

Best of LessWrong: April 2020

1. [Transportation as a Constraint](#)
2. [Choosing the Zero Point](#)
3. [Conflict vs. mistake in non-zero-sum games](#)
4. [Discontinuous progress in history: an update](#)
5. [Negative Feedback and Simulacra](#)
6. [Seemingly Popular Covid-19 Model is Obvious Nonsense](#)
7. [How Long Can People Usefully Work?](#)
8. [An Orthodox Case Against Utility Functions](#)
9. [How to evaluate \(50%\) predictions](#)
10. [Taking Initial Viral Load Seriously](#)
11. [Atari early](#)
12. [Problem relaxation as a tactic](#)
13. [Review of "Lifecycle Investing"](#)
14. [How special are human brains among animal brains?](#)
15. [Inner alignment in the brain](#)
16. [The Best Virtual Worlds for "Hanging Out"](#)
17. [AI Alignment Podcast: An Overview of Technical AI Alignment in 2018 and 2019 with Buck Shlegeris and Rohin Shah](#)
18. [The Inefficient Market Hypothesis](#)
19. [April Fools: Announcing LessWrong 3.0 - Now in VR!](#)
20. [Ethernet Is Worth It For Video Calls](#)
21. [On "COVID-19 Superspread Event Patterns and Lessons"](#)
22. [On R0](#)
23. [Coronavirus: Justified Key Insights Thread](#)
24. [The Unilateralist's "Curse" Is Mostly Good](#)
25. [College advice for people who are exactly like me](#)
26. [What are your favorite examples of distillation?](#)
27. [Why anything that can be for-profit, should be](#)
28. [Peter's COVID Consolidated Brief for 2 April](#)
29. [Treatments correlated with harm](#)
30. [Peter's COVID Consolidated Brief - 29 Apr](#)
31. [How strong is the evidence for hydroxychloroquine?](#)
32. [Hammer and Mask - Wide spread use of reusable particle filtering masks as a SARS-CoV-2 eradication strategy](#)
33. [What Surprised Me About Entrepreneurship](#)
34. [Deminatalist Total Utilitarianism](#)
35. [Law school taught me nothing](#)
36. [Ways you can get sick without human contact](#)
37. [Research on repurposing filter products for masks?](#)
38. [Where should LessWrong go on COVID?](#)
39. [Market-shaping approaches to accelerate COVID-19 response: a role for option-based guarantees?](#)
40. [The Chilling Effect of Confiscation](#)
41. [Would 2009 H1N1 \(Swine Flu\) ring the alarm bell?](#)
42. [Self-Experiment: Does Working More Hours Increase My Output?](#)
43. [The One Mistake Rule](#)
44. [In my culture: the responsibilities of open source maintainers](#)
45. [My stumble on COVID-19](#)
46. [What We Owe to Ourselves](#)
47. [My experience with the "rationalist uncanny valley"](#)
48. [Curiosity: A Greedy Feeling](#)
49. [COVID-19: List of ideas to reduce the direct harm from the virus, with an emphasis on unusual ideas](#)
50. [DeepMind team on specification gaming](#)

Best of LessWrong: April 2020

1. [Transportation as a Constraint](#)
2. [Choosing the Zero Point](#)
3. [Conflict vs. mistake in non-zero-sum games](#)
4. [Discontinuous progress in history: an update](#)
5. [Negative Feedback and Simulacra](#)
6. [Seemingly Popular Covid-19 Model is Obvious Nonsense](#)
7. [How Long Can People Usefully Work?](#)
8. [An Orthodox Case Against Utility Functions](#)
9. [How to evaluate \(50%\) predictions](#)
10. [Taking Initial Viral Load Seriously](#)
11. [Atari early](#)
12. [Problem relaxation as a tactic](#)
13. [Review of "Lifecycle Investing"](#)
14. [How special are human brains among animal brains?](#)
15. [Inner alignment in the brain](#)
16. [The Best Virtual Worlds for "Hanging Out"](#)
17. [AI Alignment Podcast: An Overview of Technical AI Alignment in 2018 and 2019 with Buck Shlegeris and Rohin Shah](#)
18. [The Inefficient Market Hypothesis](#)
19. [April Fools: Announcing LessWrong 3.0 – Now in VR!](#)
20. [Ethernet Is Worth It For Video Calls](#)
21. [On “COVID-19 Superspread Events in 28 Countries: Critical Patterns and Lessons”](#)
22. [On R0](#)
23. [Coronavirus: Justified Key Insights Thread](#)
24. [The Unilateralist’s “Curse” Is Mostly Good](#)
25. [College advice for people who are exactly like me](#)
26. [What are your favorite examples of distillation?](#)
27. [Why anything that can be for-profit, should be](#)
28. [Peter's COVID Consolidated Brief for 2 April](#)
29. [Treatments correlated with harm](#)
30. [Peter's COVID Consolidated Brief - 29 Apr](#)
31. [How strong is the evidence for hydroxychloroquine?](#)
32. [Hammer and Mask - Wide spread use of reusable particle filtering masks as a SARS-CoV-2 eradication strategy](#)
33. [What Surprised Me About Entrepreneurship](#)
34. [Deminatalist Total Utilitarianism](#)
35. [Law school taught me nothing](#)
36. [Ways you can get sick without human contact](#)
37. [Research on repurposing filter products for masks?](#)
38. [Where should LessWrong go on COVID?](#)
39. [Market-shaping approaches to accelerate COVID-19 response: a role for option-based guarantees?](#)
40. [The Chilling Effect of Confiscation](#)
41. [Would 2009 H1N1 \(Swine Flu\) ring the alarm bell?](#)
42. [Self-Experiment: Does Working More Hours Increase My Output?](#)
43. [The One Mistake Rule](#)
44. [In my culture: the responsibilities of open source maintainers](#)
45. [My stumble on COVID-19](#)

46. [What We Owe to Ourselves](#)
47. [My experience with the "rationalist uncanny valley"](#)
48. [Curiosity: A Greedy Feeling](#)
49. [COVID-19: List of ideas to reduce the direct harm from the virus, with an emphasis on unusual ideas](#)
50. [DeepMind team on specification gaming](#)

Transportation as a Constraint

Imagine it's late autumn of 332 BC. You're Alexander the Great, and your armies are marching toward Egypt from Gaza. There's just one little problem: you need to cross the Sinai peninsula - 150 miles of hot, barren desert. How will you carry food and water for the troops?



Green triangle on the left is the Nile river delta in Egypt; green chunk in the upper right is Israel. The big desert peninsula between them is the Sinai.

Option 1: carry it

A physically-active human needs about 3 lbs of food per day. (Modern hikers can probably find lighter calorie-dense foodstuffs, but we're talking ancient history here.) Water requirements vary; 5 lbs is a minimum, but the US Army Quartermaster Corps recommends 20 lbs/day when marching through a hot desert. Alexander's army crossed the desert in 7

days. Food might be reasonable, but to carry the water would mean $7*20 = 140$ lbs per person, plus 50+ lbs of armor, weapons, etc.

When I go hiking, I aim for a 20-30 lb pack. US marines are [apparently expected](#) to be able to carry 150 lbs for 9 miles - quite a bit less than the 20+ miles/day Alexander's army managed, and with no comment on how long the marine in question might need to rest afterwards. (Also, I'm not sure I trust that source - 150 lbs for 9 miles sounds unrealistic to me, and if it's true then I'm very impressed by marines.)

Suffice to say that carrying that much water across that much desert is not a realistic option, even if we drink it along the way.

Option 2: horses

A horse consumes 20 lbs of food (half of which may be forage) and 80 lbs of water per day. In exchange, it can carry about 200 lbs (surprisingly, my source claims that horses can carry more than they can pull). Of course, that 200 lbs has to include the horse's own food and water, plus whatever useful load it's carrying. So, marching through a desert, a horse can only transport $(200 \text{ lbs}) / (80 + 20 \text{ lbs/day}) = 2$ days of supplies *for itself*, and that's before whatever useful things actually need to be transported.

In other words, there's a hard upper limit on how far goods can be transported by horse without refilling supplies along the way. That limit is around 2 days travel time without any refill, 10 days if there's plenty of fresh water along the route, or 20 days if there's both water and forage. At 20 miles/day, that's 40, 200, or 400 miles. Realistically, if we want the number of horses to be reasonable, the limit is more like half that much - 20 miles, 100 miles, or 200 miles, respectively.

So horses also won't work.

Option 2.5: camels or other pack animals

Contrary to popular image, camels actually need more water than horses. They can go a couple days without, but then need to fill up all at once. They can also carry a bit more weight, but they eat more food. At the end of the day, the numbers end up quite similar.

Mules also end up with similar numbers, and cattle are generally worse.

Option 3: ships

Assuming the army marches along the coast, a supply fleet can sail alongside. At the time, a single large merchant ship could carry 400 tons - in other words, as much as about 4000 horses. Presumably the ship would cost a lot less than the horses, too.

Well then, there's our answer. Ships are clearly a vastly superior way to move goods. Range is a non-issue, capacity is far larger, and they're far cheaper. They're perfect for crossing the Sinai, which runs right along the coast anyway.

Fast forward a few years to 327 BC, and Alexander is marching his armies back from India. He plans to cross the Gedrosian desert, along the coast of modern-day Pakistan and Iran. The plan is much like the Sinai: a supply fleet will sail alongside the army. Unfortunately, neither Alexander nor his commanders know about the monsoons: across most of south Asia, the wind blows consistently southwest for half the year, and consistently northeast for the other half. There is nothing like it in the Mediterranean. And so, Alexander marches out expecting the fleet to catch up as soon as the winds turn - not realizing that the winds will not turn for months. Three quarters of his soldiers die in the desert.

Thus end the campaigns of Alexander.

Generalization

The above numbers are drawn from Donald Engels' book [Alexander the Great and the Logistics of Macedonian Army](#). But it tells us a lot more about the world than just the logistics of one particular ancient army.

First, this highlights the importance of naval dominance in premodern warfare. A fleet was a far superior supply train, capable of moving a high volume of food and water over long distance at relatively low cost. Without a fleet, transport of food became expensive at best, regular resupply became a strategic necessity, and long routes through arid terrain became altogether impassable. Destroying an enemy's fleet meant starving the army. Likewise, controlling ports wasn't just for show - without a port, feeding the army became a serious problem.

Another interesting insight into premodern warfare: away from friendly seas and rivers, the only way to keep an army fed was to either seize grain from enemies, or buy it from allies, either of whom needed to already be nearby. In Alexander's case, deals were often struck to establish supply caches along the army's intended route.

An interesting exercise: to what extent was transportation a binding constraint on the size of premodern towns/cities? (One number you may want: [Braudel](#) (pg 121) estimates that 5000 square meters of land growing wheat would provide one person-year of food, not accounting for crop rotation.) Leave a comment if you try a calculation here; I'm curious to see how other peoples' models compare to my own.

Modern Day

Today we have trains and trucks and roads, so the transportation constraint has relaxed somewhat. But here's an interesting comparison: a modern 18-wheeler in the US is legally limited to haul 40 tons, while a panamax ship could carry about 50k tons through the canal (prior to the opening of the new locks in 2016). That's a ratio of a bit over 1000 - surprisingly similar to the ship/horse ratio of antiquity, especially considering the much larger new-panamax and super-panamax ships also in use today.



Can we get a quick-and-dirty feel for tautness of the transportation constraint today? Here are a few very different angles:

- [This USDA study](#) shows rates on produce transport, typically about 7-20 cents per pound (see figure 6). The Smart & Final grocery store near me sells the cheaper produce items looked at in that study (bell peppers, cantaloupes, tomatoes, oranges) for 70-100 cents per pound, so transport alone is roughly 10-20% of the cost-to-consumer.
- What about transporting humans? [Average commute in the US](#) is ~30 minutes each way; driving is usually in the 20-30 minute range, while public transit is usually 30-50. Assuming 8 hr workdays, that means commutes are typically ~10-20% of our work-hours.
- The bureau of transportation [estimates](#) transport at 5.6% of the US economy for a very narrow measure, or 8.9% with a broader measure (though this still excludes non-market transport costs like e.g. commute time).

My interpretation: the transportation constraint becomes taut when it accounts for 10-20% of cost. If it's less than that, it usually doesn't limit production - we see plenty of goods which aren't transportation-dependent or which are higher-value-per-weight, and the transportation constraint is generally slack for those. But once transportation hits about 10-20%, people start looking for alternatives, i.e. producing the goods somewhere else or using alternative goods. Obviously this is not based on very much data, but I find it intuitively plausible.

Compared to ancient times, transportation constraints have obviously relaxed quite a lot. Yet qualitatively, the world today still does not look like a world of fully slack transportation constraints. To wrap up, let's discuss what that would look like.

Extreme Slackness

In [Material Goods as an Abundant Resource](#), we discussed the world of the duplicator - a device capable of copying any item placed on it. In such a world, material scarcity is removed as an economic constraint - all material constraints are completely slack.

What would be a corresponding sci-fi device for transportation constraints, and what would that world look like?

I suggest portals: imagine we can create pairs of devices capable of transporting things from one device to the other, across any distance, at the speed of light. (We could instead imagine teleporters, removing the need for a pre-installed device at either end, but then the entire discussion would be about security.) What does the world of the portal look like?

First, there's complete geographical decoupling of production from consumption. People have no need to live near where they work; companies can put offices and factories wherever real estate is cheap. We can enjoy miles of wilderness on the back porch and a downtown district on the front porch; a swimming pool can open right into the ocean. Buying direct from the farm or factory is standard for most material goods.

What are now tourist destinations would become options for an evening activity. Disneyworld would sell a park-hopper ticket that includes Disneyland California, Paris, and Shanghai, but the price of that ticket would be high enough to prevent the parks from becoming unpleasantly crowded - probably quite a bit more expensive than today, though possibly cheaper than today's flights to Orlando.

Obviously roads would cease to exist. Huge amounts of land would revert from asphalt to wilderness, but buildings would also be much more spread out. Buildings would be built close together more for show than for function - e.g. to provide the ambiance of a downtown or a community to those who want it. Physical life, in general, would look more like the structure of the internet rather than the structure of geography; "cities" would be clusters very spread out in space but very tightly connected via the portal network. Filter bubbles would be a much more physically tangible phenomenon.

Geographically-defined governments would likely be replaced by some other form of government - governments based around access to portal hubs/networks are one natural possibility. Security would be a priority, early on - carrying an unauthorized portal into an area would earn a facefull of high explosives. On the other hand, it would be hard to prevent a high degree of mobility between areas controlled by different governments; the implications for government behavior are conceptually similar to [seasteading](#).

The structure of space near portal networks would be different in a big-O sense; the amount of space at a distance of about r would increase exponentially, rather than like r^2 . A nuclear warhead could go off five hundred feet away and you'd feel a breeze through a fast-branching portal network. On the other hand, viruses could spread much more rapidly.

Anyway, at this point we're getting into specifics of portals, so I'll cut off the speculation. The point is: if transportation continues to get cheaper and more efficient over time, then we will converge to the world of the portal, or at least something like it. The details do matter - portals are different from teleportation or whatever might actually happen - but any method of fully relaxing transportation constraints will have qualitatively similar results, to a large extent.

Choosing the Zero Point

Summary: You can decide what state of affairs counts as neutral, and what counts as positive or negative. Bad things happen if humans do that in our natural way. It's more motivating and less stressful if, when we learn something new, we update the neutral point to [what we think the world really is like now].

A few years back, I read [an essay by Rob Bensinger](#) about vegetarianism/veganism, and it convinced me to at least eat much less meat. **This post is not about that topic.** It's about the way that essay differed, psychologically, from many others I've seen on the same topic, and the general importance of that difference.

Rob's essay referred to the same arguments I'd previously seen, but while other essays concluded with the implication "you're doing great evil by eating meat, and you need to realize what a monster you've been and immediately stop", Rob emphasized the following:

Frame animal welfare activism as an astonishingly promising, efficient, and uncrowded opportunity to do good. Scale back moral condemnation and guilt. LessWrong types can be powerful allies, but the way to get them on board is to give them opportunities to feel like munchkins with rare secret insights, not like latecomers to a not-particularly-fun party who have to play catch-up to avoid getting yelled at. It's fine to frame helping animals as *challenging*, but the challenge should be to excel and do something astonishing, not to meet a bare standard for decency.

That shouldn't have had different effects on me than other essays, but damned if it didn't.

Consider a utilitarian Ursula with a utility function U . U is defined over all possible ways the world could be, and for each of those ways it gives you a number. Ursula's goal is to maximize the expected value of U .

Now consider the utility function V , where V always equals $U + 1$. If a utilitarian Vader with utility function V is facing the same choice (in another universe) as Ursula, then because that $+1$ applies to every option equally, the right choice for Vader is the same as the right choice for Ursula. The constant difference between U and V doesn't matter for any decision whatsoever!

We represent this by saying that a utility function is only defined up to positive affine transformations. (That means you can also multiply U by any positive number and it still won't change a utilitarian's choices.)

But humans aren't perfect utilitarians, in many interesting ways. One of these is that our brains have a natural notion of outcomes that are good and outcomes that are bad, and the neutral zero point is more or less "the world I interact with every day".

So if we're suddenly told about a nearby [bottomless pit of suffering](#), what happens?

Our brains tend to hear, "Instead of the zero point where we thought we were, this claim means that we're really WAY DOWN IN THE NEGATIVE ZONE".

A few common reactions to this:

- *Denial*. "Nope nope that argument can't be true, I'm sure there's a flaw in it, we're definitely still in the normal zone"
- *Guilt*. "AAAAAHHHHH I need to drop everything and work super hard on this, I can't allow myself any distractions or any bit of happiness until this is completely fixed"
- *Despair*. "Oh no, there's no way I could get things back up to normal from here, I can't do anything, I'll just sit here and hate myself"

The thing about Rob's post is that it suggested an alternative. Instead of keeping the previous zero point and defining yourself as now being very far below it, **you can reset yourself to take the new way-the-world-is as the zero point.**

Again, this doesn't change any future choice a *utilitarian you* would make! But it does buy *human you* peace of mind. [What is true is already so](#)- the world was like this even when you didn't believe it.

The psychological benefits of this transformation:

- *Acceptance*. Is it too scary to consider the new hypothesis? No! If you accept it, you'll still start at zero, you'll just have an opportunity to do more kinds of good than you previously thought existed.
- *Relief*. Must you feel ashamed for not working your fingers to the bone? No! If you're pushing the world into the positive zone, it feels much more okay to 80-20 your efforts.
- *Hope*. Must you despair if you can't reach your old zero? No! Seen from here, this was always the world, but now you can help move it up from zero! It doesn't have to go higher than you can reach in order to be worthwhile.

A few last notes:

- I really recommend doing this for oneself first of all, and then extending it to one's efforts of persuasion.
- There are a few cases where a desperate effort is called for, but even then we can frame it as building something great that the world urgently needs.
- When it comes to personal virtue, the true neutral point for yourself shouldn't be "doing everything right", because you're consigning yourself to living in negative-land. A better neutral point is "a random person in my reference class". How are you doing relative to a typical [insert job title or credential or hobby here], in your effects on that community? Are you showing more discipline than the typical commenter on your Internet forum? That's a good starting point, and you can go a long way up from there.
- (Thanks to Isnasene for helping me realize this.) If many bad things are continuing to happen, then the zero point of "how things are right now" will inexorably lead to the world sliding into the deep negative zone. The zero point I've actually been using is "the trajectory the world would be on right now if I were replaced with a random person from my reference class". **That** is something that's within my power to make worse or better (in expectation).

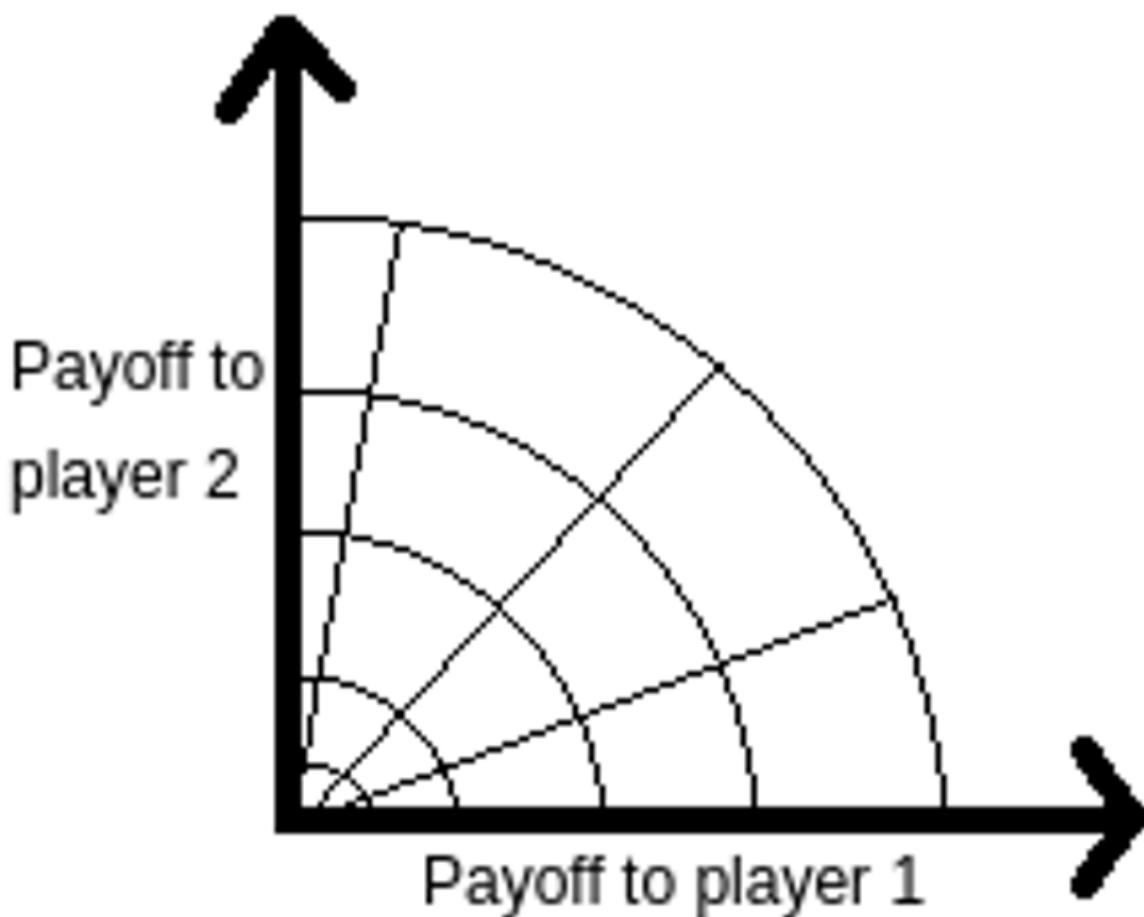
Now go forth, and make the world better than the new zero!

Conflict vs. mistake in non-zero-sum games

Summary: Whether you behave like a mistake theorist or a conflict theorist may depend more on your negotiating position in a non-zero-sum game than on your worldview.

Disclaimer: I don't really know game theory.

Plot the payoffs in a non-zero-sum two-player game, and you'll get a convex[1] set with the Pareto frontier on the top and right:



You can describe this set with two parameters: The *surplus* is how close the outcome is to the Pareto frontier, and the *allocation* tells you how much the outcome favors player 1 versus player 2. In this illustration, the level sets for surplus and allocation are depicted by concentric curves and radial lines, respectively.

It's tempting to decompose the game into two phases: A cooperative phase, where the players coordinate to maximize surplus; and a competitive phase, where the players negotiate how the surplus is allocated.

Of course, in the usual formulation, both phases occur simultaneously. But this suggests a couple of negotiation strategies where you try to make one phase happen before the other:

1. "Let's agree to maximize surplus. Once we agree to that, we can talk about allocation."
2. "Let's agree on an allocation. Once we do that, we can talk about maximizing surplus."

I'm going to provocatively call the first strategy [mistake theory](#), and the second [conflict theory](#).

Indeed, the mistake theory strategy pushes the obviously good plan of making things better off for everyone. It can frame all opposition as making the mistake of leaving surplus on the table.

The conflict theory strategy threatens to destroy surplus in order to get a more favorable allocation. Its narrative emphasizes the fact that the players can't maximize their rewards simultaneously.

Now I don't have a good model of negotiation. But intuitively, it seems that mistake theory is a good strategy if you think you'll be in a *better* negotiating position once you move to the Pareto frontier. And conflict theory is a good strategy if you think you'll be in a *worse* negotiating position at the Pareto frontier.

If you're naturally a mistake theorist, this might make conflict theory seem more appealing. Imagine [negotiating with a paperclip maximizer over the fate of billions of lives](#). Mutual cooperation is Pareto efficient, but unappealing. It's more sensible to threaten defection in order to save a few more human lives, if you can get away with it.

It also makes mistake theory seem unsavory: Apparently mistake theory is about postponing the allocation negotiation until you're in a comfortable negotiating position. (Or, somewhat better: It's about tricking the other players into cooperating before they can extract concessions from you.)

This is kind of unfair to mistake theory, which is supposed to be about educating decision-makers on efficient policies and building institutions to enable cooperation. None of that is present in this model.

But I think it describes something important about mistake theory which is usually rounded off to something like "[\[mistake theorists have\] become part of a class that's more interested in protecting its own privileges than in helping the poor or working for the good of all](#)".

The reason I'm thinking about this is that I want a theory of non-zero-sum games involving counterfactual reasoning and superrationality. It's not clear to me what superrational agents "should" do in general non-zero-sum games.

Discontinuous progress in history: an update

I. The search for discontinuities

We've been looking for historic cases of discontinuously fast technological progress, to help with reasoning about the [likelihood](#) and consequences of abrupt progress in AI capabilities. We recently finished expanding this investigation to 37 technological trends.¹ This blog post is a quick update on our findings. See [the main page on the research](#) and its outgoing links for more details.

We found [ten events](#) in history that abruptly and clearly contributed more to progress on some technological metric than another century would have seen on the previous trend.² Or as we say, we found ten events that produced 'large', 'robust' 'discontinuities'.

How we measure the size of a discontinuity (*by Rick Korzekwa*)

Another five events caused robust discontinuities of between ten and a hundred years ('moderate robust discontinuities'). And 48 more events caused some trend to depart from our best guess linear or exponential extrapolation of its past progress by at least ten years (and often a hundred), but did so in the context of such unclear past trends that this did not seem clearly remarkable.³ I call all of these departures 'discontinuities', and distinguish those that are clearly outside plausible extrapolations of the past trend, according to my judgment, as 'robust discontinuities'.⁴

Much of the data involved in this project seems at least somewhat unreliable, and the [methods](#) involve many judgments, and much ignoring of minor issues. So I would not be surprised if more effort could produce numerous small changes. However I expect the broad outlines to be correct.⁵

II. The discontinuities

Large robust discontinuities

Here is a quick list of the robust 100-year discontinuous events, which I'll describe in more detail beneath:

- The Pyramid of Djoser, 2650BC (discontinuity in [structure height trends](#))
- The SS Great Eastern, 1858 (discontinuity in [ship size trends](#))
- The first telegraph, 1858 (discontinuity in [speed of sending a 140 character message across the Atlantic Ocean](#))
- The second telegraph, 1866 (discontinuity in [speed of sending a 140 character message across the Atlantic Ocean](#))
- The Paris Gun, 1918 (discontinuity in [altitude reached by man-made means](#))
- The first non-stop transatlantic flight, in a modified WWI bomber, 1919 (discontinuity in both [speed of passenger travel across the Atlantic Ocean](#) and [speed of military payload travel across the Atlantic Ocean](#))
- The George Washington Bridge, 1931 (discontinuity in [longest bridge span](#))
- The first nuclear weapons, 1945 (discontinuity in [relative effectiveness of explosives](#))
- The first ICBM, 1958 (discontinuity in [average speed of military payload crossing the Atlantic Ocean](#))
- YBa₂Cu₃O₇ as a superconductor, 1987 (discontinuity in [warmest temperature of superconducting](#))

The Pyramid of Djoser, 2650BC

*Discontinuity in [structure height trends](#)*⁶

The Pyramid of Djoser is [considered to be](#) 'the earliest colossal stone structure' in Egypt. According to Wikipedia's data, it took seven thousand years for the tallest structures to go from five to thirteen meters tall⁷ and then suddenly the Egyptian pyramids shot up to a height of 146.5m over about a hundred years and five successively tallest pyramids.

The Pyramid of Djoser, By Charles J Sharp - Own work, from [Sharp Photography](#),
[sharpphotography](#), CC BY-SA 3.0, [Link](#)

The first of these five is the Pyramid of Djoser, standing 62.5m tall. The second one—[Meidum Pyramid](#)—is also a large discontinuity in structure height trends by our calculation, but I judge it not robust, since it is fairly unclear what the continuation of the trend should be after the first discontinuity. As is common, the more basic thing going on seems to be a change in the growth rate, and the discontinuity of the Pyramid of Djoser is just the start of it.

The Djoser discontinuity: close up on the preceding trend, cut off at the Pyramid of Djoser

A longer history of record structure heights, showing the isolated slew of pyramids

Strangely, after this spurt of progress, humanity built nothing taller than the tallest pyramid for nearly four thousand years—until [Lincoln Cathedral](#) in 1311—and nothing more than twenty percent taller than it until the Eiffel Tower in 1889.

The SS Great Eastern

Discontinuity in [ship size](#), measured in '[builder's old measurement](#)'⁸ or in displacement.

The SS *Great Eastern* was a freakishly large ship. For instance, it seems to have weighed about five times as much as any previous ship. As far as I can tell, the reason it existed is that [Isambard Kingdom Brunell](#) thought it would be good. Brunell was a 19th Century engineering hero, rated #2 greatest Briton of all time in a 2002 [BBC poll](#), who according to [Wikipedia](#), ‘revolutionised public transport and modern engineering’ and built ‘dockyards, the Great Western Railway (GWR), a series of steamships including the first propeller-driven transatlantic steamship, and numerous important bridges and tunnels’.

The SS *Great Eastern* compared to the UK Royal Navy's ships of the line, which were probably not much smaller than the largest ships overall immediately prior to the *Great Eastern*

The experimental giant sailing steamship idea doesn't seem to have gone well. The *Great Eastern* apparently never had its cargo holds filled, and ran at a deficit for years before being sold and used for laying the second telegraph cable (another source of large discontinuity—see below).⁹ It was designed for transporting passengers to the Far East, but there was never the demand.¹⁰ It was [purportedly](#) rumored to be 'cursed', and suffered various ill fortune. On its maiden voyage a boiler exploded, throwing one of the funnels into the air and killing six people.¹¹ Later it hit a rock and got a 9-foot gash, which seems to have been hard to fix because the ship was too big for standard repair methods.¹²

We don't have a whole trend for largest ships, so are using British Royal Navy [ship of the line](#) size trends as a proxy against which to compare the *Great Eastern*.¹³ This gives us discontinuities of around 400 years in both displacement and tonnage ([BOM](#)). [Added May 10: Nuño Sempere [has also investigated](#) the Great Eastern as a discontinuity, and has some nice figures comparing it to passenger and sailing vessel trends.]

The SS Great Eastern

However that is assuming we expect ship size to increase either linearly or exponentially (our usual expectation). But looking at the ship of the line trends, both displacement and cargo capacity (measured in tonnage, [BOM](#)) seemed to grow at something closer to a hyperbolic curve for some reason—apparently accelerating toward an asymptote in the late 1860s. If we had expected progress to continue this way throughout, then neither trend had any discontinuities, instead of eight or eleven of them. And supposing that overall ship size follows the same hyperbola as the military ship trends, then the *Great Eastern*'s discontinuities go from around 400 years to roughly 11 or 13 years. Which doesn't sound big, but since this was about that many years before of the asymptote of the hyperbola at which point arbitrarily large ships were theoretically expected, the discontinuities couldn't have been much bigger.

Our data ended for some reason just around the apparently impending ship size singularity of the late 1860s. But my impression is that not much happened for a while—it [apparently](#) took forty years for a ship larger than the *Great Eastern* to be built, on many measures.

I am unsure what to make of the apparently erroneous and unforced investment in the most absurdly enormous ship happening within a decade or two of the point at which trend extrapolation appears to have suggested arbitrarily large ships. Was Brunell aware of the trend? Did the forces that produced the rest of the trend likewise try to send all the players in the ship-construction economy up the asymptote, where they crashed into some yet unmet constraint? It is at least nice to have more examples of what happens when singularities are reached in the human world.

The first transatlantic telegraph

Discontinuity in [speed of sending a 140 character message across the Atlantic Ocean](#)

Until 1858, the fastest way to get a message from New York to London was by ship, and the fastest ships took over a week¹⁴. Telegraph was [used earlier](#) on land, but running it between continents was quite an undertaking. The effort to lay the a transatlantic cable failed numerous times before it became ongoingly functional.¹⁵ One of those times though, it worked for about a month, and messages were sent.¹⁶ There were celebrations in the streets.

[H.M.S. "Agamemnon" laying the Atlantic Telegraph cable in 1858. A whale crosses the line](#), R. M. Bryson, lith from a drawing by R. Dudley, 1865

[A celebration parade for the first transatlantic telegraph cable, Broadway, New York City.](#)

The telegraph [could send](#) a 98 word message in a mere 16 hours. For a message of more than about 1400 words, it would actually have been faster to send it by ship (supposing you already had it written down). So this was a big discontinuity for short messages, but not necessarily any progress at all for longer ones.

The first transatlantic telegraph cable revolutionized 140 character message speed across the Atlantic Ocean

The second transatlantic telegraph

Discontinuity in [speed of sending a 140 character message across the Atlantic Ocean](#)

After the first working transatlantic telegraph cable (see above) failed in 1858, it was another eight years before the second working cable was finished. Most of that delay was apparently for lack of support.¹⁷ and the final year seems to have been because the cable broke and the end was lost at sea after over a thousand miles had been laid, leaving the ship to return home and a new company to be established before the next try.¹⁸ Whereas it [sounds like](#) it took less than a day to go from the ship carrying the cable arriving in port, and the sending of telegraphs.

The second telegraph discontinuity: close up on the preceding trend, cut off at the second telegraph. Note that the big discontinuity of the first telegraph cable is now almost invisible.

At a glance, on Wikipedia's [telling](#), it sounds as though the perseverance of one person—[Cyrus West Field](#)—might have affected when fast transatlantic communication appeared by years. He seems to have led all five efforts, supplied substantial money himself, and ongoingly fundraised and formed new companies, even amidst a broader lack of enthusiasm after initial failures. (He was also [given a congressional gold medal](#) for establishing the transatlantic telegraph cable, suggesting the US congress also has this impression.) His actions wouldn't have affected how much of a discontinuity either telegraph was by much, but it is interesting if such a large development in a seemingly important area might have been accelerated much by a single person.

The second telegraph cable was laid by the *Great Eastern*, the discontinuously large ship of two sections ago. Is there some reason for these two big discontinuities to be connected? For instance, did one somehow cause the other? That doesn't seem plausible. The main way I can think of that the transatlantic telegraph could have caused the *Great Eastern*'s size would be if the economic benefits of being able to lay cable were anticipated and effectively subsidized the ship. I haven't heard of this being an intended use for the *Great Eastern*. And given that the first transatlantic telegraph was not laid by the *Great Eastern*, it seems unlikely that such a massive ship was strictly needed for the success of a second one at around that time, though the second cable used [was apparently around twice as heavy as the first](#). Another possibility is that some other common factor made large discontinuities more possible. For instance, perhaps it was an unusually feasible time and place for solitary technological dreamers to carry out ambitious and economically adventurous projects.

[Great Eastern again, this time at Heart's Content, Newfoundland, where it carried the end of the second transatlantic telegraph cable in 1866](#)

The first non-stop transatlantic flight

Discontinuity in both [speed of passenger travel across the Atlantic Ocean](#) and [speed of military payload travel across the Atlantic Ocean](#)

Ships were the fastest way to cross the Atlantic Ocean until the end of World War I. Passenger liners had been getting incrementally faster for about eighty years, and the fastest regular passenger liner was given a special title, '[Blue Riband](#)'. Powered heavier-than-air flight got started in 1903, but at first planes only traveled hundreds of feet, and it took time to expand that to the 1600 or so miles needed to cross the Atlantic in one hop.¹⁹

The first non-stop transatlantic flight was made shortly after the end of WWI, in 1919. The Daily Mail [had offered](#) a large cash prize, on hold during the war, and with the resumption of peace, [a slew](#) of competitors prepared to fly. [Alcock and Brown](#) were the first to do it successfully, in a modified bomber plane, taking around 16 hours, for an average speed around four times faster than the Blue Riband.

Alcock and Brown landed in Ireland, 1919

One might expect discontinuities to be especially likely in a metric like 'speed to cross the Atlantic', which involves a sharp threshold on a non-speed axis for inclusion in the speed contest. For instance if planes incrementally improved on speed and range (and cost and comfort) every year, but couldn't usefully cross the ocean at all until their range reached 1600 miles, then decades of incremental speed improvements could all hit the transatlantic speed record at once, when the range reaches that number.

Is this what happened? It looks like it. The Wright Flyer [apparently](#) had a maximum speed of 30mph. That's about the record average ocean liner speed in 1909. So if the Wright Flyer had had the range to cross the Atlantic in 1903 at that speed, it would have been about six years ahead of the ship speed trend and wouldn't have registered as a substantial discontinuity. [20](#) But because it didn't have the range, and because the speed of planes was growing faster than that of ships, in 1919 when planes could at last fly thousands of miles, they were way ahead of ships.

The transatlantic flight discontinuity: close up on the preceding trend, cut off at the first non-stop transatlantic flight.

The George Washington Bridge

Discontinuity in [longest bridge span](#)

A bridge '[span](#)' is the distance between two intermediate supports in a bridge. The history of bridge span length is not very smooth, and so arguably full of discontinuities, but the only bridge span that seems clearly way out of distribution to me is the main span of the [George Washington Bridge](#). (See below.)

The George Washington Bridge discontinuity: close up on the preceding trend, cut off at the George Washington Bridge

I'm not sure what made it so discontinuously long, but it is notably also the world's busiest motor vehicle bridge ([as of 2016](#)), connecting New York City with New Jersey, so one can imagine that it was a very unusually worthwhile expanse of water to cross. Another notable feature of it was that it was much thinner relative to its length than long suspension bridges normally were, and lacked the usual 'trusses', based on a new theory of bridge design.²¹

George Washington Bridge, [via Wikimedia Commons](#), [Photographer: Bob Jagendorf](#)

Nuclear weapons

Discontinuity in [relative effectiveness of explosives](#)

The '[relative effectiveness factor](#)' of an explosive is how much TNT you would need to do the same job.²² Pre-nuclear explosives had traversed the range of relative effectiveness factors from around 0.5 to 2 over about a thousand years, when in 1945 the first nuclear weapons came in at a relative effectiveness of [around 4500²³](#).

The nuclear weapons discontinuity: close up on the preceding trend, cut off at the first nuclear weapons

A few characteristics of nuclear weapons that could relate to their discontinuousness:

- **New physical phenomenon:** nuclear weapons are based on [nuclear fission](#), which was recently discovered, and allowed human use of nuclear energy (which exploits the strong fundamental force) whereas past explosives were based on chemical energy (which exploits the electromagnetic force). New forms of energy are rare in human history, and nuclear energy stored in a mass is characteristically much higher than chemical energy stored in it.
- **Massive investment:** the Manhattan Project, which developed the first nuclear weapons, cost around [\\$23 billion in 2018 dollars](#). This was presumably a sharp increase over previous explosives research spending.
- **Late understanding:** it looks like nuclear weapons were only understood as a possibility after it was well worth trying to develop them at a huge scale.
- **Mechanism involves a threshold:** nuclear weapons are based on nuclear chain reactions, which require a [critical mass](#) of material (how much varies by circumstance).

I discussed whether and how these things might be related to the discontinuity in 2015 [here](#) (see Gwern's comment) and [here](#).

[Preparation for the Trinity Test, the first detonation of a nuclear weapon](#)

[The trinity test explosion after 15 seconds](#)

The Paris Gun

Discontinuity in [altitude reached by man-made means](#)

The [Paris Gun](#) was the largest artillery gun in WWI, used by the Germans to bomb Paris from 75 miles away. It could shoot 25 miles into the air, whereas the previous record we know of was around 1 mile into the air (also shot by a German gun).²⁴

The Paris Gun, able to shell Paris from 75 miles away.

The Paris Gun discontinuity: close up on the preceding trend of highest altitudes reached by man-made means, cut off at the Paris Gun

I don't have much idea why the Paris Gun traveled so much higher than previous weapons. [Wikipedia](#) suggests that its goals were psychological rather than physically effective warfare:

As military weapons, the Paris Guns were not a great success: the payload was small, the barrel required frequent replacement, and the guns' accuracy was good enough for only city-sized targets. The German objective was to build a psychological weapon to attack the morale of the Parisians, not to destroy the city itself.

This might explain an unusual trade-off of distance (and therefore altitude) against features like accuracy and destructive ability. On this story, building a weapon to shoot a projectile 25 miles into the air had been feasible for some time, but wasn't worth it. This highlights the more general possibility that the altitude trend was perhaps more driven by the vagaries of demand for different tangentially-altitude-related ends than by technological progress.

The German military [apparently](#) dismantled the Paris Guns before departing, and did not comply with a Treaty of Versailles requirement to turn over a complete gun to the Allies, so the guns' capabilities are not known with certainty. However it sounds like the shells were clearly observed in Paris, and the relevant gun was clearly observed around 70 miles away, so the range is probably not ambiguous, and the altitude reached by a projectile is closely related to the range. So uncertainty around the gun probably doesn't affect our conclusions.

The first intercontinental ballistic missiles (ICBMs)

Discontinuity in [average speed of military payload crossing the Atlantic Ocean](#)

For most of history, the fastest way to send a military payload across the Atlantic Ocean was to put it on a boat or plane, much like a human passenger. So the [maximum speed of sending a military payload across the Atlantic Ocean](#) followed the [analogous passenger travel trend](#). However in August 1957, the two abruptly diverged with the [first successful test](#) of an intercontinental ballistic missile (ICBM)—the Russian [R-7 Semyorka](#). Early ICBMs traveled at around 11 thousand miles per hour, taking the minimum time to send a military payload between Moscow and New York for instance from around 14 hours to around 24 minutes.²⁵

The ICBM discontinuity: close up on the preceding trend, cut off at the first ICBM

A ‘[ballistic](#)’ missile is unpowered during most of its flight, and so follows a [ballistic trajectory](#)—the path of anything thrown into the air. Interestingly, this means that in order to go far enough to traverse the Atlantic, it has to be going a certain speed. Ignoring the curvature of the Earth or friction, this would be about 7000 knots for the shortest transatlantic distance—70% of its actual speed, and enough to be hundreds of years of discontinuity in the late 50s.²⁶ So assuming ballistic missiles crossed the ocean when they did, they had to produce a large discontinuity in the speed trend.

Does this mean the ICBM was required to be a large discontinuity? No—there would be no discontinuity if rockets were improving in line with planes, and so transatlantic rockets were developed later, or ICBM-speed planes earlier. But it means that even if the trends for rocket distance and speed are incremental and start from irrelevantly low numbers, if

they have a faster rate of growth than planes, and the threshold in distance required implies a speed way above the current record, then a large discontinuity must happen

This situation also means that you could plausibly have predicted the discontinuity ahead of time, if you were watching the trends. Seeing the rocket speed trend traveling upward faster than the plane speed trend, you could forecast that when it hit a speed that implied an intercontinental range, intercontinental weapons delivery speed would jump upward.

[An SM-65 Atlas, the first US ICBM, first launched in 1957](#) (1958 image)

YBa₂Cu₃O₇ as a superconductor

Discontinuity in [warmest temperature of superconduction](#)

When an ordinary material conducts electricity, it has some [resistance](#) (or opposition to the flow of electrons) which [takes](#) energy to overcome. The resistance can be gradually lowered by cooling the material down. For some materials though, there is a temperature threshold below which their resistance abruptly drops to zero, meaning for instance that electricity can flow through them indefinitely with no input of energy. These are '[superconductors](#)'.

