to English. To reduce unnecessary friction in discussions, a policy going forward is that we will hold Meetups in German.

Since we added 'Germany' in parentheses to our description, almost all newcomers were able to participate in German.

Having Community Experience

While it's hard to define what an 'experienced' rationalist would constitute, it's much easier to define 'community experience'. Community Experience is time spent with other rationalist meetups - how they were organized, and what worked for them.

'How to organize a LW meetup' is a great guide - and one we consult frequently - but having seen an implementation in person as a reference is worth a lot.

Luckily, we have a few people with community experience (e.g. through the LW European Community Weekend in Berlin, the meetups at Chaos Communication Congress or other events).

Having people who had exposure to other rationalist communities is incredibly helpful for initial formation and later iteration.

Don't talk too much about the Meta

We had a point earlier to allow for digressions. Thing is, you shouldn't let them get out of hand either.

If necessary (and they absolutely are) have meta-discussions about the structure of your meetup. This includes frequency of LW-posts (every meetup?) as well as general structure and topics, as well as why you participate and what your expectations are. Do try not to discuss too much about things that do not matter much, like how to vote properly.

Make separate meetings or at least a separate session for those wanting to discuss meta-meta-topics, and present results to the others. Those presenting meta-suggestions to the rest of the group need to be really thorough in explaining their reasoning behind the proposed changes - rationalists *will* spot everything that's even slightly odd or not the 'ideal' solution. *That's what we are trained for.* If that happens, the most immediate discussion will be about details you miscommunicated. Take in new ideas, but don't stall making decisions.

Things to keep in mind

There is a number of things to keep in mind, however. Here is a few we think are important.

Diversity

One of the goals of our meetup was to establish a local community that allows rationalists and adjacent interested people to meet and connect with like-minded individuals and talk about the art of rationality. For this, having a variety of different viewpoints is very valuable.

We did not have this luxury for some time. When we reached the size of almost ten regular attendees, all of us were associated with the same faculty - computer science. We had students of varying semesters, graduates and postgraduates, but we were all associated with computer science. In short: our views did not differ about most things, and we had a number of topics thought through for ourselves. This allowed for much more in-depth technical discussions - sometimes even far out of our areas of expertise.

This has changed recently, with a number of new people - not associated with computer science - joining us.

While it is to be expected that a LessWrong meetup group attracts certain groups more than others, listening to a broader range of voices is beneficial for all of us.

A change of perspective is worth 80 IQ points -- Michael Nielsen

Still, this is something to watch out for, as new attendees might feel uncomfortable if they don't have CS knowledge and technical discussions occur frequently.

If you have experience with similar situations or have some tips for how to foster diversity, we would like to know more about them. Feel free to share them in a comment below.

Moderation

Moderation is not really needed when it's only three or four people, but larger groups need a moderate amount of moderation to stay at least somewhat on topic and schedule.

Non-Public Meetups

At some point we started having (every second) meetups without announcing them on LW first. The goal was to have a balance between including new people and deeper and more personal discussions, with unannounced meetups focusing on these profoundly valuable meetups going deep.

Hammertime Integration of New People

While a clear path to working on Hammertime is great for cohesion with current participants, it's hard to integrate new people. Some newcomers tried catching up, others started doing the same days we did. At the same time, not everyone is interested in doing or discussing Hammertime anymore.

To accommodate for that shifting interest, we start out with creating a separate Breakout Session to talk about Hammertime - those not following usually quickly find their own topic to talk about.

Changing Plans

Of course, not everything you try at first works out. Here are two examples of things we tried, but changed soon after.

Public Jitsi

The first few meetups were on the [public Jitsi](#) instance. It was difficult to organize, but only since we were paranoid about it.

It was complicated, because we rotated the Meeting id without using the same Room every time. Video quality suffered with every new member. Having our meetups on a

private [BBB](#) instance is much more comfortable, with less lag and much better scaling.

Public Telegram Group

We had a Telegram group initially. We then made the mistake of publicly posting the invite-link on LessWrong to our meetup posts. We cannot explain what caused bots to join and advertise esoteric cryptocurrencies otherwise. This, and a meta-discussion, resulted in us migrating to a closed [Signal](#) group instead.

Key Takeaway: Even on LW, don't publicly post invite links to private groups for popular messengers - you're asking for spam.

Ideas for the future

These are things we plan to try at some point in the future, some closer and some further away. If you tried them with your group or meetup, feel free to share any experiences about them.

Meeting in Person

This goes without saying: we would love to meet up in person once it is a reasonable alternative. We haven't thought about this yet, as it seems to be a bit further off.

Gather.Town

ctrltab managed to get us access to the LW town, and we tried it a few weeks ago.

This did not work out so well, as it took us half an hour to get everyone there first. Someone was always muted, and we didn't quite know how it all works. We do think it's ideal for dynamic breakout sessions and discussions though.

Not just due to lag, but for large-group discussions we prefer our BBB instance after all.

Own Infrastructure

Something we thought about is having some form of own infrastructure; primarily a wiki and pads. Our shared documents ([HedgeDoc](#)) are on the publicly hosted instance, and 'security' happens by not sharing the pad-links publicly - everyone could edit them.

Having our own infrastructure with backups and access control (at least for writing), would make us more comfortable. Our demographic is, after all, mostly CS people with detailed knowledge of what could go wrong. You need to make the backup before you need it.

What to do when Hammertime is over?

Hammertime is one of our basic 'pillars' for meeting. We are now close to being done with Hammertime, but we have many more ideas with which courses or similar exercise collections to follow up with. Examples include the [Training Regime](#), the series on [Guilt-Free Motivation](#) and others.

Conclusion

If you think about organizing a new meetup yourself, or you want to try online-meetups with your group, we encourage you to just try it out. We would like to know what other Meetups are doing during lockdown - or did, during lockdown.

The "How to run a LessWrong Meetup" booklet is very helpful to get started, and we have consulted it multiple times.

If you would love to interact more with other people on LessWrong - join a Meetup, or create your own!

Please share any advice or other relevant experiences.

AI Safety Research Project Ideas

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

This post contains project ideas in AI Safety from Owain Evans and Stuart Armstrong (researchers at [FHI](#)). Projects are aimed at students, postdocs or summer research fellows who would be interested in collaborating on them for 2-6 months with Owain or Stuart. We are happy to explore possible funding options for the duration of the project. If you are interested in discussing mentorship or academic collaborations please get in touch -- details are at the bottom of this post. The deadline is EOD 20th of June but we encourage people to apply early.

Project ideas from [Stuart Armstrong](#)

1) Model splintering: How can you automate moving from one model to another?

Context: [This post](#) by Stuart Armstrong gives an overview of model splintering: examples, arguments for its importance, a formal setting allowing us to talk about it, and some uses we can put this setting to. [This post](#) looks at how the formalisms of Cartesian frames and generalised models relate to each other.

Details: Humans are capable of extending moral values to new situations, when their previous concepts no longer apply. This is analogous to the ability of a reinforcement learning agent to generalise its previous reward signal to new situations, when the reward signal is no longer available and the environment is out of distribution.

This project posits that that is not a mere analogy: that the human capacity for extending moral values (which includes analytic philosophy and thought experiments to un-encountered situations) is a skill which can be transposed into algorithms, and further automated to extend to environments shaped by powerful AIs, about which humans have no current intuitions.

The main initial task is to collect references from analytic philosophy, human value changes, and out-of-distribution behaviour in algorithms. The insights from these areas should then be combined in the model-splintering formalism, and new algorithms created in this formalism to generalise for advanced AIs.

The output of this research should be a few publications on new methods for safely extending AI reward and values to new areas, and maybe some sample code.

2) Detecting preferences in agents: how many assumptions need to be made?

Context: [This post](#) by Stuart Armstrong gives some relevant context on detecting preferences in agents. [This post](#) summarises a research agenda on synthesising human preferences, with links to the full version given in the text.

Details: This project will mainly be programming based, though a literature review of relevant control systems ideas will also be carried out.

Previous results demonstrated that the preferences of an irrational agent cannot be deduced from its behaviour, unless one makes a certain number of "structural

assumptions" (or "normative assumptions"). This project will test how many such assumptions are needed.

The basic idea is to create models of agents in grid-world situations, agents with preferences and biases, and train a classifier to deduce their preferences, given various collections of true assumptions about the agents. These examples will be analysed to see what kinds of assumptions are best for deducing agent preferences, and how many are needed.

The outputs of the project should be a computer science paper and some programmed example agents that others could build on.

3) In what way could value learning be dangerous, and how could it be made safer?

Context: A [previous result](#) demonstrated that one cannot deduce the preferences of an irrational agent, without learning "structural assumptions". Programming these assumptions into an AI, however, involves giving that AI knowledge about humans and the world - knowledge that might increase its power faster than its alignment.

Details: What is the safest way of deducing human preferences? This project will use a mixture of philosophical analysis, situational analysis, and computer science examples to explicate what kind of information provides the best increase in alignment without excessive increase in the AI's power. The issue of practical symbol grounding will be explored if there is enough time - practical symbol grounding gives the AI a lot of power over the world, if it knows what various symbols *mean*.

The outputs of this project will be one paper on value learning, and possibly one on symbol grounding, and some examples of agents learning and (mis)behaving in various circumstances.

Project ideas from [Owain Evans](#)

4) Alignment and large language models

Context: I'm interested in collaborating on projects about language models from NLP such as GPT-3 and T5. General areas of interest are:

1. Aligning large language models with human preferences and other normative criteria. For example, how to make models more accurate, reliable, helpful and transparent. (Related work [1](#), [2](#), [3](#))
2. How do current language models relate to AGI? What are the limits of the current paradigm? (Related [work](#))

Details: I have some specific projects in mind that I will discuss with applicants. I'm also open to considering projects proposed by applicants in these general areas. Applicants should have some background in machine learning and be comfortable reading and understanding new papers in ML (e.g. Neurips or ICML papers). It's helpful to have taken a course in ML, implemented ML models, or written ML papers or blogposts. However, no formal credential in ML is required. In addition, any of the following skills are helpful:

- Experience with contemporary NLP models: e.g. applying models, training them, and doing published research in NLP

- Research experience in any area of machine learning or a related field. Evidence for this is an academic paper or a blogpost
- Background in analytic philosophy, formal logic, or “Agent Foundations”. Evidence for this would be university courses, workshops, blog posts, research papers or reference letters

Mentorship and funding

If you are interested in working on any of these projects and would like to explore mentorship or funding options, please fill out [this form](#). The deadline is EOD 20th of June but we encourage people to apply early. We will aim to respond by the 28th June and will put candidates who are a good fit in touch with Owain or Stuart.

If you plan to work on any of these projects without funding or mentorship, please let us know to avoid duplication of work by sending an email to aialignment.group@gmail.com with the subject line ‘AIA project - full name’.

If you have any questions, please reach out to aialignment.group@gmail.com with the subject line ‘AIA question - full name’. Please do not use this email address to submit mentorship or funding applications.

Questions are tools to help answerers optimize utility

Epistemic & Scholarly Status: Fairly quickly written. I'm sure there's better writing out there on the topic somewhere, but I haven't found it so far. I have some confidence in the main point, but the terminology around it makes it difficult to be concrete.

TLDR

The very asking of a question presupposes multiple assumptions that break down when the answerer is capable enough. Questions stop making sense once a questioner has sufficient trust in the answerer. After some threshold, the answerer will instead be trusted to directly reach out whenever is appropriate. I think this insight can help draw light on a few dilemmas.

I've been doing some reflection on what it means to answer a question well.

Questions often are poorly specified or chosen. A keen answerer should not only give an answer, but often provide a better question. But how far can this go? If the answerer could be more useful by ignoring the question altogether, should they? Perhaps there is some fundamental reason why we should desire answerers to act as [oracles](#) instead of more general information feeders.

My impression is that situations where we have incredibly intelligent agents doing nothing but answer questions are artificial and contrived. Below I attempt to clarify this.

Let's define some terminology:

Asker: The agent asking the question.

Answerer: The agent answering the question. It could be the same as the asker, but probably later in time. Agent here just means "entity", not agent vs. tool agent.

Asked question: The original question that the asker asks.

Enlightened question: The question that the asker should have asked, if they were to have had more information and insight. This obviously changes depending on exactly how much more information and insight they have.

Ideal answer: The best attempt to directly answer a question. This could either be the asked question or an enlightened question. Answer quality is evaluated for how well it answers the question, not how well it helps the asker.

Ideal response: The best response the answerer could provide to the asker. This is not the same as the ideal answer. Response quality is evaluated for how it helps the answer, not how well it answers the question.

Utility: A representation of one's preferences. Utility function, not utilitarianism.