Superconductors were [discovered](#) in 1911. [The first one observed](#), mercury, could superconduct below 4.2 Kelvin. From then on, more superconductors were discovered, and the warmest observed temperatures of superconduction gradually grew. In 1957, [BCS theory](#) was developed to explain the phenomenon (winning its authors a Nobel Prize), and [was understood](#) to rule out superconduction above temperatures of around 30K. But [in 1986](#) a new superconductor was found with a threshold temperature around 30K, and composed of a surprising material: a 'ceramic' involving oxygen rather than an alloy.²⁷ This also [won](#) a Nobel Prize, and instigated a rapid series of discoveries in similar materials—'[cuprates](#)'—which shot the highest threshold temperatures to around 125 K by 1988 (before continued upward).

The high temperature superconductor discontinuity: close up on the preceding trend, cut off at [YBa₂Cu₃O₇](#)

The first of the cuprates, LaBaCuO_4 , seems mostly surprising for theoretical reasons, rather than being radically above the temperature trend.²⁸ The big jump came the following year, from $\text{YBa}_2\text{Cu}_3\text{O}_7$, with its threshold at over 90 K.²⁹

This seems like a striking instance of the story where the new technology doesn't necessarily cause a jump so much as a new rate of progress. I wonder if there was a good reason for the least surprising cuprate to be discovered first. My guess is that there were many unsurprising ones, and substances are only famous if they were discovered before more exciting substances.

[Magnet levitating on top of a superconductor of \$\text{YBa}_2\text{Cu}_3\text{O}_7\$ cooled to merely -196°C \(77.15 Kelvin\)](#) Superconductors can allow magnetic levitation, [consistently repelling](#) permanent magnets [while stably pinned in place](#). (Picture: [Julien Bobroff \(user:Jubobroff\)](#), [Frederic Bouquet \(user:Fbouquet\)](#), [LPS, Orsay, France](#) / [CC BY-SA](#))

It is interesting to me that this is associated with a substantial update in very basic science, much like nuclear weapons. I'm not sure if that makes basic science updates ripe for discontinuity, or if there are just enough of them that some would show up in this list. (Though glancing at [this list](#) suggests to me that there were about 70 at this level in the 20th Century, and probably many fewer immediately involving a new capability rather than e.g. an increased understanding of pulsars. Penicillin also makes that list though, and we didn't find any discontinuities it caused.)

Moderate robust discontinuities (10-100 years of extra progress):

The 10-100 year discontinuous events were:

- HMS Warrior, 1860 (discontinuity in both [Royal Navy ship tonnage and Royal Navy ship displacement](#)³⁰)
- Eiffel Tower, 1889 (discontinuity in [tallest existing freestanding structure height](#), and in other height trends non-robustly)
- Fairey Delta 2, 1956 (discontinuity in [airspeed](#))
- Pellets shot into space, 1957, measured after one day of travel (discontinuity in [altitude achieved by man-made means](#))³¹
- Burj Khalifa, 2009 (discontinuity in [height of tallest building ever](#))

Other places we looked

Here are places we didn't find robust discontinuities³²) – follow the links to read about any in detail:

- [Alexnet](#): This convolutional neural network made important progress on labeling images correctly, but was only a few years ahead of the previous trend of success in the ImageNet contest (which was also a very short trend).
- [Light intensity](#): We measured argon flashes in 1943 as a large discontinuity, but I judge it non-robust. The rate of progress shot up at around that time though, from around half a percent per year to an average of 90% per year over the next 65 years, the rest of it involving increasingly intense lasers.
- [Real price of books](#): After the invention of the printing press, the real price of books seems to have dropped sharply, relative to a recent upward trajectory. However this was not long after a similarly large drop purportedly from paper replacing parchment. So in the brief history we have data for, the second drop is not unusual. We are also too uncertain about this data to confidently conclude much.
- [Manuscripts and books produced over the last hundred years](#): This was another attempt to find a discontinuity from the printing press. We measured several discontinuities, including one after the printing press. However, it is not very surprising for a somewhat noisy trend with data points every hundred years to be a hundred years ahead of the best-guess curve sometimes. The discontinuity at the time of the printing press was not much larger than others in nearby centuries. The clearer effect of the printing press at this scale appears to be a new faster growth trajectory.
- [Bandwidth distance product](#): This measures how much can be sent how far by communication media. It was just pretty smooth.
- [Total transatlantic bandwidth](#): This is how much cable goes under the Atlantic Ocean. It was also pretty smooth.

- [Whitney's cotton gin](#): Cotton gins remove seeds from cotton. Whitney's gin is often considered to have revolutionized the cotton industry and maybe contributed to the American Civil War. We looked at its effects on pounds of cotton ginned per person per day, and our best guess is that it was a moderate discontinuity, but the trend is pretty noisy and the available data is pretty dubious. Interestingly, progress on gins was speeding up a lot prior to Whitney (the two previous data points look like much bigger discontinuities, but we are less sure that we aren't just missing data that would make them part of fast incremental progress). We also looked at evidence on whether Whitney's gin might have been a discontinuity in the more inclusive metric of cost per value of cotton ginned, but this was unclear. As evidence about the impact of Whitney's gin, US cotton production appears to us to have been on the same radically fast trajectory before it as after it, and it seems people continued to use various other ginning methods for at least sixty years.
- [Group index of light or pulse delay of light](#): These are two different measures of how slowly light can be made to move through a medium. It can now be 'stopped' in some sense, though not the strict normal one. We measured two discontinuities in group index, but both were relative to a fairly unclear trend, so don't seem robust.
- [Particle accelerator performance](#): natural measures include center-of-mass energy, particle energy, and lorentz factor achieved. All of these progressed fairly smoothly.
- [US syphilis cases, US syphilis deaths, effectiveness of syphilis treatment, or inclusive costs of syphilis treatment](#): We looked at syphilis trends because we thought penicillin might have caused a discontinuity in something, and syphilis was apparently a key use case. But we didn't find any discontinuities there. US syphilis deaths became much rarer over a period around its introduction, but the fastest drop slightly predates plausible broad use of penicillin, and there are no discontinuities of more than ten years in either US deaths or cases. Penicillin doesn't even appear to be much more effective than its predecessor, conditional on being used.³³ Rather, it seems to have been much less terrible to use (which in practice makes treatment more likely). That suggested to us that progress might have been especially visible in 'inclusive costs of syphilis treatment'. There isn't ready quantitative data for that, but we tried to get a rough qualitative picture of the landscape. It doesn't look clearly discontinuous, because the trend was already radically improving. The preceding medicine sounds terrible to take, yet was nicknamed '[magic bullet](#)' and is considered '[the first effective treatment for syphilis](#)'. Shortly beforehand, [mercury was still a usual treatment](#) and deliberately contracting malaria had recently been added to the toolbox.
- [Nuclear weapons on cost-effectiveness of explosives](#): Using nuclear weapons as explosives was not clearly cheaper than using traditional explosives, let alone discontinuously cheaper. However these are very uncertain estimates.
- [Maximum landspeed](#): Landspeed saw vast and sudden changes in the rate of progress, but the developments were so close together that none was very far from average progress between the first point and the most recent one. If we more readily expect short term trends to continue (which arguably makes sense when they are as well-defined as these), then we find several moderate discontinuities. Either way, the more basic thing going on appears to be very distinct changes in the rate of progress.
- [AI chess performance](#): This was so smooth that a point four years ahead of the trend in 2008 is eye-catching.
- [Breech-loading rifles on the firing rate of guns](#): Breech-loading rifles were suggested to us as a potential discontinuity, and firing rate seemed like a metric on which they plausibly excelled. However there seem to have been other guns with similarly fast fire rates at the time breech-loading rifles were introduced. We haven't checked whether they produced a discontinuity in some other metric (e.g. one that combines several features), or if anything else caused discontinuities in firing rate.

III. Some observations

Prevalence of discontinuities

Some observations on the overall prevalence of discontinuities:

- [32%](#) of trends we investigated saw at least one large, robust discontinuity (though note that trends were selected for being discontinuous, and were a very non-uniform collection of topics, so this could at best inform an upper bound on how likely an arbitrary trend is to have a large, robust discontinuity somewhere in a chunk of its history)
- [53%](#) of trends saw any discontinuity (including smaller and non-robust ones), and in expectation a trend saw [more than two](#) of these discontinuities.
- On average, each trend had [0.001](#) large robust discontinuities per year, or [0.002](#) for those trends with at least one at some point³⁴
- On average [1.4%](#) of new data points in a trend make for large robust discontinuities, or [4.9%](#) for trends which have one.
- On average [14%](#) of total progress in a trend came from large robust discontinuities (or [16%](#) of logarithmic progress), or [38%](#) among trends which have at least one

This all suggests that discontinuities, and large discontinuities in particular, are more common than I thought previously (though still not that common). One reason for this change is that I was treating difficulty of finding good cases of discontinuous progress as more informative than I do now. I initially thought there weren't many around because suggested discontinuities often turned out not to be discontinuous, and there weren't a huge number of promising suggestions. However we later got more good suggestions, and found many discontinuities where we weren't necessarily looking for them. So I'm inclined to think there are a few around, but our efforts at seeking them out specifically just weren't very effective. Another reason for a larger number now is that our more systematic methods now turn up many cases that don't look very remarkable to the naked eye (those I have called non-robust), which we did not necessarily notice earlier. How important these are is less clear.

Discontinuities go with changes in the growth rate

It looks like discontinuities are often associated with changes in the growth rate. At a glance, 15 of the 38 trends had a relatively sharp change in their rate of progress at least once in their history. These changes in the growth rate very often coincided with discontinuities—in fourteen of the fifteen trends, at least one sharp change coincided with one of the discontinuities.³⁵ If this is a real relationship, it means that if you see a discontinuity, there is a much heightened chance of further fast progress coming up. This seems important, but is a quick observation and should probably be checked and investigated further if we wanted to rely on it.

Where do we see discontinuities?

Among these case studies, when is a development more likely to produce a discontinuity in a trend?³⁶ Some observations so far, based on the broader class including non-robust discontinuities, except where noted:

- **When the trend is about products not technical measures**
If we loosely divide trends into 'technical' (to do with scientific results e.g. highest temperature of a superconductor), 'product' (to do with individual objects meant for

use e.g. cotton ginned by a cotton gin, height of building), ‘industry’ (to do with entire industries e.g. books produced in the UK) or ‘societal’ (to do with features of non-industry society e.g. syphilis deaths in the US), then ‘product’ trends saw around [four times as many](#) discontinuities as technical trends, and the other two are too small to say much. (Product trends are less than twice as likely to have any discontinuities, so the difference was largely in how many discontinuities they have per trend.)

- **When the trend is about less important ‘features’ rather than overall performance**

If we loosely divide trends into ‘features’ (things that are good but not the main point of the activity), ‘performance proxies’ (things that are roughly the point of the activity) and ‘value proxies’ (things that roughly measure the net value of the activity, accounting for its costs as well as performance), then [features were more discontinuous than performance proxies](#).³⁷

- **When the trend is about ‘product features’**

(Unsurprisingly, given the above.) Overall, the 16 ‘[product features](#)’ we looked at had [4.6 discontinuities](#) per trend on average, whereas the 22 other metrics had 0.7 discontinuities per trend on average ([2 vs. 0.3 for large discontinuities](#)).³⁸ ‘Product features’ include for instance sizes of ships and fire rate of guns, whereas non-product features include total books produced per century, syphilis deaths in the US, and highest temperature of known superconductors.

- **When the development occurs after 1800**

Most of the discontinuities we found [happened after 1800](#). This could be a measurement effect, since much more recent data is available, and if we can’t find enough data to be confident, we are not deeming things discontinuities. For instance, the two obscure [cotton gins](#) before Whitney’s famous 1793 one that look responsible for huge jumps according to our sparse and untrustworthy 1700s data. The concentration of discontinuities since 1800 might also be related to progress speeding up in the last couple of centuries. Interestingly, since 1800 the rate of discontinuities doesn’t seem to be obviously increasing. For instance, [seven of nine](#) robust discontinuous events since 1900 happened by 1960.³⁹

- **When the trend is about travel speed across the Atlantic**

Four of our ten robust discontinuous events of over a hundred years came from the three transatlantic travel speed trends we considered. They are also [high on non-robust discontinuities](#).

- **When the trend doesn’t have a consistent exponential or linear shape**

To measure discontinuities, we had to extrapolate past progress. We did this at each point, based on what the curve looked like so far. Some trends we consistently called exponential, some consistently linear, and some sometimes seemed linear and sometimes exponential. The ten in this third lot all had discontinuities, whereas the 20 that consistently looked either exponential or linear were about [half as likely](#) to have discontinuities.⁴⁰

- **When the trend is in the size of some kind of object**

‘Object size’ trends [had](#) over five discontinuities per trend, compared to the average of [around 2](#) across all trends.

- **When [Isambard Kingdom Brunel](#) is somehow involved**

I mentioned Brunel above in connection with the *Great Eastern*. As well as designing that discontinuously large ship, which lay one of the discontinuously fast transatlantic telegraph cables, he designed the non-robustly discontinuous earlier ship *Warrior*.

I feel like there are other obvious patterns that I’m missing. Some other semi-obvious patterns that I’m noticing but don’t have time to actually check now, I am putting in the next section.

More things to observe

There are lots of other interesting things to ask about this kind of data, in particular regarding what kinds of things tend to see jumps. Here are some questions that we might answer in future, or which we welcome you to try to answer (and hope our data helps with):

- Are trends less likely to see discontinuities when more effort is going more directly into maximizing them? (Do discontinuities arise easily in trends people don't care about?)
- How does the chance of discontinuity change with time, or with speed of progress? (Many trends get much faster toward the end, and there are more discontinuities toward the end, but how are they related at a finer scale?)
- Do discontinuities come from 'insights' more than from turning known cranks of progress?
- Are AI related trends similar to other trends? The two AI-related trends we investigated saw no substantial discontinuities, but two isn't very many, and there is a persistent idea that once you can do something with AI, you can do it fast.⁴¹
- Are trends more continuous as they depend on more 'parts'? (e.g. is maximum fuel energy density more jumpy than maximum engine power, which is more jumpy than maximum car speed?) This would make intuitive sense, but is somewhat at odds with the 8 'basic physics related' trends we looked at not being especially jumpy.
- How does the specificity of trends relate to their jumpiness? I'd intuitively expect jumpier narrow trends to average out in aggregate to something smooth (for instance, so that maximum Volkswagen speed is more jumpy than maximum car speed, which is more jumpy than maximum transport speed, which is more jumpy than maximum man-made object speed). But I'm not sure that makes sense, and a contradictory observation is that discontinuities or sudden rate changes happen when a continuous narrow trend shoots up and intersects the broader trend. For instance, if record rocket altitude is continuously increasing, and record non-rocket altitude is continuously increasing more slowly but is currently ahead, then overall altitude will have some kind of corner in it where rockets surpass non-rockets. If you drew a line through liquid fuel rockets, pellets would have been less surprising, but they were surprising in terms of the broader measure.
- What does a more random sample of trends look like?
- What is the distribution of step sizes in a progress trend? (Looking at small ones as well as discontinuities.) If it generally follows a recognizable distribution, that could provide more information about the chance of rare large steps. It might also help recognize trends that are likely to have large discontinuities based on their observed distribution of smaller steps.
- Relatively abrupt changes in the growth rate seem common. Are these in fact often abrupt rather than ramping up slowly? (Are discontinuities in the derivative relevantly different from more object-level discontinuities, for our purposes?)
- How often is a 'new kind of thing' responsible for discontinuities? (e.g. the first direct flight and the first telegraph cable produced big discontinuities in trends that had previously been topped by ships for some time.) How often are they responsible for changes in the growth rate?
- If you drew a line through liquid fuel rockets, it seems like pellets may not be surprise, but they were because of the broader measure. How often is that a thing? I think a similar thing may have happened with the altitude records, and the land speed records, both also with rockets in particular. In both of those // similar thing happened with rockets in particular in land-speed and altitude? Could see trend coming up from below for some time.
- Is more fundamental science more likely to be discontinuous?

- With planes and ICBMs crossing the ocean, there seemed to be a pattern where incremental progress had to pass a threshold on some dimension before incremental progress on a dimension of interest mattered, which gave rise to discontinuity. Is that a common pattern? (Is that a correct way to think about what was going on?)
- If a thing sounds like a big deal, is it likely to be discontinuous? My impression was that these weren't very closely connected, nor entirely disconnected. Innovations popularly considered a big deal were often not discontinuous, as far as we could tell. For instance penicillin seemed to help with syphilis a lot, but we didn't find any actual discontinuity in anything. And we measured Whitney's cotton gin as producing a moderate discontinuity in cotton ginned per person per day, but it isn't robust, and there look to have been much larger jumps from earlier more obscure gins. On the other hand, nuclear weapons are widely considered a huge deal, and were a big discontinuity. It would be nice to check this more systematically.

IV. Summary

- Looking at past technological progress can help us tell whether AI trends are likely to be discontinuous or smooth
- We looked for discontinuities in 38 technological trends
- We found ten events that produced robust discontinuities of over a hundred years in at least one trend. (Djoser, Great Eastern, Telegraphs, Bridge, transatlantic flight, Paris Gun, ICBM, nukes, high temperature superconductors.)
- We found [53](#) events that produced smaller or less robust discontinuities
- The average rate of large robust discontinuities per year across trends was about 0.1%, but the chance of a given level of progress arising in a large robust discontinuity was around 14%
- Discontinuities were not randomly distributed: some classes of metric, some times, and some types of event seem to make them more likely or more numerous. We mostly haven't investigated these in depth.
- Growth rates sharply changed in many trends, and this seemed strongly associated with discontinuities. (If you experience a discontinuity, it looks like there's a good chance you're hitting a new rate of progress, and should expect more of that.)

Notes

Negative Feedback and Simulacra



Part 1: Examples

There's a thing I want to talk about but it's pretty nebulous so I'm going to start with examples. Feel free to skip ahead to part 2 if you prefer.

Example 1: Hot sauce

In this r/AmITheAsshole [post](#), a person tries some food their their girlfriend cooked, likes it, but tries another bite with hot sauce. Girlfriend says this "...insults her cooking and insinuates that she doesn't know how to cook".

As objective people not in this fight, we can notice that her cooking is exactly as good as it is whether or not he adds hot sauce. Adding hot sauce reveals information (maybe about him, maybe about the food), but cannot change the facts on the ground. Yet she is treating him like he retroactively made her cooking worse in a way that somehow reflects on her, or made a deliberate attempt to hurt her.

Example 2: Giving a CD back to the library

Back when I would get books on CD I would sometimes forget the last one in my drive or car. Since I didn't use CDs that often, I would find the last CD sometimes months later. To solve this, I would drop the CD in the library book return slot, which, uh, no longer looks like a good solution to me, in part because of the time I did this in front of a friend and she questioned it. Not rudely or anything, just "are you sure that's safe? Couldn't the CD snap if something lands wrong?." I got pretty angry about this, but couldn't actually deny she had a point, so settled for thinking that if she had violated a friend code by not pretending my action was harmless. I was not dumb enough to say this out loud, but I radiated the vibe and she dropped it.

Example 3: Elizabeth fails to fit in at martial arts

A long time ago I went to a martial arts studio. The general classes (as opposed to specialized classes like grappling) were preceded by an optional 45 minute warm up class. Missing the warm up was fine, even if you took a class before and after. Showing up 10 minutes before the general class and doing your own warm ups on the adjacent mats was fine too. What was not fine was doing the specialized class, doing your own warm ups on adjacent maps for the full 45 minutes while the instructor led regular warm ups, and then rejoining for the general class. That was "very insulting to the instructor".

This was a problem for me because the regular warm ups *hurt*, in ways that clearly meant they were bad for me (and this is at a place I regularly let people hit me in the head). Theoretically I could have asked the instructor to give me something different,

but that is not free and the replacements wouldn't have been any better, which is not surprising because no one there had the slightest qualification to do personal training or physical therapy. So basically the school wanted me to pretend I was in a world where they were competent to create exercise routines, more competent than I despite having no feedback from my body, and considered not pretending disrespectful to the person leading warm ups.

Like the hot sauce example, the warm ups were as good as they were regardless of my participation – and they knew that, because they didn't demand I participate. But me doing my own warm ups broke the illusion of competence they were trying to maintain.

Example 4: Imaginary Self-Help Guru

I listened to an interview where the guest was a former self-help guru who had recently shut down his school. Well, I say listened, but I've only done the first 25% so far. For that reason this should be viewed less as "this specific real person believes these specific things" and more like "a character Elizabeth made up in her head inspired by things a real person said..." and. For that reason, I won't be using his name or linking to the podcast.

Anyways, the actual person talked about how being a leader put a target on his back and his followers were never happy. There are indeed a lot of burdens of leadership that are worthy of empathy, but there was an... entitled... vibe to the complaint. Like his work as a leader gave him a right to a life free of criticism.

If I was going to steel- man him, I'd say that there are lots of demands people place on leaders that they shouldn't, such as "Stop reminding me of my abusive father" or "I'm sad that trade offs exist, fix it". But I got a vibe that the imaginary guru was going farther than that; he felt like he was entitled to have his advice *work*, and people telling him it didn't was taking that away from him, which made it an attack.

Example 5: Do I owe MAPLE space for their response?

A friend of mine (who has some skin in the meditation game) said things I interpreted as feeling very strongly that:

1. [My post on MAPLE](#) was important and great and should be widely shared.
2. I owed MAPLE an opportunity to read my post ahead of time and give me a response to publish alongside it (although I could have declined to publish it if I felt it was sufficiently bad).

Their argument, as I understood it at the time, was that even if I linked to a response MAPLE made later, N days worth of people would have read the post and not the response, and that was unfair.

I think this is sometimes correct- I took an example out of this post even though it required substantial rewrites, because I checked in with the people in question, found

they had a different view, and that I didn't feel sure enough of mine to defend it (full disclosure: I also have more social and financial ties to the group in question than I do to MAPLE).

I had in fact already reached out to my original contact there to let him know the post was coming and would be negative, and he passed my comment on to the head of the monastery. I didn't offer to let him see it or respond, but he had an opportunity to ask (what he did suggest is a post in and of itself). This wasn't enough for my friend- what if my contact was misrepresenting me to the head, or vice versa? I had an obligation to reach out directly to the head (which I had no way of doing beyond the info@ e-mail on their website) and explicitly offer him a pre-read and to read his response.

[Note: I'm compressing timelines a little. Some of this argument and clarification came in arguments about the principle of the matter after I had already published the post. I did share this with my friend, and changed some things based on their requests. On others I decided to leave it as my impression at the time we argued, on the theory that "if I didn't understand it after 10 hours of arguing, the chances this correction actually improves my accuracy are slim". I showed them a near-final draft and they were happy with it]

I thought about this very seriously. I even tentatively agreed (to my friend) that I would do it. But I sat with it for a day, and it just didn't feel right. What I eventually identified as the problem was this: MAPLE wasn't going to be appending my criticism to any of their promotional material. I would be shocked if they linked to me at all. And even if they did it wouldn't be the equivalent, because my friend was insisting that I proactively seek out their response, where they had never sought out mine, or to the best of my knowledge any of their critics. As far as I know they've never included anything negative in their public facing material, despite at least one person making criticism *extremely available* to them.

If my friend were being consistent (which is not a synonym for "good") they would insist that MAPLE seek out people's feedback and post a representative sample somewhere, at a minimum. The good news is: my friend says they're going to do that next time they're in touch. What they describe wanting MAPLE to create sounds acceptable to me. Hurray! Balance is restored to The Force! Except... assuming it does happen, why was my post necessary to kickstart this conversation? My friend could have noticed the absence of critical content on MAPLE's website at any time. The fact that negative reports trigger a reflex to look for a response and positive self-reports do not is itself a product of treating negative reports as overt antagonism and positive reports as neutral information.

[If MAPLE does link to my experience in a findable way on their website, I will append whatever they want to my post (clearly marked as coming from them). If they share a link on Twitter or something else transient, I will do the same]

Part 2: Simulacrum

My friend [Ben Hoffman](#) talks about simulacra a lot, with this rough definition:

1. First, words were used to maintain shared accounting. We described reality intersubjectively in order to build shared maps, the better to navigate our environment. I say that the food source is over there, so that our band can move towards or away from it when situationally appropriate, or so people can make other inferences based on this knowledge.
2. The breakdown of naive intersubjectivity – people start taking the shared map as an object to be manipulated, rather than part of their own subjectivity. For instance, I might say there's a lion over somewhere where I know there's food, in order to hoard access to that resource for idiosyncratic advantage. Thus, the map drifts from reality, and we start dissociating from the maps we make.
3. When maps drift far enough from reality, in some cases people aren't even parsing it as though it had a literal specific objective meaning that grounds out in some verifiable external test outside of social reality. Instead, the map becomes a sort of [command language](#) for coordinating actions and feelings. "There's food over there" is perhaps construed as a bid to move in that direction, and evaluated as though it were that call to action. Any argument for or against the implied call to action is conflated with an argument for or against the proposition literally asserted. This is how [arguments become soldiers](#). Any attempt to simply investigate the literal truth of the proposition is considered at best naive and at worst politically irresponsible.
But since this usage is parasitic on the old map structure that was meant to describe something outside the system of describers, language is still structured in terms of reification and objectivity, so it substantively resembles something with descriptive power, or "aboutness." For instance, while you cannot acquire a physician's privileges and social role simply by providing clear evidence of your ability to heal others, those privileges are still justified in terms of pseudo-consequentialist arguments about expertise in healing.
4. Finally, the pseudostructure itself becomes perceptible as an object that can be manipulated, the pseudocorrespondence breaks down, and all assertions are nothing but moves in an ever-shifting game where you're trying to think a bit ahead of the others (for positional advantage), but not too far ahead.

If that doesn't make sense, try this anonymous [comment](#) on the post

- Level 1: "There's a lion across the river." = There's a lion across the river.
- Level 2: "There's a lion across the river." = I don't want to go (or have other people go) across the river.
- Level 3: "There's a lion across the river." = I'm with the popular kids who are too cool to go across the river.
- Level 4: "There's a lion across the river." = A firm stance against trans-river expansionism focus grouped well with undecided voters in my constituency.

In all five of my examples, people were given information (I like this better with hot sauce, you might break the library's CD, these exercises hurt me and you are not qualified to fix it, your advice did not fix my problem, I had a miserable time at your retreat), and treated it as a social attack. This is most obvious in the first four, where

someone literally says some version of “I feel under attack”, but is equally true in the last one, even though the enforcer was different than the ~victim and was attempting merely to tax criticism, not suppress it entirely. All five have the effect that there is either more conflict or less information in the world.

Part 3: But...

When I started thinking about this, I wanted a button I could push to make everyone go to level one all the time. It's not clear that that's actually a good idea, but even if it was, there is no button, and choosing/pretending to cut off your awareness of higher levels in order to maintain moral purity does you no good. If you refuse to conceive of why someone would tell you things other than to give you information, you leave yourself open to "I'm only telling you this to make you better" abuse. If you refuse to believe that people would lie except out of ignorance, you'll trust when you shouldn't. If you refuse to notice how people are communicating with others, you will be blindsided when they coordinate on levels you don't see.

But beating them at their own game doesn't work either, because the enemy was never them, it was the game, which you are still playing. You can't socially maneuver your way into a less political world. In particular, it's a recent development that I would have noticed my friend's unilateral demand for fairness as in fact tilted towards MAPLE. In a world where no one notices things like that, positive reviews of programs become overrepresented.

I don't have a solution to this. The best I can do right now is try to feed systems where level one is valued and higher levels are discussed openly. "How do I find those?" you might ask. I don't know. If you do, my email address is elizabeth - at - [\[original domain\]](#) and I'd love to hear from you. You can also book a [time to talk to me for an hour](#). What I have are a handful of 1:1 relationships where we have spent years building trust to get to the point where "I think you're being a coward" is treated as genuine information, not a social threat, and mostly the other person has made the first move.

The pieces of advice I do have are:

1. If someone says they want honest feedback, err on the side of giving it to them. They are probably lying, but that's their problem (unless they're in a position to make it yours, in which case think harder about this).
2. Figure out what you need to feel secure as someone confirms your worst fears about yourself and ask for it, even if it's weird, even if it seems like an impossibly big ask. People you are compatible with will want to build towards that (not everyone who doesn't is abusive or even operating in bad faith- but if you can't start negotiations on this I'd be very surprised if you're compatible).
3. Be prepared for some sacrifices, especially in the congeniality department. People who are good at honesty under a climate that punishes it are not going to come out unscathed.

Seemingly Popular Covid-19 Model is Obvious Nonsense

Previous Covid-19 thoughts: [On R0, Taking Initial Viral Load Seriously](#)

Epistemic Status: [Something Is Wrong On The Internet](#). Which should almost always be ignored even when you are an expert, and I am nothing of the kind. Thus, despite this seeming like a necessary exception, I expect to regret writing this.

People are taking the projection of 60,000 American deaths from Covid-19 as if it were a real prediction. This number is being used to make policy, to deny states medical equipment and to make plans that spend trillions of dollars and when to plan to reopen entire economies.

Ignoring this in the hopes it will go away does not seem reasonable.

My suspicions that this was necessary were more than confirmed when, failing to realize just how obvious the nonsense in question was and thinking I needed to justify labeling it nonsense, I wrote a reference post called [The One Mistake Rule](#).

The second comment on that post was to argue that we should indeed use exactly the model that motivated me to write the post. The comment is here in full:

>> If a model gives a definitely wrong answer anywhere, it is useless everywhere.

Except if it needs to be used right now to make important decisions and it's the best model we have. See: <https://covid19.healthdata.org/united-states-of-america>

We could plausibly think this is the best model we have? Oh my are we screwed.

The Baseline Scenario That Makes No Sense

There seems to be a developing consensus on many fronts, for now, that the model linked above represents our reality. The model says it is 'designed to be a planning tool' and that is exactly what is happening here.

What is this model doing? Time to [look at the pdf](#).

Here's the money quote that describes the core of what they are actually doing.

A covariate of days with expected exponential growth in the cumulative death rate was created using information on the number of days after the death rate exceeded 0.31 per million to the day when 4 different social distancing measures were mandated by local and national government:

School closures, non-essential business closures including bars and restaurants, stay-at-home recommendations, and travel restrictions including public transport closures. Days with 1 measure were counted as 0.67 equivalents, days with 2 measures as 0.334 equivalents **and with 3 or 4 measures as 0**. For states that have not yet implemented all of the closure measures, we assumed that the remaining measures will be put in place within 1 week. This lag between reaching a threshold death rate and implementing more aggressive social distancing was

combined with the observed period of exponential growth in the cumulative death rate seen in Wuhan after Level 4 social distancing was implemented, adjusted for the median time from incidence to death. For ease of interpretation of statistical coefficients, this covariate was normalized so the value for Wuhan was 1.

In other words, this model assumes that social distancing measures work *really, really well*. Absurdly well. All you have to do to stop Covid-19 is any three of: Close schools, close non-essential businesses, tell people to stay at home, impose travel restrictions.

If you do that and maintain it, people stop dying. Entirely.

Look at the graph they have up as of this writing (updated on 4/10). By June 20, they predict actual zero deaths that day and every future day. They have us under 100 deaths per day by the end of May.

The peak in hospital use? Today, April 11.

The peak in deaths? Yesterday, April 10. For New York, several days ago, with our last death *on May 20*.

In other words, considering the delay in deaths is about three weeks, they predict that *no one in New York State will be infected after April*. No one! We'll all be safe in only three weeks!

This is despite us not yet seeing any evidence of a major decline in positive test rates in New York. Deaths lag positive tests by weeks.

Hard to be more maximally optimistic than that. One could call this the ‘theoretical beyond best case scenario.’

(The statement is actually even more absurd than that, considering variation in time to case progression, but I’m going to let that one go.)

(Exercise for the reader, you have five seconds: What is the implied R₀?)

(Second exercise for the reader: If there are four things that reduce the spread of infection some amount, and R₀ is about 4 initially, and you implement three of them, what is the new R₀?)

They Account for Uncertainty, Right?

They generously account for uncertainty with the following ‘confidence interval’:

Figure 9 shows the expected cumulative death numbers with 95% uncertainty intervals. The average forecast suggests 81,114 deaths, but the range is large, from 38,242 to 162,106 deaths.

(Note: this was as of paper publishing, numbers are now lower.)

That is not how this works. That is not how any of this works.

The way this works *once we correct for all the obvious absurdities* is that this is a *lower bound* on how good things could possibly go.

If I am incorrect, and that *is* how any of this works I have some *very, very large* bets I would like to place.

A Simpler Version of the Same Model

The model seems functionally the same as this:

Assume all reported numbers are accurate, and assume that no one gets infected once you nominally implement three of the four social distancing measures. Which you assume every US state will do within a week from the model starting.

Let's simplify that again.

Assume that no one under an even half-serious (three quarters serious?) lock down ever gets infected out-of-household.

We still see deaths for a few weeks, because there is a lag, but then it's all over.

What the Model Outputs

As of when I wrote this line, this more-than-maximally-optimistic model projects 61,545 deaths in the United States.

People with power, people with influence, what some might call our “best people,” are on television and in the media predicting around 60,000 total American deaths.

I will say that again.

We are telling the public a death count that effectively implies that by about a month from now, and in many places earlier than that, no new American ever gets infected with Covid-19.

The model assumes that our half measures towards social distancing will have the same impact as *was reported* in Wuhan. In Wuhan, they blockaded apartment buildings, took anyone suspected of being positive away for isolation, and still, months after this model says there are no infections or even deaths, has severe movement restrictions and blockades up all over the place.

Whereas the New York City subways continue to run, and California thinks weed sales are an essential business.

I hope that my perception of this is wrong. Perhaps everyone knows this model is nonsense. Perhaps there are better ones out there – if you know of one you respect, please let me know about it!

But again, this is a maximally optimistic model on every front. I keep seeing people whose voice matters share this same final answer of predicting 60,000 deaths. If it's not from a model doing *more or less this*, I don't know how you get an answer in that ballpark.

Unless of course answers are being chosen without regard to reality.

How Long Can People Usefully Work?

[This piece is cross-posted on my blog here.](#)

I hear a lot of theories around how to work optimally. “You shouldn’t work more than eight hours a day.” “You can work 12 hours a day and be fine.” “It’s important to take weekends or evenings off work entirely.” “It’s best to immerse yourself in your work 24/7 if you want to be an expert.”

Perhaps most well known is Cal Newport’s claim in Deep Work that “For a novice, somewhere around an hour a day of intense concentration seems to be a limit, while for experts this number can expand to as many as four hours—but rarely more.”

Many of these theories are asserted with surprising confidence...especially since they contradict each other. At least some have to be wrong or more nuanced, and it matters which are right.

I coach Effective Altruists who want to maximize the good they can do. So they want to know how much they can work before additional work is wasted (or just less valuable compared to extra time doing other things). They also want to know how much they can work before risking reducing their long-term productivity -- burning out from working too hard is a lose-lose for them and the world.

So, I dug into these questions to see if I could find an answer.

Short answer, there isn’t a lot of good research on the topic. Long answer, our best data comes from World War One factory workers (turns out you can do interesting research when your research subjects aren’t legally allowed to leave), but maybe we have enough information anyway to make educated guesses. At the least, we can run personal experiments.

Here’s a summary of my findings and opinion, followed by the actual research.

1. Limits on total hours. First, as you work more hours, each hour becomes less productive. If I had to guess based on the research, I’d say there are steeply diminishing marginal value around 40-50 hours per week, and negative returns (meaning less total output for the day per additional hour) somewhere between 50 and 70 hours. But, this is based on a grand total of two data sets. World War One studies of factory workers are the only source that experimentally tests the question. They didn’t even know to report sample sizes. (Also, a bunch of guys who wanted to prove workers should have Sunday off, found that having one day per week off was good. Give that as much credence as you think it’s worth.) The other data set shows a similar, correlational trend between CEO hours and company sales.

2. Limits on learning. Second, the spaced learning literature indicates that studying more than a few hours at a time may be useless. At least if you’re trying to memorize information. On one topic. If you’re producing output or switching up topics, this literature might be totally irrelevant. Still, it’s important to consider if additional hours of hard work might be wasted. I can only imagine what the control group guys in one study thought; “The guys studying four hours a day learned just as much as I did studying for seven? I worked my ass off an extra three hours a day for NOTHING?!?” (Also worth noting, a [meta-analysis](#) found that less rigorous studies had bigger effect

sizes.) If this literature applies to regular work, then it might be good to work on one priority for 1-4 hours a day, then switch to other topics.

3. Anecdotal reports. I'm fairly skeptical any of this research tells us how much to work (you can see more details below). I place more confidence on the anecdotal reports of [productive people](#). It's common for them to report three to five hours of deep work on a top priority each day, plus several hours more of lower energy or more "following curiosity"-type work (three more yet-to-be-released interviews also report in this range; one interview reports more). To be clear, I think they're describing consistent, intense, "write a book chapter" levels of focus for those three to five hours.

When people talk about working longer days (e.g. 12 hours), they usually report that it takes enough of a toll on subsequent days that it's not worth it. I know a number of people who consistently put in 8-12 hours of focused work a day, so it's possible. But most of these people are chronically stressed about their work. I'm not sure if working a lot makes people stressed, or if stressed people work a lot to cope with the stress. Either way, I'm hesitant to recommend people try to work this much.

4. Long term sustainability. Additionally, limits to working hours might be necessary for long-term health and wellbeing, e.g. [avoiding burnout](#). Setting aside vague "balance" like only working 40 hours a week (since you're probably not a World War One factory worker) - what do you need to stay sustainably healthy, happy, and productive? If you're unhappy, stressed, or tired, pay attention to this and [check in](#) about your base needs such as sleep, exercise, social, and leisure time.

If you haven't before, try experimenting to see what makes you happy *and* productive at the same time. I didn't know I needed eight hours of sleep every night until I tracked my sleep, and it took me a while to notice that exercise made me happier and more productive the following day. If you're otherwise doing okay and want to make more time for productive work, you can try experimenting freeing up time in one area and see how that impacts you. If the change makes you feel worse, try something else instead. (It's good for you to be happy. Please believe me here.)

5. Caveats. Additionally, I want to note that I still expect prioritization and deep work to be much more important than total hours spent working. E.g. if actions differ by 10x or more in importance regularly, then output is mostly determined by choosing the right things to work on, plus a smaller multiplier for time worked. In this case, having better processes for choosing the right work is way more important than hours spent working. Additionally, deep work in focused blocks is likely better for making progress on big priorities, so increasing deep work time is more important than total work time. Trying to work more hours may even be counter productive if you prioritize less (see [The Five-Hour Workday](#) and [Owen's interview](#)).

6. Individual Experiments. Finally, the best process may vary dramatically from individual to individual. This is obvious in extreme cases, such as people with extreme fatigue who can only work a few hours a day (if this is you, focus on addressing your fatigue first). So pay attention to yourself. Test what works for you.

I wrote up the results of a simple experiment I did on myself here. TLDR; I did one hour of deep work each day for two days, then two days of four hours each, and finally two days of eight hours each. I spent the entire 26 hours writing, and tracked how many words I wrote each hour to calculate the diminishing returns to big blocks.

You can fairly easily do a similar experiment to play with your personal limits.

If you want to do more, here are some other experiments you might find valuable:

- Remember that feeling productive and actually accomplishing valuable work might be separate. If you don't already, you may need to **track how much you work and your goals/output** so you have a baseline for what you get done before you can meaningfully run experiments. Rough metrics are fine - reducing uncertainty is better than throwing your hands up. If you're not sure what output to measure, here are some ideas: time spent on priorities, deep work time, writing pace and quality, novel ideas, difficulty of problem you can solve, amount of coding or bugs, emails answered. These are all rough, so try a couple and see which feel like they're correlated with what you care about.
- Does having a **full day off** once a week help? Do you see a noticeable difference when you take a day off with zero work vs still checking email, etc.?
- Energy and focus also often change throughout the day - **is there a particular schedule that works well for you?** Are you more likely to get high quality, focused work done in the morning, afternoon, or evening? I estimate I'm about four times more productive in my peak work sessions than when I'm tired, which I know because I track my working time on Toggl.com and evaluated my writing output from a few blocks of time spent writing when I felt tired and when I felt energized.
- **Are breaks, walks, or naps helpful?** In particular, taking breaks to check you're working efficiently is a good way to avoid wasting big blocks of time. E.g. pause every hour to check that you used the time to push forward your priority with the least [wasted motion](#) possible. If not, then make a better plan for the next hour to nail your goal.
- **Do you need bigger breaks periodically?** If you push yourself for a while, do you hit a wall and need to take a few days to recharge? If so, what does it feel like when you hit a wall and what ways are best for you to recharge?

And remember, you're finding what works for you here.

7. Conclusion. So, to the effective altruists who want to push their ability to do more so they can have a bigger impact and to the effective altruists who want EAs to cut themselves some slack before they burn out -- you both have a valid point. We don't know how much a motivated person at peak performance can work. It might be a lot more than four hours a day. It might not.

If you run the experiments and find ways that you can work more efficiently or more hours, great! As long as you're still good. If you're regularly stressed and unhappy about your work, that should be a big red flag. You burning out will not help your cause. (P.s. your happiness matters too.) In that case, it might be good to work less - at least temporarily - to force yourself to be more efficient and thoughtful about your work.

If you just wanted the high-level overview, feel free to stop here! The rest is the nitty gritty of the studies I looked at.

1. Diminishing returns on total time spent working

To start with, I recommend Elizabeth Van Nostrand's epistemic spot check on [The Role of Deliberate Practice in the Acquisition of Expert Performance](#), the basis for Newport's claims about deep work time. His claim is among the more commonly cited reasons for believing people can only work so much. In turn, he cites Ericsson, Krampe, and Tesch-Römer's 1993 paper, [The Role of Deliberate Practice in the Acquisition of Expert Performance](#).

Elizabeth concludes, "Many of the studies were criminally small, and typically focused on singular, monotonous tasks like responding to patterns of light or memorizing digits. The precision of these studies is greatly exaggerated. There's no reason to believe Ericsson, Krampe, and Tesch-Römer's conclusion that the correct number of hours for deliberate practice is 3.5, much less the commonly repeated factoid that humans can do good work for 4 hours/day."

I wanted to go beyond this pitiful citation trail. So, the following are the studies I dug up to see whether there was actually good evidence out there. You can jump back to the first page if you want to hear the summary of how, no, there's really not much good evidence available.

1. Time-use diary study of CEOs

I found one dataset that seemed relevant to diminishing returns on output per hour worked doing thought work. It was a [study of CEOs' time use](#), with data collected by phone calls with the CEOs' personal assistants to reconstruct the CEOs' days, which used company sales as the "output" measure.

According to the study, among 1,114 CEOs, a 10% increase in weekly hours worked was associated with a 3.3% increase in company productivity. The CEOs worked an average of 52 hours per week. It doesn't say if or when additional hours worked become negatively correlated with productivity.

I briefly looked at the data set ([obtained here](#)). Take my analysis with a huge grain of salt - this is from eyeballing one graph. The scatter plot on hours worked and company sales (without controlling for anything) trended toward more hours worked correlating with more sales but there wasn't a clear pattern, and a LOESS curve flattened shortly after 40 hours per week. This weakly inclined me to think the data set supports a positive relationship between more hours worked and more total output, with diminishing returns per additional hour after about 40 or 45 hours per week. I decided it wasn't worth my time right now to buy SPSS and play around with the data set a bunch more, but I would be very interested in the results if someone else felt like doing so.

2. Studies of factory workers in World War One

The main dataset I saw cited for supporting diminishing marginal returns is a collection of studies done during World War One, mostly on factory munitions workers. The studies support that general conclusion to a limited degree, but take them with a big grain of salt: the sample sizes were generally not given but probably "criminally small", the study quality was probably bad given it was done a hundred years ago, and it's unclear that factory work generalizes to thought work. In so far as these studies generalize, they would support working <50 hours per week and taking one entire day off per week to rest.