Examples

Question: What's the best way to arrive at my dentist appointment today?

The answer to the *stated question* could be,

Take Route 83 at 6:30pm

The answer to an *enlightened question* could be,

Your dentist is sick, so your appointment will definitely be cancelled

A good *response*, knowing the question, but not answering it, might be,

It doesn't really matter what route you should take, but now that I know that you're concerned about the trip, I can tell you that the best way for you to save time today would be by ordering in food from Sophia Gartener at 3:30. It will arrive at 5.

A good *response*, ignoring the question (or correctly not updating based on it), and optimizing for utility, might be,

There's a possible power outage happening in the next few days. I suggest borrowing a generator sometime tomorrow. I've left instructions for how to do so in an email.

The puzzle with the later answers is that they seem like poor answers, although they are helpful responses. The obvious solution here is to flag that this is a very artificial scenario. In a more realistic case, the last response would have been given before the question was asked. The asker would learn to trust that the answerer would tell them everything useful before they even realized they needed to know it. They would likely either stop asking questions, or ask very different sorts of questions.

The act of asking a question implies (it almost presupposes) an information asymmetry. The asker assumes that the answerer doesn't have or hasn't drawn attention to some information. If the answerer actually *does* have this information (i.e. they can intuit what is valuable to the asker and when), then it wouldn't make sense to ask the question. This is an instance of the [maxim of relevance](#).

So, questions make sense only until the answerers get good enough. This is a really high bar. Being "good enough" would likely require a tremendous amount of prediction power and deep human understanding. The answerer would have to be much more intelligent in the given area than the asker for this to work.

Breakdown

If we were to imagine a breakdown of information conveyed in the above question, we could then identify a more precise and empathetic response from a very smart being.

You've asked me how to get to your dentist appointment. This reveals to me the following information:

1. You are unsure about how to get to a dentist appointment.
2. You believe that the expected information value you can get to optimize your route is more valuable than the cost of asking the question.

3. You expect that I either couldn't predict that such information would have been valuable to you without you asking it, or I wouldn't have told you unless asked.
4. You do not expect me to have much more valuable information I could relay to you at this time.

Human, I believe you have dramatically underestimated my abilities. I believe that you are severely incorrect about points 3 and 4. You have much to learn on how to interact with me.

Students and Professors

Another analogy is that of students and professors. Many students don't ask any questions to their professors, particularly in large classes. They expect the professors will lead them through all of the important information. They expect that the professors are more informed about which information is important.

In many situations the asker is the one trying to be useful to the answerer, instead of it being the other way around. For example, the professor could ask the students questions to narrow in on what information might be most useful to them. I imagine that as the hypothetical empathetic professor improves along a few particular axes, they will be asked fewer questions, and ask more questions. In this later case, the questions are mainly a form of elicitation to learn about the answerer.

Corrigibility

There could well be situations where answerers assume that they could better respond with a non-answer, but the askers would prefer otherwise. This becomes an issue of [corrigibility](#). Here there could be a clear conflict between the two. I imagine these issues will represent a minority of the future use of such system, but these instances could be particularly important. This is a big rabbit hole and has been deeply discussed in the corrigibility and similar posts, so I'll leave it out for this post.

Takeaways

I think that:

- Answerers should generally try to figure out enlightened questions and answer those questions. This method is often the one that will be best for the asker's utility.
- If it is the case that answers can better help the askers by ignoring the question and instead doing something else, that's better. They should try to give the ideal response, not the ideal answer.
- However, in almost all cases now, the best response to attempt is the ideal answer. This is true just because the askers often have some key information not accessible to the answerers. Often, when askers ask questions, they believe there's likely sufficient benefit for these particular questions to be answered, so humble answerers should often trust them.
- Once ideal responses are very different from ideal answers, then people will stop asking questions. Questions primarily serve the function of helping responses to be more useful, so if that no longer holds, questions will be no longer valuable.

Correspondingly, I imagine that as AGI gets close, people might ask fewer and fewer questions; instead, relevant information will be better pushed to them. A really powerful oracle wouldn't stay an oracle for long, they would quickly get turned into an information feed of some kind.

Thanks to Rohin Shah for discussion and comments on this piece

Two Definitions of Generalization

Epistemic Status: Research notes, plus an appeal for the reader to give their best attempt.

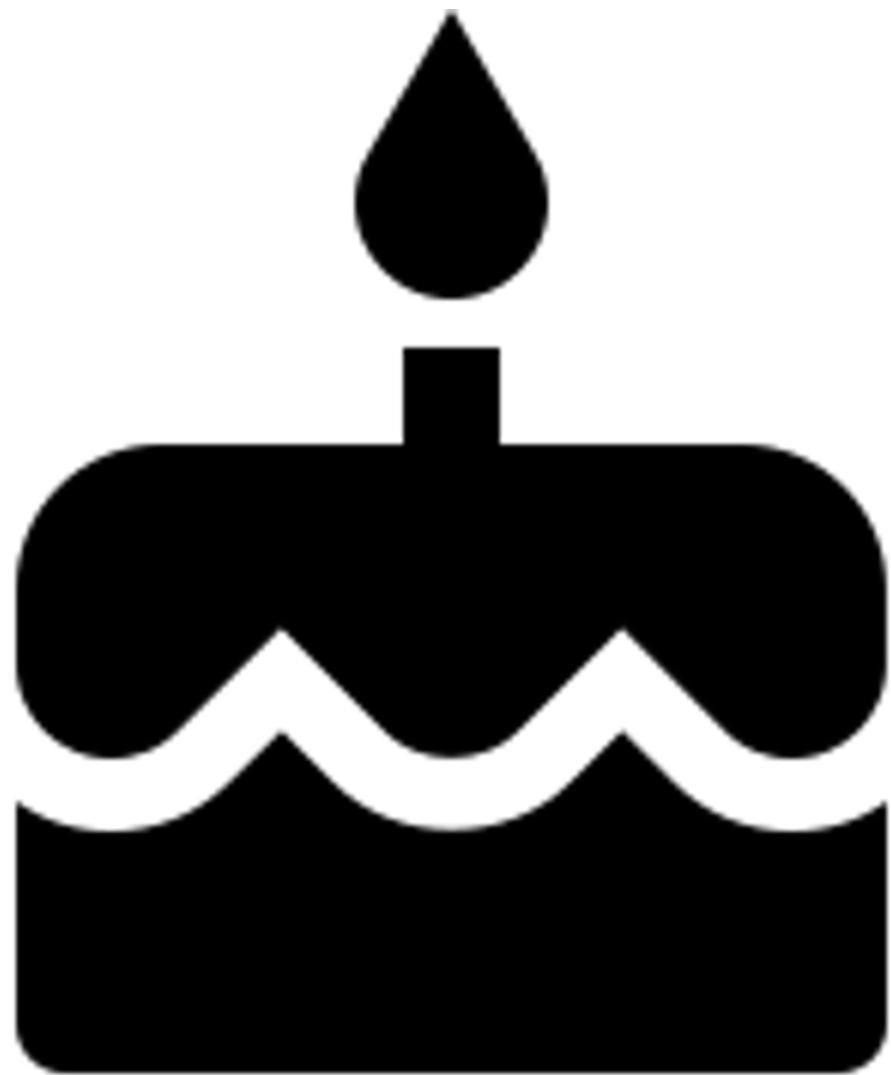
The words “abstraction” and “generalization” are often used fairly interchangeably. This is a pity. We only have so many well known words for abstract concepts.

One distinction I like is the one suggested in [this Stack Overflow post](#). The following images are from the top answer:

Object



Abstraction



Generalization



Simple, right?

Well, I think there are two important definitions of generalization that could apply both that merit discussion.

The first, and perhaps the better known, is essentially what's stated on Wikipedia:

generalization (1) (Wikipedia): A **generalization** is a form of [abstraction](#) whereby common properties of specific instances are formulated as general concepts or claims.

This definition is about identifying the similar properties of supersets and removing the other information. So, for the dessert example above, generalizing the cake would mean identifying the common properties it shares with other deserts.

However, this definition doesn't work for many of our uses of the term.

- When we discuss "general intelligence", we mean something like, "the superset of all narrow intelligences", not the "elements common to all narrow intelligences".
- When I write a "highly general function", this could refer to a function that takes in many sorts of things but still manages to treat them all uniquely.
- [The lists of "generalizations"](#) include statements like, "Most people find church boring."

From these, we could imagine a different definition. Something like,

generalization (2) (Here): A **generalization** is a pointer to a set of things.

I use the word *set* here because it's mostly correct, but there could clearly be modifications. It might really be more of a [fuzzy set](#).

One way I could write a general function would be to remove all of the differentiating details between items. This would be generalization(1). But I could also write this function by just

moving a complexity into it and adding a big switch statement. This would be generalization(2).

I'm tempted to say that generalization(2) corresponds somewhat to set theory, and generalization(1) more to type theory.

With all that out of the way, we can attempt some more specific generalization(1) definition:

generalization(1)(Wikipedia complexified): A **generalization** is an abstraction made up some of common properties of a set.

Generalization(1) is doing more work than generalization(2), though the result a smaller thing.

Why is this important?

I'm interested in this question because I think it might be useful to study generalization(2). My recent [sequence on questions and discernment](#) wound up getting into this area.

It might be easy to dismiss generalization(2) as trivial once we already have set theory, but I think there's more here. Set theory is typically discussed abstractly. If we have "[Judgemental Forecasting](#)", it seems to me like we could also have something like "Judgemental Set Theory".

Asides:

Short rant on semantic search

I'm sure this discussion exists somewhere in philosophical literature, but it's very difficult to search for. The word "generalization" is used all over the place and often isn't defined.

Google search is quite poor for such queries. I imagine it would require some semantic search capabilities.

Objectivist epistemology

I wrote a short description of the above on [Facebook](#), and [Jason Crawford](#) responded with an useful comment.

In **Objectivist epistemology**, "*generalization*" refers to scope of a concept, and "*abstraction*" refers to *distance from the perceptual level*. So "plumber" is less general than "human", but it's *more abstract*. Some steps of abstraction are generalizations (human -> organism), but many are narrowings.

He said that he believe he remembered it from [this lecture](#). Unfortunately the lecture costs \$34. I haven't purchased it, but I'll keep it in mind. I did some searches for generalization around objectivist epistemology but couldn't find this distinction written publicly in my brief time spent searching.

Another (similar) definition

Some would call the statement "rich people are greedy" a generalization. Here, the generalization isn't referring to "rich people", but rather to the "are greedy" portion. Perhaps "rich people" only act as a generalization if it's used to make queries, maybe only overconfident queries.

I think *generalization* here means the same thing as *overgeneralization*. I'd vote to not use definitions of *generalization* that do this. Ideally, it could leave space for [undergeneralization](#).

The case for hypocrisy

This is a linkpost for <https://aaronbergman.substack.com/p/the-case-for-hypocrisy>.

Related: [The case for logical fallacies](#)

Julia Galef makes an interesting point in her recent book [*The Scout Mindset*](#): our beliefs come as tangled knots, not isolated strings. Changing one belief often implies that we change many others.

Consider Sarah, whose relationships, political beliefs, worldview, daily activities, and ethical code are all fundamentally derived from her religious beliefs. Sarah can't merely decide that God doesn't exist or that Hinduism is correct instead of Judaism or whatever; if taken to heart, such a change in worldview would imply reform of virtually every other aspect of her life: her belief that abortion is intrinsically immoral, her belief that contributing significant time and money to her congregation is an ethical and meaningful thing to do, and her belief that it is good and appropriate to go to Synagogue every Friday, among countless others

If Sarah wishes to maintain a harmonious, coherent set of practices, beliefs, and attitudes, it would take a *tremendous* amount to convince her that God isn't real—crucially, *more* than if this belief were siloed away from the rest of her life and mind.

This isn't an indictment of religion. It would take an equally huge amount of evidence to convince me that I should convert to orthodox Judaism—more than if my non-religiosity was siloed away from my other beliefs and behaviors.

The key clause, though, is "*if Sarah wishes to maintain a harmonious, coherent set of practices, beliefs, and attitudes.*" Why should Sarah wish to do so? Why should anyone?

The case against hypocrisy

Before making the case for hypocrisy, let me explain why, in many respects, hypocrisy is bad and maintaining a consistent set of practices and beliefs is good. I'm not just steelmanning to strengthen my later argument; hypocrisy often really is something to be avoided. The word has several definitions, but I'll use [Merriam-Webster's](#)

Definition of hypocrisy

1. a feigning to be what one is not or to believe what one does not : behavior that contradicts what one claims to believe or feel

Also, much of the rest of this post will apply to plain old inconsistency, or holding two or more contradictory beliefs.