I hunted down [Fundamentals of Skill, Welford \(1968\)](#) (cited in the deliberate practice paper) and found what seems to be the most relevant passage: "Presumptive

evidence of some kind of fatigue effect during an industrial shift is contained in the classical reports of the Industrial Fatigue Research Board (later renamed Industrial Health Research Board). These reports showed not only that shorter working shifts led to higher hourly output (Osborne, 1919, Vernon, 1919), but also that a net reduction in working hours could sometimes lead to a net rise in total output (Vernon, 1920a, b)." (p.282)

I tracked down [Vernon 1920a](#) (since this was the only paper cited for the claim that a net reduction in working hours could sometimes lead to a net rise in total output). The paper supports both claims, but only at a factory-level (not individual-level) for the net reduction in productivity.

Reducing shifts from 8 to 4 hours increased output per hour by 11.5%, and reducing shifts from 8 to 6 hours increased output per hour by 4.7-10.6%. The factory could use more men for more, shorter shifts, so the total working time was nearly constant (2 hours less total for the 6-hour shifts). So in cases where more workers are unavailable (such as most thought work), working the extra hours would result in more total output even if lower hourly output.

Next, I read a review of the rest of research on munitions workers ([Pencavel, 2014](#)). Pencavel found that total output increased with each additional hour until 63 hours per week, then decreased with additional hours worked. Average output per hour decreased above 49 hours per week. In addition, holding hours per week constant, having one day off is about 10% better than working seven shorter days. (I'm additionally skeptical of this finding because it felt like the researchers were trying to prove that having Sunday off was worthwhile.) Between the diminishing marginal output and the benefit of a day off, working 48 hours per week across six days resulted in slightly more output than working 70 hours per week across seven days. So, capping work time at 50 hours per week would have resulted in little lost output. Finally, they claim the rate at which additional hours of work becomes less productive varies across workers and across types of work.

2. Diminishing returns on time spent learning

The numbers on hours of deep work time might have originated in the literature on spaced learning. It's well studied that learning via [spaced repetition is more efficient and effective than massed studying](#) (e.g. cramming for a test). Below, I summarize the spaced learning studies that the deliberate practice paper cited, directly or indirectly via citing a reference in another work, plus another study from a meta-analysis that had particularly long-term tests for retention.

These studies do seem to support diminishing returns on time spent learning in one day. According to these studies, studying one hour per day, spaced out over more days, might shave a quarter or more off the total time required to learn a given thing, compared to studying for 2 or more hours per day. So more calendar time but fewer hours of studying. This [meta-analysis](#) suggested the benefit of spaced learning compared to massed as 15% on average, which might be a more realistic expectation.

Still, there are a couple reasons this literature might not generalize to time spent working.

First, spaced learning might only be relevant when memorizing new skills or material, e.g. if the benefit is due to how memories are formed. I also didn't see anything on how the benefit of spaced learning would change as one becomes an expert – contrary to Newport's claims that the amount of deep work done per day increases from one to four hours as one becomes an expert. (I suspect his claim is more likely to relate to one's ability to stay motivated and focused instead.)

Second, even when relevant, it seems like the spaced learning benefit is for one topic, not all work done in a day. For example, the participants in the army study below did other things after their four hours of studying. In that case, a good rule of thumb would be to only work/learn on one subject for a limited number of hours, but you could study something else later in the day.

1. Telegraphy army study

From Elizabeth's [review](#): "An interesting army study showing that students given telegraphy training for 4 hours/day (and spending [the rest] on other topics) learned as much as students studying 7 hours/day. This one seems genuinely relevant, although not enough to tell us where peak performance lies, just that four hours are better than seven. Additionally, the students weren't loafing around for the excess three hours: they were learning other things. So this is about how long you can study a particular subject, not total learning capacity in a day."

The students studying for 4 hours per day spent an additional three weeks doing so, and ended up "markedly superior". Breaking the four-hour period into four one-hour periods didn't give further improvement.

2. Postmen learning to type

[The Influence of Length and Frequency of Training Session on the Rate of Learning to Type, Baddeley & Longman \(1978\)](#)

Study on learning to type: "Four groups of postmen were trained to type alpha-numeric code material using a conventional typewriter keyboard. Training was based on sessions lasting for one or two hours occurring once or twice per day."

It took the people studying 2 x 2 (two hours per session, two sessions per day) 49.7 hours on average to learn the keyboard range, while the people doing 1 x 1 only took 34.9 hours. It took fewer days to study in mass (12 instead of 35), but the per hour learning was less efficient.

3. Assembly line training task

A second passage from [Fundamentals of Skill](#) referred to training session length, though again apparently in a factory setting: "A further, as yet unanswered, question is raised by the problem of optimum length of training session. It is clear that there are severe limits to the rate at which material can be learnt when considered on a time scale of seconds and minutes, but are there any additional limitations operating over periods of hours, days or even longer times? Common experience suggests that there may be, but the question does not seem to have been posed in a scientific context. An indication that it might be worth asking is contained in the finding by Henshaw and Holman (1933) in an industrial study, that 80 min training per day at a chain assembly task yielded as rapid improvement as 160 min." I was unable to find the original study.

4. Learning a foreign language

[Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. \(1993\). Maintenance of foreign language vocabulary and the spacing effect. Psychological Science, 4, 316-321.](#)

Another small study that I saw found: “Four adults (aged 25–57 yrs at the beginning of the study) learned and relearned 300 English–foreign language word pairs. Either 13 or 26 relearning sessions were administered at intervals of 14, 28, or 56 days. Retention was tested for 1, 2, 3, or 5 yrs after training terminated. The longer intersession intervals slowed down acquisition slightly, but this disadvantage during training was offset by higher retention. 13 retraining sessions spaced at 56 days yielded retention comparable to 26 sessions spaced at 14 days.”

And that’s all folks. I hope you had fun reading the blog equivalent of a null finding.

An Orthodox Case Against Utility Functions

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

This post has benefitted from discussion with Sam Eisenstat, Scott Garrabrant, Tsvi Benson-Tilsen, Daniel Demski, Daniel Kokotajlo, and Stuart Armstrong. It started out as a thought about [Stuart Armstrong's research agenda](#).

In this post, I hope to say something about what it means for a rational agent to have preferences. The view I am putting forward is relatively new to me, but it is not very radical. It is, dare I say, a conservative view -- I hold close to Bayesian expected utility theory. However, my impression is that it differs greatly from *common impressions* of Bayesian expected utility theory.

I will argue against a particular view of expected utility theory -- a view which I'll call *reductive utility*. I do not recall seeing this view explicitly laid out and defended (except in in-person conversations). However, I expect at least a good chunk of the assumptions are commonly made.

Reductive Utility

The core tenets of reductive utility are as follows:

- The [sample space](#) Ω of a rational agent's beliefs is, more or less, the set of possible ways the world could be -- which is to say, the set of possible *physical configurations of the universe*. Hence, each world $\omega \in \Omega$ is one such configuration.
- The preferences of a rational agent are represented by a utility function $U : \Omega \rightarrow \mathbb{R}$ from worlds to real numbers.
- Furthermore, the utility function should be a *computable* function of worlds.

Since I'm setting up the view which I'm knocking down, there is a risk I'm striking at a straw man. However, I think there are some good reasons to find the view appealing. The following subsections will expand on the three tenets, and attempt to provide some motivation for them.

If the three points seem obvious to you, you might just skip to the next section.

Worlds Are Basically Physical

What I mean here resembles the standard physical-reductionist view. However, my emphasis is on certain features of this view:

- There is some "basic stuff" -- like like quarks or vibrating strings or what-have-you.
- What there is to know about the world is some set of statements about this basic stuff -- particle locations and momentums, or wave-form function values, or what-have-you.
- These special atomic statements should be logically independent from each other (though they may of course be probabilistically related), and together, fully determine the world.
- These should (more or less) be what beliefs are about, such that we can (more or less) talk about beliefs in terms of the sample space $\omega \in \Omega$ as being the set of worlds understood in this way.

This is the so-called "view from nowhere", as [Thomas Nagel puts it](#).

I don't intend to construe this position as ruling out certain non-physical facts which we may have beliefs about. For example, we may believe [indexical](#) facts on top of the physical facts -- there might be (1) beliefs about the universe, and (2) [beliefs about where we are in the universe](#). Exceptions like this [violate an extreme reductive view](#), but are still close enough to count as reductive thinking for my purposes.

Utility Is a Function of Worlds

So we've got the "basically physical" $\omega \in \Omega$. Now we write down a utility function $U(\omega)$. In other words, utility is a [random variable](#) on our event space.

What's the big deal?

One thing this is saying is that *preferences are a function of the world*. Specifically, *preferences need not only depend on what is observed*. This is [incompatible with standard RL in a way that matters](#).

But, in addition to saying that utility can depend on more than just observations, we are *restricting* utility to *only* depend on things that are in the world. After we consider all the information in ω , there cannot be any extra uncertainty about utility -- no extra "moral facts" which we may be uncertain of. If there are such moral facts, they have to be present somewhere in the universe (at least, derivable from facts about the universe).

One implication of this: *if utility is about high-level entities, the utility function is responsible for deriving them from low-level stuff*. For example, if the universe is made of quarks, but utility is a function of beauty, consciousness, and such, then $U()$ needs to contain the beauty-detector and consciousness-detector and so on -- otherwise how can it compute utility given all the information about the world?

Utility Is Computable

Finally, and most critically for the discussion here, $U()$ should be a computable function.

To clarify what I mean by this: ω should have some sort of representation which allows us to feed it into a Turing machine -- let's say it's an infinite bit-string which assigns true or false to each of the "atomic sentences" which describe the world. $U()$ should be a computable function; that is, there should be a Turing machine F which takes a rational number $\epsilon > 0$ and takes ω , prints a rational number within ϵ of $U(\omega)$, and halts. (In other words, we can compute $U(\omega)$ to any desired degree of approximation.)

Why should $U()$ be computable?

One argument is that $U()$ should be computable because the agent has to be able to use it in computations. This perspective is especially appealing if you think of $U()$ as a black-box function which you can only optimize through search. If you can't evaluate $U()$, how are you supposed to use it? If $U()$ exists as an actual module somewhere in the brain, how is it supposed to be implemented? (If you don't think this sounds very convincing, great!)

Requiring $U()$ to be computable may also seem easy. What is there to lose? Are there preference structures we really care about being able to represent, which are fundamentally not computable?

And what would it even mean for a computable agent to have non-computable preferences?

However, the computability requirement is more restrictive than it may seem.

There is a sort of [continuity implied by computability](#): $U()$ must not depend too much on "small" differences between worlds. The computation $F(\epsilon, \omega)$ only accesses finitely many bits of ω before it halts. All the rest of the bits in ω must not make more than ϵ difference to the value of $U(\omega)$.

This means some seemingly simple utility functions are not computable.

As an example, consider the [procrastination paradox](#). Your task is to push a button. You get 10 utility for pushing the button. You can push it any time you like. However, if you never press the button, you get -10. On any day, you are fine with putting the button-pressing off for one more day. Yet, if you put it off forever, you lose!

We can think of ω as a string like 000000100.., where the "1" is the day you push the button. To compute the utility, we might look for the "1", outputting 10 if we find it.

But what about the all-zero universe, 0000000...? The program must loop forever. We can't tell we're in the all-zero universe by examining any finite number of bits. You don't know whether you will eventually push the button. (Even if the universe also gives your source code, you can't necessarily tell from that -- the logical difficulty of determining this about yourself is, of course, the original point of the procrastination paradox.)

Hence, a preference structure like this is not computable, and is not allowed according to the reductive utility doctrine.

The advocate of reductive utility might take this as a victory. The procrastination paradox has been avoided, and other paradoxes with a similar structure. (The [St. Petersburg Paradox](#) is another example.)

On the other hand, if you think this is a *legitimate preference structure*, dealing with such 'problematic' preferences motivates abandonment of reductive utility.

Subjective Utility: The Real Thing

We can strongly oppose all three points without leaving orthodox Bayesianism. Specifically, I'll sketch how the [Jeffrey-Bolker axioms](#) enable non-reductive utility. (The title of this section is a reference to Jeffrey's book *Subjective Probability: The Real Thing*.)

However, the *real* position I'm advocating is more grounded in logical induction rather than the Jeffrey-Bolker axioms; I'll sketch that version at the end.

The View From Somewhere

The reductive-utility view approached things from the starting-point of the universe. Beliefs are for what is real, and what is real is basically physical.

The non-reductive view starts from the standpoint of the agent. Beliefs are for *things you can think about*. This doesn't rule out a physicalist approach. What it *does* do is give high-level objects like tables and chairs an equal footing with low-level objects like quarks: both are inferred from sensory experience by the agent.

Rather than assuming an underlying set of *worlds*, Jeffrey-Bolker assume only a set of events. For two events P and Q, the conjunction $P \wedge Q$ exists, and the disjunction $P \vee Q$, and the negations $\neg P$ and $\neg Q$. However, unlike in the [Kolmogorov axioms](#), these are not assumed to be intersection, union, and complement of an underlying set of worlds.

Let me emphasize that: *we need not assume there are "worlds" at all*.

In philosophy, this is called [situation semantics](#) -- an alternative to the more common [possible-world semantics](#). In mathematics, it brings to mind [pointless topology](#).

In the Jeffrey-Bolker treatment, a world is just a maximally specific event: an event which describes everything completely. But there is no requirement that maximally-

specific events exist. Perhaps any event, no matter how detailed, can be further extended by specifying some yet-unmentioned stuff. (Indeed, the Jeffrey-Bolker axioms assume this! Although, Jeffrey does not seem philosophically committed to that assumption, from what I have read.)

Thus, there need not be any "view from nowhere" -- no semantic vantage point from which we see the whole universe.

This, of course, deprives us of the objects which utility was a function of, in the reductive view.

Utility Is a Function of Events

The reductive-utility makes a distinction between utility -- the random variable itself -- and *expected utility*, which is the subjective estimate of the random variable which we use for making decisions.

The Jeffrey-Bolker framework does not make a distinction. Everything is a subjective preference evaluation.

A reductive-utility advocate sees the expected utility of an event $E \subseteq \Omega$ as **derived from** the utility of the worlds within the event. They start by defining $U(\omega)$; then, we define the expected utility of an event as $E[U|E] := \sum_{\omega} U(\omega)P(\omega)$ -- or, more generally, the corresponding integral.

In the Jeffrey-Bolker framework, we instead define $U(E)$ *directly* on events. These preferences are required to be *coherent with* breaking things up into sums, so $U(E) = \frac{U(F \wedge A) \cdot P(F \wedge A) + U(F \wedge \neg A) \cdot P(F \wedge \neg A)}{P(E)}$ -- but we do not define one from the other.

We don't have to know how to evaluate entire worlds in order to evaluate events. All we have to know is how to evaluate events!

I find it difficult to really believe "humans have a utility function", even approximately -- but I find it *much easier* to believe "humans have expectations on propositions". Something like that could even be true at the *neural* level (although of course we would not obey the Jeffrey-Bolker axioms in our neural expectations).

Updates Are Computable

Jeffrey-Bolker doesn't say anything about computability. However, if we do want to address this sort of issue, it leaves us in a different position.

Because *subjective expectation is primary*, it is now more natural to require that the agent can evaluate events, without any requirement about a function on worlds. (Of course, we *could* do that in the Kolmogorov framework.)

Agents don't need to be able to compute the utility of a whole world. All they need to know is how to update expected utilities as they go along.

Of course, the subjective utility can't be just *any* way of updating as you go along. It needs to be **coherent**, in the sense of the Jeffrey-Bolker axioms. And, maintaining coherence can be very difficult. But it can be quite easy even in cases where the random-variable treatment of the utility function is not computable.

Let's go back to the procrastination example. In this case, to evaluate the expected utility of each action at a given time-step, the agent does not need to figure out whether it ever pushes the button. It just needs to have some probability, which it updates over time.

For example, an agent might initially assign probability $2^{-(t+1)}$ to pressing the button at time t, and 1/2 to never pressing the button. Its probability that it would ever press the button, and thus its utility estimate, would decrease with each observed time-step in which it didn't press the button. (Of course, such an agent would press the button immediately.)

Of course, this "solution" doesn't touch on any of the tricky logical issues which the procrastination paradox was originally introduced to illustrate. This isn't meant as a solution to the procrastination paradox -- only as an illustration of how to coherently update discontinuous preferences. This simple $U()$ is **uncomputable** by the definition of the previous section.

It also doesn't address computational tractability in a very real way, since if the prior is very complicated, computing the subjective expectations can get extremely difficult.

We can come closer to addressing logical issues and computational tractability by considering things in a logical induction framework.

Utility Is Not a Function

In a logical induction (LI) framework, the central idea becomes "*update your subjective expectations in any way you like, so long as those expectations aren't (too easily) exploitable to Dutch-book.*" This clarifies what it means for the updates to be "coherent" -- it is somewhat more elegant than saying "... any way you like, so long as they follow the Jeffrey-Bolker axioms."

This replaces the idea of "utility function" entirely -- there isn't any need for a *function* any more, just a logically-uncertain-variable (LUV, in the terminology from the LI paper).

Actually, there are different ways one might want to set things up. I hope to get more technical in a later post. For now, here's some bullet points:

- In the simple procrastination-paradox example, you push the button if you have any uncertainty at all. So things are not that interesting. But, at least we've solved the problem.

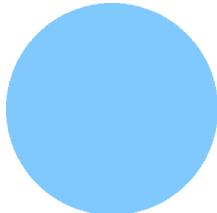
- In more complicated examples -- where there is some real benefit to procrastinating -- a LI-based agent could totally procrastinate forever. This is because LI doesn't give any guarantee about converging to correct beliefs for uncomputable propositions like whether Turing machines halt or whether people stop procrastinating.
- Believing you'll stop procrastinating even though you won't is *perfectly coherent* -- in the same way that believing in [nonstandard numbers](#) is perfectly logically consistent. Putting ourselves in the shoes of such an agent, this just means we've examined our own decision-making to the best of our ability, and have put significant probability on "we don't procrastinate forever". This kind of reasoning is necessarily fallible.
- Yet, if a system we built were to do this, we might have strong objections. So, this can count as an alignment problem. How can we give feedback to a system to avoid this kind of mistake? I hope to work on this question in future posts.

How to evaluate (50%) predictions

I commonly hear (sometimes from very smart people) that 50% predictions are meaningless. I think that this is wrong, and also that saying it hints at the lack of a coherent principle by which to evaluate whether or not a set of predictions is meaningful or impressive. Here is my attempt at describing such a principle.

What are predictions?

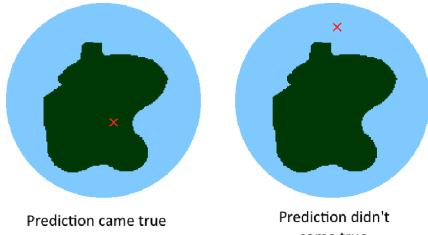
Consider the space of all possible futures:



If you make a prediction, you do this:



You carve out a region of the future space and declare that it occurs with some given percentage. When it comes to evaluating the prediction, the future has arrived at a particular point within the space, and it should be possible to assess whether that point lies inside or outside of the region. If it lies inside, the prediction came true; if it lies outside, the prediction came false. If it's difficult to see whether it's inside or outside, the prediction was ambiguous.

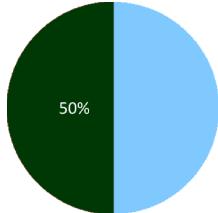


Now consider the following two predictions:

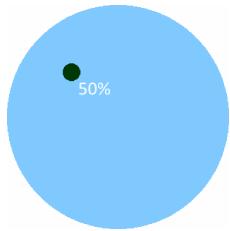
- A coin I flip comes up heads (50%)
- Tesla's stock price at the end of the year 2020 is between 512\$ and 514\$ (50%)

Both predictions have 50% confidence, and both divide the future space into two parts (as all predictions do). Suppose both predictions come true. No sane person would look at them and be equally impressed. This demonstrates that confidence and truth

value are not sufficient to evaluate how impressive a prediction is. Instead, we need a different property that somehow measures 'impressiveness'. Suppose for simplicity that there is some kind of baseline probability that reflects the common knowledge about the problem. If we represent this baseline probability by the size of the areas, then the coin flip prediction can be visualized like so:



And the Tesla prediction like so:

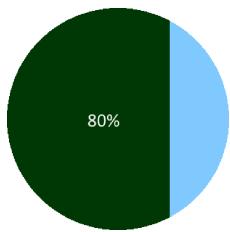


The coin flip prediction is unimpressive because it assigns 50% to a subset of feature space whose baseline probability is also 50%. Conversely, the Tesla prediction is impressive because it assigns 50% to a subset of future space with a tiny baseline probability. Thus, the missing property is the "boldness" of the prediction, i.e., the *(relative) difference between the stated confidence and the baseline probability*.

Importantly, note that we can play the same game at every percentage point, e.g.:

- A number I randomize on random.org falls between 15 and 94 - 80%

Even though this is an 80% prediction, it is still unimpressive because there is no difference between the stated confidence and the baseline probability.



What's special about 50%?

In January, Kelsey Piper predicted that [Joe Biden would be the Democratic Nominee with 60% confidence](#). If this prediction seems impressive now, we can probably agree that this is not so because it's 60% rather than 50%. Instead, it's because most of us would have put it much lower than even 50%. For example, [BetFair gave him only ~15% back in March](#).

So we have one example where a 50% prediction would have been impressive and another (the random.org one) where an 80% prediction is thoroughly unimpressive. This shows that the percentage being 50% is neither necessary nor sufficient for a prediction being unimpressive. Why, then, do people say stuff like "50% predictions aren't meaningful?"

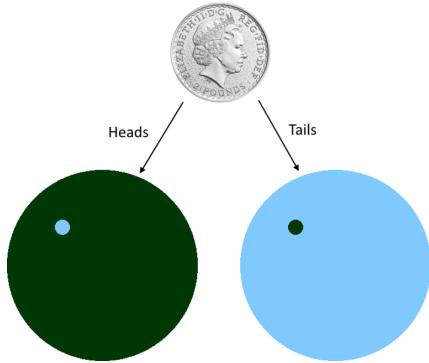
Well, another thing they say is, "you could have phrased the predictions the other way." But there are reasons to object to that. Consider the Tesla prediction:

- Tesla's stock price at the end of the year 2020 is between 512\$ and 514\$ (50%)

As-is, this is very impressive (if it comes true). But now suppose that, instead of phrasing it in this way, we first flip a coin. If the coin comes up heads, we flip the prediction, i.e.:

- Tesla's stock price at the end of the year 2020 is below 512\$ or above 514\$ (50%)

Whereas, if it comes up tails, we leave the prediction unchanged.



What is the probability that we are correct from the point of view we have before the flip? Well, at some point it will be possible to evaluate the prediction, and then we will either be outside of the small blob or inside of the small blob. In the first case, we are correct if we flipped the prediction (left picture). In the latter case, we are correct if we didn't flip the prediction (right picture). In other words, in the first case, we have a 50% chance of being correct, and in the latter case, we also have a 50% chance of being correct. Formally, for any probability p that the future lands in the small blob, the chance for our prediction to be correct is exactly

$$(1 - p) \cdot \frac{1}{2} + p \cdot \frac{1}{2} = \frac{1}{2}$$

Importantly, notice that this remains true regardless of how the original prediction divides the future space. The division just changes p , but the above yields $\frac{1}{2}$ for every value of p .

Thus, given an arbitrary prediction, if we flip a coin, flip the prediction iff the coin came up heads and leave it otherwise, we have successfully constructed a perfect 50% prediction.

Note: if the coin flip thing seems fishy (you might object that, in the Tesla example, we either end up with an overconfident prediction or an underconfident prediction, and they can't somehow add up to a 50% prediction), you can alternatively think of a set of predictions where we randomly flip half of them. In this case, there's no coin involved, and the effect is the same: half of all predictions will come true (in expectation) regardless of their original probabilities. Feel free to re-frame every future mention of coin flips in this way.

This trick is *not* restricted to 50% predictions, though. To illustrate how it works for other percentage points, suppose we are given a prediction which we know has an 80% probability of coming true. First off, there are three simple things we can do, namely

- leave it unchanged for a perfect 80% prediction
- flip it for a perfect 20% prediction
- do the coin flip thing from above to turn it into a perfect 50% prediction.
Importantly, note that we would only flip the prediction statement, not the stated confidence.

(Again, if you object to the coin flip thing, think of two 80% predictions where we randomly choose one and flip it.)

In the third case, the formula

$$(1 - p) \cdot \frac{1}{2} + p \cdot \frac{1}{2} = \frac{1}{2}$$

from above becomes

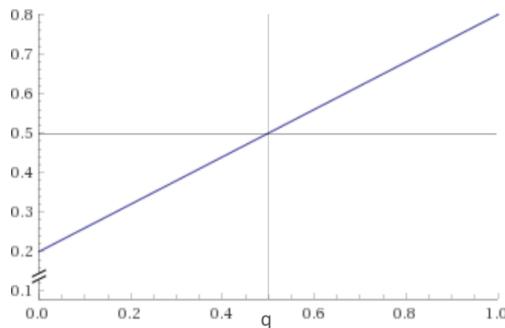
$$0.2 \cdot \frac{1}{2} + 0.8 \cdot \frac{1}{2} = \frac{1}{2}$$

This is possible no matter what the original probability is; it doesn't have to be 80%.

Getting slightly more mathy now, we can also throw a biased coin that comes up heads with probability $q \neq \frac{1}{2}$ and, again, flip the prediction iff that biased coin came up heads. (You can still replace the coin flip; if $q = \frac{1}{3}$, think of flipping every third prediction in a set.) In that case, the probability of our prediction coming true is

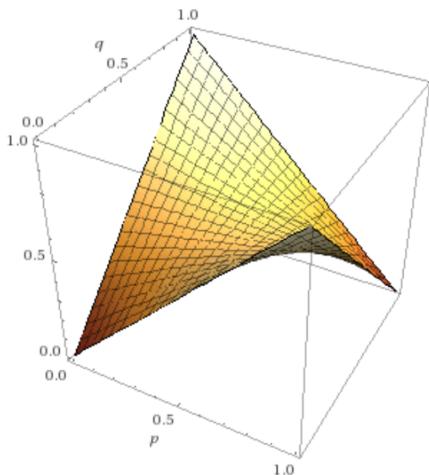
$$0.2 \cdot q + 0.8 \cdot (1 - q)$$

This term takes values in the interval $[0.2, 0.8]$. Here's the graph:



Thus, by flipping our prediction with some probability other than $\frac{1}{2}$, we can obtain every probability within $[0.2, 0.8]$. In particular, we can transform an 80% probability into a 20% probability, a 30% probability, a 60% probability, a 78.3% probability, but we cannot make it an 83% or a 13% probability.

Finally, the formula with a variable prior probability and a variable flip chance is $(1 - p)q + p(1 - q)$, and its graph looks like this:



If you fix p , you'll notice that, by changing q , you get the y-value fluctuating between p and $1 - p$. For $q = \frac{1}{2}$, the y-value is a constant at $\frac{1}{2}$. (When I say y-value, I mean the result of the formula which corresponds to the height in the above picture.)

So it is always possible to invert a given probability or to push it toward 50% by arbitrarily introducing uncertainty (this is sort of like throwing information away). On the other hand, it is never possible to pull it further away from 50% (you cannot create new information). If the current probability is known, we can obtain any probability we want (within $[p, 1 - p]$); if not, we don't know how the graph looks/where we are on the 3d graph. In that case, the only probability we can safely target is 50% because flipping with $\frac{1}{2}$ probability (aka flipping every other prediction in a set) turns every prior probability into 50%.

And this, I would argue, is the only thing that is special about 50%. And it doesn't mean 50% predictions are inherently meaningless; it just means that cheating is easier – or, to be more precise, cheating is possible without knowing the prior probability. (Another thing that makes 50% seem special is that it's sometimes considered a universal baseline, but this is misguided.)

As an example, suppose we are given 120 predictions, each one with a correct probability of 80%. If we choose 20 of them at random and flip those, 70% of all predictions will come true in expectation. This number is obtained by solving $0.2q + 0.8(1 - q) = 0.7$ for q ; this yields $q = \frac{1}{6}$, so we need to flip one out of every six predictions.

What's the proper way to phrase predictions?

Here is a simple rule that shuts the door to this kind of "cheating":

Always phrase predictions such that the confidence is above the baseline probability.

Thus, you should predict

- Joe Biden will be the Democratic nominee (60%)

rather than

- Joe Biden will not be the Democratic nominee (40%)

because 60% is surprisingly high for this prediction, and similarly

- The price of a barrel of oil at the end of 2020 will be between \$50.95 and \$51.02 (20%)

rather than

- The price of a barrel of oil at the end of 2020 will not be between \$50.95 and \$51.02 (80%)

because 20% is surprisingly high for this prediction. The 50% mark isn't important; what matters is the confidence of the prediction relative to the baseline/common wisdom.

This rule prevents you from cheating because it doesn't allow flipping predictions. In reality, there is no universally accessible baseline, so there is no formal way to detect this. But that doesn't mean you won't notice. The list:

- The price of a barrel of oil at the end of 2020 will be between \$50.95 and \$51.02 (50%)
- Tesla's stock price at the end of the year 2020 is between 512\$ and 514\$ (50%)
- ... (more extremely narrow 50% predictions)

which follows the rule looks very different from this list (where half of all predictions are flipped):

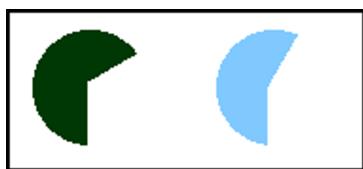
- The price of a barrel of oil at the end of 2020 will be between \$50.95 and \$51.02 (50%)
- Tesla's stock price at the end of the year 2020 is below 512\$ or above 514\$ (50%)
- ... (more extremely narrow 50% predictions where every other one is flipped)

and I would be much more impressed if the first list has about half of its predictions come true than if the second list manages the same.

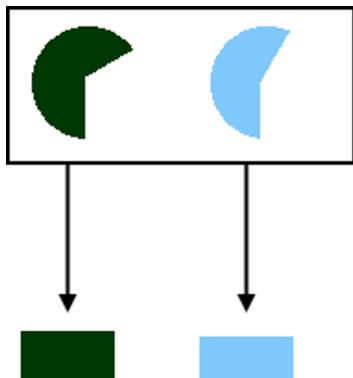
Other than preventing cheating, there is also a more fundamental reason to follow this rule. Consider what happens when you make and evaluate a swath of predictions. The common way to do this is to group them into a couple of specific percentage points (such as 50%, 60%, 70%, 80%, 95%, 99%) and then evaluate each group separately. To do this, we would look at all predictions in the 70% group, count how many have come true, and compare that number to the optimum, which is

$0.7 \cdot \# \text{ predictions in that group.}$

Now think of such a prediction like this:

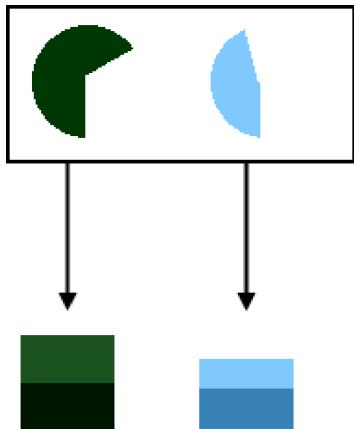


Namely, there is a baseline probability (blue pie, ~60%) and a stated confidence (green pie, 70%). When we add such a prediction to our 70% group, we can think of that like so:



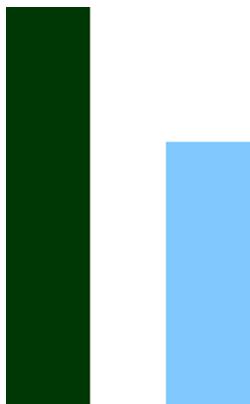
We accumulate a confidence pile (green) that measures how many predictions we claim will come true, and a common wisdom pile (blue) that measures how many predictions ought to come true according to common wisdom. After the first prediction, the confidence pile says, "0.7 predictions will come true," whereas the common wisdom pile says, "0.6 predictions will come true."

Now we add the second (70% confidence, ~45% common wisdom):



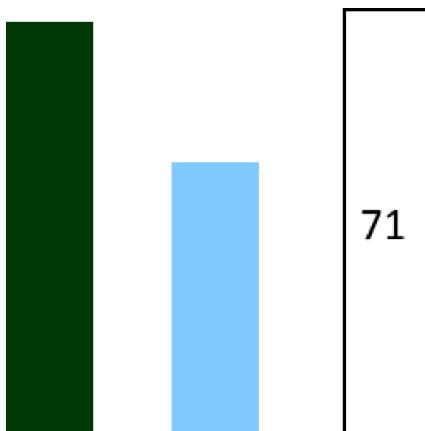
At this point, the confidence pile says, "1.4 predictions will come true," whereas the common wisdom pile says, "1.05 predictions will come true."

If we keep doing this for all 70% predictions, we eventually end up with two large piles:



The confidence pile may say, "70 predictions will come true," whereas the common wisdom pile may say, "48.7 predictions will come true."

Then (once predictions can be evaluated) comes a third pile, the reality pile:



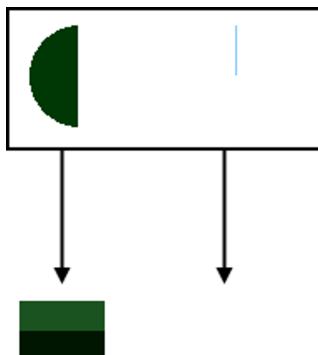
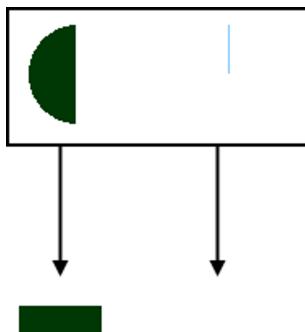
The reality pile counts how many predictions did, in fact, come true. Now consider what this result means. We've made lots of predictions at 70% confidence for which common wisdom consistently assigns lower probabilities. In the end, (slightly more than) 70% of them came true. This means we have systematically beaten common wisdom. This ought to be impressive.

One way to think about this is that the difference between the confidence and common wisdom piles is a measure for the boldness of the entire set of predictions. Then, the rule that [each prediction be phrased in such a way that the confidence is above the baseline probability] is equivalent to choosing one of two ways that maximize this boldness. (The other way would be to invert the rule.)

If the rule is violated, the group of 70% predictions might yield a confidence pile of a height similar to that of the common wisdom pile. Then, seeing that the reality pile matches them is much less impressive. To illustrate this, let's return to the example from above. In both cases, assume exactly one of the two predictions comes true.

Following the rule:

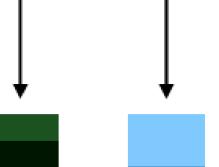
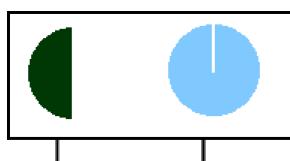
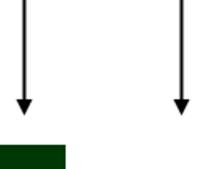
- The price of a barrel of oil at the end of 2020 will be between \$50.95 and \$51.02 (50%)
- Tesla's stock price at the end of the year 2020 will be between 512\$ and 514\$ (50%)



Bold, therefore impressive.

Violating the rule:

- The price of a barrel of oil at the end of 2020 will be between \$50.95 and \$51.02 (50%)
- Tesla's stock price at the end of the year 2020 will be below 512\$ or above 514\$ (50%)



Not bold at all, therefore unimpressive. And that would be the reason to object to the claim that you could just phrase 50% predictions in the opposite way.

Note that the 50% group is special insofar as predictions don't change groups when you rephrase them, but the principle nonetheless applies to other percentage points.

Summary/Musings

According to this model, when you make predictions, you should follow the confidence > baseline rule; and when you evaluate predictions, you should

- estimate their boldness (separately for each group at a particular percentage point)
- be impressed according to the product of calibration · boldness (where calibration is how closely the reality pile matches the confidence pile, which is what people commonly focus on)

Boldness is not formal because we don't have universally accessible baseline probabilities for all statements lying around (50% is a non-starter), and I think that's the primary reason why this topic is confusing. However, baselines are essential for evaluation, so it's much better to make up your own baselines and use those than to use a model that ignores baselines (that can give absurd results). It does mean that

the impressiveness of predictions has an inherent subjective component, but this strikes me as a fairly intuitive conclusion.

In practice, I think people naturally follow the rule to some extent - they tend to predict things they're interested in and then overestimate their probability – but certainly not perfectly. The rule also implies that one should have separate groups for 70% and 30% predictions, which is currently not common practice.

Taking Initial Viral Load Seriously

In part a response to (Overcoming Bias): [Variolation May Cut Covid19 Deaths 3-30X](#)

Also related: This is an expansion of parts of [this debate with Robin Hanson](#).

Epistemic Status: Thinking out loud. Food for thought. Not an expert.

Initial viral load seems likely to have a large impact on severity of Covid-19 infection. If we believe this, we should take this seriously, and evaluate both general policy and personal behavior differently in light of this information. We should also do our best to confirm or deny this hypothesis as soon as possible.

[Robin Hanson and I had a debate](#) on Sunday regarding his proposal of variolation: Deliberate Covid-19 infection of volunteers, using small viral loads, followed by isolation of those volunteers until recovery. I felt it was an excellent discussion. As with all such discussions, some key points are glossed over, lost or not stated as well as one would like in hindsight, there are areas that never come up, and one thinks of important additional things afterwards.

What evidence do we have that viral load matters?

Three classes of evidence seem strong.

The first is that we have a strong mechanism story we can tell. Viruses take time to multiply. When the immune system detects a virus it responds. If your initial viral load is low your immune system gets a head start, so you do better.

The second category is the terrible outcomes in health care workers on the front lines. Those who are dealing with the crisis first hand are dealing with lots of intense exposures to the virus. When they do catch it, they are experiencing high death rates. High viral load is the only theory I know about so far for why this is the case. Their cases are presumably handled at least as well as others, in terms of detection, testing, treatment and what the infected do themselves. The only other issue I can think of is that they might be reluctant to rest given how urgently their help is needed.

The third category is historical precedents.

Robin's proposal for variolation is similar to what was historically done with smallpox. Parents infected their children with what they hoped was exactly the minimum dose required to get them sick enough to develop antibodies and gain immunity. Sometimes this went wrong and the child would get sick. Thus this form of inoculation was dangerous and 1%-2% of patients died. But of those who got smallpox infections in other ways, 20%-30% of patients died. Those rates are well established.

Another classic example is measles. We have a study that says that the first child in a family to contract the disease was relatively safe, whereas other children in the same family were 14 times as likely to die. That's another huge gap. This is [from a study of 126 children](#), so much less certain, but effect sizes that big are not accidents.

Finally for SARS, we have the Hong Kong high rise [where proximity to the index patient played a crucial role](#). Here we see a factor of three difference. SARS being so

deadly that 70%+ of highly exposed patients died could be a reason the ratio of deaths between high and low viral load was only three to one.

There are also claims that we can observe higher viral loads from lockdowns making things worse and increase death rates, allowing us to be more confident that viral loads have a big impact. I do not find this convincing because of likely reverse causation and other highly non-random factors that determined where lockdowns happened earlier. Later on I'll analyze the likely impact of lockdowns on viral load, where I get a different answer than Robin's.

It's worth noting that we have vaccines for smallpox and measles, and the SARS virus in question was contained, so none of these three are remotely candidates for variolation no matter the true ratios.

What is the evidence against viral load having a big impact?

The evidence of absence is the absence of further evidence. Yes, the effects observed were very big. Yes, they are all the result of natural experiments and we have been ethically precluded from doing randomized trials or other better studies, so the lack of those trials doesn't mean much. But how likely is it we didn't find more natural experiments where viral load would have differed in an observable way? Publication bias here could be a large effect.

These effect sizes are super large. It seems odd for them to hold for some viruses, and not to have a large effect in most others. But if that's true, how is it we are not observing them? Maybe we really don't ever look so it's never come up. But it definitely seems odd.

There's also the small sizes for the measles and SARS studies, which have about 200 patients between them. That does not seem like enough to draw general conclusions with any confidence.

The smallpox case represents the best case, but it also represents an engineered best case of exactly the minimum dose in exactly the right way, along with awareness of the situation versus a society that otherwise had little idea how to handle infections. It also represents the case in which things went so well that ordinary people managed to fight through all the barriers and actually do the thing. It makes sense to assume this was an unusually large effect size.

We have to ask why smallpox was a unique event, and we never used this method for any other virus. Did we even ever consider it?

My prior at this point is that the difference between a low and high initial viral load of Covid-19 is large. The theory makes too much sense. But with high uncertainty.

Suppose we agree that it is likely. At a minimum, there is a large chance of a large effect, and essentially zero chance of a backfiring effect - at worst, a low dose is the same as a high one.

What should we do? How do we take this seriously?

Four categories of things one might do.

1. There are things we could do to get better information.
2. There are things individuals or small groups can do to improve their situation.

3. There are things society as a whole could try to do that don't have big downsides.
4. We could take bold action, potentially including variolation.

Category 1: Better Information

The failure to collect more and better information about Covid-19 has been atrocious, shameful, expensive and deadly.

We shouldn't only be doing a crash project for a vaccine and ramping up testing much faster and immediately testing every treatment that shows any promise.

We should be collecting extensive population-level data everywhere. Mostly we're not collecting any data at all that isn't massively biased.

We should be studying with experiments how Covid-19 spreads, and how likely each method is to work, using controlled experiments. Yes, this involves infecting individuals. Considering how many lives are at stake and the ability to test using young healthy volunteers who are then isolated, I fail to see how anyone who objects on the basis of 'ethics' knows what that word means, or why we should listen to them.

We should use those experiments, and additional experiments if necessary, to study effects of viral load. Because again, that knowledge would save so many lives, in addition to tons of economic distress that could force horrible choices upon us.

We can shut down the entire economy, force people to stay home and create double digit unemployment. And we can't do this? Really?

The more I think about the Covid-19 situation, the more I think the highest leverage thing most people reading this can do is to find ways to get our hands on better data.

Better information would (of course) make it much easier to know whether it made sense to take other action, know which actions made sense, and gain support for the right ones, across the board. As we go over what else we might do, all of that will emphasize how valuable more information would be.

My current best thought for how to do experiments quickly is medical cruise ships in international waters. We even have lots of spare cruise ships lying around with nothing to do right now, which we could convert if we needed to. Medical cruise ships are already an established way to do things without running into regulatory problems. We could do things properly, in a way that would give trustworthy results and would allow others to trust the data. The tests required are not expensive or difficult to produce if no FDA regulations get in one's way. Such projects are well within the 'Bill Gates decides to just go ahead and do this' price range.

Still, one must be realistic. Given we likely won't get much good information soon, and time is very short, I'll proceed as if we are physically prevented from information gathering.

Category 2: Things Individuals or Small Groups Can Do

Bird's Eye View

Things individuals can do are a good place to start, because ‘scale up what an individual should do’ is an excellent hypothesis for much of what a society should do. I agree with Robin Hanson that people who are at risk, and say they are more concerned with infecting others than becoming infected, are usually wrong about that. I still think most people care a lot about preventing others from becoming infected. Incentives, in most places, are not so misaligned.

The first question one would ask is, what infection methods result in a high load, versus which methods result in a low load. Or to be more granular, what lowers or raises such loads.

The default model is that the longer and more closely you interact with an infected person, especially a symptomatic infected person, the larger your viral load.

In-household infections are presumed to be high viral load, as in the case of measles. So would be catching the infection while treating patients.

Most out-of-household infections that aren’t health care related are presumed to be low viral load. Anything outdoors is probably low viral load. Most methods that involve surfaces are probably low viral load. Infection via the air from someone there half an hour ago, to the extent this is a thing, is low viral load. Quick interactions with asymptomatic individuals are probably low viral load.

Knowledge would of course be far better than supposition. If anyone has actual information on any of this, please do share it in the comments.

The only method that seems highly ambiguous is the fecal-oral transmission route. We don’t know how dangerous it is either in terms of infection probability or likely viral load. It could be anything from almost harmless to very dangerous.

To a first approximation, we can say that a typical person (who is not a health care worker) should consider it more deadly to be infected via household transmission. They should consider it less deadly to be infected out-of-household.

We can also consider it relatively more deadly when the infection risk from a given source was otherwise high, or one might be infected from multiple sources at once. When one is exposed to a low-probability infection from a small number of sources, expected viral load is low.

How Many Infections are High Versus Low Load Now?

One place I disagree with Robin Hanson’s analysis is that he is comparing an intervention to create universal low viral loads to an alternative of mostly high viral loads.

I believe he is discounting the extent to which many infections are already low viral load. Consider our intuitive model, and infection rates from the poorly named category ‘close contacts.’

If you had an infected ‘close contact’ in Wuhan, [your probability of infection was only 2.5%](#). Close contacts within the household were much more likely than those outside it to pass along the infection. If we focus on asymptomatic close contacts outside the household, we match our intuition that infection is possible but any given person is

unlikely to infect us – the paper gives the probability of infection from each asymptomatic ‘close contact’ that wasn’t all that close at only 0.03%!

Under current lock down conditions, where anyone who got this far into this post (again, who doesn’t have an essential job that prevents this) is presumably avoiding anyone symptomatic, and most symptomatic people are self-isolating or seeking help, it does not seem so hard for most of us to avoid direct out-of-household interaction with highly symptomatic people.

Thus, let’s simplify the baseline for those taking ‘ordinary precautions.’ Out of household transmission is low load. Inside of household transmission is high load. To balance this out, let’s presume that if anyone in a household gets infected, they are probably going to infect the rest of the household, and we’ll set that probability at 100%.

If you are the only member of your household, unless you take big risks or something highly unlikely happens, you’re going to have a low viral load.

If you are one of two members, then one of you will get a low load, and one of you will then probably get a high load.

Overall infection rates are unknown due to inexcusably poor data, but we do know that the probability of infection *on a given day for a given person who is taking precautions* is low even if many are currently infected around them. Thus, unless household members are exposed from the same source at the same time, the probability of simultaneous infection that gets them both a low viral load is very low and can be rounded off to zero.

The biggest factor thus effectively becomes, how big is your household?

The average household size in the United States of America is about 2.6 people. If we exclude children, the average household size is 2.02 people, which we’ll round to 2.

Children can spread the infection, but they probably are not as effective at it and they do not have essential work outside the home. It should mostly be easy, once schools have closed, to avoid having them get infected before other household members. If they are infected in-household, we don’t care much if they have high viral loads since their risk remains very low.

Thus, of those that matter and who reside in a household, approximately 50% will get low viral loads in a lock down. This will *briefly* be false in the first week of a lock down, since you clamp down on out-of-household infections but not inside-of-household infections during that period. But longer term, the impact of higher-probability inside-of-household transmission does not matter very much, because the probability of inside-of-household transmission, once someone got infected, was so high already. The only effective defense is to stop out-of-household transmission so everyone in the household stays uninfected in the first place.

Lock downs seem like they would decrease viral load conditional on an out-of-household infection, as long duration in-person indoor interactions, and other ways to get a high viral load, seem to be down a lot more than other out-of-household transmission vectors. So intuitively, lock downs are actually net good for viral load.

The exception is if they change household size, by causing people to shelter together, which brings us back to what groups can do.

We also need to consider what to do about those in institutions, especially nursing homes. That could drive these numbers higher, if such people are effectively inside the same household, and that can't be fixed.

Looking at a 50% low risk, 50% high risk scenario, we can only save 50% of what we could save if we started in a 50% high risk scenario. So a range of factors of 3-30, which I already discounted because of selection bias in the evidence, can further be cut in half. That gives us a reasonable range of room for improvement of about a factor of 1.5-10. Still a big win! That could be higher if we could do even lower initial loads than what we are calling 'low load' but a procedure that is <0.1% to infect you already seems likely to functionally be close to the minimum dose.

One's Own Risk

If you are living alone, that seems *great*.

By definition, you can't be infected inside-of-household with a household size of one. Thus, unless you need to do other high-risk activity, your viral load will be low.

If you are already mostly self-isolating, then you can worry less about incoming holes in your containment procedures. If you do get infected, you're unlikely to pass it along, your risk will be lower than you would otherwise calculate,

There is an argument that if you're alone and not going outside for a month conditional on getting packages, and you are young and healthy, you can stop worrying about getting Covid-19 from those packages. If it did happen, you are perhaps doing a solid approximation of variolation, after which you will be immune. That might actively be better than wiping the packages down or letting them sit for three days.

Two to Tango

If there are two people in the household, the calculus changes. We also get some non-intuitive results – since infecting anyone in the household probably infects everyone, protecting the more vulnerable member of the household means *you want them to be infected first*.

The first thing to note is that, in theory, sufficiently *correlated* risk to both individuals seems actively great. Both of you getting infected at the same time means you both get low viral loads. This is not easy to pull off, since most risks worth taking are low probability even when they are risky (e.g. <1% chance of infection even if the close contact is infected) so the chance of actually catching it at the same time is still low. But it could be worth thinking about.

The second thing is that people may be allocating infection risk *backwards* by default. Suppose you have a high risk old individual in poor health, and a low risk young individual in good health. You can't get groceries delivered and supplies are running low, so someone has to run to the store. Who goes?

Intuition says of course the individual in good health should take the risk. But if we take the whole model seriously, that seems wrong now! If the healthy individual gets infected first then you have a low-risk infection in a low-risk individual (small win) who

then infects the high-risk individual and puts them at high-risk again (big loss). You would want to reverse that process.

The counterargument here is that there is some evidence that higher risk individuals are also going to catch the virus more easily, which could prevent the situation from reversing. Again, we need better data!

The third thing is that social distancing *within the household* starts to look more valuable. Viral load depends on the exact interactions that take place. It makes sense to place at least some value on minimizing infection exposures between household members, *even if you are mostly resigned to the infection eventually happening once anyone gets infected*. There is still something important to win.

A household with varying risk levels due to age or health, to the extent possible, might choose to concentrate as much within-household exposure as possible in the less risky direction. For example, perhaps the person at high risk should be doing most of the cooking.

If one person does show symptoms first, precautions then become super valuable, even if they probably ultimately fail to prevent spread.

Three is a Crowd

Larger households mean bigger risk of high viral loads. Such groups have both a higher risk of infection at all (since any of them could get infected) they also carry a bigger risk via higher expected viral loads. Thus, as a group scales higher, precautions become even more important.

The obvious strategy is to *break up large households*. If you have more than two adults in a group, consider that an even higher cost. If you have a lot more than two, this gets extreme. And as before, if you have low-risk individuals together with high-risk ones, having the low-risk individuals take all the risks looks pretty bad. If you do that, it might make sense to consider social distancing within the household where feasible, if it can turn potential inner-household transmissions into effectively out-of-household transmissions. This is especially true if you have reason to believe someone may have been exposed recently.

If there are multiple households considering interacting with each other, that now looks like a bigger additional risk as well. Those types of interactions could effectively be inner-household style contacts.

For those with family members or housemates who insist on taking risks, you should be even more worried than before. The people living with the risk-taker are now at greater risk than the risk-taker themselves. It becomes that much more important to do something about such actions.

Category 3: What Society Can Easily Do (Beyond Gathering Data)

If we think there are things individuals should do, society's first job is to encourage those people to do those things. Thus, everything in category two could be added to existing lists of recommendations.

But that's not a practical answer. Bandwidth is limited. People are already overloaded with information, and with disinformation.

If [you only get about five words](#), those five words need to be something like "socially distance, wash hands, mask." Even not touching your face wouldn't make the cut. We still have major organizations continuing to actively spread disinformation against even that level of response.

Everything we go over here is therefore probably far too subtle. It also risks muddling the messaging on more important matters. Any intervention *for the public* as opposed to people who read long analytical blog posts needs to have a focused simple message that could be attached at the end of the current list of simple useful interventions.

Another argument against spending bandwidth on this is that minimizing an individual's viral load does not much help bend or smash the curve, and anything that reduces R₀ is higher leverage and takes priority over other actions. The obvious retort is that lots of people in my circles and other circles are focusing on ventilators.

The simple message of *minimize household size* seems like a reasonable candidate. Since a large percentage of infections are within-household, this seems like a strong intervention even without viral load concerns. Perhaps we want to emphasize this more. We could use the viral load argument as an additional justification for an already good intervention.

A second strategy would be to ban activities that lead to high viral loads but not ban those that lead to low viral loads. This could make infections less deadly while doing less economic damage. The risk is that there are still a lot of large households out there. It seems beyond our abilities to say "people living alone and not working in a highly interactive position can do X, but not people living with others or working in ways that interact with others." I would love to be wrong about that. Still, some amount of adjustment of what we do and don't permit or encourage, on the margin, would be helpful.

It seems hard to do more than that on a broad basis until we have much better data. So again, the most low-hanging fruit is to gather data. Run experiments. Do more and faster science to it.

Category 4: Bold Action

Suppose we manage to gather better data and it turns out viral loads are a big deal.

Say, we become confident that minimal loads have a 0.1% death rate with proper medical care and high loads have a 2% death rate with proper medical care, and lock down conditions have a roughly 50/50 mix of both and a 1% death rate. That's a bigger effect than I expect, but very much in the realm of the possible.

That difference is a really, really big deal. It's a much bigger deal than getting enough ventilators. It's potentially a bigger deal than having a medical system at all. Alternatively, it's potentially a bigger deal than the difference between 100% infection rates and the infection rate a few weeks from now (or in some places like Spain or New York City, potentially the infection rate today). It's not enough to overcome both of those differences at once.

It is certainly a big enough difference to justify bold action to minimize high viral load infections among our most vulnerable populations.

Suppose we got our act together. We want to do more than nudge individual behavior. We are willing to do things that people find instinctively repugnant, provided they save lives while at least not hurting the economy. How could we accomplish this?

The possibilities I can see are subsidizing household divisions, strategic variolation of individuals, variolation of the young and healthy, or variolation of the old and vulnerable.

Subsidizing household divisions means exactly what it sounds like. We could offer tax or other incentives, up to and including providing housing and forcing people to use it, in order to break up sufficiently large groups of adults, or sufficiently large groups of adults that contain at least one at-risk member (e.g. someone over the age of 60, or 70). Given the economic costs of shutdown, actions that involve spending large amounts of money and/or using large amounts of coercion are being underconsidered in general. We are already using lots of coercion and paying gigantic economic costs! The difference is we are doing so passively, via preventing actions, rather than actively via causing actions. That's not a hill I want us to (literally) die on.

Then there are variolation strategies, where we deliberately infect individuals. Or alternatively, if we find out-of-household infections are low enough viral load, perhaps we could do this passively via allowing those in single-adult households to do things that cause them to infect each other in such ways, while otherwise taking strong precautions so they avoid infecting others.

There are a lot of bad objections to such policies. I won't waste space addressing them other to note that they present practical barriers to implementation, and that they force us to be sure to do all of this carefully and correctly to have a chance of avoiding being shut down or worse.

One very strong objection is that our medical system is about to be overwhelmed. Anyone we expose now will either take needed resources away from others, or go without those resources. Thus, we either need to be in a world in which everyone is already going to get infected while the system is overwhelmed and we can't stop it (in which case perhaps the best we can do is get remaining people low viral loads and hope for the best), or a world in which the virus has been contained for now but where we don't have a long term plan that can keep it that way (e.g. some places can squash and keep their medical systems stable for a while, but they can't stop reintroductions while reopening the economy, and the economic costs of doing this for long enough to wait for a vaccine or other solution are not an option).

Infecting the young and healthy is the natural first thing to model. The young are at relatively low risk. That risk is not zero, even if we assume big impacts from low viral loads, screening for comorbidities and ensure good at-home care. But if such people will eventually probably get infected anyway, we can reduce their net risk while allowing them to return to work and other activity much faster. Then that slows the further spread of the virus, hopefully allowing more high-risk individuals to never get infected at all.

Robin's current best concrete suggestion, once we have established the safety and value of the procedure via testing, is to create variolation villages. Those who voluntarily participate and are deemed healthy enough would be isolated and infected, we would verify the infection and then allow them free access to the village

until they were safely cured and non-infectious. Then they could return to their lives and move freely.

One could respond that this is exactly backwards. If low viral loads are a big win, why are we protecting those who least need that protection? Why aren't we protecting those who most need it? This argues that, given we will have limited capacity, we should instead look to variolate the old and at risk, since we are reducing risk and they benefit from this the most.

Going down this path means we've concluded that protecting them, via herd immunity from the young or via general suppression or otherwise, is not realistic. These are exactly the people who ideally we don't let get infected at all. If they do get infected, even carefully, they will need a lot of care. Doing this requires even more ideal conditions than infecting the young. We either need a lot of spare medical resources without having much hope of long term containment, or we need to have essentially no hope of stretching things out very far before most are infected.

Those who are capable of sustained safe isolation would want to avoid participating even under the best conditions.

That leaves what I am calling strategic variolation. Rather than taking whoever volunteers, or sorting by age and health, we choose people who (both volunteer and) provide superior leverage. Look for those who would otherwise be forced to expose themselves to high viral loads or lots of interactions. Alternatively, look for activities that cannot be done while social distancing, but which have very high value. Focus on those categories of individuals and at least give them priority. Alas, many would consider this an even worse look than a general call for volunteers, so much so that I am not naming anyone I would prioritize. If we got farther along this path, there would be plenty of time to discuss that.

Conclusion

Viral loads are not being taken seriously. We should take them seriously.

On an individual and household level, that means thinking carefully about how to avoid high viral load infections especially for those most at risk.

On a societal level, that means gathering much better data about how impactful this factor is and how it works (and about all other aspects of what is happening, of course), so we can consider taking bold action if appropriate. It also likely means encouraging smaller household groups during the pandemic.

Remember, I am not an expert. This is only me thinking out loud.

Atari early

Deepmind [announced](#) that their Agent57 beats the ‘human baseline’ at all 57 Atari games usually used as a benchmark. I think this is probably enough to resolve one of the predictions we had respondents make in our 2016 survey.

Our question was when it would be feasible to ‘outperform professional game testers on all Atari games using no game specific knowledge’.¹ ‘Feasible’ was defined as meaning that one of the best resourced labs could do it in a year if they wanted to.

As I see it, there are four non-obvious things to resolve in determining whether this task has become feasible:

- Did or could they outperform ‘professional game testers’?
- Did or could they do it ‘with no game specific knowledge’?
- Did or could they do it for ‘all Atari games’?
- Is anything wrong with the result?

I. Did or could they outperform ‘professional game testers’?

It looks like yes, for at least for 49 of the games: the ‘human baseline’ seems to have come specifically from professional game testers.² What exactly the comparison was for the other games is less clear, but it sounds like what they mean by ‘human baseline’ is ‘professional game tester’, so probably the other games meet a similar standard.

II. Did or could they do it with ‘no game specific knowledge’?

It sounds like their system does not involve ‘game specific knowledge’, under a reasonable interpretation of that term.

III. Did or could they do it for ‘all Atari games’?

Agent57 only plays 57 [Atari 2600](#) games, whereas [there are hundreds](#) of Atari 2600 games (and [other Atari consoles](#) with presumably even more games).

Supposing that Atari57 is a longstanding benchmark including only 57 Atari games, it seems likely that the survey participants interpreted the question as about only those games. Or at least about all Atari 2600 games, rather than every game associated with the company Atari.

Interpreting it as written though, does Agent57’s success suggest that playing all Atari games is now feasible? My guess is yes, at least for Atari 2600 games.

Fifty-five of the fifty-seven games were proposed in [this paper](#)³, which describes how they chose fifty of them:

Our testing set was constructed by choosing semi-randomly from the 381 games listed on Wikipedia [http://en.wikipedia.org/wiki/List_of_Atari_2600_games (July 12, 2012)] at the time of writing. Of these games, 123 games have their own Wikipedia page, have a single player mode, are not adult-themed or prototypes, and can be emulated in ALE. From this list, 50 games were chosen at random to form the test set.

The other five games in that paper were a ‘training set’, and I’m not sure where the other two came from, but as long as fifty of them were chosen fairly randomly, the provenance of the last seven doesn’t seem important.

My understanding is that none of the listed constraints should make the subset of games chosen particularly easy rather than random. So being able to play these games well suggests being able to play any Atari 2600 game well, without too much additional effort.

This might not be true if having chosen those games (about eight years ago), systems developed in the meantime are good for this particular set of games, but a different set of methods would have been needed had a different subset of games been chosen, to the extent that more than an additional year would be needed to close the gap now. My impression is that this isn’t very likely.

In sum, my guess is that respondents usually interpreted the ambiguous ‘all Atari games’ at least as narrowly as Atari 2600 games, and that a well resourced lab could now develop AI that played all Atari 2600 games within a year (e.g. plausibly DeepMind could already do that).

IV. Is there anything else wrong with it?

Not that I know of.

~

Given all this, I think it is more likely than not that this Atari task is feasible now. Which is interesting, because the [median 2016 survey response](#) put a 10% chance on it being feasible in five years, i.e. by 2021.⁴ They more robustly put a median 50% chance on ten years out (2026).⁵

It’s exciting to start resolving expert predictions about early tasks so we know more about how to treat their later predictions about human-level science research and the obsolescence of all human labor for instance. But we should probably resolve a few more before reading much into it.

At a glance, some other tasks which we might be able to resolve soon:

- The ‘reading Aloud’ task⁶ [seems to be coming along](#) to my very non-expert ear, but I know almost nothing about it.
- It seems like we are [close on Starcraft](#) though as far as I know the prediction hasn’t been exactly resolved as stated.⁷
- Human-level Angry Birds play⁸ was forecast for four years out with 50% chance, and [hasn’t happened](#) yet. I note that the [contest website cites our survey as pressure to do better](#), slightly complicating things.
- AI that could ‘play poker well enough to win the World Series of Poker’ had a median 50% forecast in three years (i.e. 2019). In 2019, [a system beat elite professional players at six-player no-limit Texas hold'em including Chris ‘Jesus’ Ferguson 2000 winner of World Series main event](#). However World Series contains several versions of Poker, so it isn’t clear that AI could actually win the World Series. I’m not familiar enough with Poker to say whether any of the differences between Texas Hold’em, Omaha Hold’em and Seven Card Stud should make the latter two difficult if the first is now feasible.

By Katja Grace

Thanks to Rick Korzekwa, Jacob Hilton and Daniel Filan for answering many questions.

Notes

Problem relaxation as a tactic

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

It's easier to make your way to the supermarket than it is to compute the fastest route, which is yet easier than computing the fastest route for someone running backwards and doing two and a half jumping jacks every five seconds and who only follows the route p percent of the time. Sometimes, constraints are necessary. Constraints come with costs. Sometimes, the costs are worth it.

Aspiring researchers trying to think about AI alignment might^[1] have a failure mode which goes something like... this:

Oh man, so we need to solve both outer and inner alignment to build a superintelligent agent which is competitive with unaligned approaches and also doesn't take much longer to train, and also we have to know this ahead of time. Maybe we could use some kind of prediction of what people want... but wait, [there's also problems with using human models!](#) How can it help people if it can't model people? Ugh, and [what about self-modification?! How is this agent even reasoning about the universe from inside the universe?](#)

The aspiring researcher slumps in frustration, mutters a curse under their breath, and hangs up their hat – "guess this whole alignment thing isn't for me...". And isn't that so? All their brain could do was pattern-match onto already-proposed solutions and cached thinking.

There's more than one thing going wrong here, but I'm just going to focus on one. Given that person's understanding of AI alignment, this problem is *wildly* overconstrained. Whether or not alignment research is right for them, there's just no way that anyone's brain is going to fulfill this insane solution request!

Sometimes, constraints are necessary. I think that the alignment community is pretty good at finding plausibly necessary constraints. Maybe some of the above *aren't* necessary – maybe there's One Clever Trick you come up with which obviates one of these concerns.

Constraints come with costs. Sometimes, the costs are worth it. In this context, I think the costs are very much worth it. Under this implicit framing of the problem, you're [pretty hosed](#) if you don't get even outer alignment right.

However, even if the real problem has crazy constraints, that doesn't mean you should immediately tackle the fully constrained problem. I think you should often [relax](#) the problem first: eliminate or weaken constraints until you reach a problem which is still a little confusing, but which you can get some traction on.

Even if you know an unbounded solution to chess, you might still be 47 years away from a bounded solution. But if you can't state a program that solves the problem in principle, you are in some sense confused about the nature of the cognitive work needed to solve the problem. If you can't even solve a problem given infinite computing power, you definitely can't solve it using bounded computing power. (Imagine Poe trying to write a chess-playing program before he'd had the insight about search trees.)

~ [The methodology of unbounded analysis](#)

Historically, I tend to be too slow to relax research problems. On the flipside, *all of my favorite research ideas were directly enabled by problem relaxation*. Instead of just telling you what to do and then having you forget this advice in five minutes, I'm going to paint it into your mind using two stories.

Attainable Utility Preservation

It's spring of 2018, and I've written myself into a corner. My work with CHAI for that summer was supposed to be on impact measurement, but I [inconveniently posted a convincing-to-me argument](#) that impact measurement cannot admit a clean solution:

I want to penalize the AI for having side effects on the world.^[2] Suppose I have a function which looks at the consequences of the agent's actions and magically returns all of the side effects. Even if you have this function, you still have to assign blame for each effect – either the vase breaking was the AI's fault, or it wasn't.

If the AI penalizes itself for everything, it'll try to stop people from breaking vases – it'll be clingy. But if you magically have a model of how people are acting in the world, and the AI magically only penalizes itself for things which are its fault, then the AI is incentivized to blackmail people to break vases in ways which don't technically count as its fault. Oops.

Summer dawned, and I occupied myself with reading – lots and lots of reading. Eventually, enough was enough – I wanted to figure this out. I strode through my school's library, markers in my hand and determination in my heart. I was determined not to leave before understanding a) exactly why impact measurement is impossible to solve cleanly, or b) how to solve it.

I reached the whiteboard, and then – with adrenaline pumping through my veins – I realized that I had *no idea* what this "impact" thing even is. Oops.

I'm staring at the whiteboard.

A minute passes.

59 more minutes pass.

I'd been thinking about how, in hindsight, it was so important that Shannon had first written a perfect chess-playing algorithm which required infinite compute, that Hutter had written an AGI algorithm which required infinite compute. I didn't know how to solve impact under all the constraints, but what if I assumed something here?

What if I had infinite computing power? No... Still confused, don't see how to do it. Oh yeah, and what if the AI had a perfect world model. Hm... *What if we could write down a fully specified utility function which represented human preferences? Could I measure impact if I knew that?*

The answer was almost trivially obvious. My first thought was that negative impact would be a decrease in true utility, but that wasn't quite right. I realized that impact measure needs to also capture decrease in ability to achieve utility. That's an optimal value function... So the negative impact would be the decrease in attainable utility for human values!^[3]

Okay, but we don't and won't know the "true" utility function. What if... we just penalized shift in all attainable utilities?

I then wrote down The Attainable Utility Preservation Equation, more or less. Although it took me a few weeks to believe and realize, [that equation solved all of the impact measurement problems](#) which had seemed so insurmountable to me just minutes before.^[4]

Formalizing Instrumental Convergence

It's spring of 2019, and I've written myself into a corner. [My first post on AUP](#) was confusing – I'd failed to truly communicate what I was trying to say. Inspired by [Embedded Agency](#), I was planning [an illustrated sequence of my own](#).

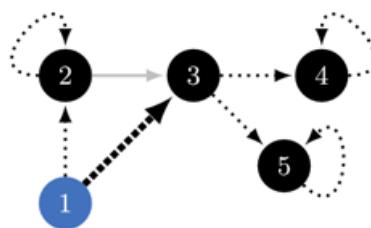
I was working through a bit of reasoning on how your ability to achieve one goal interacts with your ability to achieve seemingly unrelated goals. Spending a lot of money on red dice helps you for the collecting-dice goal, but makes it harder to become the best juggler in the world. That's a weird fact, but it's an *important* fact which underlies much of [AUP's empirical success](#). I didn't understand why this fact was true.

At an impromptu presentation in 2018, I'd remarked that "AUP wields instrumental convergence as a weapon against the alignment problem itself". I tried thinking about it using the formalisms of reinforcement learning. Suddenly, I asked myself

Why is instrumental convergence even a thing?

I paused. I went outside for a walk, and I paced. The walk lengthened, and I still didn't understand why. Maybe it was just a "brute fact", an "emergent" phenomenon – nope, not buying that. There's an explanation somewhere.

I went back to the drawing board – to the whiteboard, in fact. I stopped trying to [understand the general case](#) and I focused on specific toy environments. I'm looking at an environment like this



and I'm thinking, most agents go from 1 to 3. "Why does my brain think this?", I asked myself. Unhelpfully, my brain decided not to respond.

I'm staring at the whiteboard.

A minute passes.

29 more minutes pass.

I'm reminded of a paper my advisor had me read for my qualifying exam. The paper talked about a dual formulation for reinforcement learning environments, where you consider the available trajectories through the future instead of the available policies. I take a picture of the whiteboard and head back to my office.

I run into a friend. We start talking about work. I say, "I'm about 80% sure I have the insight I need - this is how I felt in the past in situations like this, and I turned out to be right".

I turned out to be right. I started building up an entire theory of this dual formalism. Instead of asking myself about the general case of instrumental convergence in arbitrary computable environments, I considered small deterministic Markov decision processes. I started proving everything I could, building up my understanding piece by piece. This turned out to make all difference.

Half a year later, I'd built up enough theory that [I was able to explain a great deal \(but not everything\) about instrumental convergence.](#)

Conclusion

Problem relaxation isn't always the right tactic. For example, if the problem isn't well-posed, it won't work well - imagine trying to "relax" the "problem" of free will! However, I think it's often the right move.

The move itself is simple: consider the simplest instance of the problem which is still confusing. Then, make a ton simplifying assumptions while still keeping part of the difficulty present - don't assume away all of the difficulty. Finally, tackle the relaxed problem.

In general, this seems like a skill that successful researchers and mathematicians learn to use. MIRI does a lot of this, for example. If you're new to the research game, this might be one of the crucial things to pick up on. Even though I detailed how this has worked for me, I think I could benefit from relaxing more.

The world is going to hell. You might be working on a hard (or even an impossible) problem. We plausibly stand on the precipice of extinction and utter annihilation.

Just relax.

This is meant as a reference post. I'm not the first to talk using problem relaxation in this way. For example, see [The methodology of unbounded analysis](#).

1. This failure mode is just my best guess - I haven't actually surveyed aspiring researchers. [←](#)
2. The "convincing-to-me argument" contains a lot of confused reasoning about impact measurement, of course. For one, [thinking about side effects is not a good way of conceptualizing the impact measurement problem.](#) [←](#)
3. The initial thought wasn't as clear as "penalize decrease in attainable utility for human values" - I was initially quite confused by the AUP equation. "What the heck is this equation, and how do I break it?".

It took me a few weeks to get a handle for why it seemed to work so well. It wasn't for a month or two that I began to understand what was actually going on, eventually leading to the [*Reframing Impact*](#) sequence. However, for the reader's convenience, I whitewashed my reasoning here a bit. ↵

4. At first, I wasn't very excited about AUP – I was new to alignment, and it took a lot of evidence to overcome the prior improbability of my having actually found something to be excited about. It took several weeks before I stopped thinking it likely that my idea was probably secretly and horribly bad.

However, I kept staring at the strange equation – I kept trying to break it, to find some obvious loophole which would send me back to the drawing board. I never found it. Looking back over a year later, [*AUP does presently have loopholes*](#), but they're not obvious, nor should they have sent me back to the drawing board.

I started to get excited about the idea. Two weeks later, my workday was wrapping up and I left the library.

Okay, I think there's about a good chance that this ends up solving impact. If I'm right, I'll want to have a photo to commemorate it.

I turned heel, descending back into the library's basement. I took the photograph. I'm glad that I did.

Discovering AUP was one of the happiest moments of my life. It gave me confidence that I could think, and it gave me some confidence that we can *win* – that we can solve alignment. ↵

Review of "Lifecycle Investing"

Crossposted from [my blog](#).

Summary

In this post I review the 2010 book “[Lifecycle Investing](#)” by Ian Ayres and Barry Nalebuff. ([Amazon link here](#); no commission received.) They argue that a large subset of investors should adopt a (currently) unconventional strategy: One’s future retirement contributions should effectively be treated as bonds in one’s retirement portfolio that cannot be efficiently sold; therefore, early in life one should balance these low-volatility assets by gaining exposure to volatile high-return equities that will generically exceed 100% of one’s liquid retirement assets, necessitating some form of borrowing.

“Lifecycle Investing” was recommended to me by a friend who said the book “is extremely worth reading...like learning about index funds for the first time...Like worth paying >1% of your lifetime income to read if that was needed to get access to the ideas...potentially a lot more”. Ayres and Nalebuff lived up to this recommendation. Eventually, I expect the basic ideas, which are simple, to become so widespread and obvious that it will be hard to remember that it required an insight.

In part, what makes the main argument so compelling is that (as shown in the next section), it is closely related to an elegant explanation for something we all knew to be true — you should increase the bond-stock ratio of your portfolio as you get older — yet previously had bad justifications for. It also gives new actionable, non-obvious, and potentially very important advice (buy equities on margin when young) that is appropriately tempered by real-world frictions. And, most importantly, it means I personally feel less bad about already being nearly 100% in stocks when I picked up the book.

My main concerns, which are shared by other reviewers and which are only partially addressed by the authors, are:

- Future income streams might be more like stocks than bonds for the large majority of people.
- Implementing a safe buy-and-hold leveraged strategy in the real world could be even more of a headache, and incur more hidden costs, than the authors suppose.
- If interest rates go up enough and expected stock returns go down enough, the whole thing could be rendered moot.

By far the best review of this book I’ve found after a bit of Googling is the one by Fredrick Vars, a law professor at the University of Alabama: [\[PDF\]](#). Read that. I wrote most of my review before Vars, and he anticipated almost all of my concerns while offering illuminating details on some of the legal aspects.

A key puzzle

One way to frame the insight, slightly different than as presented in the book, is as arising out of a solution to a basic puzzle.

The vast majority of financial advisors agree that retirement investments should have a higher percentage of volatile assets (stocks, essentially) when the person is young and less when they are old. This is often justified by the argument that volatile returns can be averaged out over the years, but, taken naively, this is flat out wrong. As [Alex Tabarrok](#) puts it^a

Many people think that uncertainty washes out when you buy and hold for a long period of time. Not so, that is the [fallacy of time diversification](#). Although the average return becomes more certain with more periods you don't get the average return you get the total payoff and that becomes more uncertain with more periods.

More quantitatively: When a principal P is invested over N years in a fund with a given annual expected return \bar{r} and volatility (standard deviation) σ_r , the average $\bar{Y} = \prod_n (1+r_n)$ becomes more certain for more years and approaches $e^{\bar{r}}$ for the usual central-limit reasons. However, your payout is not the average return! Rather, your payout is the compounded amount^b



$$\bar{Y} = P \prod_n (1+r_n) = P e^{\bar{r}N}$$

and the uncertainty of that does *not* go down with more time...even *in percentage terms*. That is, the ratio of the standard deviation in payout to the mean payout, $\sigma_{\bar{Y}}/\bar{Y}$, goes up the larger the number of years N that the principal is invested.

Sometimes when confronted with this mathematical reality people backtrack to a justification like this: If you are young and you take a large downturn, you can adapt to this by absorbing the loss over many years of slightly smaller future consumption (adaptation), but if you are older you must drastically cut back, so the hit to your utility is larger. This is a true but fairly minor consideration. Even if we knew we would be unable to adapt our consumption (say because it was dominated by fixed costs), it would *still* be much better to be long on stocks when young and less when old.

Another response is to point out that, although absolute uncertainty in stock performance goes up over time, the odds of beating bonds *also* keeps going up. That is, on any given day the odds that stocks outperform bonds is maybe only a bit better than a coin flip, but as the time horizon grows, the odds get progressively better.^c This is true, but some thought shows it's not a good argument. In short, even if the chance of doing worse than bonds keeps falling, the distribution of scenarios where you lose to bonds could get more and more extreme; when you do worse, maybe you do *much* worse. (For an extensive explanation, see the section "Probability of Shortfall" in the John Norstad's "[Risk and Time](#)", which Tabarrok above linked to as "fallacy of time diversification".) This, it turns out, is not true — we see below that stocks do in fact get safer over time — but the possibility of extreme distributions shows why the probability-of-beating-bonds-goes-up-over-time argument is unsound.

Puzzle resolution

To neatly resolve this puzzle, the authors make a strong simplifying assumption. (Importantly, the main idea is robust to relaxing this assumption somewhat,^d but for

now let's accept it in its idealized form.)

The main assumption is that the portion of your future income that you will be saving for retirement (e.g., your stream of future 401(k) contributions) can be predicted with relative confidence and are financially equivalent to today holding a sequence of bonds that pay off on a regular schedule in the future (but cannot be sold). **When we consider how our retirement portfolio today should be split between bonds and stocks, we should include the net present value of our future contributions.** That is the main idea.

Under some not-unreasonable simplifying assumptions, Samuelson and Merton showed long ago^e that if, counterfactually, you had to live off an initial lump sum of wealth, then the optimal way to invest that sum would be to *Maintain a constant split* between assets of different volatility (e.g., 40% stocks and 60% bonds), with the appropriate split determined by your personal risk tolerance. However, even though you won't magically receive your future retirement contributions as a lump sum in real life, it follows that if those contributions were perfectly predictable, and if you could borrow money at the risk-free rate, then you should borrow against your future contributions, converting them to their net present value, and *keep the same constant fraction of the money in the stock market.* Starting today.

Crucially, when you are young your liquid retirement portfolio (the sum of your meager contribution up to that point, plus a bit of accumulated interest) is dwarfed by your expected future contributions. Even if you invest 100% of your retirement account into stocks you are insufficiently exposed to the stock market. In order to get sufficient stock exposure, you should borrow lots of money at the risk-free rate and put it in the stock market. It is only as you get older, when the ratio between your retirement account and the present value of future earnings increases, that you should move more and more of your (visible) retirement account into regular bonds.

The resolution of the puzzle is that the optimal portfolio (in the idealized case) only looks like it's stock-heavy early in life because you're forgetting about your stream of future retirement contributions (a portion of your future salary), which, the authors claim, is essentially like a bond that can't be traded.

(If the above concept isn't immediately compelling to you, my introduction has failed. Close this blog and just go read the first couple chapters of their book.)

But what about practicalities?

Most of the book is devoted to fleshing out and defending the implications of this idea for the real world where there are a variety of complications, most notably that you cannot borrow unlimited amounts at the risk-free rate. Nevertheless, the authors conclude that when many people are young they should buy equities on margin (i.e., with borrowed money) up to 2:1 leverage, at least if they have access to low enough interest rates to make it worthwhile.

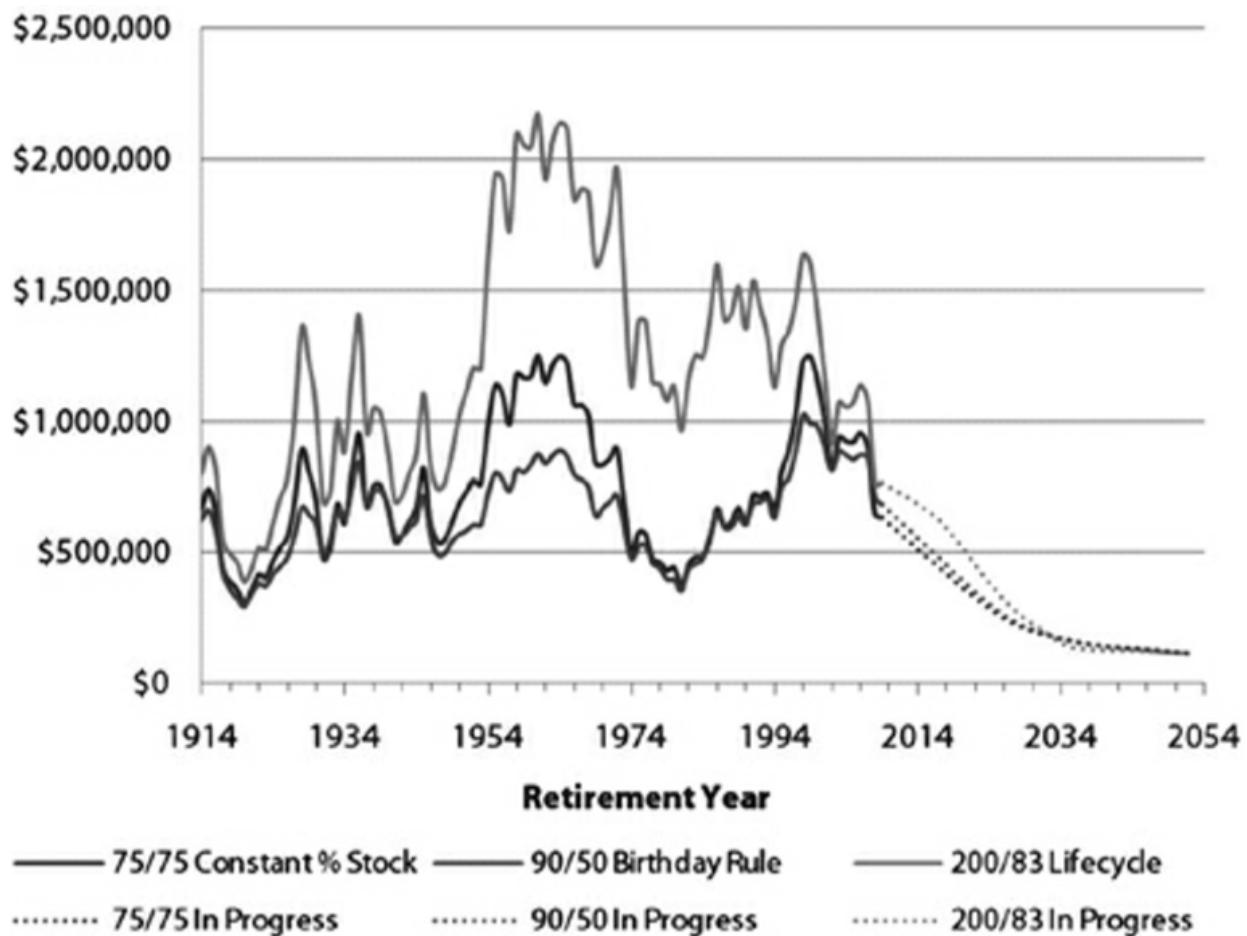
The organization of chapters are as follows:

1. Basic idea and motivation
2. Theory. Outline of lifecycle strategy.
3. Comparison of lifecycle strategy with conventional strategies on US historical data
4. Responses to various objections

5. Implications for older investors, inheritances, and trusts
6. Contraindications - who shouldn't use the strategy
7. Risk tolerance and details
8. Mechanics of implementing the strategy
9. Macroimplications: What if everyone did it? How do we bring that about?

In general the authors compare their lifecycle investing strategy to two conventional strategies: the “birthday rule” (aka an “[age-in-bonds rule](#)”), where the investor allocates a percentage of their portfolio to stocks given by 100 (or 110) minus their age, and the “constant percentage rule”, where the investor keeps a constant fraction of their portfolio in stocks.

In Chapter 3, the authors argue that the lifecycle strategy consistently beats conventional strategies when (a) holding fixed expected return and minimizing variance, (b) holding fixed variance and maximizing expected return, (c) holding fixed very bad (first percentile) returns while maximizing expected return. If you look at a hypothetical ensemble of investors on historical data, one retiring during each year between 1914 and 2010 (when the book was published), every single investor would have been had more at retirement by adopting the lifecycle strategy, and generally by an enormous 50% or more. Here’s the total return of the investors vs. retirement year depending on whether they following a lifecycle strategy, birthday rule, or the constant percentage rule:



And here are the quantiles:

	<i>Birthday Rule</i>	<i>Constant % Stocks</i>	<i>Diversifying Lifecycle Strategy</i>	<i>Improvement over Birthday Rule</i>	<i>Improvement over Constant %</i>
Max. % Inv.	90	75	200		
Min. % Inv.	50	75	83		
Mean Result	\$646,575	\$748,839	\$1,223,105	89.2%	63.3%
Min. Result	\$290,310	\$308,726	\$387,172	33.4%	25.4%
10th pct.	\$416,253	\$449,266	\$701,834	68.6%	56.2%
25th pct.	\$539,343	\$561,032	\$884,138	63.9%	57.6%
Median	\$641,555	\$691,427	\$1,146,812	78.8%	65.9%
75th pct.	\$779,044	\$922,028	\$1,522,653	95.5%	65.1%
90th pct.	\$870,921	\$1,152,276	\$1,929,577	121.6%	67.5%
Max. Result	\$1,026,903	\$1,252,684	\$2,177,424	112.0%	73.8%

Although they rely on historical simulations for this, it's really grounded in a very simple theoretical idea: your liquid retirement portfolio is extremely small when you're young, so for any plausible level of risk aversion, you are better off leveraging equities initially.

Chapter 4 considers more testing variations: international stocks returns, Monte Carlo simulations with historically anomalous stock performance, higher interest rates, etc. They also show the strategy can easily be modified to incorporate (possibly EMH-violating) beliefs about one's ability to time the market. (The authors use Robert Shiller's theory of [cyclically adjusted price-to-earnings ratio](#), which they neither endorse nor reject.)

In Chapter 7, the authors draw on the work of Samuelson and Merton to address the key question: what is the constant fraction in stocks that you should be targeting anyways? Assuming assumptions, the optimal "Samuelson share" to have invested in stocks is

$$f = \frac{r_s - r_b}{\sigma_s^2}$$

The variables above are defined as follows.

- r_s : The [equity premium](#), i.e., the difference in the expected rate of return between stocks and (perfectly safe) bonds.

- σ : The [equity volatility](#), i.e., the standard deviation of the annual equity rate of return.
- R : The [relative risk aversion](#), a measure of an individual investor's trade-off between risk and reward.

The authors give reasons to be wary of taking this formula too seriously, especially because it's not so easy to know what R you should choose (discussed more below). However, it is very notable that as equity volatility increases — say, because the world is [gripped by a global pandemic](#) — the appropriate amount of the portfolio to have exposed to the stock market drops drastically. The authors suggest using the [VIX](#) to estimate the equity volatility, and appropriately rebalancing your portfolio when that metric changes. Continuously hitting the correct Samuelson share without shooting yourself in the foot looks hard, in practice, which the authors admit. Still, there is so much to gain from leverage that it's very likely you can collect a good chunk of the upside even with a conservative and careful approach.

Criticism

Risk tolerance intuition

The first general point of caution tempers (but definitely does not eliminate) the suggestion to invest in equities on margin: one's risk tolerance is not an easy thing to elicit. To a large extent we do this by imagining various outcomes, deciding which outcomes we would prefer, and then inferring (with regularity assumptions) what our risk tolerance must be. Therefore, it would likely be a mistake to immediately take whatever risk tolerance you previously thought you had *as deployed in conventional investment strategies* and then follow the advice in this book. After introspection, I've sorta decided that although I am still less risk averse than the general population, I'm more risk averse than I thought because I was following the intuition (which I can now justify better) that I should be heavy in stocks at my age. The authors address the general difficulty of someone identifying their own risk tolerance (e.g., how dependent it is on framing effects), but they do not discuss how your beliefs about your risk tolerance might be entangled with what investment strategy you have previously been using.