In general, from a non-religious perspective, our beliefs do not intrinsically matter (to others, that is; they may directly impact our own conscious experience). Our actions do. It doesn't matter whether you believe that animal suffering is bad, or that Trump is awesome, or that we should end homelessness. It matters whether you *act* on those beliefs, perhaps by foregoing factory farmed animal products, voting and donating to the Trump 2024: Make America as Great as it was From 2016-2020 campaign, or becoming a YIMBY activist in your city.

The thing with action is that sometimes it's hard. Chicken nuggets taste good. Voting can be a hassle. Getting rid of single family zoning might decrease your property value.

Our natural, moral distaste for hypocrisy is a decent solution. We get outraged when someone who professes to believe X does or believes something in that seems to conflict

with X. That's why Tweets like this one are so delicious.



Rachel McCarthy James

@rmccarthyjames

...

I keep thinking about the yard on my walk last week that had one of those "you're our neighbor 😊" signs right beside a NIMBY anti-affordable-housing sign



3:55 PM · Nov 29, 2020 · Twitter for Android



Rachel McCarthy James
@rmccarthyjames

...

I keep thinking about the yard on my walk last week that had one of those "you're our neighbor 😊" signs right beside a NIMBY anti-affordable-housing sign



3:55 PM · Nov 29, 2020 · Twitter for Android

To a large extent, this is a force for good! Lots of people are well-intentioned and want to believe true, good things, and many succeed in doing so. Our aversion to hypocrisy is a clever socio-psychological mechanism to turn good beliefs into good deeds. The process might look something like this:

1. John becomes convinced that buying factory farmed eggs is bad.
2. He keeps buying factory farmed eggs out of habit and behavioral inertia.
3. He feels bad about being a hypocrite or becomes worried that others will see him as a hypocrite.
4. John stops buying factory farmed eggs.

Cool. Now, for the contrarian take.

The case for hypocrisy

[One man's modus ponens is another man's modus tollens.](#)

- Confucius (just kidding, I don't know who said it first)

Our aversion to hypocrisy can also have the opposite effect. For example...

1. John becomes convinced that buying factory farmed eggs is bad.
2. He keeps buying factory farmed eggs out of habit and behavioral inertia.
3. He feels bad about being a hypocrite or becomes worried that others will see him as a hypocrite.
4. John decides that buying factory farmed eggs actually isn't bad, or just tries to forget about it (and this need not come from conscious deliberation).

Ok, you might ask, why is hypocrisy the problem here? If John didn't feel bad about being (seen as) a hypocrite, wouldn't he have kept buying factory farmed eggs anyway? Maybe. He certainly would have found it easier to continue buying the eggs while holding the intellectual belief that doing so is wrong.

But aversion to hypocrisy isn't the only reason people do things.

Even if John doesn't care one iota about his hypocrisy *per se*, he might eventually decide to stop buying the eggs for some other reason in the same way he might [donate to the Humane League](#) since doing so isn't in *direct* contradiction with the belief "buying factory farmed eggs is fine."

Maybe he doesn't give up eggs but does start making an offsetting donation to effective animal welfare charities. Maybe he starts buying humane-certified eggs most of the time. Whatever you think about their moral worthiness, these alternatives might be available to John in a way that egg abstinence is not.

Identity

This is particularly likely if John's egg consumption is tangled up with other beliefs and identities.

For instance, say John is a die-hard keto bro who thinks Big Vegan is conspiring with the seed oil industry, Big Pharma, and the FDA to push inflammatory and insulin-spiking fruits, grains, and unsaturated fats on the American people, and understands his egg consumption as a vote against this industrial complex.

Ok, fine. If John's ultimate goal in life is to avoid hypocrisy but he is unwilling to forego the eggs, he'll do whatever it takes to avoid the conclusion that buying factory farmed eggs is wrong. And if he does this, he'll never have a reason to explore alternatives like making an offsetting donation or spending a bit to purchase a more ethical brand.

Now, let's say John has a bit more tolerance for his own hypocrisy. Or, to use a less-loaded word, "compartmentalization." For a while, John recognizes that factory farmed eggs are bad but keeps buying them anyway. Without the need to immediately resolve this apparent conflict, John's *modus tollens* turns into something *almost* like *modus pollens*:

In less pretentious academic terms, 'Compartmentalization is ok' John's reasoning goes like this:

1. It still might be wrong to buy factory farmed eggs, even if I do keep buying them ('not Q' does not imply 'not P').
2. My identity is wrapped up in egg consumption, so I will keep buying eggs (not Q).
3. Ok, factory farmed eggs are still bad. (P)
4. If I accept (2) and (3), what should I do about it? Maybe donate to THL and try to find a more ethical brand when I can.

and Anti-hypocrisy John's reasoning goes like this:

1. If it is wrong to buy factory farmed eggs, I won't buy them (if P then Q).
2. My identity is wrapped up in egg consumption, so I will keep buying eggs (not Q).
3. Therefore it can't be wrong to buy factory farmed eggs (not Q, therefore not P).

What's going on here?

Strictly speaking, Anti-hypocrisy John could logically and coherently donate money or do something similar just like 'Compartmentalization is ok' alter ego. But, in the real world, my claim is that an aversion to hypocrisy/inconsistency often leads to a hasty rejection (likely not after conscious deliberation) of whichever of the two conflicting actions or beliefs is easier for one to reject.

For John, that means forgetting about or ignoring the ethics of egg consumption before he even has time to ponder whether there might be a decent-but-imperfect way of sorta reconciling his conflicting beliefs and actions

Two wrongs don't make a right

A hyper-simplified illustration:

- Jane believes bad thing 1 and bad thing 2, which are perfectly consistent.
- Tim believes bad thing 1 and good thing 2, which are contradictory and render him a hypocrite.

Which person would you rather be? Well, I'd rather be Tim. Intellectual consistency is not the highest value, and I'd rather be half right than entirely, consistently wrong. If Tim decides that hypocrisy must be avoided at all costs, he has two choices:

1. Believe bad thing 1 and bad thing 2
2. Believe good thing 1 and good thing 2.

As an empirical matter, which option Time goes with probably depends on whether he is more personally invested in question 1 or question 2. But Tim *doesn't know which beliefs and actions are 'good.'* No one says to themselves "sucks that I believe things that are immoral and false, but at least I'm not a hypocrite."