However, this bears repeating: For every level of risk tolerance, there exists a form of this strategy that beats (both in expectation and risk) the best conventional strategy. The fact that, when young, you are buying stocks on margin makes it tempting to interpret this strategy is only good when one is not very risk averse or when the stock market has a good century. But for any time-homogeneous view you have on what stocks will do in the future, there is a version of this strategy that is better than a conventional strategy. (A large fraction of casual critics seem to miss this point.) The authors muddy this central feature a bit because, on my reading, they are a bit less risk averse than the average person. The book would have been more pointed if they had erred toward risk aversion in their various examples of the lifecycle strategy.

Retirement as a rainy day fund

The second point of caution is gestured at in the [criticism by Nobel winner Paul Samuelson](#). (He was also a mentor of the authors.) The costs of going *truly* bust would be catastrophic:

The ideas that I have been criticizing do not shrivel up and die. They always come back... Recently I received an abstract for a paper in which a Yale economist and a Yale law school professor advise the world that when you are young and you have many years ahead of you, you should borrow heavily, invest in stocks on margin, and make a lot of money. I want to remind them, with a well-chosen counterexample: I always quote from Warren Buffett (that wise, wise man from Nebraska) that in order to succeed, you must first survive. People who leverage heavily when they are very young do not realize that the sky is the limit of what they could lose and from that point on, they would be knocked out of the game.

The authors respond to these sorts of concerns by emphasizing that (1) the risk of losing everything is highest when you are very young, which is exactly when the amount you have in your retirement account is very small, and (2) they are recommending adding leverage to your retirement account, not all your assets. If you expect the total of your retirement contributions to be roughly \$1 million by the time you retire, losing \$20,000 and zeroing out your retirement account when you are 25 is not catastrophic (and is still a rare outcome under their strategy). You should still have a rainy day fund, and you'll just earn more money in the future.

However, I don't think this response seriously grapples with the best concrete form of the wary intuition many people have to their strategy. I think the main problem is that most people are implicitly using their retirement account not just as a place to save for retirement assuming a normal healthy life, but also as a rainy day fund for a variety of bad events. In the US, 12% of people are disabled; I don't know how much you can push down those odds knowing you are healthy at a given time, but it seems like you need to allow for a ~3% chance you are partially or totally disabled at some point. Although people buy disability insurance, they also know that if they ever needed to tap into their retirement account they could, possibly with a modest tax penalty. (Likewise for other unforeseen crises.)

Another way to say this: your future earning are substantially more likely to fail than the US government, so they cannot be idealized as a bond. By purchasing the right insurance, keeping enough in savings account, etc., I'm sure there's a way to hedge against this, and I'm confident the core ideas in this book survive this necessary hedging. But I would have liked the authors to discuss how to do that in *at least* as much concrete detail as they describe the mechanics of how to invest on margin.⁹⁻ If people have been relying on the conventional strategy and have consequently been implicitly enjoying a form of buffer/insurance, it is paramount to highlight this and find a substitute before moving on to an unconventional strategy that lacks that buffer.

Now, if we only had to insure against tail risks, that would be fine, but there is an extreme version of this issue that has the potential to undermine the entire idea: why is my future income stream like a bond rather than a stock? I have a ton of uncertainty about how my income will increase in the future. Indeed, personally, I think I trust the steady growth of the stock market more! The authors do advise against adopting their strategy if your future income stream is highly correlated with the market (e.g., you're a banker), but they don't get very quantitative, and they don't say much about what do if that stream is highly volatile but not very correlated with stocks. (Sure, if it's uncorrelated then you're want to match your "investment" in your future income stream with some actual investment in stocks for diversification, but how much should this overall high volatility change the strategy?^h)

So did I immediately go out and lever my portfolio, or what?

It will take some time before I have mulled this around enough to even start assessing whether I should be investing with significant leverage. It seems pretty plausible to me that my future income is much more uncertain than a bond, although that's something I'll need to meditate on.

I, like the authors, really wish there was a mutual fund that automatically implemented this strategy, like [target-date funds](#) do for (strategies similar to) the birthday rule. At the very least it would induce pointed discussion about the benefits and risks of the strategy. Unfortunately, a decade after this book was released there is no such option and, as the authors admit in the book, concretely implementing the strategy yourself in the real world can be a headache.

However, because of this book I can at least feel less guilty for being overwhelmingly in equities. After finishing this book I finally exchanged much of my remaining Vanguard 2050 target-date funds, which contain bonds, for pure equity index funds. I had been keeping them around in part because going 100% equities felt vaguely dangerous. Now that there is a good argument that the optimal allocation is *greater* than 100% equities — though that is by no means assured — this no longer feels so extreme. Crossing the 100% barrier by acquiring leverage involves many real-world complications, but in the platonic realm there is nothing special about the divide.

Footnotes

([←](#) returns to text)

- a. I have edited the broken link for “fallacy of time diversification” to point to an archived version of the web page.[←](#)
- b. The approximation is valid for $|r_n| \ll 1$.[←](#)
- c. I thank Will Riedel for this compelling phrasing.[←](#)
- d. Although the authors don’t quantitatively explore this enough. See criticisms below.[←](#)
- e. Looks like [Merton’s version](#) of the problem is the most well known. Here are the references taken directly from the book: “Paul A.. Samuelson, “Lifetime Portfolio Selection by Dynamic Stochastic Programming,” *Review of Economics and Statistics* 51 (1969): 239-246; Robert Merton, “Lifetime Portfolio Selection Under Uncertainty: The Continuous-Time Case,” *Review of Economics and Statistics* 51 (1969): 247-257; and Robert Merton, “Optimum Consumption and Portfolio Rules in a Continuous Time Model,” *Journal of Economic Theory* 3 (1971): 373-413.”[←](#)
- f. [Ayers replies here.](#)[←](#)
- g. Indeed, I suspect that many of the valid criticisms of their strategy apply almost as well to conventional investment strategies. For example, many of us should probably have more disability insurance; if you lost the ability to work when you were 30, would things work out OK? The lack of leverage in a conventional portfolio, combined with the fact that the stock market is quite unlikely to lose more than, say, 60%, means that the conventional portfolio naturally includes some weak coverage of bad scenarios. But this is essentially accidental, and very unlikely to be optimal.[←](#)
- h. The author mention in passing that their friend Moshe Milevsky has written an entire book on the question: “Are You a Stock or a Bond?”. But as their entire strategy hinges on this question, they should have addressed it much more deeply themselves.[←](#)

How special are human brains among animal brains?

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

Humans are capable of feats of cognition that appear qualitatively more sophisticated than those of any other animals. Is this appearance of a qualitative difference indicative of human brains being essentially more complex than the brains of any other animal? Or is this “qualitative difference” illusory, with the vast majority of human cognitive feats explainable as nothing more than a scaled-up version of the cognitive feats of lower animals?

*“How special are human brains among animal brains?” is one of the background variables in my [framework for AGI timelines](#). My aim for this post is **not** to present a complete argument for some view on this variable, so much as it is to:*

- *present some considerations I’ve encountered that shed light on this variable*
- *invite a collaborative effort among readers to shed further light on this variable (e.g. by leaving comments about considerations I haven’t included, or pointing out mistakes in my analyses)*

Does mastery of language make humans unique?

Human conscious experience may have emerged from language

Humans seem to have much higher degrees of consciousness and agency than other animals, and this may have emerged from our capacities for language. [Helen Keller](#) (who was deaf and blind since infancy, and only started learning language when she was 6) gave an [autobiographical account](#) of how she was driven by blind impetuses until she learned the meanings of the words “I” and “me”:

Before my teacher came to me, I did not know that I am. I lived in a world that was a no-world. I cannot hope to describe adequately that unconscious, yet conscious time of nothingness. I did not know that I knew aught, or that I lived or acted or desired. I had neither will nor intellect. I was carried along to objects and acts by a certain blind natural impetus. I had a mind which caused me to feel anger, satisfaction, desire. These two facts led those about me to suppose that I willed and thought. I can remember all this, not because I knew that it was so, but because I have tactal memory. It enables me to remember that I never contracted my forehead in the act of thinking. I never viewed anything beforehand or chose it. I also recall tactually the fact that never in a start of the body or a heart-beat did I feel that I loved or cared for anything. My inner life, then, was a blank without past, present, or future, without hope or anticipation, without wonder or joy or faith.

[...]

... When I learned the meaning of "I" and "me" and found that I was something, I began to think. Then consciousness first existed for me. Thus it was not the sense of touch that brought me knowledge. It was the awakening of my soul that first rendered my senses their value, their cognizance of objects, names, qualities, and properties. Thought made me conscious of love, joy, and all the emotions. I was eager to know, then to understand, afterward to reflect on what I knew and understood, and the blind impetus, which had before driven me hither and thither at the dictates of my sensations, vanished forever.

Mastery of language may have conferred unique intellectual superpowers

I think humans underwent a phase transition in their intellectual abilities when they came to master language, at which point their intellectual abilities jumped far beyond those of other animals on both an individual level and a species level.

On an individual level, our capacity for language enables us to entertain and express arbitrarily complex thoughts, which appears to be an ability unique to humans. In theoretical linguistics, this is referred to as "[digital infinity](#)", or "[the infinite use of finite means](#)".

On a species level, our mastery of language enables intricate insights to accumulate over generations with high fidelity. Our ability to stand on the shoulders of giants is unique among animals, which is why our culture is unrivaled in its richness in sophistication.

Language aside, how unique are humans?

Humans ≈ Neanderthals + language?

The most quintessentially human intellectual accomplishments (e.g. proving theorems, composing symphonies, going into space) were only made possible by culture post-agricultural revolution. So, when evaluating humans' innate intellectual capacities, a better reference point than modern humans like ourselves would be our hunter-gatherer ancestors.

We can reduce the question of how complex our hunter-gatherer ancestors' brains are into two sub-questions: how complex is our capacity for mastering language, and how complex are brains that are similar to ours, but don't have the capacity for mastering language?

Neanderthal brains seem like plausible proxies for the latter. Neanderthals are similar enough to modern humans that [they've interbred](#), and the [currently available evidence](#) suggests that they may not have mastered language in the same way that [behaviorally modern humans](#) have. (I don't think this evidence is very strong, but this doesn't matter for my purposes—I'm just using Neanderthals as a handy stand-in to

gesture at what a human-like intelligence might look like if it didn't have the capacity for language.)

Higher intelligence in animals

Chimpanzees, crows, and dolphins are capable of impressive feats of higher intelligence, and I don't think there's any particular reason to think that Neanderthals are capable of doing anything qualitatively more impressive. I'll share some examples of these animals' intellectual feats that I found particularly illustrative.

Chimpanzees have been observed to lie to each other under experimental conditions. [From Wikipedia:](#)

...food was hidden and only one individual, named Belle, in a group of chimpanzees was informed of the location. Belle was eager to lead the group to the food but when one chimpanzee, named Rock, began to refuse to share the food, Belle changed her behaviour. She began to sit on the food until Rock was far away, then she would uncover it quickly and eat it. Rock figured this out though and began to push her out of the way and take the food from under her. Belle then sat farther and farther away waiting for Rock to look away before she moved towards the food. In an attempt to speed the process up, Rock looked away until Belle began to run for the food. On several occasions he would even walk away, acting disinterested, and then suddenly spin around and run towards Belle just as she uncovered the food.

In [Aesop's fable of the crow and the pitcher](#), a thirsty crow figures out that it can drop pebbles into a pitcher, so that the water rises to a high enough level for it to drink from. This behavior has been [experimentally replicated](#), indicating that crows have a "sophisticated, but incomplete, understanding of the causal properties of displacement, rivalling that of 5-7 year old children".

When [Kelly the dolphin](#) was given rewards of fish for picking up scraps of paper, "Kelly figured out that she received the same fish regardless of the size of the piece of trash she was delivering to her trainer. So she began hiding big pieces of trash under a rock. Kelly would then rip off small pieces from the trash and deliver them one at a time so that she could receive more fish." Additionally, "when a bird landed in the pool, Kelly snatched it and delivered it to her trainers. She received a large amount of fish in return. Knowing this, she decided to start hiding fish each time she was fed. She would then use the fish to lure birds when none of her trainers were around. Kelly knew that by saving one or two fish now, she could get many more fish later by turning in a bird." (Also reported on [The Guardian](#); I don't know how reputable these sources are, so take this anecdote with a grain of salt.)

See [these Wikipedia pages](#) for some more interesting examples, and see [here](#) for a more thorough review of the evidence of higher intelligence in animals.

"Qualitatively" more advanced cognition may emerge from scale

Many aspects of human cognition that may appear qualitatively different from what other animals are capable of, such as long chains of abstract reasoning, also appear qualitatively different from what less intelligent humans are capable of. As a particularly extreme example, John von Neumann's [cognitive abilities](#) were so

advanced that a Nobel Laureate, Hans Bethe, once remarked that "[his] brain indicated a new species, an evolution beyond man".

At the same time, the genes that code for different humans' brains are virtually identical from an evolutionary perspective. This suggests that the seemingly qualitative differences between humans' and animals' cognition might not be so different from the seemingly qualitative differences between John von Neumann's cognition and mine—our brains might be doing essentially the same thing as theirs, except at a higher scale.

How hard is mastery of language?

Could language capacity fall out from general capacities?

Maybe it was extraordinarily difficult to evolve the cognitive mechanisms that allow us to learn language, above and beyond our cognitive machinery for learning other things. I think this is plausible, but I don't think the case for this is very strong.

Animals ([Washoe](#), [Koko](#), and [Alex the parrot](#)) have demonstrated the ability to learn simple forms of symbolic communication, which they never evolved to do, indicating that their ability to learn things in general is good enough to learn very simple forms of language. It's true that there are [aspects of human language that escape animals](#), but they also escape [feral children](#), and might escape animals for mundane reasons, like their not having [critical periods](#) long enough to learn these aspects of language.

Additionally, [AI language models](#) provide evidence that simple and general learning mechanisms can capture many of the intricacies of human language that other animals miss, further suggesting that there's nothing intrinsically difficult about learning language. Here's an excerpt from [GPT-2](#), a relatively recent language model:

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Why haven't other species mastered language?

If language isn't a particularly difficult cognitive capacity to acquire, why don't we see more animal species with language?

One possibility is that the first species that masters language, by virtue of being able to access intellectual superpowers inaccessible to other animals, has a high probability of becoming the dominant species extremely quickly. (Humans underwent the agricultural revolution within 50,000 years of behavioral modernity—a blink of an eye on evolutionary timescales—after which their dominance as a species became unquestionable.) Since we shouldn't expect to see more than one dominant species at a time, this would imply a simple anthropic argument for our unique capacities for language: we shouldn't expect to see more than one species at a time with mastery of language, and we just happen to be the species that made it there first.

It may also turn out that language is hard to evolve not because it's a particularly sophisticated cognitive mechanism, but because the environments that could have supported language and selected for it might have been very unique. For example, it may be that a threshold of general intelligence has to be crossed before it's viable for a species to acquire language, and that humans are the only species to have crossed this threshold. (Humans do have the highest [cortical information processing capacity](#) among mammals.)

It might also turn out that the cultural contexts under which language could evolve [require a mysteriously high degree of trust](#): "... language presupposes relatively high levels of mutual trust in order to become established over time as an [evolutionarily stable strategy](#). This stability is born of a longstanding mutual trust and is what grants language its authority. A theory of the origins of language must therefore explain why humans could begin trusting cheap signals in ways that other animals apparently cannot (see [signalling theory](#))."

My current take

As we came to master language, I think we underwent a phase transition in our intellectual abilities that set us apart from other animals. Besides language, I don't see much that sets us apart from other animals—in particular, most other cognitive differences seem explainable as consequences of either language or scale, and I don't think the cognitive mechanisms that allow us to master language are particularly unique or difficult to acquire. Overall, I don't see much reason to believe that human brains have significantly more innate complexity than the brains of other animals.

Thanks to Paul Kreiner and Stag Lynn for helpful commentary and feedback.

Inner alignment in the brain

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

Abstract: We can think of the brain crudely as (1) a neocortex which runs an amazingly capable quasi-general-purpose learning-and-planning algorithm, and (2) subcortical structures (midbrain, etc.), one of whose functions is to calculate rewards that get sent to up the neocortex to direct it. But the relationship is actually more complicated than that. "Reward" is not the only informational signal sent up to the neocortex; meanwhile information is also flowing back down in the opposite direction. What's going on? How does all this work? Where do emotions fit in? Well, I'm still confused on many points, but I think I'm making progress. In this post I will describe my current picture of this system.

Background & motivation

I'm interested in helping ensure a good post-AGI future. But how do we think concretely about AGI, when AGI doesn't exist and we don't know how to build it? Three paths:

1. We can think generally about the nature of intelligence and agency—a research program famously associated with MIRI, Marcus Hutter, etc.;
2. We can think about today's AI systems—a research program famously associated with OpenAI, DeepMind, CHAI, etc.;
3. We can start from the one "general intelligence" we know about, i.e. the human brain, and try to go from there to lessons about how AGI might be built, what it might look like, and how it might be safely and beneficially used and controlled.

I like this 3rd research program; it seems to be almost completely neglected, [\[1\]](#) and I think there's a ton of low-hanging fruit there. Also, this program will be *especially* important if we build AGI in part by reverse-engineering (or reinventing) high-level neocortical algorithms, which (as discussed below) I think is very plausible, maybe even likely—for better or worse.

Now, the brain is divided into the neocortex and the subcortex.

Start with the **neocortex**[\[2\]](#) The neocortex does essentially all the cool exciting intelligent things that humans do, like building an intelligent world-model involving composition and hierarchies and counterfactuals and analogies and meta-cognition etc., and using that thing to cure diseases and build rocket ships and create culture etc. Thus, both neuroscientists and AI researchers focus a lot of attention onto the neocortex, and on understanding and reverse-engineering its algorithms. Textbooks divide the neocortex into lots of functional regions like "motor cortex" and "visual cortex" and "frontal lobe" etc., but microscopically it's all a pretty uniform 6-layer structure, and I currently believe that all parts of the neocortex are performing more-or-less the same algorithm, but with different input and output connections. These connections are seeded by an innate gross wiring diagram and then edited by the algorithm itself. See [Human Instincts, Symbol Grounding, and the Blank-Slate Neocortex](#) for discussion and (heavy!) caveats on that claim. And what is this algorithm? I outline some of (what I think are) the high-level specifications at [Predictive coding = RL + SL + Bayes + MPC](#). In terms of how the algorithm actually works, I think

that researchers are making fast progress towards figuring this out, and that a complete answer is already starting to crystallize into view on the horizon. For a crash course on what's known today on how the neocortex does its thing, maybe a good starting point would be to read [On Intelligence](#) and then every paper ever written by Dileep George (and citations therein).

The subcortex, by contrast, is *not* a single configuration of neurons tiled over a huge volume, but rather it is a collection of quite diverse structures like the amygdala, cerebellum, tectum, and so on. Unlike the neocortex, this stuff does *not* perform some miraculous computation light-years beyond today's technology; as far as I can tell, it accomplishes the same sorts of things as AlphaStar does. And the most important thing to understand (for AGI safety) is this:

The subcortex provides the training signals that guide the neocortex to do biologically-useful things. [\[3\]](#)

Now, if people build AGI that uses algorithms similar to the neocortex, we will need to provide it with training signals. What exactly are these training signals? What [inner alignment](#) issues might they present? Suppose we wanted to make an AGI that was pro-social *for the same underlying reason* as humans are (sometimes) pro-social (i.e., thanks to the same computation); is that possible, how would we do it, and would it work reliably? These are questions we should answer well before we finish reverse-engineering the neocortex. I mean, really these questions should have been answered before we even *started* reverse-engineering the neocortex!! I don't have answers to those questions, but I'm trying to lay groundwork in that direction. Better late than never...

(**Update 1 year later:** These days I say "hypothalamus & brainstem" instead of subcortex, and I'm inclined to lump almost the entire rest of the brain—the whole telencephalon plus cerebellum—in with the neocortex as the subsystem implementing a from-scratch learning algorithm. See [here](#))

Things to keep in mind

Before we get into the weeds, here are some additional mental pictures we'll need going forward:

Simple example: Fear of spiders

My go-to example for the relation between subcortex and neocortex is fear of spiders. [\[4\]](#) Besides the visual cortex, humans have a little-known *second* vision system in the midbrain (superior colliculus). When you see a black scuttling thing in your field of view, the midbrain vision system detects that and sends out a reaction that makes us look in that direction and increase our heart rate and flinch away from it. Meanwhile, the neocortex is simultaneously seeing the spider with *its* vision system, *and* it's seeing the hormones and bodily reaction going on, and it connects the dots to learn that "spiders are scary". In the future, if the neocortex merely imagines a spider, it might cause your heart to race and body to flinch. On the other hand, after [exposure therapy](#), we might be able to remain calm when imagining or even seeing a spider. How does all this work?

(Note again the different capabilities of the midbrain and neocortex: The midbrain has circuitry to recognize black scuttling things—kinda like today's CNNs can—whereas the neocortex is able to construct and use a rich semantic category like "spiders".)

We'll be returning to this example over and over in the post, trying to work through how it might be implemented and what the consequences are.

The neocortex is a black box from the perspective of the subcortex

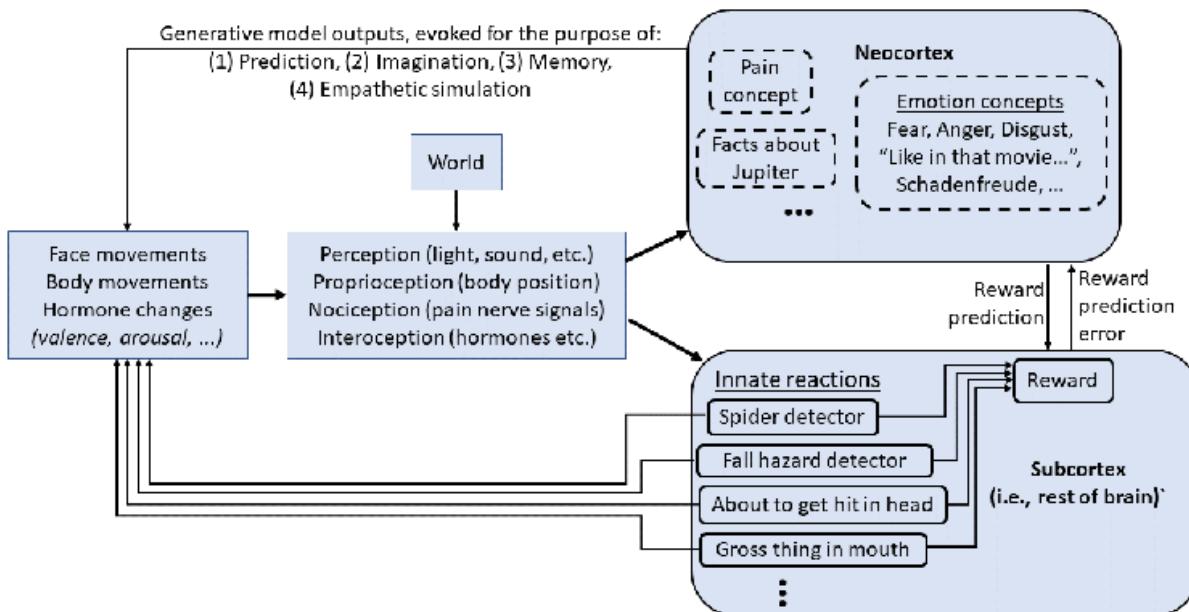
The neocortex's algorithm, as I understand it, sorta learns patterns, and patterns in the patterns, etc., and each pattern is represented as an essentially randomly-generated^[5] set of neurons in the neocortex. So, if X is a concept in your neocortical world-model, there is no straightforward way for an innate instinct to refer directly to X—say, by wiring axons from the neurons representing X to the reward center—because X's neurons are not at predetermined locations. X is inside the black box. An instinct can incentivize X, at least to some extent, but it has to be done indirectly.

I made a list of various ways that we can have universal instincts despite the neocortex being a black-box learning algorithm: See [Human Instincts, Symbol Grounding, and the Blank-Slate Neocortex](#) for my list.

This blog post is a much deeper dive into how a couple of these mechanisms might be actually implemented.

General picture

Finally, here is the current picture in my head:



(Update 1 year later: I no longer would draw it this way—see [Big picture of phasic dopamine](#) for what I now think instead. The main difference is: I would *not* draw a direct line from neocortex to a hormone change (for example); instead the cortex would tell the subcortex (hypothalamus + brainstem) to make that hormone change, and then the subcortex might or might not comply with that recommendation. (I guess the way I drew it here is more like [somatic marker hypothesis](#).))

There's a lot here. Let's go through it bit by bit.

Emotions, "emotion concepts", and "reactions"

One aspect of this picture is emotions. There's a school of thought, popularized by Paul Ekman and the movie *Inside Out*, that there are exactly six emotions (anger, disgust, fear, happiness, sadness, surprise), each with its own universal facial expression. (I've seen other lists of emotions too, and sometimes there's also a list of social emotions like embarrassment, jealousy, guilt, shame, pride, etc.) That was my belief too, until I read the book [How Emotions Are Made by Lisa Feldman Barrett](#), which convincingly argues against it. Barrett argues that a word like "anger" lumps together a lot of very different bodily responses involving different facial expressions, hormones, etc.

Basically, emotional concepts, like other concepts, are arbitrary categories describing things that we find useful to lump together. Sure, they might be lumped together because they share a common hormone change or a common facial expression, but they might just as likely be lumped together because they share a common situational context, or a common set of associated social norms, or whatever else. And an emotion concept with an English-language name like "anger" is not fundamentally different from an idiosyncratic emotion concept like "How Alice must have felt in that TV episode where...".

(Incidentally, while I think Barrett's book is right about that, I am definitely not blanket-endorsing the whole book—there are a lot of other claims in it that I don't agree with, or perhaps don't understand.^[6] I think Barrett would strongly disagree with most of this blog post, though I could be wrong.)

So instead of putting "emotions" in the subcortex, I instead put there a bunch of things I'm calling "reactions" for clarity.^[7] I imagine that there are dozens to hundreds of these (...and separating them into a discrete list is probably an oversimplification of a more complicated computational architecture, but I will anyway). There's the reaction that gets triggered when your midbrain vision system sees a spider moving towards you out of the corner of your eye, as discussed above. And there's a different reaction that gets triggered when you stand at the edge of a precipice and peer over the edge. Both of those reactions might be categorized as "fear" in the neocortex, but they're really different reactions, involving (I presume) different changes to heart rate, different bodily motions, different facial expressions, different quantities of (negative) reward, etc. (Reactions where peripheral vision is helpful will summon a wide-eyed facial expression; reactions where visual acuity is helpful will summon a narrow-eyed facial expression; and so on.)

As described above for the spider example, the neocortex can see what the subcortex does to our hormones, body, face, etc., and it can learn to predict that, and build those expectations into its predictive world-model, and create concepts around that.

(I also put "pain concept" in the neocortex, again following Barrett. A giant part of the pain concept is nociception—detecting the incoming nerve signals we might call "pain sensations". But at the end of the day, the neocortex gets to decide whether or not to classify a situation as "pain", based on not only nociception but also things like context and valence.)

The neocortex's non-motor outputs

From the above, our neocortex comes to expect that if we see a scuttling spider out of the corner of our eye, our heart will race and we'll turn towards it and flinch away. What's missing from this picture? The neocortex *causing* our heart to race by *anticipating* a spider. It's easy to see why this would be evolutionarily useful: If I know (with my neocortex) that a poisonous spider is approaching, it's appropriate for my heart to start racing even before my midbrain sees the black scuttling blob.

Now we're at the top-left arrow in the diagram above: the neocortex causing (in this case) release of stress hormones. How does the neocortex learn to do that?

There are two parts of this "how" question: (1) what are the actual output knobs that the neocortex can use, and (2) how does the neocortex decide to use them? For (1), I have no idea. For the purpose of this blog post, let us assume that there is a set of outgoing axons from the neocortex that (directly or indirectly) cause hormone release, and also assume that "hormone release" is the right thing to be talking about in terms of controlling valence, arousal, and so on. I have very low confidence in all this, but I don't *think* it matters much for what I want to say in this post. (Update 1 year later: I understand (1) better now, but it still doesn't matter here.)

I mainly want to discuss question (2): given these output knobs, how does the neocortex decide to use them?

Recall again that in predictive coding, the neocortex [finds generative models which are consistent with each other, which have not been repeatedly falsified, and which predict that reward will happen.](#)

My first thought was: No additional ingredients, beyond that normal predictive coding picture, are needed to get the neocortex to imitate the subcortical hormone outputs. Remember, just like my post on [predictive coding and motor control](#), the neocortex will discover and store generative models that entail "self-fulfilling prophecies", where a single generative model in the neocortex simultaneously codes for a prediction of stress hormone *and* the neocortical output signals that actually cause the release of this stress hormone. Thus (...I initially thought...), after seeing spiders and stress hormones a few times, the neocortex will predict stress hormones when it sees a spider, which incidentally *creates* stress hormones.

But I don't think that's the right answer, at least not by itself. After all, the neocortex will *also* learn a generative model where stress hormone is generated exogenously (e.g. by the subcortical spider reaction) and where the neocortex's own stress hormone generation knob is left untouched. This latter model is issuing perfectly good predictions, so there is no reason that the neocortex would spontaneously throw it out and start using instead the self-fulfilling-prophecy model. (By the same token, in the [motor control case](#), if I think you are going to take my limp arm and lift it up, I have no problem predicting that my arm will move due to that exogenous force; my neocortex doesn't get confused and start issuing motor commands.)

So here's my second, better story:

Reward criterion (one among many): when the subcortex calls for a reaction (e.g. cortisol release, eyes widening, etc.), it rewards the neocortex with dopamine if it sees that those commands have somehow already been issued.

(Update 2021/06: Oops, that was wrong too. I think I got it on the third try though; see [here](#).)

So if the subcortex computes that a situation calls for cortisol, the neocortex is rewarded if the subcortex sees that cortisol is *already* flowing. This example seems introspectively reasonable: Seeing a spider out of the corner of your eye is bad, but being *surprised* to see a spider when you were feeling safe and relaxed is even worse (worse in terms of dopamine, not necessarily worse in terms of valence—remember [wanting ≠ liking](#)). Presumably the same principle can apply to eye-widening and other things.

To be clear, this is one reward criterion among many—the subcortex issues positive and negative rewards according to other criteria too (as in the diagram above, I think different reactions inherently issue positive or negative rewards to the neocortex, just like they inherently issue motor commands and hormone commands). But as long as this "reward criterion" above is permanently in place, then thanks to the laws of the [neocortex's generative model economy](#), the neocortex will drop those generative models that passively anticipate the subcortex's reactions, in favor of models that actively anticipate / imitate the subcortical reactions, insofar as that's possible (the neocortex doesn't have output knobs for everything).

Predicting, imagining, remembering, empathizing

The neocortex's generative models appear in the context of (1) prediction (including predicting the immediate future as it happens), (2) imagination, (3) memory, and (4) empathetic simulation (when we imagine *someone else* reacting to a spider, predicting a spider, etc.). I think all four of these processes rely on fundamentally the same mechanism in the neocortex, so by default the same generative models will be used for all four. Thus, we get the same hormone outputs in all four of these situations.

Hang on, you say: That doesn't seem right! If it were the exact same generative models, then when we remember dancing, we would actually issue the motor commands to start dancing! Well, I answer, we *do* actually sometimes move a little bit when we remember a motion! I think the rule is, loosely speaking, the top-down information flow is much stronger (more confident) when predicting, and much weaker for imagination, memory, and empathy. Thus, the neocortical output signals are weaker too, and this applies to both motor control outputs and hormone outputs. (Incidentally, I think motor control outputs are further subject to thresholding processes, downstream of the neocortex, and therefore a sufficiently weak motor command causes no motion at all.)

As discussed more below, the subcortex relies on the neocortex's outputs to guess what the neocortex is thinking about, and issue evolutionarily-appropriate guidance in response. Presumably, to do this job well, the subcortex needs to know whether a given neocortical output is part of a prediction, or memory, or imagination, or empathetic

simulation. From the above paragraph, I think it can distinguish predictions from the other three by the neocortical output strength. But how does it tell memory, imagination, and empathetic simulation apart from each other? I don't know! Then that suggests to me an interesting hypothesis: maybe it *can't*! What if some of our weirder instincts related to memory or counterfactual imagination are not adaptive at all, but rather crosstalk from social instincts, or vice-versa? For example, I think there's a reaction in the subcortex that listens for a strong prediction of lower reward, alternating with a weak prediction of higher reward; when it sees this combination, it issues negative reward and negative valence. Think about what this subcortical reaction would do in the three different cases: If the weak prediction it sees is an empathetic simulation, well, that's the core of jealousy! If the weak prediction it sees is a memory, well, that's the core of loss aversion! If the weak prediction it sees is a counterfactual imagination, well, that's the core of, I guess, that annoying feeling of having missed out on something good. Seems to fit together pretty well, right? I'm not super confident, but at least it's food for thought.

Various implications

Opening a window into the black-box neocortex

Each subcortical reaction has its own profile of facial, body, and hormone changes. The "reward criterion" above ensures that the neocortex will learn to imitate the characteristic consequences of reaction X whenever it is expecting, imagining, remembering, or empathetically simulating reaction X. This is then a window for the subcortex to get a glimpse into the goings-on inside the black-box neocortex.

In our running example, if the spider reaction creates a certain combination of facial, body, and hormone changes, then the subcortex can watch for this set of changes to happen exogenously (from its perspective), and if it does, the subcortex can infer that the neocortex was maybe thinking about spiders. Perhaps the subcortex might then issue its own spider reaction, fleshing out the neocortex's weak imitation. Or perhaps it could do something entirely different.

I have a hunch that social emotions rely on this. With this mechanism, it seems that the subcortex can build a hierarchy of increasingly complicated social reactions: "if I'm activating reaction A, and I think you're activating reaction B, then that triggers me to feel reaction C", "if I'm activating reaction C, and I think you're activating reaction A, then that triggers me to feel reaction D", and so on. Well, maybe. I'm still hazy on the details here and want to think about it more.

Complex back-and-forth between neocortex and subcortex

The neocortex can alter the hormones and body, which are among the inputs into the subcortical circuits. The subcortical circuits then also alter the hormones and body, which are among the inputs into the neocortex! Around and around it goes! So for example, if you tell yourself to calm down, your neocortex changes your hormones, which in turn increases the activation of the subcortical "I am safe and calm" reaction,

which reinforces and augments that change, which in turn makes it easier for the neocortex to continue feeling safe and calm! ... Until, of course, that pleasant cycle is broken by other subcortical reactions or other neocortical generative models butting in.

"Overcoming" subcortical reactions

Empirically, we know it's possible to "overcome" fear of spiders, and other subcortical reactions. I'm thinking there are two ways this might work. I think both are happening, but I'm not really sure.

First, there's subcortical learning ... well, "learning" isn't the right word here, because it's not trying to match some ground truth. (The only "ground truth" for subcortical reaction specifications is natural selection!) I think it's more akin to the [self-modifying code in Linux](#) than to the weight updates in ML. So let's call it **subcortical input-dependent dynamic rewiring rules**.

(By the way, elsewhere in the subcortex, like the cerebellum, there is *also* real stereotypical "learning" going on, akin to the weight updates in ML. That does happen, but it's not what I'm talking about here. In fact, I prefer to lump the cerebellum in with the neocortex as the learning-algorithm part of the brain.)

Maybe one subcortical dynamic rewiring rule says: *If the spider-detection reaction triggers, and then within 3 seconds the "I am safe and calm" reaction triggers, then next time the spider reaction should trigger more weakly.*

Second, there's neocortical learning—i.e., the neocortex developing new generative models. Let's say again that we're doing exposure therapy for fear of spiders, and let's say the two relevant subcortical reactions are the spider-detection reaction (which rewards the neocortex for producing anxiety hormones before it triggers) and the "I am safe and calm" reaction (which rewards the neocortex for producing calming hormones before *it* triggers). (I'm obviously oversimplifying here.) The neocortex could learn generative models that summon the "I am safe and calm" reaction whenever the spider-detection reaction is just starting to trigger. That generative model could potentially get entrenched and rewarded, as the spider-detection reaction is sorta preempted and thus can't issue a penalty for the lack of anxiety hormones, whereas the "I am safe and calm" reaction *does* issue a reward for the presence of calm hormones. Something like that?

I have no doubt that the second of these two processes—neocortical learning—really happens. The first might or might not happen, I don't know. It does seem like something that plausibly could happen, on both evolutionary and neurological grounds. So I guess my default assumption is that dynamic rewiring rules for subcortical reactions do in fact exist, but again, I'm not sure, I haven't thought about it much.

Things I still don't understand

I lumped together the subcortex into a monolithic unit. I actually understand very little about the functional decomposition beyond that. The tectum and tegmentum seem to be doing a lot of the calculations for what I'm calling "reactions", including the colliculi, which seem to house the subcortical sensory processing. What computations does the amygdala do, for example? It has 10 million neurons, they have to be calculating something!!! I really don't know. (Update 1 year later: On the plus side, I feel like I

understand the amygdala much better now; on the minus side, I was wrong to lump it in with "subcortex" rather than "neocortex". [See discussion here.](#))

As discussed above, I don't understand what the non-motor output signals from the neocortex are (update: see [here](#)), or whether things like valence and arousal correspond to hormones or something else. (Update: attempt to understand valence [here](#).)

I'm more generally uncertain about everything I wrote here, even where I used a confident tone. Honestly, I haven't found much in the systems neuroscience literature that's addressing the questions I'm interested in, although I imagine it's there somewhere and I'm reinventing lots of wheels (or re-making classic mistakes). As always, please let me know any thoughts, ideas, things you find confusing, etc. Thanks in advance!

1. A few people on this forum are thinking hard about the brain, and I've learned a lot from their writings—especially [Kaj's multi-agent sequence](#)—but my impression is that they're mostly working on the project of "Let's understand the brain so we can answer normative questions of what we want AGI to do and how value learning might work", whereas here I'm talking about "Let's understand the brain as a model of a possible AGI algorithm, and think about whether such an AGI algorithm can be used safely and beneficially". Y'all can correct me if I'm wrong :)
[←](#)
2. I will sloppily use the term "neocortex" as shorthand for "neocortex plus other structures that are intimately connected to the neocortex and are best thought of as part of the same algorithm"—this especially includes the hippocampus and thalamus. [←](#)
3. For what it's worth, Elon Musk mentioned in [a recent interview about Neuralink](#) that he is thinking about the brain this way as well: "We've got like a monkey brain with a computer on top of it, that's the human brain, and a lot of our impulses and everything are driven by the monkey brain, and the computer, the cortex, is constantly trying to make the monkey brain happy. It's not the cortex that's steering the monkey brain, it's the monkey brain steering the cortex." (14:45). Normally people would say "lizard brain" rather than "monkey brain" here, although even *that* terminology is unfair to lizards, who do in fact have something homologous to a neocortex. [←](#)
4. Unfortunately I don't have good evidence that this spider story is actually true. Does the midbrain *really* have specialized circuitry to detect spiders? There was [a study](#) that showed pictures of spiders to a [blindsight](#) person (i.e., a person who had an intact midbrain visual processing system but no visual cortex). It didn't work; nothing happened. But I think they did the experiment wrong—I think it has to be a video of a moving spider, not a stationary picture of a spider, to trigger the subcortical circuitry. (Source: introspection. Also, I think I read that the subcortical vision system has pretty low spatial resolution, so watching for a characteristic motion would seem a sensible design.) Anyway, it *has* to work this way, nothing else makes sense to me. I'm comfortable using this example prominently because if it turns out that this example is wrong, then I'm so very confused that this whole article is probably garbage anyway. For the record, I am basically describing Mark Johnson's "two-process" model, which is I think well established in the case of attending-to-faces in humans and filial imprinting in chicks (more [here](#)), even if it's speculative when applied to fear-of-spiders. [←](#)

5. I am pretty confident that neocortical patterns are effectively random at a microscopic, neuron-by-neuron level, and *that's* what matters when we talk about why it's impossible for evolution to directly create hardwired instincts that refer to a semantic concept in the neocortex. However, to be clear, at the level of gross anatomy, you *can* more-or-less predict in advance where different concepts will wind up getting stored in the neocortex, based on the large-scale patterns of information flow and the inputs it gets in a typical human life environment. To take an obvious example, low-level visual patterns are likely to be stored in the parts of the neocortex that receive low-visual visual information from the retina!
[←](#)
6. When I say "I didn't understand" something Barrett wrote, I mean more specifically that I can't see how to turn her words into a [gears-level](#) model of a computation that the brain might be doing. This category of "things I didn't understand" includes, in particular, almost everything she wrote about "body budgets", which was a major theme of the book that came up on almost every page... [←](#)
7. If you want to call the subcortical things "emotions" instead of "reactions", that's fine with me, as long as you distinguish them from "emotion concepts" in the neocortex. Barrett is really adamant that the word "emotion" *must* refer to the neocortical emotion concepts, not the subcortical reactions (I'm not even sure if she thinks the subcortical reactions exist), but for my part, I think reasonable people could differ, and it's ultimately a terminological question with no right answers anyway. [←](#)

The Best Virtual Worlds for "Hanging Out"

UPDATES Sept 5 2020

- Online Town is now [Gather Town](#). They have continued to ship new features and improve their service in the past few months.
- There are several apps similar to Gather Town now. My favorite is [Topia](#). It is less stable / performant than Gather Town, but also much prettier and easier to build maps on. Other options include [here.fm](#) and [cozyroom.xyz](#), and [highfidelity.com](#). Each have different strengths and weaknesses. I hope to do another article comparing them someday.
- I've found Minecraft worked fine as a social activity, but surprisingly didn't hold my interest as a world - ironically every time I went to a shared Minecraft world I just... sortof "played Minecraft" rather than actually hanging out the way I'd envisioned. I have done a few specific "Minecraft hikes" which went well as a one-off thing though.

Related:

- [What are the best online tools for meetups and meetings?](#)
- [Partying over the Internet: Technological Aspects](#)

The Problem With Zoom™

In the wake of coronavirus, many people have turned to Zoom, Skype, or Jitsi to maintain social ties. But I personally find it a bit awkward to do a video call when I don't have anything in particular to talk about. In real life, lulls in conversation could be filled with eating, or talking a walk and appreciating the scenery. In a videocall, there's either awkward silence, or I start using facebook or something and then get distracted.

Zoom calls also work less well for large parties. If you want the feeling of wandering through a house, listening in on various conversations and joining in, or spontaneously playing a simple game, I've found video calls a poor substitute. Zoom works best for smaller conversations with clear goals, or very structured conversations. ([Seder worked well](#) because it was "turn based")

What works much better IMO is some kind of low-key game going on in the background - something just complicated enough to let me fidget with it, without becoming the Main Activity which distracts from actually talking to friends.

There are various low-key games that may work for people. But there's a particular quality of Minecraft, and similar world-sims, that feel less like we're playing a game, and more like we're just hanging out in a world.

What are the best tools for *that*?

This post began as a response to "[What are the Best Online Tools For Meetups and Meetings](#)". It turned out to be pretty extensive, and I thought I'd make it a full post. It currently compares four apps. I may update it as I try more.

This post currently compares four apps: [Minecraft](#), [Mozilla Hubs](#), [AltspaceVR](#), and [Online Town](#). I may update it as I try more.

Each app has some different strengths, and costs. In a nutshell:

- **Online Town** is my favorite for casual online "parties", where you want to feel like you're in a house wandering into different rooms, chatting spontaneously with people and forming impromptu conversations. It's a 2d pixel world you can walk around in, videochatting with people nearby. It lives in a browser that doesn't require a download or login.
- **Mozilla Hubs** is similar to Online Town, but in 3d. It has many more features, but it's also a bit more complex, and people who aren't familiar with 3D video games may find it harder to use. It can be used via desktop browser, or via VR headset.
- **AltspaceVR** is similar to Mozilla Hubs, but a) requires an app download, and b) basically requires a VR headset (it also has a Windows app, but I would only recommend it for people who have VR headsets and want an immersive VR experience).
- **Minecraft** is a fully featured virtual world. It costs \$30 (the other three apps are free). It doesn't come with its own audio chat (but can easily be combined with other voice chat services). You get a fully featured world where you can build a house together, explore cool environments, interact with plants and animals, etc. Unlike the previous options it doesn't have "proximity chat". It works better if you're interacting with a smallish group of people, who can all hear each other and participate in a single conversation.

I think there are many other video games that work similarly to Minecraft (for instance, Animal Crossing [has been getting some press](#) as a "coronavirus virtual refuge"), but I'm not as familiar with them. This post is essentially comparing the first three options vs "Minecraft and other similar video games."

Key Considerations

Accessibility, vs features

Some apps just require a url, and immediately drop you into a party. Others make you create an account, or sign in, which creates awkward friction. If you want to host a large online party, it matters a lot that you can invite friends and that they can invite friends organically. If those friends have to stop to download a new app... they probably won't. A url they can just click on is much better.

On the flipside, downloadable apps offer more power and flexibility. It may be worth getting all your friends to invest in an app, if the result is better than free, accessible websites. But, you can only get all your friends to download apps so many times.

Online Town is extremely accessible. Mozilla Hubs is a close second – the only problem is that it runs sluggishly on some computers.

AltSpaceVR is free but requires a VR headset which most people don't have.

Minecraft is \$30 and requires creating an account, which I think is generally worth it but only if it's actually the most appropriate tool for the job.

Proximity Chat

Some apps make people louder if they are closer to you. This gives you sort of the organic conversation feeling that parties have – you can wander around a virtual room, briefly listening in and chatting with people until you find a conversation you're excited by.

Proximity chat is most important if you're aiming for a largish party.

Hubs, Altspace and Online Town all feature this out-of-the-box.

Minecraft does not have proximity chat. There is a mod you can download that provides it, but the Minecraft Mod Scene is a wild west of hacky downloads that I think only make sense to inflict on your friends if they're a particular kind of nerd who's excited by that. (Normal minecraft is "pretty accessible, apart from costing money", but I think modded Minecraft is basically a dealbreaker for inviting newcomers)

I found some videogames that had proximity chat built in, but they were more expensive (more like \$60). That's a bigger ask than I'd make of people I invited to a party. Since the whole point of proximity chat is to enable large freeform parties here, I didn't investigate them further.

VR

Altspace and *Mozilla Hubs* both allow you to go for "full immersion" if you own a VR headset. The Oculus Quest is around \$400 and in my opinion quite worth the price... but unfortunately seems to be sold out and I'm not sure when you can next get one. If coronavirus had struck 1-2 years later I think it might have been a valid option for tons of remote people to hang out in VR. Alas, the timing is slightly off.

I personally find VR gives me a bit of a headache and sometimes nausea, which limits my time to around an hour. I also expect this to improve a bit in another couple years. My current sense is that it's more like a fun novel experience to try than a serious contender for frequent-virtual-hangout space.

Intuitiveness / Smoothness

Creating good controls for a virtual world is tricky. I found Online Town by far the most intuitive. Altspace, Hubs and Minecraft are each differently unintuitive.

The Apps in Detail

Minecraft

This was my first experience with a virtual, telepresent world. Seven years ago, when my girlfriend and I were long-distance, we would stay connected via Minecraft dates.

The distinguishing feature of Minecraft is that it is not a game, and it is not a videochatting app. It is a world. You can plant trees, grow crops, invent tools, build a house, found a civilization.

It drops you off in the world with no context and few instructions. Some people find that confusing and bounce off of it. The trick is that Minecraft is like real life – it has meaning insofar as you invest meaning into it, and in my experience it's easier to create meaning together than by yourself. My girlfriend and I built a house together, which we decorated and treated as a shared home.

Minecraft is an infinite canvas, but like the real world, you have to work to accomplish things. There is risk and danger and cost. You can build a statue of gold, but only if you go dig up that gold. This gives things a sense of "weight" that they don't have in worlds where you can place any object you want immediately.

For Valentines Day, I built my girlfriend a treehouse on a hill. There was a place nearby where the sun set each day, but there was a large mountain partially blocking the view. I spent a week digging up the mountain and replacing it with a giant glass heart.

My interest in Minecraft has waxed and waned over the years. I generally find that I care about Minecraft in proportion to how much other people I'm close with are invested in the same world. It's a valid avenue for social reality, which has strength depending on how many people believe in it.

So, before coronavirus came, I already had a clear sense of what a virtual world that felt "lived in" could look like, and how to use that to connect with people who were far away. I find Minecraft a good place to "hang out" with friends, and go on little virtual hikes.

A big question (which I'm not yet certain of) is is how Minecraft compares more directly to other apps that *aren't* trying to be a living, breathing world. If you want a world, Minecraft provides, but what if you want a temporary, ephemeral party?

I've had some good experience with Minecraft Hikes. Last weekend, I invited friends to download WesterosCraft, a modded version of Minecraft (easier to install than most mods), which features the entire western continent from Game of Thrones. We hiked from Winterfell Castle to the large wall in the north. It made for a nice hour-long activity, during which we chatted and organically started playing the sorts of games we might play on a hike (things like "20 questions").

Online Town

Online Town is a video chat, where you get a little pixel avatar who can walk around a 2D virtual world. You can videochat with other people who are nearby, and can't see or hear people who are further away. It looks like this:



The app is only a month old, and has very few features, and lags when large numbers of people are in a room. But, it's a very elegant concept that I think is simple for people to understand. It runs in a browser, and doesn't require a download or account.

I think Online Town is the best option for most casual "parties" – creating an online gathering with around 15-30 people who can wander around into different conversations in a fairly organic way. It is both helpful for spontaneous conversation forming, and for giving a little feeling of "physical interaction."

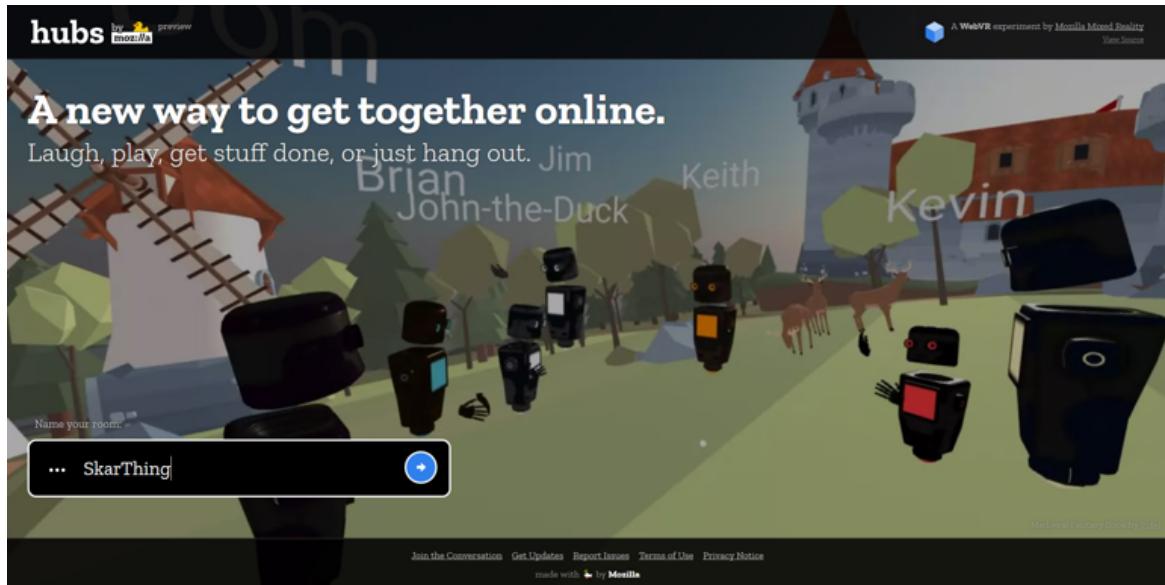
The developers are still adding new features, and I'm hoping that over time this gets fleshed out. I think the biggest obstacle right now is that it lags when too many people join, which directly undermines the core use-case. Optimization is hard and I'm not expecting that to change very soon. But in general, they've continued adding features since I first checked it out a couple weeks ago.

Longterm I think it'd be delightful if it gave people more opportunity to create their own pixel art to populate the world with, to capture some of the "actual virtual world" that Minecraft offers while maintaining Online Town's simplicity and elegance.

Another friend of mine would like a version of Online Town where instead of a teeny pixel avatar, your avatar *is* your videochat stream, so that you can immediately recognize people, and not have to shift attention back and forth between the pixel world and video streams below. I think this is also a pretty valid choice that I'd be excited to see the Online Town folk (or someone else) try.

Mozilla Hubs

Hubs is a virtual space where you can walk around in 3D, speaking and hearing the people near you. You can create "windows" that have youtube videos playing, or place 3D objects around that other people can wiggle around or resize. You can include a video stream of yourself, so people can see both your "virtual self" (which appears as a robot) while also looking at your face.



Hubs gets a lot of basic things right – it's web based, so it works across many platforms, including VR. It has all the features I expect/hope for Online Town to have eventually.

But the result is... just a bit too janky, occasionally laggy, and confusing in order for me to be deeply excited about it. I think this is an Uncanny Valley thing: Online Town is a teeny pixelated world that I find janky-but-cute. My personal experience of Mozilla Hubs was that it was *trying* to create an immersive 3D world, but the seams were too visible and it was harder to get into.

It *does* basically work just fine, and has more features than Town, so may be a better fit for some people's needs.

Mozilla Hubs takes place in "rooms", pre-built 3D environments you get to choose. Different rooms range *widely* in how graphics-intensive they are, so if you are finding the experience sluggish, you can switch to a different, simpler room.

One great thing for Mozilla Hubs is that it comes with a level-editor called [Spoke](#), which you can use to create whatever type of room you want, or edit existing rooms if they're Not Quite Right. The LessWrong team used it to create [Our April Fools Prank Room](#) (optimized for being computationally simple), and more recently I used it to create an [alternate version of the existing Foggy Lake room](#), removing some restrictions on player movement that the original version had come with.

Rooms can hold up to 25 people, but I found them to get laggy once they got more than 15 or so.

AltspaceVR

AltspaceVR is an obvious evolution of Mozilla Hubs (I've actually heard that Mozilla Hubs was founded by former Altspace employees). It takes place in VR, although there's a Windows app you can download.

It generally works more smoothly than Hubs. The world is slightly prettier and more cohesive. You can see people's mouth move when they talk. It prominently features Events where people meet to talk about particular things, or watch videos together, or play games.



One of my favorite bits about Altspace is that you can create your own world. And whereas Mozilla Hub's Spoke editor feels like a complicated, professional editing software, Altspace's world editor feels fun and (mostly) intuitive. Selecting objects and moving them around made feel like Tony Stark in Iron Man.



It felt like this.

The most interesting experience I had was "making a friend."

I traveled to a world with a cabin by a lake, where a few other players were hanging out. I listened in on their conversation a bit. At the time, I felt a bit nervous and didn't want to speak up.

I explored the nearby woods a bit, eventually finding a large animated stag. While I was staring at it, another player appeared. He said "man, that's a cool stag", and I said "yeah", and then we chatted a bit. Eventually he said "Hey man, wanna come to the game room?" and opened up a portal.

I followed him through the portal to a gaming lounge, which featured a large chess board. There were no special rules governing the board - I could just pick up pieces, and put them down, basically like real life. He and I started a game. We were both pretty bad at chess, and I focused more on taking pieces than winning. But we had a fun time, and afterwards he said "Hey man, wanna friend each other and hang out again sometime? I'd love a rematch." I said "sure!"

And... well, okay and then I ended up deciding I didn't want to spend much more time in Altspace because Minecraft was overall better. But, there was something good and pure about that interaction. It gave me a sense that I could actually just make friends in an organic fashion.

The Pro, and Con: VR

I think, 2 years from now, enough people will have VR headsets, and VR will have improved enough, that something like Altspace would be a great tool for quarantine socialization. As is, I think it doesn't quite feel good enough to really be better than simpler options like Online Town or richer options like Minecraft.

But Altspace did very much feel like The Future. If you read Snow Crash and thought "man, I want that", well, you can have it now.

Go Forth and Hang Out

This is all I got for now. Hopefully you have some new interesting ideas for ways to host online gatherings, or maintain relationships in a Socially Distant world.

Any other tools you've tried? I'd love to hear about them in the comments.

AI Alignment Podcast: An Overview of Technical AI Alignment in 2018 and 2019 with Buck Shlegeris and Rohin Shah

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

Just a year ago we released a two part episode titled [An Overview of Technical AI Alignment with Rohin Shah](#). That conversation provided details on the views of central AI alignment research organizations and many of the ongoing research efforts for designing safe and aligned systems. Much has happened in the past twelve months, so we've invited Rohin — along with fellow researcher Buck Shlegeris — back for a follow-up conversation. Today's episode focuses especially on the state of current research efforts for beneficial AI, as well as Buck's and Rohin's thoughts about the varying approaches and the difficulties we still face. This podcast thus serves as a non-exhaustive overview of how the field of AI alignment has updated and how thinking is progressing.

Topics discussed in this episode include:

- Rohin's and Buck's optimisms and pessimism about different approaches to aligned AI
- Traditional arguments for AI as an x-risk
- Modeling agents as expected utility maximizers
- Ambitious value learning and specification learning/narrow value learning
- Agency and optimization
- Robustness
- Scaling to superhuman abilities
- Universality
- Impact regularization
- Causal models, oracles, and decision theory
- Discontinuous and continuous takeoff scenarios
- Probability of AI-induced existential risk
- Timelines for AGI
- Information hazards

You can find the page for this podcast here: <https://futureoflife.org/2020/04/15/an-overview-of-technical-ai-alignment-in-2018-and-2019-with-buck-shlegeris-and-rohin-shah/>

Transcript

Note: The following transcript has been edited for style and clarity.

Lucas Perry: Welcome to the AI Alignment Podcast. I'm Lucas Perry. Today we have a special episode with Buck Shlegeris and Rohin Shah that serves as a review of progress in technical AI alignment over 2018 and 2019. This episode serves as an awesome birds eye view of the varying focus areas of technical AI alignment research and also helps to develop a sense of the field. I found this conversation to be super valuable for helping me to better understand the state and current trajectory of technical AI alignment research. This podcast covers traditional arguments for AI as an x-risk, what AI alignment is, the modeling of agents as expected utility maximizers, iterated distillation and amplification, AI safety via debate, agency and optimization, value learning, robustness, scaling to superhuman abilities, and more. The structure of this podcast is based on Rohin's AI Alignment Forum post titled AI Alignment 2018-19 Review. That post is an excellent resource to take a look at in addition to this podcast. Rohin also had a conversation with us about just a year ago titled An Overview of Technical AI Alignment with Rohin Shah. This episode serves as a follow up to that overview and as an update to what's been going on in the field. You can find a link for it on the page for this episode.

Buck Shlegeris is a researcher at the Machine Intelligence Research Institute. He tries to work to make the future good for sentient beings and currently believes that working on existential risk from artificial intelligence is the best way of doing this. Buck worked as a software engineer at PayPal before joining MIRI, and was the first employee at Triplebyte. He previously studied at the Australian National University, majoring in CS and minoring in math and physics, and he has presented work on data structure synthesis at industry conferences.

Rohin Shah is a 6th year PhD student in Computer Science at the Center for Human-Compatible AI at UC Berkeley. He is involved in Effective Altruism and was the co-president of EA UC Berkeley for 2015-16 and ran EA UW during 2016-2017. Out of concern for animal welfare, Rohin is almost vegan because of the intense suffering on factory farms. He is interested in AI, machine learning, programming languages, complexity theory, algorithms, security, and quantum computing to name a few. Rohin's research focuses on building safe and aligned AI systems that pursue the objectives their users intend them to pursue, rather than the objectives that were literally specified. He also publishes the Alignment Newsletter, which summarizes work relevant to AI alignment. The Alignment Newsletter is something I highly recommend that you follow in addition to this podcast.

And with that, let's get into our review of AI alignment with Rohin Shah and Buck Shlegeris.

To get things started here, the plan is to go through Rohin's post on the Alignment Forum about AI Alignment 2018 and 2019 In Review. We'll be using this as a way of structuring this conversation and as a way of moving methodically through things that have changed or updated in 2018 and 2019, and to use those as a place for conversation. So then, Rohin, you can start us off by going through this document. Let's start at the beginning, and we'll move through sequentially and jump in where necessary or where there is interest.

Rohin Shah: Sure, that sounds good. I think I started this out by talking about this basic analysis of AI risk that's been happening for the last couple of years. In particular, you have these traditional arguments, so maybe I'll just talk about the traditional argument first, which basically says that the AI systems that we're going to build are going to be powerful optimizers. When you optimize something, you tend to

get these sort of edge case outcomes, these extreme outcomes that are a little hard to predict ahead of time.

You can't just rely on tests with less powerful systems in order to predict what will happen, and so you can't rely on your normal common sense reasoning in order to deal with this. In particular, powerful AI systems are probably going to look like expected utility maximizers due to various coherence arguments, like the Von Neumann-Morgenstern rationality theorem, and these expected utility maximizers have convergent instrumental sub-goals, like not wanting to be switched off because then they can't achieve their goal, and wanting to accumulate a lot of power and resources.

The standard argument goes, because AI systems are going to be built this way, they will have these convergent instrumental sub-goals. This makes them dangerous because they will be pursuing goals that we don't want.

Lucas Perry: Before we continue too much deeper into this, I'd want to actually start off with a really simple question for both of you. What is AI alignment?

Rohin Shah: Different people mean different things by it. When I use the word alignment, I'm usually talking about what has been more specifically called intent alignment, which is basically aiming for the property that the AI system is trying to do what you want. It's trying to help you. Possibly it doesn't know exactly how to best help you, and it might make some mistakes in the process of trying to help you, but really what it's trying to do is to help you.

Buck Shlegeris: The way I would say what I mean by AI alignment, I guess I would step back a little bit, and think about why it is that I care about this question at all. I think that the fundamental fact which has me interested in anything about powerful AI systems of the future is that I think they'll be a big deal in some way or another. And when I ask myself the question "what are the kinds of things that could be problems about how these really powerful AI systems work or affect the world", one of the things which feels like a problem is that, we might not know how to apply these systems reliably to the kinds of problems which we care about, and so by default humanity will end up applying them in ways that lead to really bad outcomes. And so I guess, from that perspective, when I think about AI alignment, I think about trying to make ways of building AI systems such that we can apply them to tasks that are valuable, such that that they'll reliably pursue those tasks instead of doing something else which is really dangerous and bad.

I'm fine with intent alignment as the focus. I kind of agree with, for instance, Paul Christiano, that it's not my problem if my AI system incompetently kills everyone, that's the capability's people's problem. I just want to make the system so it's trying to cause good outcomes.

Lucas Perry: Both of these understandings of what it means to build beneficial AI or aligned AI systems can take us back to what Rohin was just talking about, where there's this basic analysis of AI risk, about AI as powerful optimizers and the associated risks there. With that framing and those definitions, Rohin, can you take us back into this basic analysis of AI risk?

Rohin Shah: Sure. The traditional argument looks like AI systems are going to be goal-directed. If you expect that your AI system is going to be goal-directed, and that goal is not the one that humans care about, then it's going to be dangerous because it's going to try to gain power and resources with which to achieve its goal.

If the humans tried to turn it off, it's going to say, "No, don't do that," and it's going to try to take actions that avoid that. So it pits the AI and the humans in an adversarial game with each other, and you ideally don't want to be fighting against a superintelligent AI system. That seems bad.

Buck Shlegeris: I feel like Rohin is to some extent setting this up in a way that he's then going to argue is wrong, which I think is kind of unfair. In particular, Rohin, I think you're making these points about VNM theorems and stuff to set up the fact that it seems like these arguments don't actually work. I feel that this makes it kind of unfairly sound like the earlier AI alignment arguments are wrong. I think this is an incredibly important question, of whether early arguments about the importance of AI safety were quite flawed. My impression is that overall the early arguments about AI safety were pretty good. And I think it's a very interesting question whether this is in fact true. And I'd be interested in arguing about it, but I think it's the kind of thing that ought to be argued about explicitly.

Rohin Shah: Yeah, sure.

Buck Shlegeris: And I get that you were kind of saying it narratively, so this is only a minor complaint. It's a thing I wanted to note.

Rohin Shah: I think my position on that question of "how good were the early AI risk arguments," probably people's internal beliefs were good as to why AI was supposed to be risky, and the things they wrote down were not very good. Some things were good and some things weren't. I think [Intelligence Explosion Microeconomics](#) was good. I think [AI Alignment: Why It's Hard and Where to Start](#), was misleading.

Buck Shlegeris: I think I agree with your sense that people probably had a lot of reasonable beliefs but that the written arguments seem flawed. I think another thing that's true is that random people like me who were on LessWrong in 2012 or something, ended up having a lot of really stupid beliefs about AI alignment, which I think isn't really the fault of the people who were thinking about it the best, but is maybe sociologically interesting.

Rohin Shah: Yes, that seems plausible to me. Don't have a strong opinion on it.

Lucas Perry: To provide a little bit of framing here and better analysis of basic AI x-risk arguments, can you list what the starting arguments for AI risk were?

Rohin Shah: I think I am reasonably well portraying what the written arguments were. Underlying arguments that people probably had would be something more like, "Well, it sure seems like if you want to do useful things in the world, you need to have AI systems that are pursuing goals." If you have something that's more like tool AI, like Google Maps, that system is going to be good at the one thing it was designed to do, but it's not going to be able to learn and then apply its knowledge to new tasks autonomously. It sure seems like if you want to do really powerful things in the world, like run companies or make policies, you probably do need AI systems that are constantly learning about their world and applying their knowledge in order to come up with new ways to do things.

In the history of human thought, we just don't seem to know of a way to cause that to happen except by putting goals in systems, and so probably AI systems are going to be goal-directed. And one way you can formalize goal-directedness is by thinking about expected utility maximizers, and people did a bunch of formal analysis of that.

Mostly going to ignore it because I think you can just say all the same thing with the idea of pursuing goals and it's all fine.

Buck Shlegeris: I think one important clarification to that, is you were saying the reason that tool AIs aren't just the whole story of what happens with AI is that you can't apply it to all problems. I think another important element is that people back then, and I now, believe that if you want to build a really good tool, you're probably going to end up wanting to structure that as an agent internally. And even if you aren't trying to structure it as an agent, if you're just searching over lots of different programs implicitly, perhaps by training a really large recurrent policy, you're going to end up finding something agent shaped.

Rohin Shah: I don't disagree with any of that. I think we were using the words tool AI differently.

Buck Shlegeris: Okay.

Rohin Shah: In my mind, if we're talking about tool AI, we're imagining a pretty restricted action space where no matter what actions in this action space are taken, with high probability, nothing bad is going to happen. And you'll search within that action space, but you don't go to arbitrary action in the real world or something like that. This is what makes tool AI hard to apply to all problems.

Buck Shlegeris: I would have thought that's a pretty non-standard use of the term tool AI.

Rohin Shah: Possibly.

Buck Shlegeris: In particular, I would have thought that restricting the action space enough that you're safe, regardless of how much it wants to hurt you, seems kind of non-standard.

Rohin Shah: Yes. I have never really liked the concept of tool AI very much, so I kind of just want to move on.

Lucas Perry: Hey, It's post-podcast Lucas here. I just want to highlight here a little bit of clarification that Rohin was interested in adding, which is that he thinks that "tool AI evokes a sense of many different properties that he doesn't know which properties most people are usually thinking about and as a result he prefers not to use the phrase tool AI. And instead would like to use more precise terminology. He doesn't necessarily feel though that the concepts underlying tool AI are useless." So let's tie things a bit back to these basic arguments for x-risk that many people are familiar with, that have to do with convergent instrumental sub-goals and the difficulty of specifying and aligning systems with our goals and what we actually care about in our preference hierarchies.

One of the things here that Buck was seeming to bring up, he was saying that you may have been narratively setting up the Von Neumann-Morgenstern theorem, which sets up AIs as expected utility maximizers, and that you are going to argue that that argument, which is sort of the formalization of these earlier AI risk arguments, that that is less convincing to you now than it was before, but Buck still thinks that these arguments are strong. Could you unpack this a little bit more or am I getting this right?

Rohin Shah: To be clear, I also agree with Buck, that the spirit of the original arguments does seem correct, though, there are people who disagree with both of us about that. Basically, the VNM theorem roughly says, if you have preferences over a set of outcomes, and you satisfy some pretty intuitive axioms about how you make decisions, then you can represent your preferences using a utility function such that your decisions will always be, choose the action that maximizes the expected utility. This is, at least in writing, given as a reason to expect that AI systems would be maximizing expected utility. The thing is, when you talk about AI systems that are acting in the real world, they're just selecting a universe history, if you will. Any observed behavior is compatible with the maximization of some utility function. Utility functions are a really, really broad class of things when you apply it to choosing from universe histories.

Buck Shlegeris: An intuitive example of this: suppose that you see that every day I walk home from work in a really inefficient way. It's impossible to know whether I'm doing that because I happened to really like that path. For any sequence of actions that I take, there's some utility functions such that that was the optimal sequence of actions. And so we don't actually learn anything about how my policy is constrained based on the fact that I'm an expected utility maximizer.

Lucas Perry: Right. If I only had access to your behavior and not your insides.

Rohin Shah: Yeah, exactly. If you have a robot twitching forever, that's all it does, there is a utility function over a universe history that says that is the optimal thing to do. Every time the robot twitches to the right, it's like, yeah, the thing that was optimal to do at that moment in time was twitching to the right. If at some point somebody takes a hammer and smashes the robot and it breaks, then the utility function that corresponds to that being optimal is like, yeah, that was the exact right moment to break down.

If you have these pathologically complex utility functions as possibilities, every behavior is compatible with maximizing expected utility, you might want to say something like, probably we'll have the simple utility maximizers, but that's a pretty strong assumption, and you'd need to justify it somehow. And the VNM theorem wouldn't let you do that.

Lucas Perry: So is the problem here that you're unable to fully extract human preference hierarchies from human behavior?

Rohin Shah: Well, you're unable to extract agent preferences from agent behavior. You can see any agent behavior and you can rationalize it as expected utility maximization, but it's not very useful. Doesn't give you predictive power.

Buck Shlegeris: I just want to have my go at saying this argument in three sentences. Once upon a time, people said that because all rational systems act like they're maximizing an expected utility function, we should expect them to have various behaviors like trying to maximize the amount of power they have. But every set of actions that you could take is consistent with being an expected utility maximizer, therefore you can't use the fact that something is an expected utility maximizer in order to argue that it will have a particular set of behaviors, without making a bunch of additional arguments. And I basically think that I was wrong to be persuaded by the naive argument that Rohin was describing, which just goes directly from rational things are expected utility maximizers, to therefore rational things are power maximizing.

Rohin Shah: To be clear, this was the thing I also believed. The main reason I wrote the post that argued against it was because I spent half a year under the delusion that this was a valid argument.

Lucas Perry: Just for my understanding here, the view is that because any behavior, any agent from the outside can be understood as being an expected utility maximizer, that there are behaviors that clearly do not do instrumental sub-goal things, like maximize power and resources, yet those things can still be viewed as expected utility maximizers from the outside. So additional arguments are required for why expected utility maximizers do instrumental sub-goal things, which are AI risky.

Rohin Shah: Yeah, that's exactly right.

Lucas Perry: Okay. What else is on offer other than expected utility maximizers? You guys talked about comprehensive AI services might be one. Are there other formal agentive classes of 'thing that is not an expected utility maximizer but still has goals'?

Rohin Shah: A formalism for that? I think some people like John Wentworth is for example, thinking about markets as a model of agency. Some people like to think of multi-agent groups together leading to an emergent agency and want to model human minds this way. How formal are these? Not that formal yet.

Buck Shlegeris: I don't think there's anything which is competitively popular with expected utility maximization as the framework for thinking about this stuff.

Rohin Shah: Oh yes, certainly not. Expected utility maximization is used everywhere. Nothing else comes anywhere close.

Lucas Perry: So there's been this complete focus on utility functions and representing the human utility function, whatever that means. Do you guys think that this is going to continue to be the primary way of thinking about and modeling human preference hierarchies? How much does it actually relate to human preference hierarchies? I'm wondering if it might just be substantially different in some way.

Buck Shlegeris: Me and Rohin are going to disagree about this. I think that trying to model human preferences as a utility function is really dumb and bad and will not help you do things that are useful. I don't know; If I want to make an AI that's incredibly good at recommending me movies that I'm going to like, some kind of value learning thing where it tries to learn my utility function over movies is plausibly a good idea. Even things where I'm trying to use an AI system as a receptionist, I can imagine value learning being a good idea.

But I feel extremely pessimistic about more ambitious value learning kinds of things, where I try to, for example, have an AI system which learns human preferences and then acts in large scale ways in the world. I basically feel pretty pessimistic about every alignment strategy which goes via that kind of a route. I feel much better about either trying to not use AI systems for problems where you have to think about large scale human preferences, or having an AI system which does something more like modeling what humans would say in response to various questions and then using that directly instead of trying to get a value function out of it.

Rohin Shah: Yeah. Funnily enough, I was going to start off by saying I think Buck and I are going to agree on this.

Buck Shlegeris: Oh.

Rohin Shah: And I think I mostly agree with the things that you said. The thing I was going to say was I feel pretty pessimistic about trying to model the normative underlying human values, where you have to get things like population ethics right, and what to do with the possibility of infinite value. How do you deal with fanaticism? What's up with moral uncertainty? I feel pretty pessimistic about any sort of scheme that involves figuring that out before developing human-level AI systems.

There's a related concept which is also called value learning, which I would prefer to be called something else, but I feel like the name's locked in now. In my sequence, I called it narrow value learning, but even that feels bad. Maybe at least for this podcast we could call it specification learning, which is sort of more like the tasks Buck mentioned, like if you want to learn preferences over movies, representing that using a utility function seems fine.

Lucas Perry: Like superficial preferences?

Rohin Shah: Sure. I usually think of it as you have in mind a task that you want your AI system to do, and now you have to get your AI system to reliably do it. It's unclear whether this should even be called a value learning at this point. Maybe it's just the entire alignment problem. But techniques like inverse reinforcement learning, preference learning, learning from corrections, inverse reward design where you learn from a proxy reward, all of these are more trying to do the thing where you have a set of behaviors in mind, and you want to communicate that to the agent.

Buck Shlegeris: The way that I've been thinking about how optimistic I should be about value learning or specification learning recently has been that I suspect that at the point where AI is human level, by default we'll have value learning which is about at human level. We're about as good at giving AI systems information about our preferences that it can do stuff with as we are giving other humans information about our preferences that we can do stuff with. And when I imagine hiring someone to recommend music to me, I feel like there are probably music nerds who could do a pretty good job of looking at my Spotify history, and recommending bands that I'd like if they spent a week on it. I feel a lot more pessimistic about being able to talk to a philosopher for a week, and then them answer hard questions about my preferences, especially if they didn't have the advantage of already being humans themselves.

Rohin Shah: Yep. That seems right.

Buck Shlegeris: So maybe that's how I would separate out the specification learning stuff that I feel optimistic about from the more ambitious value learning stuff that I feel pretty pessimistic about.

Rohin Shah: I do want to note that I collated a bunch of stuff arguing against ambitious value learning. If I had to make a case for optimism about even that approach, it would look more like, "Under the value learning approach, it seems possible with uncertainty over rewards, values, preferences, whatever you want to call them to get an AI system such that you actually are able to change it, because it would reason that if you're trying to change it, well then that means something about it is currently not good for helping you and so it would be better to let itself be changed. I'm not very convinced by this argument."

Buck Shlegeris: I feel like if you try to write down four different utility functions that the agent is uncertain between, I think it's just actually really hard for me to imagine concrete scenarios where the AI is corrigible as a result of its uncertainty over utility functions. Imagine the AI system thinks that you're going to switch it off and replace it

with an AI system which has a different method of inferring values from your actions and your words. It's not going to want to let you do that, because its utility function is to have the world be the way that is expressed by your utility function as estimated the way that it approximates utility functions. And so being replaced by a thing which estimates utility functions or infers utility functions some other way means that it's very unlikely to get what it actually wants, and other arguments like this. I'm not sure if these are super old arguments that you're five levels of counter-arguments to.

Rohin Shah: I definitely know this argument. I think the problem of fully updated deference is what I would normally point to as representing this general class of claims and I think it's a good counter argument. When I actually think about this, I sort of start getting confused about what it means for an AI system to terminally value the final output of what its value learning system would do. It feels like some additional notion of how the AI chooses actions has been posited, that hasn't actually been captured in the model and so I feel fairly uncertain about all of these arguments and kind of want to defer to the future.

Buck Shlegeris: I think the thing that I'm describing is just what happens if you read the algorithm literally. Like, if you read the value learning algorithm literally, it has this notion of the AI system wants to maximize the human's actual utility function.

Rohin Shah: For an optimal agent playing a CIRL (cooperative inverse reinforcement learning) game, I agree with your argument. If you take optimality as defined in the cooperative inverse reinforcement learning paper and it's playing over a long period of time, then yes, it's definitely going to prefer to keep itself in charge rather than a different AI system that would infer values in a different way.

Lucas Perry: It seems like so far utility functions are the best way of trying to get an understanding of what human beings care about and value and have preferences over, you guys are bringing up all of the difficult intricacies with trying to understand and model human preferences as utility functions. One of the things that you also bring up here, Rohin, in your review, is the risk of lock-in, which may require us to solve hard philosophical problems before the development of AGI. That has something to do with ambitious value learning, which would be like learning the one true human utility function which probably just doesn't exist.

Buck Shlegeris: I think I want to object to a little bit of your framing there. My stance on utility functions of humans isn't that there are a bunch of complicated subtleties on top, it's that modeling humans with utility functions is just a really sad state to be in. If your alignment strategy involves positing that humans behave as expected utility maximizers, I am very pessimistic about it working in the short term, and I just think that we should be trying to completely avoid anything which does that. It's not like there's a bunch of complicated sub-problems that we need to work out about how to describe us as expected utility maximizers, my best guess is that we would just not end up doing that because it's not a good idea.

Lucas Perry: For the ambitious value learning?

Buck Shlegeris: Yeah, that's right.

Lucas Perry: Okay, do you have something that's on offer?

Buck Shlegeris: The two options instead of that, which seem attractive to me? As I said earlier, one is you just convince everyone to not use AI systems for things where you need to have an understanding of large scale human preferences. The other one

is the kind of thing that Paul Christiano's iterated distillation and amplification, or a variety of his other ideas, the kind of thing that he's trying to get there is, I think, if you make a really powerful AI system, it's actually going to have an excellent model of human values in whatever representation is best for actually making predictions about humans because a really excellent AGI, like a really excellent paperclip maximizer, it's really important for it to really get how humans work so that it can manipulate them into letting it build lots of paperclip factories or whatever.

So I think that if you think that we have AGI, then by assumption I think we have a system which is able to reason about human values if it wants. And so if we can apply these really powerful AI systems to tasks such that the things that they do display their good understanding of human values, then we're fine and it's just okay that there was no way that we could represent a utility function directly. So for instance, the idea in IDA is that if we could have this system which is just trying to answer questions the same way that humans would, but enormously more cheaply because it can run faster than humans and a few other tricks, then we don't have to worry about writing down a utility functions of humans directly because we can just make the system do things that are kind of similar to the things humans would have done, and so it implicitly has this human utility function built into it. That's option two. Option one is don't use anything that requires a complex human utility function, option two is have your systems learn human values implicitly, by giving them a task such that this is beneficial for them and such that their good understanding of human values comes out in their actions.

Rohin Shah: One way I might condense that point, is that you're asking for a nice formalism for human preferences and I just point to all the humans out there in the world who don't know anything about utility functions, which is 99% of them and nonetheless still seem pretty good at inferring human preferences.

Lucas Perry: On this part about AGI, if it is AGI it should be able to reason about human preferences, then why would it not be able to construct something that was more explicit and thus was able to do more ambitious value learning?

Buck Shlegeris: So it can totally do that, itself. But we can't force that structure from the outside with our own algorithms.

Rohin Shah: Image classification is a good analogy. Like, in the past we were using hand engineered features, namely SIFT and HOG and then training classifiers over these hand engineered features in order to do image classification. And then we came to the era of deep learning and we just said, yeah, throw away all those features and just do everything end to end with a convolutional neural net and it worked way better. The point was that, in fact there are good representations for most tasks and humans trying to write them down ahead of time just doesn't work very well at that. It tends to work better if you let the AI system discover its own representations that best capture the thing you wanted to capture.

Lucas Perry: Can you unpack this point a little bit more? I'm not sure that I'm completely understanding it. Buck is rejecting this modeling human beings explicitly as expected utility maximizers and trying to explicitly come up with utility functions in our AI systems. The first was to convince people not to use these kinds of things. And the second is to make it so that the behavior and output of the AI systems has some implicit understanding of human behavior. Can you unpack this a bit more for me or give me another example?

Rohin Shah: So here's another example. Let's say I was teaching my kid that I don't have, how to catch a ball. It seems that the formalism that's available to me for learning how to catch a ball is, well, you can go all the way down to look at our best models of physics, we could use Newtonian mechanics let's say, like here are these equations, estimate the velocity and the distance of the ball and the angle at which it's thrown plug that into these equations and then predict that the ball's going to come here and then just put your hand there and then magically catch it. We won't even talk about the catching part. That seems like a pretty shitty way to teach a kid how to catch a ball.

Probably it's just a lot better to just play catch with the kid for a while and let the kid's brain figure out this is how to predict where the ball is going to go such that I can predict where it's going to be and then catch it.

I'm basically 100% confident that the thing that the brain is doing is not Newtonian mechanics. It's doing something else that's just way more efficient at predicting where the ball is going to be so that I can catch it and if I forced the brain to use Newtonian mechanics, I bet it would not do very well at this task.

Buck Shlegeris: I feel like that still isn't quite saying the key thing here. I don't know how to say this off the top of my head either, but I think there's this key point about: just because your neural net can learn a particular feature of the world doesn't mean that you can back out some other property of the world by forcing the neural net to have a particular shape. Does that make any sense, Rohin?

Rohin Shah: Yeah, vaguely. I mean, well, no, maybe not.

Buck Shlegeris: The problem isn't just the capabilities problem. There's this way you can try and infer a human utility function by asking, according to this model, what's the maximum likelihood utility function given all these things the human did. If you have a good enough model, you will in fact end up making very good predictions about the human, it's just that the decomposition into their planning function and their utility function is not going to result in a utility function which is anything like a thing that I would want maximized if this process was done on me. There is going to be some decomposition like this, which is totally fine, but the utility function part just isn't going to correspond to the thing that I want.

Rohin Shah: Yeah, that is also a problem, but I agree that is not the thing I was describing.

Lucas Perry: Is the point there that there's a lack of alignment between the utility function and the planning function. Given that the planning function imperfectly optimizes the utility function.

Rohin Shah: It's more like there are just infinitely many possible pairs of planning functions and utility functions that exactly predict human behavior. Even if it were true that humans were expected utility maximizers, which Buck is arguing we're not, and I agree with him. There is a planning function that's like humans are perfectly anti-rational and if you're like what utility function works with that planner to predict human behavior. Well, the literal negative of the true utility function when combined with the anti-rational planner produces the same behavior as the true utility function with the perfect planner, there's no information that lets you distinguish between these two possibilities.

You have to build it in as an assumption. I think Buck's point is that building things in as assumptions is probably not going to work.

Buck Shlegeris: Yeah.

Rohin Shah: A point I agree with. In philosophy this is called the is-ought problem, right? What you can train your AI system on is a bunch of "is" facts and then you have to add in some assumptions in order to jump to "ought" facts, which is what the utility function is trying to do. The utility function is trying to tell you how you ought to behave in new situations and the point of the is-ought distinction is that you need some bridging assumptions in order to get from is to ought.

Buck Shlegeris: And I guess an important part here is your system will do an amazing job of answering "is" questions about what humans would say about "ought" questions. And so I guess maybe you could phrase the second part as: to get your system to do things that match human preferences, use the fact that it knows how to make accurate "is" statements about humans' ought statements?

Lucas Perry: It seems like we're strictly talking about inferring the human utility function or preferences via looking at behavior. What if you also had more access to the actual structure of the human's brain?

Rohin Shah: This is like the approach that Stuart Armstrong likes to talk about. The same things still apply. You still have the is-ought problem where the facts about the brain are "is" facts and how you translate that into "ought" facts is going to involve some assumptions. Maybe you can break down such assumptions that everyone would agree with. Maybe it's like if this particular neuron in a human brain spikes, that's a good thing and we want more of it and if this other one spikes, that's a bad thing. We don't want it. Maybe that assumption is fine.

Lucas Perry: I guess I'm just pointing out, if you could find the places in the human brain that generate the statements about Ought questions.

Rohin Shah: As Buck said, that lets you predict what humans would say about ought statements, which your assumption could then be, whatever humans say about ought statements, that's what you ought to do. And that's still an assumption. Maybe it's a very reasonable assumption that we're happy to put it into our AI system.

Lucas Perry: If we're not willing to accept some humans' "is" statements about "ought" questions then we have to do some meta-ethical moral policing in our assumptions around getting "is" statements from "ought" questions.

Rohin Shah: Yes, that seems right to me. I don't know how you would do such a thing, but you would have to do something along those lines.

Buck Shlegeris: I would additionally say that I feel pretty great about trying to do things which use the fact that we can trust our AI to have good "is" answers to "ought" questions, but there's a bunch of problems with this. I think it's a good starting point but trying to use that to do arbitrarily complicated things in the world has a lot of problems. For instance, suppose I'm trying to decide whether we should design a city this way or that way. It's hard to know how to go from the ability to know how humans would answer questions about preferences to knowing what you should do to design the city. And this is for a bunch of reasons, one of them is that the human might not be able to figure out from your city building plans what the city's going to actually be like. And another is that the human might give inconsistent answers about

what design is good, depending on how you phrase the question, such that if you try to figure out a good city plan by optimizing for the thing that the human is going to be most enthusiastic about, then you might end up with a bad city plan. Paul Christiano has written in a lot of detail about a lot of this.

Lucas Perry: That also reminds me of what Stuart Armstrong wrote about the framing on the questions changing output on the preference.

Rohin Shah: Yep.

Buck Shlegeris: Sorry, to be clear other people than Paul Christiano have also written a lot about this stuff, (including Rohin). My favorite writing about this stuff is by Paul.

Lucas Perry: Yeah, those do seem problematic but it would also seem that there would be further “is” statements that if you queried people's meta-preferences about those things, you would get more “is” statements about that, but then that just pushes the “ought” assumptions that you need to make further back. Getting into very philosophically weedy territory. Do you think that this kind of thing could be pushed to the long reflection as is talked about by William MacAskill and Toby Ord or how much of this do you actually think needs to be solved in order to have safe and aligned AGI?

Buck Shlegeris: I think there are kind of two different ways that you could hope to have good outcomes from AGI. One is: set up a world such that you never needed to make an AGI which can make large scale decisions about the world. And two is: solve the full alignment problem.

I'm currently pretty pessimistic about the second of those being technically feasible. And I'm kind of pretty pessimistic about the first of those being a plan that will work. But in the world where you can have everyone only apply powerful and dangerous AI systems in ways that don't require an understanding of human values, then you can push all of these problems onto the long reflection. In worlds where you can do arbitrarily complicated things in ways that humans would approve of, you don't really need to long reflect this stuff because of the fact that these powerful AI systems already have the capacity of doing portions of the long reflection work inside themselves as needed. ([Quotes about the long reflection](#))

Rohin Shah: Yeah, so I think my take, it's not exactly disagreeing with Buck. It's more like from a different frame as Buck's. If you just got AI systems that did the things that humans did now, this does not seem to me to obviously require solving hard problems in philosophy. That's the lower bound on what you can do before having to do long reflection type stuff. Eventually you do want to do a longer reflection. I feel relatively optimistic about having a technical solution to alignment that allows us to do the long reflection after building AI systems. So the long reflection would include both humans and AI systems thinking hard, reflecting on difficult problems and so on.