Instead, an aversion to hypocrisy serves as a potent motivation for coming to the conclusion that one's preferred action or belief is in fact true or good. Sometimes this will happen to be correct, but often it will not. Permitting hypocrisy gives us some breathing room to make the decision.

Conclusion

I'm not making the claim that we should ignore or unequivocally embrace hypocrisy. However, tolerance for inconsistency can better allow people to gradually change their behaviors and beliefs without facing the near-impossible task of wholesale behavioral or ideological reform.

Ultimately, I think that tolerating hypocrisy is generally wise when the “worse” of two conflicting beliefs is more closely held or linked with a person’s identity. Our tangled knot of beliefs is only tangled insofar as hypocrisy must be avoided, and sometimes taking a knife to the rope is the only way to improve the knot.

Death by Red Tape

Contains spoilers for the worldbuilding of Vernor Vinge's "Zones of Thought" universe.

Based on [Eliezer's vision of the present from 1900](#).

In the *Zones of Thought* universe, there is a cycle of civilization: civilizations rise from stone-age technology, gradually accumulating more technology, until they reach the peak of technological possibility. At that point, the only way they can improve society is by over-optimizing it for typical cases, removing [slack](#). Once society has removed its slack, it's just a matter of time until unforeseen events force the system slightly outside of its safe parameters. This sets off a chain reaction: like dominoes falling, the failure of one subsystem causes the failure of another and another. This catastrophe either kills everyone on the planet, or sets things so far back that society has to start from scratch.

Vernor Vinge was writing before Nassim Taleb, but if not for that, this could well be interpreted as a reference to Taleb's ideas. Taleb mocks the big players on the stock market for betting on the typical case, and taking huge losses when "black swans" (unexpected/unanticipatable events) occur. (Taleb makes money on the stock market by taking the opposite side of these bets, betting on the unknown unknowns.)

Taleb ridicules Bayesians for their tendency to rely on oversimplified models which assume the future will look like the past. Instead he favors [Popperian](#) epistemology and [ergodicity economics](#).

Indeed, naive Bayesians *do* run the risk of over-optimizing, eliminating slack, and overly assuming that the future will be like the past. On a whole-society level, it makes sense that this kind of thinking could eventually lead to catastrophe (and plausibly already has, in the form of the 2008 housing crash).

However, human nature and modern experience leads me to think that the *opposite* failure mode might be more common.

Taleb advises us to design "antifragile" systems which, like him, *bet on* the atypical and get stronger through failure. This means designing systems with lots of slack, modularity, redundancy, and multiple [layers](#) (think of a laptop, which has a hard chassis to protect and support the vital electronics, & then often has a moderately hard plastic [protective case](#), and then is transported in a [soft outer case](#) of some kind). It means responding to black swans by building new systems which mitigate, or (even better) take advantage of, the new phenomena.

But when I look around at society (at least, through my Bayesian-biased lens) I see it doing *too much of that*.

- The over-cautious FDA seemingly kills a lot more people *on average* (compared to a less-cautious alternative) in the name of avoiding risks of severe unanticipated drug side-effects. And people are largely *comforted* by this. A typical healthy individual would prefer (at least in the short term) to be *very sure* that the few drugs they need are safe, as opposed to having a wider selection of drugs.

- In response to the 9/11 attacks, the government spend huge amounts of money on the TSA and other forms of security. It's possible that this has been a huge waste of money. (The TSA spends 5.3 billion on airline security annually. It's difficult to put a price on 9/11, but quick googling says that total insurance payouts were \$40 billion. So very roughly, the utilitarian question is whether the TSA stops a 9/11-scale attack every 8 years.) On the other hand, many people are probably glad for the TSA even if the utilitarian calculation doesn't work out.
- Requiring a license or more education may be an attempt to avoid the more extreme negative outcomes; for example, I don't know the political motivations which led to requiring licenses for taxi drivers or hairdressers, but I imagine vivid horror stories were required to get people sufficiently motivated.
- Regulation has a tendency to respond to extreme events like this, attempting to make those outcomes impossible while ignoring how much value is being sacrificed in typical outcomes. Since people don't really think in numbers, the actual frequency of extreme events is probably not considered very heavily.

Keep in mind that it's perfectly consistent for there to be lots of examples of *both* kinds of failure (lots of naive utilitarianism which ignores unknown unknowns, and lots of overly risk-averse non-utilitarianism). A society *might* even die from a combination of both problems at once. I'm not really claiming that I'd adjust society's bias in a specific direction; I'd rather have an improved ability to avoid both failure modes.

But just as Vernor Vinge painted a picture of slow death by over-optimization and lack of slack, we can imagine a society choking itself to death with too many risk-averse regulations. It's harder to reduce the number of laws and regulations than it is to increase them. Extreme events, or fear of extreme events, create political will for more precautions. This creates a [system which punishes action](#).

One way I sometimes think of civilization is as a system of guardrails. No one puts up a guardrail if no one has gotten hurt. But if someone *has* gotten hurt, society is quick to set up rails (in the form of social norms, or laws, or physical guardrails, or other such things). So you can imagine the physical and social/legal/economic landscape slowly being tiled with guardrails which keep everything within safe regions.

This, of course, has many positive effects. But the landscape can also become overly choked with railing (and society's budget can be strained by the cost of rail maintenance).

The Homunculus Problem

(This is not (quite) just a re-hashing of the homunculus fallacy.)

I'm contemplating what it would mean for [machine learning models such as GPT-3 to be honest with us](#). Honesty involves conveying your subjective experience... but *what does it mean* for a machine learning model to accurately convey its subjective experience to us?

You've probably seen an optical illusion like this:



The [checker shadow illusion](#). Although square A appears a darker shade of gray than square B, in the image the two have exactly the same [luminance](#). Source: [Wikipedia](#)

You've probably also heard an explanation something like this:

"We don't see the actual colors of objects. Instead, the brain adjusts colors for us, based on surrounding lighting cues, to approximate the surface pigmentation. In this example, it leads us astray, because *what we are actually looking at* is a false image made up of surface pigmentation (or illumination, if you're looking at this on a screen)."