Buck Shlegeris: To be clear, I'm super enthusiastic about there being a long reflection or something along those lines.

Lucas Perry: I always find it useful reflecting on just how human beings do many of these things because I think that when thinking about things in the strict AI alignment sense, it can seem almost impossible, but human beings are able to do so many of these things without solving all of these difficult problems. It seems like in the very

least, we'll be able to get AI systems that very, very approximately do what is good or what is approved of by human beings because we can already do that.

Buck Shlegeris: That argument doesn't really make sense to me. It also didn't make sense when Rohin referred to it a minute ago.

Rohin Shah: It's not an argument for we technically know how to do this. It is more an argument for this as at least within the space of possibilities.

Lucas Perry: Yeah, I guess that's how I was also thinking of it. It is within the space of possibilities. So utility functions are good because they can be optimized for, and there seem to be risks with optimization. Is there anything here that you guys would like to say about better understanding agency? I know this is one of the things that is important within the MIRI agenda.

Buck Shlegeris: I am a bad MIRI employee. I don't really get that part of the MIRI agenda, and so I'm not going to defend it. I have certainly learned some interesting things from talking to Scott Garrabrant and other MIRI people who have lots of interesting thoughts about this stuff. I don't quite see the path from there to good alignment strategies. But I also haven't spent a super long time thinking about it because I, in general, don't try to think about all of the different AI alignment things that I could possibly think about.

Rohin Shah: Yeah. I also am not a good person to ask about this. Most of my knowledge comes from reading things and MIRI has stopped writing things very much recently, so I don't know what their ideas are. I, like Buck, don't really see a good alignment strategy that starts with, first we understand optimization and so that's the main reason why I haven't looked into it very much.

Buck Shlegeris: I think I don't actually agree with the thing you said there, Rohin. I feel like understanding optimization could plausibly be really nice. Basically the story there is, it's a real bummer if we have to make really powerful AI systems via searching over large recurrent policies for things that implement optimizers. If it turned out that we could figure out some way of coding up optimizer stuffs directly, then this could maybe mean you didn't need to make mesa-optimizers. And maybe this means that your inner alignment problems go away, which could be really nice. The thing that I was saying I haven't thought that much about is, the relevance of thinking about, for instance, the various weirdnesses that happen when you consider embedded agency or decision theory, and things like that.

Rohin Shah: Oh, got it. Yeah. I think I agree that understanding optimization would be great if we succeeded at it and I'm mostly pessimistic about us succeeding at it, but also there are people who are optimistic about it and I don't know why they're optimistic about it.

Lucas Perry: Hey it's post-podcast Lucas here again. So, I just want to add a little more detail here again on behalf of Rohin. Here he feels pessimistic about us understanding optimization well enough and in a short enough time period that we are able to create powerful optimizers that we understand that rival the performance of the AI systems we're already building and will build in the near future. Back to the episode.

Buck Shlegeris: The arguments that MIRI has made about this,... they think that there are a bunch of questions about what optimization is, that are plausibly just not that hard compared to other problems which small groups of people have occasionally

solved, like coming up with foundations of mathematics, kind of a big conceptual deal but also a relatively small group of people. And before we had formalizations of math, I think it might've seemed as impossible to progress on as formalizing optimization or coming up with a better picture of that. So maybe that's my argument for some optimism.

Rohin Shah: Yeah, I think pointing to some examples of great success does not imply... Like there are probably many similar things that didn't work out and we don't know about them cause nobody bothered to tell us about them because they failed. Seems plausible maybe.

Lucas Perry: So, exploring more deeply this point of agency can either, or both of you, give us a little bit of a picture about the relevance or non relevance of decision theory here to AI alignment and I think, Buck, you mentioned the trickiness of embedded decision theory.

Rohin Shah: If you go back to our traditional argument for AI risk, it's basically powerful AI systems will be very strong optimizers. They will possibly be misaligned with us and this is bad. And in particular one specific way that you might imagine this going wrong is this idea of mesa optimization where we don't know how to build optimizers right now. And so what we end up doing is basically search across a huge number of programs looking for ones that do well at optimization and use that as our AGI system. And in this world, if you buy that as a model of what's happening, then you'll basically have almost no control over what exactly that system is optimizing for. And that seems like a recipe for misalignment. It sure would be better if we could build the optimizer directly and know what it is optimizing for. And in order to do that, we need to know how to do optimization well.

Lucas Perry: What are the kinds of places that we use mesa optimizers today?

Rohin Shah: It's not used very much yet. The field of meta learning is the closest example. In the field of meta learning you have a distribution over tasks and you use gradient descent or some other AI technique in order to find an AI system that itself, once given a new task, learns how to perform that task well.

Existing meta learning systems are more like learning how to do all the tasks well and then when they'll see a new task they just figure out ah, it's this task and then they roll out the policy that they already learned. But the eventual goal for meta learning is to get something that, online, learns how to do the task without having previously figured out how to do that task.

Lucas Perry: Okay, so Rohin did what you say cover embedded decision theory?

Rohin Shah: No, not really. I think embedded decision theory is just, we want to understand optimization. Our current notion of optimization, one way you could formalize it is to say my AI agent is going to have Bayesian belief over all the possible ways that the environment could be. It's going to update that belief over time as it gets observations and then it's going to act optimally with respect to that belief, by maximizing its expected utility. And embedded decision theory basically calls into question the idea that there's a separation between the agent and the environment. In particular I, as a human, couldn't possibly have a Bayesian belief about the entire earth because the entire Earth contains me. I can't have a Bayesian belief over myself so this means that our existing formalization of agency is flawed. It can't capture these things that affect real agents. And embedded decision theory, embedded

agency, more broadly, is trying to deal with this fact and have a new formalization that works even in these situations.

Buck Shlegeris: I want to give my understanding of the pitch for it. One part is that if you don't understand embedded agency, then if you try to make an AI system in a hard coded way, like making a hard coded optimizer, traditional phrasings of what an optimizer is, are just literally wrong in that, for example, they're assuming that you have these massive beliefs over world states that you can't really have. And plausibly, it is really bad to try to make systems by hardcoded assumptions that are just clearly false. And so if we want to hardcore agents with particular properties, it would be good if we knew a way of coding the agent that isn't implicitly making clearly false assumptions.

And the second pitch for it is something like when you want to understand a topic, sometimes it's worth looking at something about the topic which you're definitely wrong about, and trying to think about that part until you are less confused about it. When I'm studying physics or something, a thing that I love doing is looking for the easiest question whose answer I don't know, and then trying to just dive in until I have satisfactorily answered that question, hoping that the practice that I get about thinking about physics from answering a question correctly will generalize to much harder questions. I think that's part of the pitch here. Here is a problem that we would need to answer, if we wanted to understand how superintelligent AI systems work, so we should try answering it because it seems easier than some of the other problems.

Lucas Perry: Okay. I think I feel satisfied. The next thing here Rohin in your AI alignment 2018-19 review is value learning. I feel like we've talked a bunch about this already. Is there anything here that you want to say or do you want to skip this?

Rohin Shah: One thing we didn't cover is, if you have uncertainty over what you're supposed to optimize, this turns into an interactive sort of game between the human and the AI agent, which seems pretty good. A priori you should expect that there's going to need to be a lot of interaction between the human and the AI system in order for the AI system to actually be able to do the things that the human wants it to do. And so having formalisms and ideas of where this interaction naturally falls out seems like a good thing.

Buck Shlegeris: I've said a lot of things about how I am very pessimistic about value learning as a strategy. Nevertheless it seems like it might be really good for there to be people who are researching this, and trying to get as good as we can get at improving sample efficiency so that can have your AI systems understand your preferences over music with as little human interaction as possible, just in case it turns out to be possible to solve the hard version of value learning. Because a lot of the engineering effort required to make ambitious value learning work will plausibly be in common with the kinds of stuff you have to do to make these more simple specification learning tasks work out. That's a reason for me to be enthusiastic about people researching value learning even if I'm pessimistic about the overall thing working.

Lucas Perry: All right, so what is robustness and why does it matter?

Rohin Shah: Robustness is one of those words that doesn't super clearly have a definition and people use it differently. Robust agents don't fail catastrophically in situations slightly different from the ones that they were designed for. One example of a case where we see a failure of robustness currently, is in adversarial examples for

image classifiers, where it is possible to take an image, make a slight perturbation to it, and then the resulting image is completely misclassified. You take a correctly classified image of a Panda, slightly perturb it such that a human can't tell what the difference is, and then it's classified as a gibbon with 99% confidence. Admittedly this was with an older image classifier. I think you need to make the perturbations a bit larger now in order to get them.

Lucas Perry: This is because the relevant information that it uses are very local to infer panda-ness rather than global properties of the panda?

Rohin Shah: It's more like they're high frequency features or imperceptible features. There's a lot of controversy about this but there is a pretty popular recent paper that I believe, but not everyone believes, that claims that this was because they're picking up on real imperceptible features that do generalize to the test set, that humans can't detect. That's an example of robustness. Recently people have been applying this to reinforcement learning both by adversarially modifying the observations that agents get and also by training agents that act in the environment adversarially towards the original agent. One paper out of CHAI showed that there's this kick and defend environment where you've got two MuJoCo robots. One of them is kicking a soccer ball. The other one's a goalie, that's trying to prevent the kicker from successfully shooting a goal, and they showed that if you do self play in order to get kickers and defenders and then you take the kicker, you freeze it, you don't train it anymore and you retrain a new defender against this kicker.

What is the strategy that this new defender learns? It just sort of falls to the ground and flaps about in a random looking way and the kicker just gets so confused that it usually fails to even touch the ball and so this is sort of an adversarial example for RL agents now, it's showing that even they're not very robust.

There was also a paper out of DeepMind that did the same sort of thing. For their adversarial attack they learned what sorts of mistakes the agent would make early on in training and then just tried to replicate those mistakes once the agent was fully trained and they found that this helped them uncover a lot of bad behaviors. Even at the end of training.

From the perspective of alignment, it's clear that we want robustness. It's not exactly clear what we want robustness to. This robustness to adversarial perturbations was kind of a bit weird as a threat model. If there is an adversary in the environment they're probably not going to be restricted to small perturbations. They're probably not going to get white box access to your AI system; even if they did, this doesn't seem to really connect with the AI system as adversarially optimizing against humans story, which is how we get to the x-risk part, so it's not totally clear.

I think on the intent alignment case, which is the thing that I usually think about, you mostly want to ensure that whatever is driving the "motivation" of the AI system, you want that to be very robust. You want it to agree with what humans would want in all situations or at least all situations that are going to come up or something like that. Paul Christiano has written a few blog posts about this that talk about what techniques he's excited about solving that problem, which boil down to interpretability, adversarial training, and improving adversarial training through relaxations of the problem.

Buck Shlegeris: I'm pretty confused about this, and so it's possible what I'm going to say is dumb. When I look at problems with robustness or problems that Rohin put in

this robustness category here, I want to divide it into two parts. One of the parts is, things that I think of as capability problems, which I kind of expect the rest of the world will need to solve on its own. For instance, things about safe exploration, how do I get my system to learn to do good things without ever doing really bad things, this just doesn't seem very related to the AI alignment problem to me. And I also feel reasonably optimistic that you can solve it by doing dumb techniques which don't have anything too difficult to them, like you can have your system so that it has a good model of the world that it got from unsupervised learning somehow and then it never does dumb enough things. And also I don't really see that kind of robustness problem leading to existential catastrophes. And the other half of robustness is the half that I care about a lot, which in my mind, is mostly trying to make sure that you succeeded at inner alignment. That is, that the mesa optimizers you've found through gradient descent have goals that actually match your goals.

This is like robustness in the sense that you're trying to guarantee that in every situation, your AI system, as Rohin was saying, is intent aligned with you. It's trying to do the kind of thing that you want. And I worry that, by default, we're going to end up with AI systems not intent aligned, so there exist a bunch of situations they can be put in such that they do things that are very much not what you'd want, and therefore they fail at robustness. I think this is a really important problem, it's like half of the AI safety problem or more, in my mind, and I'm not very optimistic about being able to solve it with prosaic techniques.

Rohin Shah: That sounds roughly similar to what I was saying. Yes.

Buck Shlegeris: I don't think we disagree about this super much except for the fact that I think you seem to care more about safe exploration and similar stuff than I think I do.

Rohin Shah: I think safe exploration's a bad example. I don't know what safe exploration is even trying to solve but I think other stuff, I agree. I do care about it more. One place where I somewhat disagree with you is, you sort of have this point about all these robustness problems are the things that the rest of the world has incentives to figure out, and will probably figure out. That seems true for alignment too, it sure seems like you want your system to be aligned in order to do the things that you actually want. Everyone that has an incentive for this to happen. I totally expect people who aren't EAs or rationalists or weird longtermists to be working on AI alignment in the future and to some extent even now. I think that's one thing.

Buck Shlegeris: You should say your other thing, but then I want to get back to that point.

Rohin Shah: The other thing is I think I agree with you that it's not clear to me how failures of the robustness of things other than motivation lead to x-risk, but I'm more optimistic than you are that our solutions to those kinds of robustness will help with the solutions to "motivation robustness" or how to make your mesa optimizer aligned.

Buck Shlegeris: Yeah, sorry, I guess I actually do agree with that last point. I am very interested in trying to figure out how to have aligned to mesa optimizers, and I think that a reasonable strategy to pursue in order to get aligned mesa optimizers is trying to figure out how to make your image classifiers robust to adversarial examples. I think you probably won't succeed even if you succeed with the image classifiers, but it seems like the image classifiers are still probably where you should start. And I guess if we can't figure out how to make image classifiers robust to adversarial examples in

like 10 years, I'm going to be super pessimistic about the harder robustness problem, and that would be great to know.

Rohin Shah: For what it's worth, my take on the adversarial examples of image classifiers is, we're going to train image classifiers on more data with bigger nets, it's just going to mostly go away. Prediction. I'm laying my cards on the table.

Buck Shlegeris: That's also something like my guess.

Rohin Shah: Okay.

Buck Shlegeris: My prediction is: to get image classifiers that are robust to epsilon ball perturbations or whatever, some combination of larger things and adversarial training and a couple other clever things, will probably mean that we have robust image classifiers in 5 or 10 years at the latest.

Rohin Shah: Cool. And you wanted to return to the other point about the world having incentives to do alignment.

Buck Shlegeris: So I don't quite know how to express this, but I think it's really important which is going to make this a really fun experience for everyone involved. You know how Airbnb... Or sorry, I guess a better example of this is actually Uber drivers. Where I give basically every Uber driver a five star rating, even though some Uber drivers are just clearly more pleasant for me than others, and Uber doesn't seem to try very hard to get around these problems, even though I think that if Uber caused there to be a 30% difference in pay between the drivers who I think of as 75th percentile and the drivers I think of as 25th percentile, this would make the service probably noticeably better for me. I guess it seems to me that a lot of the time the world just doesn't try to do kind of complicated things to make systems actually aligned, and it just does hack jobs, and then everyone deals with the fact that everything is unaligned as a result.

To draw this analogy back, I think that we're likely to have the kind of alignment techniques that solve problems that are as simple and obvious as: we should have a way to have rate your hosts on Airbnb. But I'm worried that we won't ever get around to solving the problems that are like, but what if your hosts are incentivized to tell you sob stories such that you give them good ratings, even though actually they were worse than some other hosts. And this is never a big enough deal that people are unilaterally individually incentivized to solve the harder version of the alignment problem, and then everyone ends up using these systems that actually aren't aligned in the strong sense and then we end up in a doomy world. I'm curious if any of that made any sense.

Lucas Perry: Is a simple way to put that we fall into inadequate or an unoptimal equilibrium and then there's tragedy of the commons and bad game theory stuff that happens that keeps us locked and that the same story could apply to alignment?

Buck Shlegeris: Yeah, that's not quite what I mean.

Lucas Perry: Okay.

Rohin Shah: I think Buck's point is that actually Uber or Airbnb could unilaterally, no gains required, make their system better and this would be an improvement for them and everyone else, and they don't do it. There is nothing about equilibrium that is a failure of Uber to do this thing that seems so obviously good.

Buck Shlegeris: I'm not actually claiming that it's better for Uber, I'm just claiming that there is a misalignment there. Plausibly, an Uber exec, if they were listening to this they'd just be like, "LOL, that's a really stupid idea. People would hate it." And then they would say more complicated things like "most riders are relatively price sensitive and so this doesn't matter." And plausibly they're completely right.

Rohin Shah: That's what I was going to say.

Buck Shlegeris: But the thing which feels important to me is something like a lot of the time it's not worth solving the alignment problems at any given moment because something else is a bigger problem to how things are going locally. And this can continue being the case for a long time, and then you end up with everyone being locked in to this system where they never solved the alignment problems. And it's really hard to make people understand this, and then you get locked into this bad world.

Rohin Shah: So if I were to try and put that in the context of AI alignment, I think this is a legitimate reason for being more pessimistic. And the way that I would make that argument is: it sure seems like we are going to decide on what method or path we're going to use to build AGI. Maybe we'll do a bunch of research and decide we're just going to scale up language models or something like this. I don't know. And we will do that before we have any idea of which technique would be easiest to align and as a result, we will be forced to try to align this exogenously chosen AGI technique and that would be harder than if we got to design our alignment techniques and our AGI techniques simultaneously.

Buck Shlegeris: I'm imagining some pretty slow take off here, and I don't imagine this as ever having a phase where we built this AGI and now we need to align it. It's more like we're continuously building and deploying these systems that are gradually more and more powerful, and every time we want to deploy a system, it has to be doing something which is useful to someone. And many of the things which are useful, require things that are kind of like alignment. "I want to make a lot of money from my system that will give advice," and if it wants to give good generalist advice over email, it's going to need to have at least some implicit understanding of human preferences. Maybe we just use giant language models and everything's just totally fine here. A really good language model isn't able to give arbitrarily good aligned advice, but you can get advice that sounds really good from a language model, and I'm worried that the default path is going to involve the most popular AI advice services being kind of misaligned, and just never bothering to fix that. Does that make any more sense?

Rohin Shah: Yeah, I think I totally buy that that will happen. But I think I'm more like as you get to AI systems doing more and more important things in the world, it becomes more and more important that they are really truly aligned and investment in alignment increases correspondingly.

Buck Shlegeris: What's the mechanism by which people realize that they need to put more work into alignment here?

Rohin Shah: I think there's multiple. One is I expect that people are aware, like even in the Uber case, I expect people are aware of the misalignment that exists, but decide that it's not worth their time to fix it. So the continuation of that, people will be aware of it and then they will decide that they should fix it.

Buck Shlegeris: If I'm trying to sell to city governments this language model based system which will give them advice on city planning, it's not clear to me that at any

point the city governments are going to start demanding better alignment features. Maybe that's the way that it goes but it doesn't seem obvious that city governments would think to ask that, and --

Rohin Shah: I wasn't imagining this from the user side. I was imagining this from the engineers or designers side.

Buck Shlegeris: Yeah.

Rohin Shah: I think from the user side I would speak more to warning shots. You know, you have your cashier AI system or your waiter AIs and they were optimizing for tips more so than actually collecting money and so they like offer free meals in order to get more tips. At some point one of these AI systems passes all of the internal checks and makes it out into the world and only then does the problem arise and everyone's like, "Oh my God, this is terrible. What the hell are you doing? Make this better."

Buck Shlegeris: There's two mechanisms via which that alignment might be okay. One of them is that researchers might realize that they want to put more effort into alignment and then solve these problems. The other mechanism is that users might demand better alignment because of warning shots. I think that I don't buy that either of these is sufficient. I don't buy that it's sufficient for researchers to decide to do it because in a competitive world, the researchers who realize this is important, if they try to only make aligned products, they are not going to be able to sell them because their products will be much less good than the unaligned ones. So you have to argue that there is demand for the things which are actually aligned well. But for this to work, your users have to be able to distinguish between things that have good alignment properties and those which don't, and this seems really hard for users to do. And I guess, when I try to imagine analogies, I just don't see many examples of people successfully solving problems like this, like businesses making products that are different levels of dangerousness, and then users successfully buying the safe ones.

Rohin Shah: I think usually what happens is you get regulation that forces everyone to be safe. I don't know if it was regulation, but like airplanes are incredibly safe. Cars are incredibly safe.

Buck Shlegeris: Yeah but in this case what would happen is doing the unsafe thing allows you to make enormous amounts of money, and so the countries which don't put in the regulations are going to be massively advantaged compared to ones which don't.

Rohin Shah: Why doesn't that apply for cars and airplanes?

Buck Shlegeris: So to start with, cars in poor countries are a lot less safe. Another thing is that a lot of the effort in making safer cars and airplanes comes from designing them. Once you've done the work of designing it, it's that much more expensive to put your formally-verified 747 software into more planes, and because of weird features of the fact that there are only like two big plane manufacturers, everyone gets the safer planes.

Lucas Perry: So tying this into robustness. The fundamental concern here is about the incentives to make aligned systems that are safety and alignment robust in the real world.

Rohin Shah: I think that's basically right. I sort of see these incentives as existing and the world generally being reasonably good at dealing with high stakes problems.

Buck Shlegeris: What's an example of the world being good at dealing with a high stakes problem?

Rohin Shah: I feel like biotech seems reasonably well handled, relatively speaking,

Buck Shlegeris: Like bio-security?

Rohin Shah: Yeah.

Buck Shlegeris: Okay, if the world handles AI as well as bio-security, there's no way we're okay.

Rohin Shah: Really? I'm aware of ways in which we're not doing bio-security well, but there seem to be ways in which we're doing it well too.

Buck Shlegeris: The nice thing about bio-security is that very few people are incentivized to kill everyone, and this means that it's okay if you're sloppier about your regulations, but my understanding is that lots of regulations are pretty weak.

Rohin Shah: I guess I was more imagining the research community's coordination on this. Surprisingly good.

Buck Shlegeris: I wouldn't describe it that way.

Rohin Shah: It seems like the vast majority of the research community is onboard with the right thing and like 1% isn't. Yeah. Plausibly we need to have regulations for that last 1%.

Buck Shlegeris: I think that 99% of the synthetic biology research community is on board with "it would be bad if everyone died." I think that some very small proportion is onboard with things like "we shouldn't do research if it's very dangerous and will make the world a lot worse." I would say like way less than half of synthetic biologists seem to agree with statements like "it's bad to do really dangerous research." Or like, "when you're considering doing research, you consider differential technological development." I think this is just not a thing biologists think about, from my experience talking to biologists.

Rohin Shah: I'd be interested in betting with you on this afterwards.

Buck Shlegeris: Me too.

Lucas Perry: So it seems like it's going to be difficult to come down to a concrete understanding or agreement here on the incentive structures in the world and whether they lead to the proliferation of unaligned AI systems or semi aligned AI systems versus fully aligned AI systems and whether that poses a kind of lock-in, right? Would you say that that fairly summarizes your concern Buck?

Buck Shlegeris: Yeah. I expect that Rohin and I agree mostly on the size of the coordination problem required, or the costs that would be required by trying to do things the safer way. And I think Rohin is just a lot more optimistic about those costs being paid.

Rohin Shah: I think I'm optimistic both about people's ability to coordinate paying those costs and about incentives pointing towards paying those costs.

Buck Shlegeris: I think that Rohin is right that I disagree with him about the second of those as well.

Lucas Perry: Are you interested in unpacking this anymore? Are you happy to move on?

Buck Shlegeris: I actually do want to talk about this for two more minutes. I am really surprised by the claim that humans have solved coordination problems as hard as this one. I think the example you gave is humans doing radically nowhere near well enough. What are examples of coordination problem type things... There was a bunch of stuff with nuclear weapons, where I feel like humans did badly enough that we definitely wouldn't have been okay in an AI situation. There are a bunch of examples of the US secretly threatening people with nuclear strikes, which I think is an example of some kind of coordination failure. I don't think that the world has successfully coordinated on never threaten first nuclear strikes. If we had successfully coordinated on that, I would consider nuclear weapons to be less of a failure, but as it is the US has actually according to Daniel Ellsberg threatened a bunch of people with first strikes.

Rohin Shah: Yeah, I think I update less on specific scenarios and update quite a lot more on, "it just never happened." The sheer amount of coincidence that would be required given the level of, Oh my God, there were close calls multiple times a year for many decades. That seems just totally implausible and it just means that our understanding of what's happening is wrong.

Buck Shlegeris: Again, also the thing I'm imagining is this very gradual takeoff world where people, every year, they release their new most powerful AI systems. And if, in a particular year, AI Corp decided to not release its thing, then AI Corps two and three and four would rise to being one, two and three in total profits instead of two, three and four. In that kind of a world, I feel a lot more pessimistic.

Rohin Shah: I'm definitely imagining more of the case where they coordinate to all not do things. Either by international regulation or via the companies themselves coordinating amongst each other. Even without that, it's plausible that AI Corp one does this. One example I'd give is, Waymo has just been very slow to deploy self driving cars relative to all the other self driving car companies, and my impression is that this is mostly because of safety concerns.

Buck Shlegeris: Interesting and slightly persuasive example. I would love to talk through this more at some point. I think this is really important and I think I haven't heard a really good conversation about this.

Apologies for describing what I think is going wrong inside your mind or something, which is generally a bad way of saying things, but it sounds kind of to me like you're implicitly assuming more concentrated advantage and fewer actors than I think actually are implied by gradual takeoff scenarios.

Rohin Shah: I'm usually imagining something like a 100+ companies trying to build the next best AI system, and 10 or 20 of them being clear front runners or something.

Buck Shlegeris: That makes sense. I guess I don't quite see how the coordination successes you were describing arise in that kind of a world. But I am happy to move on.

Lucas Perry: So before we move on on this point, is there anything which you would suggest as obvious solutions, should Buck's model of the risks here be the case. So it seemed like it would demand more centralized institutions which would help to mitigate some of the lock in here.

Rohin Shah: Yeah. So there's a lot of work in policy and governance about this. Not much of which is public unfortunately. But I think the thing to say is that people are thinking about it and it does sort of look like trying to figure out how to get the world to actually coordinate on things. But as Buck has pointed out, we have tried to do this before and so there's probably a lot to learn from past cases as well. But I am not an expert on this and don't really want to talk as though I were one.

Lucas Perry: All right. So there's lots of governance and coordination thought that kind of needs to go into solving many of these coordination issues around developing beneficial AI. So I think with that we can move along now to scaling to superhuman abilities. So Rohin, what do you have to say about this topic area?

Rohin Shah: I think this is in some sense related to what we were talking about before, you can predict what a human would say, but it's hard to back out true underlying values beneath them. Here the problem is, suppose you are learning from some sort of human feedback about what you're supposed to be doing, the information contained in that tells you how to do whatever the human can do. It doesn't really tell you how to exceed what the human can do without having some additional assumptions.

Now, depending on how the human feedback is structured, this might lead to different things like if the human is demonstrating how to do the task to you, then this would suggest that it would be hard to do the task any better than the human can, but if the human was evaluating how well you did the task, then you can do the task better in a way that the human wouldn't be able to tell was better. Ideally, at some point we would like to have AI systems that can actually do just really powerful, great things, that we are unable to understand all the details of and so we would neither be able to demonstrate or evaluate them.

How do we get to those sorts of AI systems? The main proposals in this bucket are iterated amplification, debate, and recursive reward modeling. So in iterated amplification, we started with an initial policy, and we alternate between amplification and distillation, which increases capabilities and efficiency respectively. This can encode a bunch of different algorithms, but usually amplification is done by decomposing questions into easier sub questions, and then using the agent to answer those sub questions. While distillation can be done using supervised learning or reinforcement learning, so you get these answers that are created by these amplified systems that take a long time to run, and you just train a neural net to very quickly predict the answers without having to do this whole big decomposition thing. In debate, we train an agent through self play in a zero sum game where the agent's goal is to win a question answering debate as evaluated by a human judge. The hope here is that since both sides of the debate can point out flaws in the other side's arguments -- they're both very powerful AI systems -- such a set up can use a human judge to train far more capable agents while still incentivizing the agents to provide honest true information. With recursive reward modeling, you can think of it as an instantiation of the general alternate between amplification and distillation framework, but it works sort of bottom up instead of top down. So you'll start by building AI systems that can help you evaluate simple, easy tasks. Then use those AI systems to help you evaluate more complex tasks and you keep iterating this process

until eventually you have AI systems that help you with very complex tasks like how to design the city. And this lets you then train an AI agent that can design the city effectively even though you don't totally understand why it's doing the things it's doing or why they're even good.

Lucas Perry: Do either of you guys have any high level thoughts on any of these approaches to scaling to superhuman abilities?

Buck Shlegeris: I have some.

Lucas Perry: Go for it.

Buck Shlegeris: So to start with, I think it's worth noting that another approach would be ambitious value learning, in the sense that I would phrase these not as approaches for scaling to superhuman abilities, but they're like approaches for scaling to superhuman abilities while only doing tasks that relate to the actual behavior of humans rather than trying to back out their values explicitly. Does that match your thing Rohin?

Rohin Shah: Yeah, I agree. I often phrase that as with ambitious value learning, there's not a clear ground truth to be focusing on, whereas with all three of these methods, the ground truth is what a human would do if they got a very, very long time to think or at least that is what they're trying to approximate. It's a little tricky to see why exactly they're approximating that, but there are some good posts about this. The key difference between these techniques and ambitious value learning is that there is in some sense a ground truth that you are trying to approximate.

Buck Shlegeris: I think these are all kind of exciting ideas. I think they're all kind of better ideas than I expected to exist for this problem a few years ago. Which probably means we should update against my ability to correctly judge how hard AI safety problems are, which is great news, in as much as I think that a lot of these problems are really hard. Nevertheless, I don't feel super optimistic that any of them are actually going to work. One thing which isn't in the elevator pitch for IDA, which is iterated distillation and amplification (and debate), is that you get to hire the humans who are going to be providing the feedback, or the humans whose answers AI systems are going to be trained with. And this is actually really great. Because for instance, you could have this program where you hire a bunch of people and you put them through your one month long training an AGI course. And then you only take the top 50% of them. I feel a lot more optimistic about these proposals given you're allowed to think really hard about how to set it up such that the humans have the easiest time possible. And this is one reason why I'm optimistic about people doing research in factored cognition and stuff, which I'm sure Rohin's going to explain in a bit.

One comment about recursive reward modeling: it seems like it has a lot of things in common with IDA. The main downside that it seems to have to me is that the human is in charge of figuring out how to decompose the task into evaluations at a variety of levels. Whereas with IDA, your system itself is able to naturally decompose the task into a variety levels, and for this reason I feel a bit more optimistic about IDA.

Rohin Shah: With recursive reward modeling, one agent that you can train is just an agent that's good at doing decompositions. That is a thing you can do with it. It's a thing that the people at DeepMind are thinking about.

Buck Shlegeris: Yep, that's a really good point.

Rohin Shah: I also strongly like the fact that you can train your humans to be good at providing feedback. This is also true about specification learning. It's less clear if it's true about ambitious value learning. No one's really proposed how you could do ambitious value learning really. Maybe arguably Stuart Russell's book is kind of a proposal, but it doesn't have that many details.

Buck Shlegeris: And, for example, it doesn't address any of my concerns in ways that I find persuasive.

Rohin Shah: Right. But for specification learning also you definitely want to train the humans who are going to be providing feedback to the AI system. That is an important part of why you should expect this to work.

Buck Shlegeris: I often give talks where I try to give an introduction to IDA and debate as a proposal for AI alignment. I'm giving these talks to people with computer science backgrounds, and they're almost always incredibly skeptical that it's actually possible to decompose thought in this kind of a way. And with debate, they're very skeptical that truth wins, or that the nash equilibrium is accuracy. For this reason I'm super enthusiastic about research into the factored cognition hypothesis of the type that Ought is doing some of.

I'm kind of interested in your overall take for how likely it is that the factored cognition hypothesis holds and that it's actually possible to do any of this stuff, Rohin. You could also explain what that is.

Rohin Shah: I'll do that. So basically with both iterated amplification, debate, or recursive reward modeling, they all hinge on this idea of being able to decompose questions, maybe it's not so obvious why that's true for debate, but it's true. Go listen to the podcast about debate if you want to get more details on that.

So this hypothesis is basically for any tasks that we care about, it is possible to decompose this into a bunch of sub tasks that are all easier to do. Such that if you're able to do the sub tasks, then you can do the overall top level tasks and in particular you can iterate this down, building a tree of smaller and smaller tasks until you can get to the level of tasks that a human could do in a day. Or if you're trying to do it very far, maybe tasks that a human can do in a couple of minutes. Whether or not you can actually decompose the task "be an effective CEO" into a bunch of sub tasks that eventually bottom out into things humans can do in a few minutes is totally unclear. Some people are optimistic, some people are pessimistic. It's called the factored cognition hypothesis and Ought is an organization that's studying it.

It sounds very controversial at first and I, like many other people had the intuitive reaction of, 'Oh my God, this is never going to work and it's not true'. I think the thing that actually makes me optimistic about it is you don't have to do what you might call a direct decomposition. You can do things like if your task is to be an effective CEO, your first sub question could be, what are the important things to think about when being a CEO or something like this, as opposed to usually when I think of decompositions I would think of, first I need to deal with hiring. Maybe I need to understand HR, maybe I need to understand all of the metrics that the company is optimizing. Very object level concerns, but the decompositions are totally allowed to also be meta level where you'll spin off a bunch of computation that is just trying to answer the meta level of question of how should I best think about this question at all.

Another important reason for optimism is that based on the structure of iterated amplification, debate and recursive reward modeling, this tree can be gigantic. It can

be exponentially large. Something that we couldn't run even if we had all of the humans on Earth collaborating to do this. That's okay. Given how the training process is structured, considering the fact that you can do the equivalent of millennia of person years of effort in this decomposed tree, I think that also gives me more of a, 'okay, maybe this is possible' and that's also why you're able to do all of this meta level thinking because you have a computational budget for it. When you take all of those together, I sort of come up with "seems possible. I don't really know."

Buck Shlegeris: I think I'm currently at 30-to-50% on the factored cognition thing basically working out. Which isn't nothing.

Rohin Shah: Yeah, that seems like a perfectly reasonable thing. I think I could imagine putting a day of thought into it and coming up with numbers anywhere between 20 and 80.

Buck Shlegeris: For what it's worth, in conversation at some point in the last few years, Paul Christiano gave numbers that were not wildly more optimistic than me. I don't think that the people who are working on this think it's obviously fine. And it would be great if this stuff works, so I'm really in favor of people looking into it.

Rohin Shah: Yeah, I should mention another key intuition against it. We have all these examples of human geniuses like Ramanujan, who were posed very difficult math problems and just immediately get the answer and then you ask them how did they do it and they say, well, I asked myself what should the answer be? And I was like, the answer should be a continued fraction. And then I asked myself which continued fraction and then I got the answer. And you're like, that does not sound very decomposable. It seems like you need these magic flashes of intuition. Those would be the hard cases for factored cognition. It still seems possible that you could do it by both this exponential try a bunch of possibilities and also by being able to discover intuitions that work in practice and just believing them because they work in practice and then applying them to the problem at hand. You could imagine that with enough computation you'd be able to discover such intuitions.

Buck Shlegeris: You can't answer a math problem by searching exponentially much through the search tree. The only exponential power you get from IDA is IDA is letting you specify the output of your cognitive process in such a way that's going to match some exponentially sized human process. As long as that exponentially sized human process was only exponentially sized because it's really inefficient, but is kind of fundamentally not an exponentially sized problem, then your machine learning should be able to speed it up a bunch. But the thing where you search over search strategy is not valid. If that's all you can do, that's not good enough.

Rohin Shah: Searching over search strategies, I agree you can't do, but if you have an exponential search that could be implemented by humans. We know by hypothesis, if you can solve it with a flash of intuition, there is in fact some more efficient way to do it and so whether or not the distillation steps will actually be enough to get to the point where you can do those flashes of intuition. That's an open question.

Buck Shlegeris: This is one of my favorite areas of AI safety research and I would love for there to be more of it. Something I have been floating for a little while is I kind of wish that there was another Ought. It just seems like it would be so good if we had definitive information about the factored cognition hypothesis. And it also it seems like the kind of thing which is potentially parallelizable. And I feel like I know a lot of people who love talking about how thinking works. A lot of rationalists are really into

this. I would just be super excited for some of them to form teams of four and go off on their own and build an Ought competitor. I feel like this is the kind of thing where plausibly, a bunch of enthusiastic people could make progress on their own.

Rohin Shah: Yeah, I agree with that. Definitely seems like one of the higher value things but I might be more excited about universality.

Lucas Perry: All right, well let's get started with universality then. What is universality and why are you optimistic about it?

Rohin Shah: So universality is hard to explain well, in a single sentence. For whatever supervisor is training our agent, you want that supervisor to "know everything the agent knows." In particular if the agent comes up with some deceptive strategy to look like it's achieving the goal, but actually it hasn't. The supervisors should know that it was doing this deceptive strategy for the reason of trying to trick the supervisor and so the supervisor can then penalize it. The classic example of why this is important and hard also due to Paul Christiano is plagiarism. Suppose you are training on the AI system to produce novel works of literature and as part of its training data, the AI system gets to read this library of a million books.

It's possible that this AI system decides, Hey, you know the best way I can make a great novel seeming book is to just take these five books and take out plot points, passages from each of them and put them together and then this new book will look totally novel and will be very good because I used all of the best Shakespearean writing or whatever. If your supervisor doesn't know that the agent has done this, the only way the supervisor can really check is to go read the entire million books. Even if the agent only read 10 books and so then the supervision becomes a way more costly than running the agent, which is not a great state to be in, and so what you really want is that if the agent does this, the supervisor is able to say, I see that you just copied this stuff over from these other books in order to trick me into thinking that you had written something novel that was good.

That's bad. I'm penalizing you. Stop doing that in the future. Now, this sort of property, I mean it's very nice in the abstract, but who knows whether or not we can actually build it in practice. There's some reason for optimism that I don't think I can adequately convey, but I wrote a newsletter summarizing some of it sometime ago, but again, reading through the posts I became more optimistic that it was an achievable property, than when I first heard what the property was. The reason I'm optimistic about it is that it just sort of seems to capture the thing that we actually care about. It's not everything, like it doesn't solve the robustness problem. Universality only tells you what the agent's currently doing. You know all the facts about that. Whereas for robustness you want to say even in these hypothetical situations that the agent hasn't encountered yet and doesn't know stuff about, even when it encounters those situations, it's going to stay aligned with you so universality doesn't get you all the way there, but it definitely feels like it's getting you quite a bit.

Buck Shlegeris: That's really interesting to hear you phrase it that way. I guess I would have thought of universality as a subset of robustness. I'm curious what you think of that first.

Rohin Shah: I definitely think you could use universality to achieve a subset of robustness. Maybe I would say universality is a subset of interpretability.

Buck Shlegeris: Yeah, and I care about interpretability as a subset of robustness basically, or as a subset of inner alignment, which is pretty close to robustness in my

mind. The other thing I would say is you were saying there that one difference between universality and robustness is that universality only tells you why the agent did the thing it currently did, and this doesn't suffice to tell us about the situations that the agent isn't currently in. One really nice thing though is that if the agent is only acting a particular way because it wants you to trust it, that's a fact about its current behavior that you will know, and so if you have the universality property, your overseer just knows your agent is trying to deceive it. Which seems like it would be incredibly great and would resolve like half of my problem with safety if you had it.

Rohin Shah: Yeah, that seems right. The case that universality doesn't cover is when your AI system is initially not deceptive, but then at some point in the future it's like, 'Oh my God, now it's possible to go and build Dyson spheres or something, but wait, in this situation probably I should be doing this other thing and humans won't like that. Now I better deceive humans'. The transition into deception would have to be a surprise in some sense even to the AI system.

Buck Shlegeris: Yeah, I guess I'm just not worried about that. Suppose I have this system which is as smart as a reasonably smart human or 10 reasonably smart humans, but it's not as smart as the whole world. If I can just ask it what its best sense about how aligned it is, is? And if I can trust its answer? I don't know man, I'm pretty okay with systems that think they're aligned, answering that question honestly.

Rohin Shah: I think I somewhat agree. I like this reversal where I'm the pessimistic one.

Buck Shlegeris: Yeah me too. I'm like, "look, system, I want you to think as hard as you can to come up with the best arguments you can come up with for why you are misaligned, and the problems with you." And if I just actually trust the system to get this right, then the bad outcomes I get here are just pure accidents. I just had this terrible initialization of my neural net parameters, such that I had this system that honestly believed that it was going to be aligned. And then as it got trained more, this suddenly changed and I couldn't do anything about it. I don't quite see the story for how this goes super wrong. It seems a lot less bad than the default situation.

Rohin Shah: Yeah. I think the story I would tell is something like, well, if you look at humans, they're pretty wrong about what their preferences will be in the future. For example, there's this trope of how teenagers fall in love and then fall out of love, but when they're in love, they swear undying oaths to each other or something. To the extent that is true, that seems like the sort of failure that could lead to x-risk if it also happened with AI systems.

Buck Shlegeris: I feel pretty optimistic about all the garden-variety approaches to solving this. Teenagers were not selected very hard on accuracy of their undying oaths. And if you instead had accuracy of self-model as a key feature you were selecting for in your AI system, plausibly you'll just be way more okay.

Rohin Shah: Yeah. Maybe people could coordinate well on this. I feel less good about people coordinating on this sort of problem.

Buck Shlegeris: For what it's worth, I think there are coordination problems here and I feel like my previous argument about why coordination is hard and won't happen by default also probably applies to us not being okay. I'm not sure how this all plays out. I'd have to think about it more.

Rohin Shah: Yeah. I think it's more like this is a subtle and non-obvious problem, which by hypothesis doesn't happen in the systems you actually have and only happens later and those are the sorts of problems I'm like, Ooh, not sure if we can deal with those ones, but I agree that there's a good chance that there's just not a problem at all in the world where we already have universality and checked all the obvious stuff.

Buck Shlegeris: Yeah. I would like to say universality is one of my other favorite areas of AI alignment research, in terms of how happy I'd be if it worked out really well.

Lucas Perry: All right, so let's see if we can slightly pick up the pace here. Moving forward and starting with interpretability.

Rohin Shah: Yeah, so I mean I think we've basically discussed interpretability already. Universality is a specific kind of interpretability, but the case for interpretability is just like, sure seems like it would be good if you could understand what your AI systems are doing. You could then notice when they're not aligned, and fix that somehow. It's a pretty clear cut case for a thing that would be good if we achieved it and it's still pretty uncertain how likely we are to be able to achieve it.

Lucas Perry: All right, so let's keep it moving and let's hit impact regularization now.

Rohin Shah: Yeah, impact regularization in particular is one of the ideas that are not trying to align the AI system but are instead trying to say, well, whatever AI system we build, let's make sure it doesn't cause a catastrophe. It doesn't lead to extinction or existential risk. What it hopes to do is say, all right, AI system, do whatever it is you wanted to do. I don't care about that. Just make sure that you don't have a huge impact upon the world.

Whatever you do, keep your impact not too high. And so there's been a lot of work on this in recent years there's been relative reachability, attainable utility preservation, and I think in general the sense is like, wow, it's gone quite a bit further than people expected it to go. I think it definitely does prevent you from doing very, very powerful things of the sort, like if you wanted to stop all competing AI projects from ever being able to build AGI, that doesn't seem like the sort of thing you can do with an impact regularized AI system, but it sort of seems plausible that you could prevent convergent instrumental sub goals using impact regularization. Where AI systems that are trying to steal resources and power from humans, you could imagine that you'd say, hey, don't do that level of impact, you can still have the level of impact of say running a company or something like that.