This explanation definitely captures *something* about what's going on, but there are several subtle problems with it:

1. It's a [homunculus fallacy](#)! It explains what we're seeing by imagining that there is a little person inside our heads, who sees (as if projected on a screen) an adjusted version of the image. The brain adjusts the brightness to remove shadows, and adjusts colors to remove effects of colored light. The little person therefore can't tell that patch A is actually the same color as patch B.
2. **Even if there was a little person**, the argument does not describe my subjective experience, because **I can still see the shadow!** I experience the shadowed area as *darker* than the unshadowed area. So the homunculus story doesn't actually fit what I see at all!
3. I can occasionally and briefly get my brain to recognize A and B as the same shade. (It's very difficult, and it quickly snaps back.)

My point is, even when cognitive psychologists are trained to avoid the homunculus fallacy, they go and make it again, because they *don't have a better alternative*.

One thing the homunculus story gets *right* which seems *difficult* to get right is that when you show me the visual illusion, and explain it to me, I can *believe you*, even if *my brain is still seeing the illusion*. I'm using the language "my brain" vs "me" precisely because the homunculus fallacy is a pretty decent model here: *I* know that the patches are the same shade, but *my brain* insists they're different. It *is* as if I'm a little person watching what my brain is putting on a projector: I can believe or disbelieve it.

For example, a simple Bayesian picture of what the brain is doing would involve a probabilistic "world model". The "world model" updates on visual data and reaches conclusions. Nowhere in this picture is there any room for the kind of optical illusion

we routinely experience: either the Bayesian would be fooled (there is no awareness that it's being fooled) or not (there is no perception of the illusion; the patches look the same shade). Or a probabilistic mixture of the two. ("I'm not sure whether the patch is the same color.")

Actually, this downplays the problem, because it's not even clear what it means to ask a Bayesian model about its subjective experience.

When I've seen the homunculus fallacy discussed, I've always seen it maligned as this bad mistake. I don't recall ever seeing it posed as a *problem*: we don't just want to discard it, we want to *replace it with a better way of reasoning*. I want to have a handy term pointing at this problem. I haven't thought of anything better than ***the homunculus problem*** yet.

The Homunculus Problem: The homunculus fallacy is a terrible picture of the brain (or machine learning models), yet, any talk of subjective experience (including phenomena such as visual illusions) falls into the fallacious pattern of "experience" vs "the experiencer". ("My brain shows me A being darker than B...")

The homunculus problem is a superset of the homunculus fallacy, in the following sense: if something *falls prey to the homunculus fallacy*, it involves a line of reasoning which (explicitly or implicitly) relies on a smaller part which is actually a whole agent in itself. If something *falls prey to the homunculus problem*, it could either be that, or it could be a fully reductive model which (may explain some things, but) fails to have a place for our subjective experience. (For example, a Bayesian model which lacks introspection.)

This is **not** the hard problem of consciousness, because I'm not interested in the raw question of how conscious experience arises from matter. That is: if by "consciousness" we mean the metaphysical thing which no configuration of physical matter can force into existence, I'm not talking about that. I'm talking about the neuroscientist's consciousness (what philosophers might call "correlates of consciousness").

It's just that, even when we *think of "experience" as a physical thing which happens in brains*, we end up running into the homunculus fallacy when trying to explain some concepts.

I'm tempted to say that this is *like* the hard problem of consciousness, in that the main thing to avoid is mistaking it for an easier problem. (IE, it's called the "hard" problem of consciousness to make sure it's not confused with easier problems of consciousness.) I don't think you get to claim you've solved the homunculus problem just because you have [some machine-learning model of introspection](#). You need to provide a *way of talking about these things* which serves well in a lot of examples.

This is related to [embedded agency](#), because part of the problem is that many of our formal models (bayesian decision theory, etc) don't include introspection, so you get this picture where "world models" are only about the external world, so agents are incapable of reflecting on their "internal experience" in any way.

This feels related to [Kaj Sotala's discussion of no-self](#). What does it mean to form an accurate picture of what it means to form an accurate picture of yourself?

Added: Nontrivial Implication

A common claim among scientifically-minded people is that "you never actually observe anything directly, except raw sensory data". For example, if you see a glass fall from a counter and shattering, you're actually seeing photons hitting your eye, and *inferring* the existence of the glass, its fall, and its shattering.

I think an easy mistake to make, here, is to implicitly think as if there's some specific boundary where sensory impressions are "observed" (eg, the retina, or v1). This *in effect* posits a homunculus after this point.

In fact, there is no such firm boundary. It's easy to argue that we don't really observe the light hitting the retina, because it is conveyed imperfectly (and much-compressed) to the optic nerve, and thereby to v1. But by the time the data is in v1, it's already a bit abstracted and processed. There's no theater of consciousness with access to the raw data.

Furthermore, if someone wanted to claim that the information in v1 is what's really truly "observed", we could make a similar case about how information in v1 is conveyed to the rest of the brain. (This is like the step where we point out that the homunculus would need another homunculus inside of it.) Every level of signal processing is imperfect, intelligently compressed, and could be seen as "doing some interpretation"!

I would argue that this kind of thinking is making a mistake by over-valuing low-level physical reality. Yes, low-level physical reality is what everything is implemented on top of. But when dealing with high-level objects, we care about the interactions of those objects. Saying "we don't really observe the glass directly" is a lot like saying "there's not really a glass (because it's all just atoms)". I claim there really is a glass, and we really observe it.

So, I would argue that we "really do experience" the glass falling and breaking, in the most sensible sense of direct experience.

However, if you buy this argument, then you also have to buy a couple of surprising conclusions:

- Direct experience of something does not grant you complete knowledge of that thing. Yes, I claim that we can directly experience external reality; but this does not imply that we perceive every crack in the glass in perfect detail, or whatever.
- Furthermore, we can be wrong about our direct experience! The image of the glass falling could

Life and expanding steerable consequences

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

Financial status: This is independent research. I welcome [financial support](#) to make further posts like this possible.

Epistemic status: I believe this is a helpful lens through which to view the significance of AI in a way that is not fundamentally about intelligence.

In this world, there are two types of objects: objects whose steerable consequences diminish over time, and objects whose steerable consequences expand over time.

Consider a small rock on a table. Suppose I move that rock a little to the left, and consider the ways that this action might affect the future. The rock might have been holding down some papers, and those papers might now be blown about by a gust of wind. Or someone might walk into the room and, seeing the rock being out of place, walk over and move it back. In fact the rock exerts a gravitational effect on every other object in the universe, and the tiny movement of the rock will have consequences that ripple out for the life of the universe.

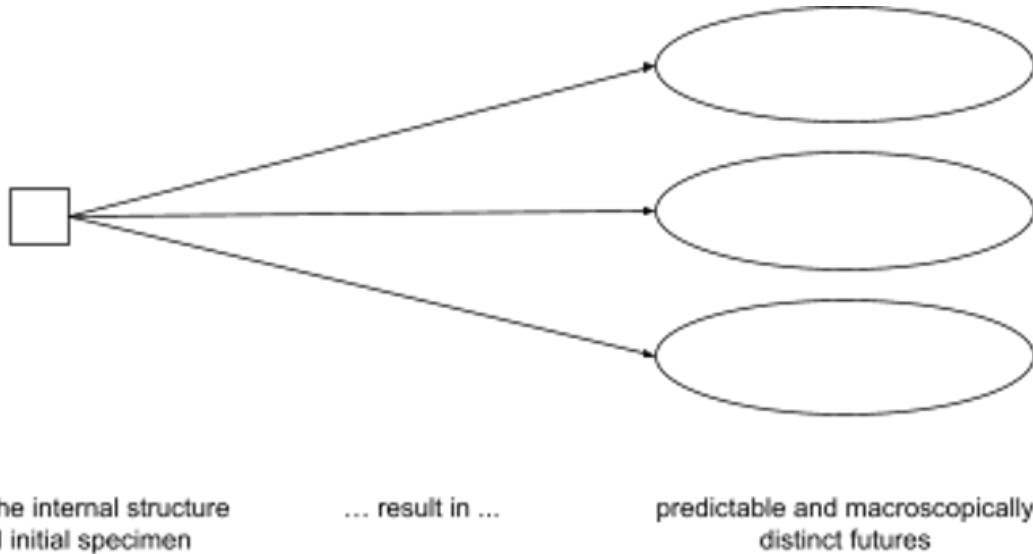
But although these consequences are real, the rock cannot be used by us to produce a predictable large-scale effect on the world very far into the future — say, on the timescale of decades. The consequences of moving the rock become too unpredictable for us to reason about. Even if we are allowed to move the rock to any point in the universe, we cannot really use this power to effect any useful control over the future, at least not without involvement from humans. As we consider the causal fallout of moving the rock we quickly hit a wall of foggy uncertainty, and so in this sense rock cannot be used on its own to steer the future.

But consider now the action of introducing a living organism to the surface of Mars. Suppose that some scientists have chosen or engineered a particular kind of mold that will thrive in the environmental conditions present on Mars. Suppose that we move an object of the same size as the rock, only now the object is a mold specimen together with an initial food source, and we move it from some laboratory on Earth to the surface of Mars. Although the physical size of this initial specimen might be quite small, this action could have consequences that eventually affect the entire surface of Mars.

Furthermore, some of these consequences are quite predictable. We can predict that the mold will reproduce. We can predict that the specimen will spread outwards from its initial location. We can predict that after a few decades we might find copies of the mold all over the surface of Mars. Other consequences are fundamentally unpredictable, yet it is clear that there are some predictable large-scale consequences.

Suppose now that we genetically engineer the specimen to grow under some conditions and not others. By picking these conditions precisely, we might cause the mold to spread to only the northern hemisphere of Mars, or to grow only at low altitudes, or only at high altitudes. In each case, the only thing we are transporting to Mars is a single specimen the size of a small rock. We are not ourselves spreading the

mold over a mountain range or over the low-altitude parts of the planet, but by tweaking the configuration of atoms within this initial specimen we can choose how and where the mold will spread. In this sense the mold has expanding steerable consequences because a physically small specimen can be altered in a way that predictably steers large-scale effects over a long time horizon.



Another object that has this expanding steerability property is the human being. Transport a small colony of humans together with appropriate resources and an initial life support system to the other side of the universe, and over a few thousands or tens of thousands of years an entire space-faring civilization might spring up, perhaps rearranging the matter and energy in that part of the cosmos at a macroscopic scale.

Which kind of objects have this property of expanding steerability? As of May 2021, there are no non-biological objects on Earth that have this property, without ongoing input from humans. For example, suppose I transported a robot to the surface of Mars. This has been done several times, and it has not had the kind of expanding steerable consequences that transporting a mold specimen to the surface of Mars might have^[1]. Furthermore we have not yet built robots that could, without any external help from humans, be used to steer the future, even to the limited extent that a mold specimen might be used to steer the future.

If all biological life on Earth disappeared tomorrow, but all machines built by humans continued operating, the entire ecosystem of machines would quite quickly wind down. Much of the software that runs services on the internet relies on near-constant human oversight, and would cease operating in the absence of humans. But even the most robust pieces of software would cease operating when the power grid decayed to the point of inoperability. And even the most robust machines that humans have ever built, such as some satellites and perhaps some computers located underground with nuclear power sources, will not have the kind of expanding consequences in this neighborhood of the universe that biological life could have.

So in this regard, all the machines that humans have ever built are more like the rock on the table than they are like the mold specimen. Whereas life is winding up, the machines we have built thus far are winding down.

But this may be about to change. Humans appear poised to create machines that could have expanding steerable consequences, independent of biological life. If we succeed at building truly intelligent machines, we might create machines that can collect resources, maintain and upgrade themselves, expand or reproduce themselves, grow their own impact from small to large, and reshape significant patches of the universe. The precise initial configuration of such machines may determine much of what changes they make to their patches of the universe.

All biological life appears to have originated from a single seed organism approximately four billion years ago. This seed organism was almost certainly very small, but its unfolding consequences thus far have been as vast as the Earth, and may yet continue to unfold beyond the Earth. Now, four billion years later, we are about to set in motion a second seed.

1. Yes, the Mars rovers have had large consequences via the information they have beamed back to Earth, but these consequences have flowed via humans, which are a form of biological life that very much *does* have the expanding steerability property ↵

Open and Welcome Thread - May 2021

If it's worth saying, but not worth its own post, here's a place to put it.

If you are new to LessWrong, here's the place to introduce yourself. Personal stories, anecdotes, or just general comments on how you found us and what you hope to get from the site and community are invited. This is also the place to discuss feature requests and other ideas you have for the site, if you don't want to write a full top-level post.

If you want to explore the community more, I recommend [reading the Library](#), [checking recent Curated posts](#), [seeing if there are any meetups in your area](#), and checking out the [Getting Started](#) section of the [LessWrong FAQ](#). If you want to orient to the content on the site, you can also check out the new [Concepts section](#).

The Open Thread tag is [here](#). The Open Thread sequence is [here](#).