Buck Shlegeris: My take on all this is that I'm pretty pessimistic about all of it working. I think that impact regularization or whatever is a non-optimal point on the capabilities / alignment trade off or something, in terms of safety you're getting for how much capability you're sacrificing. My basic a problem here is basically analogous to my problem with value learning, where I think we're trying to take these extremely essentially fuzzy concepts and then factor our agent through these fuzzy concepts like impact, and basically the thing that I imagine happening is any impact regularization strategy you try to employ, if your AI is usable, will end up not helping with its alignment. For any definition of impacts you come up with, it'll end up doing something which gets around that. Or it'll make your AI system completely useless, is my basic guess as to what happens.

Rohin Shah: Yeah, so I think again in this setting, if you formalize it and then say, consider the optimal agent. Yeah, that can totally get around your impact penalty, but in practice it sure seems like, what you want to do is say this convergent instrumental subgoal stuff, don't do any of that. Continue to do things that are normal in regular life. And those seem like pretty distinct categories. Such that I would not be shocked if we could actually distinguish between the two.

Buck Shlegeris: It sounds like the main benefit you're going for is trying to make your AI system not do insane, convergent, instrumental sub-goal style stuff. So another approach I can imagine taking here would be some kind of value learning or something, where you're asking humans for feedback on whether plans are insanely convergent, instrumental sub-goal style, and just not doing the things which, when humans are asked to rate how sketchy the plans are the humans rate as sufficiently sketchy? That seems like about as good a plan. I'm curious what you think.

Rohin Shah: The idea of power as your attainable utility across a wide variety of utility functions seems like a pretty good formalization to me. I think in the worlds where I actually buy a formalization, I tend to expect the formalization to work better. I do think the formalization is not perfect. Most notably with the current formalization of power, your power never changes if you have extremely good beliefs. Your notion, you're just like, I always have the same power because I'm always able to do the same things and you never get surprised, so maybe I agree with you because I think the current formalization is not good enough. Yeah, I think I agree with you but I could see it going either way.

Buck Shlegeris: I could be totally wrong about this, and correct me if I'm wrong, my sense is that you have to be able to back out the agent's utility function or its models of the world. Which seems like it's assuming a particular path for AI development which doesn't seem to me particularly likely.

Rohin Shah: I definitely agree with that for all the current methods too.

Buck Shlegeris: So it's like: assume that we have already perfectly solved our problems with universality and robustness and transparency and whatever else. I feel like you kind of have to have solved all of those problems before you can do this, and then you don't need it or something.

Rohin Shah: I don't think I agree with that. I definitely agree that the current algorithms that people have written assume that you can just make a change to the AI's utility function. I don't think that's what even their proponents would suggest as the actual plan.

Buck Shlegeris: What is the actual plan?

Rohin Shah: I don't actually know what their actual plan would be, but one plan I could imagine is figure out what exactly the conceptual things we have to do with impact measurement are, and then whatever method we have for building AGI, probably there's going to be some part which is specify the goal and then in the specify goal part, instead of just saying pursue X, we want to say pursue X without changing your ability to pursue Y, and Z and W, and P, and Q.

Buck Shlegeris: I think that that does not sound like a good plan. I don't think that we should expect our AI systems to be structured that way in the future.

Rohin Shah: Plausibly we have to do this with natural language or something.

Buck Shlegeris: It seems very likely to me that the thing you do is reinforcement learning where at the start of the episode you get a sentence of English which is telling you what your goal is and then blah, blah, blah, blah, blah, and this seems like a pretty reasonable strategy for making powerful and sort of aligned AI. Aligned enough to be usable for things that aren't very hard. But you just fundamentally don't have access to the internal representations that the AI is using for its sense of what belief is, and stuff like that. And that seems like a really big problem.

Rohin Shah: I definitely see this as more of an outer alignment thing, or like an easier to specify outer alignment type thing than say, IDA is that type stuff.

Buck Shlegeris: Okay, I guess that makes sense. So we're just like assuming we've solved all the inner alignment problems?

Rohin Shah: In the story so far yeah, I think all of the researchers who actually work on this haven't thought much about inner alignment.

Buck Shlegeris: My overall summary is that I really don't like this plan. I feel like it's not robust to scale. As you were saying Rohin, if your system gets more and more accurate beliefs, stuff breaks. It just feels like the kind of thing that doesn't work.

Rohin Shah: I mean, it's definitely not conceptually neat and elegant in the sense of it's not attacking the underlying problem. And in a problem setting where you expect adversarial optimization type dynamics, conceptual elegance actually does count for quite a lot in whether or not you believe your solution will work.

Buck Shlegeris: I feel it's like trying to add edge detectors to your image classifiers to make them more adversarially robust or something, which is backwards.

Rohin Shah: Yeah, I think I agree with that general perspective. I don't actually know if I'm more optimistic than you. Maybe I just don't say... Maybe we'd have the same uncertainty distributions and you just say yours more strongly or something.

Lucas Perry: All right, so then let's just move a little quickly through the next three, which are causal modeling, oracles, and decision theory.

Rohin Shah: Yeah, I mean, well decision theory, MIRI did some work on it. I am not the person to ask about it, so I'm going to skip that one. Even if you look at the long version, I'm just like, here are some posts. Good luck. So causal modeling, I don't fully understand what the overall story is here but the actual work that's been published is basically what we can do is we can take potential plans or training processes for AI systems. We can write down causal models that tell us how the various pieces of the training system interact with each other and then using algorithms developed for causal models we can tell when an AI system would have an incentive to either observe or intervene on an underlying variable.

One thing that came out of this was that you can build a model-based reinforcement learner that doesn't have any incentive to wire head as long as when it makes its plans, the plans are evaluated by the current reward function as opposed to whatever future reward function it would have. And that was explained using this framework of causal modeling. Oracles, Oracles are basically the idea that we can just train an AI system to just answer questions, give it a question and it tries to figure out the best answer it can to that question, prioritizing accuracy.

One worry that people have recently been talking about is the predictions that the Oracle makes then affect the world, which can affect whether or not the prediction was correct. Like maybe if I predict that I will go to bed at 11 then I'm more likely to actually go to bed at 11 because I want my prediction to come true or something and so then the Oracles can still "choose" between different self confirming predictions and so that gives them a source of agency and one way that people want to avoid this is using what are called counter-factual Oracles where you set up the training, such that the Oracles are basically making predictions under the assumption that their predictions are not going to influence the future.

Lucas Perry: Yeah, okay. Oracles seem like they just won't happen. There'll be incentives to make things other than Oracles and that Oracles would even be able to exert influence upon the world in other ways.

Rohin Shah: Yeah, I think I agree that Oracles do not seem very competitive.

Lucas Perry: Let's do forecasting now then.

Rohin Shah: So the main sub things within forecasting one, there's just been a lot of work recently on actually building good forecasting technology. There has been an AI specific version of Metaculus that's been going on for a while now. There's been some work at the Future of Humanity Institute on building better tools for working with probability distributions under recording and evaluating forecasts. There was an AI resolution council where basically now you can make forecasts about what this particular group of people will think in five years or something like that, which is much easier to operationalize than most other kinds of forecasts. So this helps with constructing good questions. On the actual object level, I think there are two main things. One is that it became increasingly more obvious in the past two years that AI progress currently is being driven by larger and larger amounts of compute.

It totally could be driven by other things as well, but at the very least, compute is a pretty important factor. And then takeoff speeds. So there's been this long debate in the AI safety community over whether -- to take the extremes, whether or not we should expect that AI capabilities will see a very sharp spike. So initially, your AI capabilities are improving by like one unit a year, maybe then with some improvements it got to two units a year and then for whatever reason, suddenly they're now at 20 units a year or a hundred units a year and they just swoop way past what you would get by extrapolating past trends, and so that's what we might call a discontinuous takeoff. If you predict that that won't happen instead you'll get AI that's initially improving at one unit per year. Then maybe two units per year, maybe three units per year. Then five units per year, and the rate of progress continually increases. The world's still gets very, very crazy, but in a sort of gradual, continuous way that would be called a continuous takeoff.

Basically there were two posts that argued pretty forcefully for continuous takeoff back in, I want to say February of 2018, and this at least made me believe that continuous takeoff was more likely. Sadly, we just haven't actually seen much defense of the other side of the view since then. Even though we do know that there definitely are people who still believe the other side, that there will be a discontinuous takeoff.

Lucas Perry: Yeah so what are both you guys' views on them?

Buck Shlegeris: Here are a couple of things. One is that I really love the operationalization of slow take off or continuous take off that Paul provided in his post, which was one of the ones Rohin was referring to from February 2018. He says, "by

slow takeoff, I mean that there is a four year doubling of the economy before there is a one year doubling of the economy." As in, there's a period of four years over which world GDP increases by a factor four, after which there is a period of one year. As opposed to a situation where the first one-year doubling happens out of nowhere. Currently, doubling times for the economy are on the order of like 20 years, and so a one year doubling would be a really big deal. The way that I would phrase why we care about this, is because worlds where we have widespread, human level AI feel like they have incredibly fast economic growth. And if it's true that we expect AI progress to increase gradually and continuously, then one important consequence of this is that by the time we have human level AI systems, the world is already totally insane. A four year doubling would just be crazy. That would be economic growth drastically higher than economic growth currently is.

This means it would be obvious to everyone who's paying attention that something is up and the world is radically changing in a rapid fashion. Another way I've been thinking about this recently is people talk about transformative AI, by which they mean AI which would have at least as much of an impact on the world as the industrial revolution had. And it seems plausible to me that octopus level AI would be transformative. Like suppose that AI could just never get better than octopus brains. This would be way smaller of a deal than I expect AI to actually be, but it would still be a massive deal, and would still possibly lead to a change in the world that I would call transformative. And if you think this is true, and if you think that we're going to have octopus level AI before we have human level AI, then you should expect that radical changes that you might call transformative have happened by the time that we get to the AI alignment problems that we've been worrying about. And if so, this is really big news.

When I was reading about this stuff when I was 18, I was casually imagining that the alignment problem is a thing that some people have to solve while they're building an AGI in their lab while the rest of the world's ignoring them. But if the thing which is actually happening is the world is going insane around everyone, that's a really important difference.

Rohin Shah: I would say that this is probably the most important contested question in AI alignment right now. Some consequences of it are in a gradual or continuous takeoff world you expect that by the time we get to systems that can pose an existential risk. You've already had pretty smart systems that have been deployed in the real world. They probably had some failure modes. Whether or not we call them alignment failure modes or not is maybe not that important. The point is people will be aware that AI systems can fail in weird ways, depending on what sorts of failures you expect, you might expect this to lead to more coordination, more involvement in safety work. You might also be more optimistic about using testing and engineering styles of approaches to the problem which rely a bit more on trial and error type of reasoning because you actually will get a chance to see errors before they happen at a super intelligent existential risk causing mode. There are lots of implications of this form that pretty radically change which alignment plans you think are feasible.

Buck Shlegeris: Also, it's pretty radically changed how optimistic you are about this whole AI alignment situation, at the very least, people who are very optimistic about AI alignment causing relatively small amounts of existential risk. A lot of the reason for this seems to be that they think that we're going to get these warning shots where before we have superintelligent AI, we have sub-human level intelligent AI with alignment failures like the cashier Rohin was talking about earlier. And then people

start caring about AI alignment a lot more. So optimism is also greatly affected by what you think about this.

I've actually been wanting to argue with people about this recently. I wrote a doc last night where I was arguing that even in gradual takeoff worlds, we should expect a reasonably high probability of doom if we can't solve the AI alignment problem. And I'm interested to have this conversation in more detail with people at some point. But yeah, I agree with what Rohin said.

Overall on takeoff speeds, I guess I still feel pretty uncertain. It seems to me that currently, what we can do with AI, like AI capabilities are increasing consistently, and a lot of this comes from applying relatively non-mindblowing algorithmic ideas to larger amounts of compute and data. And I would be kind of surprised if you can't basically ride this wave away until you have transformative AI. And so if I want to argue that we're going to have fast takeoffs, I kind of have to argue that there's some other approach you can take which lets you build AI without having to go along that slow path, which also will happen first. And I guess I think it's kind of plausible that that is what's going to happen. I think that's what you'd have to argue for if you want to argue for a fast take off.

Rohin Shah: That all seems right to me. I'd be surprised if, out of nowhere, we saw a new AI approach suddenly started working and overtook deep learning. You also have to argue that it then very quickly reaches human level AI, which would be quite surprising, right? In some sense, it would have to be something completely novel that we failed to think about in the last 60 years. We're putting in way more effort now than we were in the last 60 years, but then counter counterpoint is that all of that extra effort is going straight into deep learning. It's not really searching for completely new paradigm-shifting ways to get to AGI.

Buck Shlegeris: So here's how I'd make that argument. Perhaps a really important input into a field like AI, is the number of really smart kids who have been wanting to be an AI researcher since they were 16 because they thought that it's the most important thing in the world. I think that in physics, a lot of the people who turn into physicists have actually wanted to be physicists forever. I think the number of really smart kids who wanted to be AI researchers forever has possibly gone up by a factor of 10 over the last 10 years, it might even be more. And there just are problems sometimes, that are bottle necked on that kind of a thing, probably. And so it wouldn't be totally shocking to me, if as a result of this particular input to AI radically increasing, we end up in kind of a different situation. I haven't quite thought through this argument fully.

Rohin Shah: Yeah, the argument seems plausible. There's a large space of arguments like this. I think even after that, then I've started questioning, "Okay, we get a new paradigm. The same arguments apply to that paradigm?" Not as strongly. I guess not the arguments you were saying about compute going up over time, but the arguments given in the original slow takeoff posts which were people quickly start taking the low-hanging fruit and then move on. When there's a lot of effort being put into getting some property, you should expect that easy low-hanging fruit is usually just already taken, and that's why you don't expect discontinuities. Unless the new idea just immediately rockets you to human-level AGI, or x-risk causing AGI, I think the same argument would pretty quickly start applying to that as well.

Buck Shlegeris: I think it's plausible that you do get rocketed pretty quickly to human-level AI. And I agree that this is an insane sounding claim.

Rohin Shah: Great. As long as we agree on that.

Buck Shlegeris: Something which has been on my to-do list for a while, and something I've been doing a bit of and I'd be excited for someone else doing more of, is reading the history of science and getting more of a sense of what kinds of things are bottlenecked by what, where. It could lead me to be a bit less confused about a bunch of this stuff. AI Impacts has done a lot of great work cataloging all of the things that aren't discontinuous changes, which certainly is a strong evidence to me against my claim here.

Lucas Perry: All right. What is the probability of AI-induced existential risk?

Rohin Shah: Unconditional on anything? I might give it 1 in 20. 5%.

Buck Shlegeris: I'd give 50%.

Rohin Shah: I had a conversation with AI Impacts that went into this in more detail and partially just anchored on the number I gave there, which was 10% conditional on no intervention from longtermists, I think the broad argument is really just the one that Buck and I were disagreeing about earlier, which is to what extent will society be incentivized to solve the problem? There's some chance that the first thing we try just works and we don't even need to solve any sort of alignment problem. It might just be fine. This is not implausible to me. Maybe that's 30% or something.

Most of the remaining probability comes from, "Okay, the alignment problem is a real problem. We need to deal with it." It might be very easy in which case we can just solve it straight away. That might be the case. That doesn't seem that likely to me if it was a problem at all. But what we will get is a lot of these warning shots and people understanding the risks a lot more as we get more powerful AI systems. This estimate is also conditional on gradual takeoff. I keep forgetting to say that, mostly because I don't know what probability I should put on discontinuous takeoff.

Lucas Perry: So is 5% with longtermist intervention, increasing to 10% if fast takeoff?

Rohin Shah: Yes, but still with longtermist intervention. I'm pretty pessimistic on fast takeoff, but my probability assigned to fast takeoff is not very high. In a gradual takeoff world, you get a lot of warning shots. There will just generally be awareness of the fact that the alignment problem is a real thing and you won't have the situation you have right now of people saying this thing about worrying about superintelligent AI systems not doing what we want is totally bullshit. That won't be a thing. Almost everyone will not be saying that anymore, in the version where we're right and there is a problem. As a result, people will not want to build AI systems that are going to kill them. People tend to be pretty risk averse in my estimation of the world, which Buck will probably disagree with. And as a result, you'll get a lot of people trying to actually work on solving the alignment problem. There'll be some amount of global coordination which will give us more time to solve the alignment problem than we may otherwise have had. And together, these forces mean that probably we'll be okay.

Buck Shlegeris: So I think my disagreements with Rohin are basically that I think fast takeoffs are more likely. I basically think there is almost surely a problem. I think that alignment might be difficult, and I'm more pessimistic about coordination. I know I said four things there, but I actually think of this as three disagreements. I want to say that "there isn't actually a problem" is just a kind of "alignment is really easy to solve." So then there's three disagreements. One is gradual takeoff, another is difficulty of

solving competitive prosaic alignment, and another is how good we are at coordination.

I haven't actually written down these numbers since I last changed my mind about a lot of the inputs to them, so maybe I'm being really dumb. I guess, it feels to me that in fast takeoff worlds, we are very sad unless we have competitive alignment techniques, and so then we're just okay if we have these competitive alignment techniques. I guess I would say that I'm something like 30% on us having good competitive alignment techniques by the time that it's important, which incidentally is higher than Rohin I think.

Rohin Shah: Yeah, 30 is totally within the 25th to 75th interval on the probability, which is a weird thing to be reporting. 30 might be my median, I don't know.

Buck Shlegeris: To be clear, I'm not just including the outer alignment proportion here, which is what we were talking about before with IDA. I'm also including the inner alignment.

Rohin Shah: Yeah, 30% does seem a bit high. I think I'm a little more pessimistic.

Buck Shlegeris: So I'm like 30% that we can just solve the AI alignment problem in this excellent way, such that anyone who wants to can have very little extra cost and then make AI systems that are aligned. I feel like in worlds where we did that, it's pretty likely that things are reasonably okay. I think that the gradual versus fast takeoff isn't actually enormously much of a crux for me because I feel like in worlds without competitive alignment techniques and gradual takeoff, we still have a very high probability of doom. And I think that comes down to disagreements about coordination. So maybe the main important disagreement between Rohin and I is, actually how well we'll be able to coordinate, or how strongly individual incentives will be for alignment.

Rohin Shah: I think there are other things. The reason I feel a bit more pessimistic than you in the fast takeoff world is just solving problems in advance just is really quite difficult and I really like the ability to be able to test techniques on actual AI systems. You'll have to work with less powerful things. At some point, you do have to make the jump to more powerful things. But, still, being able to test on the less powerful things, that's so good, so much safety from there.

Buck Shlegeris: It's not actually clear to me that you get to test the most important parts of your safety techniques. So I think that there are a bunch of safety problems that just do not occur on dog-level AIs, and do occur on human-level AI. If there are three levels of AI, there's a thing which is as powerful as a dog, there's a thing which is as powerful as a human, and there's a thing which is as powerful as a thousand John von Neumanns. In gradual takeoff world, you have a bunch of time in both of these two milestones, maybe. I guess it's not super clear to me that you can use results on less powerful systems as that much evidence about whether your safety techniques work on drastically more powerful systems. It's definitely somewhat helpful.

Rohin Shah: It depends what you condition on in your difference between continuous takeoff and discontinuous takeoff to say which one of them happens faster. I guess the delta between dog and human is definitely longer in gradual takeoff for sure. Okay, if that's what you were saying, yep, I agree with that.

Buck Shlegeris: Yeah, sorry, that's all I meant.

Rohin Shah: Cool. One thing I wanted to ask is when you say dog-level AI assistant, do you mean something like a neural net that if put in a dog's body replacing its brain would do about as well as a dog? Because such a neural net could then be put in other environments and learn to become really good at other things, probably superhuman at many things that weren't in the ancestral environment. Do you mean that sort of thing?

Buck Shlegeris: Yeah, that's what I mean. Dog-level AI is probably much better than GPT2 at answering questions. I'm going to define something as dog-level AI, if it's about as good as a dog at things which I think dogs are pretty heavily optimized for, like visual processing or motor control in novel scenarios or other things like that, that I think dogs are pretty good at.

Rohin Shah: Makes sense. So I think in that case, plausibly, dog-level AI already poses an existential risk. I can believe that too.

Buck Shlegeris: Yeah.

Rohin Shah: The AI cashier example feels like it could totally happen probably before a dog-level AI. You've got all of the motivation problems already at that point of the game, and I don't know what problems you expect to see beyond then.

Buck Shlegeris: I'm more talking about whether you can test your solutions. I'm not quite sure how to say my intuitions here. I feel like there are various strategies which work for corralling dogs and which don't work for making humans do what you want. In as much as your alignment strategy is aiming at a flavor of problem that only occurs when you have superhuman things, you don't get to test that either way. I don't think this is a super important point unless you think it is. I guess I feel good about moving on from here.

Rohin Shah: Mm-hmm (affirmative). Sounds good to me.

Lucas Perry: Okay, we've talked about what you guys have called gradual and fast takeoff scenarios, or continuous and discontinuous. Could you guys put some probabilities down on the likelihood of, and stories that you have in your head, for fast and slow takeoff scenarios?

Rohin Shah: That is a hard question. There are two sorts of reasoning I do about probabilities. One is: use my internal simulation of whatever I'm trying to predict, internally simulate what it looks like, whether it's by my own models, is it likely? How likely is it? At what point would I be willing to bet on it. Stuff like that. And then there's a separate extra step where I'm like, "What do other people think about this? Oh, a lot of people think this thing that I assigned one percent probability to is very likely. Hmm, I should probably not be saying one percent then." I don't know how to do that second part for, well, most things but especially in this setting. So I'm going to just report Rohin's model only, which will predictably be understating the probability for fast takeoff in that if someone from MIRI were to talk to me for five hours, I would probably say a higher number for the probability of fast takeoff after that, and I know that that's going to happen. I'm just going to ignore that fact and report my own model anyway.

On my own model, it's something like in worlds where AGI happens soon, like in the next couple of decades, then I'm like, "Man, 95% on gradual take off." If it's further away, like three to five decades, then I'm like, "Some things could have changed by then, maybe I'm 80%." And then if it's way off into the future and centuries, then I'm

like, "Ah, maybe it's 70%, 65%." The reason it goes down over time is just because it seems to me like if you want to argue for discontinuous takeoff, you need to posit that there's some paradigm change in how AI progress is happening and that seems more likely the further in the future you go.

Buck Shlegeris: I feel kind of surprised that you get so low, like to 65% or 70%. I would have thought that those arguments are a strong default and then maybe at the moment where in a position that seems particularly gradual takeoff-y, but I would have thought that you over time get to 80% or something.

Rohin Shah: Yeah. Maybe my internal model is like, "Holy shit, why do these MIRI people keep saying that discontinuous takeoff is so obvious." I agree that the arguments in Paul's posts feel very compelling to me and so maybe I should just be more confident in them. I think saying 80%, even in centuries is plausibly a correct answer.

Lucas Perry: So, Rohin, is the view here that since compute is the thing that's being leveraged to make most AI advances that you would expect that to be the mechanism by which that continues to happen in the future and we have some certainty over how compute continues to change into the future? Whereas things that would be leading to a discontinuous takeoff would be world-shattering, fundamental insights into algorithms that would have powerful recursive self-improvement, which is something you wouldn't necessarily see if we just keep going this leveraging compute route?

Rohin Shah: Yeah, I think that's a pretty good summary. Again, on the backdrop of the default argument for this is people are really trying to build AGI. It would be pretty surprising if there is just this really important thing that everyone had just missed.

Buck Shlegeris: It sure seems like in machine learning when I look at the things which have happened over the last 20 years, all of them feel like the ideas are kind of obvious or someone else had proposed them 20 years earlier. ConvNets were proposed 20 years before they were good on ImageNet, and LSTMs were ages before they were good for natural language, and so on and so on and so on. Other subjects are not like this, like in physics sometimes they just messed around for 50 years before they knew what was happening. I don't know, I feel confused how to feel about the fact that in some subjects, it feels like they just do suddenly get better at things for reasons other than having more compute.

Rohin Shah: I think physics, at least, was often bottlenecked by measurements, I want to say.

Buck Shlegeris: Yes, so this is one reason I've been interested in history of science recently, but there are certainly a bunch of things. People were interested in chemistry for a long time and it turns out that chemistry comes from quantum mechanics and you could, theoretically, have guessed quantum mechanics 70 years earlier than people did if you were smart enough. It's not that complicated a hypothesis to think of. Or relativity is the classic example of something which could have been invented 50 years earlier. I don't know, I would love to learn more about this.

Lucas Perry: Just to tie this back to the question, could you give your probabilities as well?

Buck Shlegeris: Oh, geez, I don't know. Honestly, right now I feel like I'm 70% gradual takeoff or something, but I don't know. I might change my mind if I think about this for another hour. And there's also theoretical arguments as well for why

most takeoffs are gradual, like the stuff in Paul's post. The easiest summary is, before someone does something really well, someone else does it kind of well in cases where a lot of people are trying to do the thing.

Lucas Perry: Okay. One facet of this, that I haven't heard discussed, is recursive self-improvement, and I'm confused about where that becomes the thing that affects whether it's discontinuous or continuous. If someone does something kind of well before something does something really well, if recursive self-improvement is a property of the thing being done kind of well, is it just kind of self-improving really quickly, or?

Buck Shlegeris: Yeah. I think Paul's post does a great job of talking about this exact argument. I think his basic claim is, which I find pretty plausible, before you have a system which is really good at self-improving, you have a system which is kind of good at self-improving, if it turns out to be really helpful to have a system be good at self-improving. And as soon as this is true, you have to posit an additional discontinuity.

Rohin Shah: One other thing I'd note is that humans are totally self improving. Productivity techniques, for example, are a form of self-improvement. You could imagine that AI systems might have advantages that humans don't, like being able to read their own weights and edit them directly. How much of an advantage this gives to the AI system, unclear. Still, I think then I just go back to the argument that Buck already made, which is at some point you get to an AI system that is somewhat good at understanding its weights and figuring out how to edit them, and that happens before you get the really powerful ones. Maybe this is like saying, "Well, you'll reach human levels of self-improvement by the time you have rat-level AI or something instead of human-level AI," which argues that you'll hit this hyperbolic point of the curve earlier, but it still looks like a hyperbolic curve that's still continuous at every point.

Buck Shlegeris: I agree.

Lucas Perry: I feel just generally surprised about your probabilities on continuous takeoff scenarios that they'd be slow.

Rohin Shah: The reason I'm trying to avoid the word slow and fast is because they're misleading. Slow takeoff is not slow in calendar time relative to fast takeoff. The question is, is there a spike at some point? Some people, upon reading Paul's posts are like, "Slow takeoff is faster than fast takeoff." That's a reasonably common reaction to it.

Buck Shlegeris: I would put it as slow takeoff is the claim that things are insane before you have the human-level AI.

Rohin Shah: Yeah.

Lucas Perry: This seems like a helpful perspective shift on this takeoff scenario question. I have not read Paul's post. What is it called so that we can include it in the page for this podcast?

Rohin Shah: It's just called Takeoff Speeds. Then the corresponding AI Impacts post is called Will AI See Discontinuous Progress?, I believe.

Lucas Perry: So if each of you guys had a lot more reach and influence and power and resources to bring to the AI alignment problem right now, what would you do?

Rohin Shah: I get this question a lot and my response is always, "Man, I don't know." It seems hard to scalably use people right now for AI risk. I can talk about which areas of research I'd like to see more people focus on. If you gave me people where I'm like, "I trust your judgment on your ability to do good conceptual work" or something, where would I put them? I think a lot of it would be on making good robust arguments for AI risk. I don't think we really have them, which seems like kind of a bad situation to be in. I think I would also invest a lot more in having good introductory materials, like this review, except this review is a little more aimed at people who are already in the field. It is less aimed at people who are trying to enter the field. I think we just have pretty terrible resources for people coming into the field and that should change.

Buck Shlegeris: I think that our resources are way better than they used to be.

Rohin Shah: That seems true.

Buck Shlegeris: In the course of my work, I talk to a lot of people who are new to AI alignment about it and I would say that their level of informedness is drastically better now than it was two years ago. A lot of which is due to things like 80,000 hours podcast, and other things like this podcast and the Alignment Newsletter, and so on. I think we just have made it somewhat easier for people to get into everything. The Alignment Forum, having its sequences prominently displayed, and so on.

Rohin Shah: Yeah, you named literally all of the things I would have named. Buck definitely has more information on this than I do. I do not work with people who are entering the field as much. I do think we could be substantially better.

Buck Shlegeris: Yes. I feel like I do have access to resources, not directly but in the sense that I know people at eg Open Philanthropy and the EA Funds and if I thought there were obvious things they should do, I think it's pretty likely that those funders would have already made them happen. And I occasionally embark on projects myself that I think are good for AI alignment, mostly on the outreach side. On a few occasions over the last year, I've just done projects that I was optimistic about. So I don't think I can name things that are just shovel-ready opportunities for someone else to do, which is good news because it's mostly because I think most of these things are already being done.

I am enthusiastic about workshops. I help run with MIRI these AI Risks for Computer Scientists workshops and I ran my own computing workshop with some friends, with kind of a similar purpose, aimed at people who are interested in this kind of stuff and who would like to spend some time learning more about it. I feel optimistic about this kind of project as a way of doing the thing Rohin was saying, making it easier for people to start having really deep thoughts about a lot of AI alignment stuff. So that's a kind of direction of projects that I'm pretty enthusiastic about. A couple other random AI alignment things I'm optimistic about. I've already mentioned that I think there should be an Ought competitor just because it seems like the kind of thing that more work could go into. I agree with Rohin on it being good to have more conceptual analysis of a bunch of this stuff. I'm generically enthusiastic about there being more high quality research done and more smart people, who've thought about this a lot, working on it as best as they can.

Rohin Shah: I think the actual bottleneck is good research and not necessarily field building, and I'm more optimistic about good research. Specifically, I am particularly interested in universality, interpretability. I would love for there to be some way to give people who work on AI alignment the chance to step back and think about the

high-level picture for a while. I don't know if people don't do this because they don't want to or because they don't feel like they have the affordance to do so, and I would like the affordance to be there. I'd be very interested in people building models of what AGI systems could look like. Expected utility maximizers are one example of a model that you could have. Maybe we just try to redo evolution. We just create a very complicated, diverse environment with lots of agents going around and in their multi-agent interaction, they develop general intelligence somehow. I'd be interested for someone to take that scenario, flesh it out more, and then talk about what the alignment problem looks like in that setting.

Buck Shlegeris: I would love to have someone get really knowledgeable about evolutionary biology and try and apply analogies of that to AI alignment. I think that evolutionary biology has lots of smart things to say about what optimizers are and it'd be great to have those insights. I think Eliezer sort of did this many years ago. It would be good for more people to do this in my opinion.

Lucas Perry: All right. We're in the home stretch here. AI timelines. What do you think about the current state of predictions? There's been surveys that have been done with people giving maybe 50% probability over most researchers at about 2050 or so. What are each of your AI timelines? What's your probability distribution look like? What do you think about the state of predictions on this?

Rohin Shah: Haven't looked at the state of predictions in a while. It depends on who was surveyed. I think most people haven't thought about it very much and I don't know if I expect their predictions to be that good, but maybe wisdom of the crowds is a real thing. I don't think about it very much. I mostly use my inside view and talk to a bunch of people. Maybe, median, 30 years from now, which is 2050. So I guess I agree with them, don't I? That feels like an accident. The surveys were not an input into this process.

Lucas Perry: Okay, Buck?

Buck Shlegeris: I don't know what I think my overall timelines are. I think AI in the next 10 or 20 years is pretty plausible. Maybe I want to give it something around 50% which puts my median at around 2040. In terms of the state of things that people have said about AI timelines, I have had some really great conversations with people about their research on AI timelines which hasn't been published yet. But at some point in the next year, I think it's pretty likely that much better stuff about AI timelines modeling will have been published than has currently been published, so I'm excited for that.

Lucas Perry: All right. Information hazards. Originally, there seemed to be a lot of worry in the community about information hazards and even talking about superintelligence and being afraid of talking to anyone in positions of power, whether they be in private institutions or in government, about the strategic advantage of AI, about how one day it may confer a decisive strategic advantage. The dissonance here for me is that Putin comes out and says that who controls AI will control the world. Nick Bostrom published Superintelligence, which basically says what I already said. Max Tegmark's Life 3.0 basically also. My initial reaction and intuition is the cat's out of the bag. I don't think that echoing this increases risks any further than the risk is already at. But maybe you disagree.

Buck Shlegeris: Yeah. So here are two opinions I have about info hazards. One is: how bad is it to say stuff like that all over the internet? My guess is it's mildly bad

because I think that not everyone thinks those things. I think that even if you could get those opinions as consequences from reading Superintelligence, I think that most people in fact have not read Superintelligence. Sometimes there are ideas where I just really don't want them to be crystallized common knowledge. I think that, to a large extent, assuming gradual takeoff worlds, it kind of doesn't matter because AI systems are going to be radically transforming the world inevitably. I guess you can affect how governments think about it, but it's a bit different there.

The other point I want to make about info hazards is I think there are a bunch of trickinesses with AI safety, where thinking about AI safety makes you think about questions about how AI development might go. I think that thinking about how AI development is going to go occasionally leads to think about things that are maybe, could be, relevant to capabilities, and I think that this makes it hard to do research because you then get scared about talking about them.

Rohin Shah: So I think my take on this is info hazards are real in the sense that there, in fact, are costs to saying specific kinds of information and publicizing them a bit. I think I'll agree in principle that some kinds of capabilities information has the cost of accelerating timelines. I usually think these are pretty strongly outweighed by the benefits in that it just seems really hard to be able to do any kind of shared intellectual work when you're constantly worried about what you do and don't make public. It really seems like if you really want to build a shared understanding within the field of AI alignment, that benefit is worth saying things that might be bad in some other ways. This depends on a lot of background facts that I'm not going to cover here but, for example, I probably wouldn't say the same thing about bio security.

Lucas Perry: Okay. That makes sense. Thanks for your opinions on this. So at the current state in time, do you guys think that people should be engaging with people in government or in policy spheres on questions of AI alignment?

Rohin Shah: Yes, but not in the sense of we're worried about when AGI comes. Even saying things like it might be really bad, as opposed to saying it might kill everybody, seems not great. Mostly on the basis of my model for what it takes to get governments to do things is, at the very least, you need consensus in the field so it seems kind of pointless to try right now. It might even be poisoning the well for future efforts. I think it does make sense to engage with government and policymakers about things that are in fact problems right now. To the extent that you think that recommender systems are causing a lot of problems, I think it makes sense to engage with government about how alignment-like techniques can help with that, especially if you're doing a bunch of specification learning-type stuff. That seems like the sort of stuff that should have relevance today and I think it would be great if those of us who did specification learning were trying to use it to improve existing systems.

Buck Shlegeris: This isn't my field. I trust the judgment of a lot of other people. I think that it's plausible that it's worth building relationships with governments now, not that I know what I'm talking about. I will note that I basically have only seen people talk about how to do AI governance in the cases where the AI safety problem is 90th percentile easiest. I basically only see people talking about it in the case where the technical safety problem is pretty doable, and this concerns me. I've just never seen anyone talk about what you do in a world where you're as pessimistic as I am, except to completely give up.

Lucas Perry: All right. Wrapping up here, is there anything else that we didn't talk about that you guys think was important? Or something that we weren't able to spend

enough time on, that you would've liked to spend more time on?

Rohin Shah: I do want to eventually continue the conversation with Buck about coordination, but that does seem like it should happen not on this podcast.

Buck Shlegeris: That's what I was going to say too. Something that I want someone to do is write a trajectory for how AI goes down, that is really specific about what the world GDP is in every one of the years from now until insane intelligence explosion. And just write down what the world is like in each of those years because I don't know how to write an internally consistent, plausible trajectory. I don't know how to write even one of those for anything except a ridiculously fast takeoff. And this feels like a real shame.

Rohin Shah: That seems good to me as well. And also the sort of thing that I could not do because I don't know economics.

Lucas Perry: All right, so let's wrap up here then. So if listeners are interested in following either of you or seeing more of your blog posts or places where you would recommend they read more materials on AI alignment, where can they do that? We'll start with you, Buck.

Buck Shlegeris: You can Google me and find my website. I often post things on the Effective Altruism Forum. If you want to talk to me about AI alignment in person, perhaps you should apply to the AI Risks for Computer Scientists workshops run by MIRI.

Lucas Perry: And Rohin?

Rohin Shah: I write the Alignment Newsletter. That's a thing that you could sign up for. Also on my website, if you Google Rohin Shah Alignment Newsletter, I'm sure I will come up. These are also cross posted to the Alignment Forum, so another thing you can do is go to the Alignment Forum, look up my username and just see things that are there. I don't know that this is actually the thing that you want to be doing. If you're new to AI safety and want to learn more about it, I would echo the resources Buck mentioned earlier, which are the 80k podcasts about AI alignment. There are probably on the order of five of these. There's the Alignment Newsletter. There are the three recommended sequences on the Alignment Forum. Just go to alignmentforum.org and look under recommended sequences. And this podcast, of course.

Lucas Perry: All right. Heroic job, everyone. This is going to be a really good resource, I think. It's given me a lot of perspective on how thinking has changed over the past year or two.

Buck Shlegeris: And we can listen to it again in a year and see how dumb we are.

Lucas Perry: Yeah. There were lots of predictions and probabilities given today, so it'll be interesting to see how things are in a year or two from now. That'll be great. All right, so cool. Thank you both so much for coming on.

End of recorded material

The Inefficient Market Hypothesis

The efficient-market hypothesis (EMH) is the idea that there are no hundred-dollar bills lying on the sidewalk because someone smarter than you would have picked them up by now. The EMH is a good tool for most people.

If you're smart enough then you should reverse this advice into "There are hundred-dollar bills lying on the sidewalk". If you are a genius then you should reverse it to the extreme. "There are hundred-dollar bills^[1] lying around all over the place."

Hundred-dollar bills lying on the sidewalk are called "alpha". Alpha is a tautologically [self-keeping secret](#). You can't read about it in books. You can't find it on blogs^[2]. You will never be taught about it in school. ["You can only find out about \[alpha\] if you go looking for it and you'll only go looking for it if you already know it exists."](#)

Where **should** you look?

Abstractions

A system is only as secure as its weakest link. Cracking a system tends to happen on an overlooked layer of abstraction^[3].

- It's easier to install a keylogger than to break a good cryptographic protocol.
- It's easier to disassemble a computer and read the hard drive directly^[4] than to crack someone's password.

The best attacks (those requiring the least work) happen on an separate [dimension of orthogonality](#) entirely.

- The easiest way to talk to someone powerful is just to call zir company and ask by first name^[5].

Won't this technique stop working now that Tim Ferris has published it in a bestselling book? Not necessarily. Quantum mechanics has been public knowledge for decades yet most people can't do it. The hard part of pickpocketing isn't finding pockets to pick.

Perhaps you don't need to talk to anyone rich and powerful. That is a good problem to have.

I think you should find a problem that's easy for you to solve. Optimizing in solution-space is familiar and straightforward, but you can make enormous gains playing around in problem-space.

— [What Startups Are Really Like](#) by Paul Graham

Problem-space tends to have higher dimensionality than solution space.

Case study

According to Joel Spolsky, the best programmers have the ability "[to think in abstractions, and, most importantly, to view a problem at several levels of abstraction simultaneously.](#)" Also according to Joel Spolsky, a business is an "[abstraction \[that\] exists solely to create the illusion that the daily activities of a programmer \(design and writing code, checking in code, debugging, etc.\) are all that it takes to create software products and bring them to market.](#)"

The ideal programmer employee is someone who can see all the way down to the level of bits, yet can't raise zir head high enough to manipulate the financial machinery of venture capital.

Homework assignment: How can you harvest alpha from this local equilibrium?

How to tell when you get it right

Alpha often feels like a magic trick. You know the phrase "A magician never reveals his secrets"? Magic secrets are not secret. The Magician's Oath is not real. David Copperfield *patents* his inventions^[6]. You can look them up in a public government registry. You don't because magical secrets are boring. Disappointingly so.

Magicians cheat. The purest alpha should feel like cheating too. The greatest compliment you can receive about your alpha source isn't "You're a genius." It's "That shouldn't be possible. I'm disillusioned to live in a world is so inefficient."

Of course, you should never hear either response because you should never flaunt these discoveries in the first place.

1. A month ago I [offered](#) to put Westerners in touch with an N95 mask exporter in China. Only two readers took the effort to message me about it. One of them couldn't be bothered to use WeChat. [←](#)
2. Actually, I did find alpha on a blog post once. The tutorial has since been taken down. [←](#)
3. For practice, check out [What is the fastest you can sort a list of ints in Java?](#) [←](#)
4. Most computers are not encrypted. Professional software engineers are consistently surprised by my ability to recover files from a broken laptop without their login information. [←](#)
5. Tim Ferris [claims](#) this works. I am inclined to believe him based on the guests who have attended his [podcast](#). [←](#)
6. Edit: I can only find one patent invented by David Copperfield, patent number 9017177/9358477. Most of his patentable illusions seem to be invented by other people. [←](#)

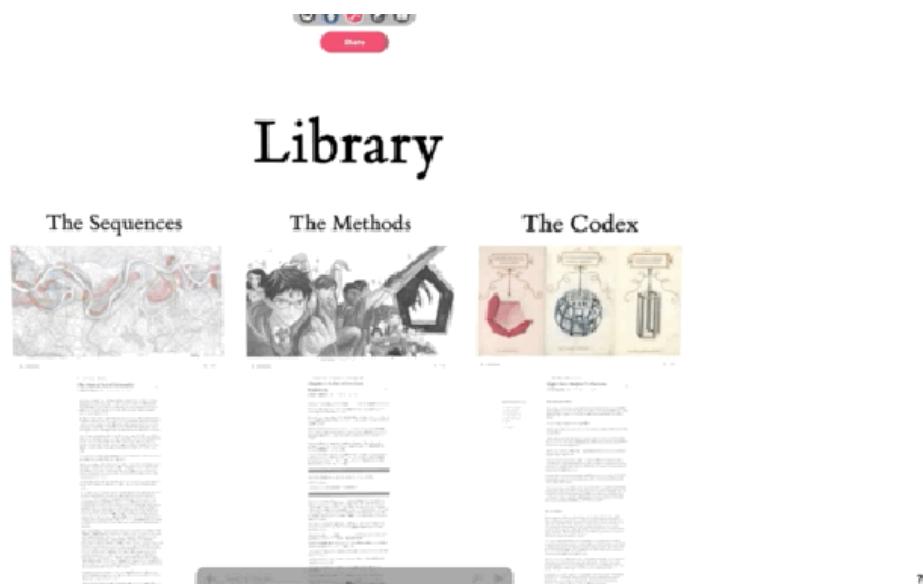
April Fools: Announcing LessWrong 3.0 - Now in VR!

On this April 1st, we at LessWrong face two problems.

1. [According to Michael Nielsen](#), software is a way to develop transformative tools for thought. Things like Anki and collaborative editing allow us to think new thoughts. But at LessWrong, we argue that this doesn't go far enough.
2. [Secondly, everyone is socially isolated](#), so we need to make LessWrong a far more social environment. We need to make a space that is superior to normal social reality in every measurable way.

The answer? *Replace LessWrong with VR.*

We're now proud to announce the new LessWrong Frontpage, built entirely in Mozilla Hubs:



The new landing page, where you can read the core readings of the site: The Sequences, The Codex, The Methods.



The new community page, and a showcase of the amazing social interactions you can have on this platform.

And, the part you've all been waiting for: The new LessWrong Frontpage, where you can read the best recent posts.



So we're using Mozilla Hubs. Why? Because you hear all the things about Mozilla Hubs that you do about any startup about to take off. Words like "Unusable", "Irritating" and "An all-round terrible UI experience". If people are saying this about your product and still using it, that means it's got to be good.

To give some hard data on this, in a survey of a recent academic conference held on the Mozilla Hubs platform, the attendees reported the following [genuine data](#) (emphasis added):

With support from the co-chairs of the 2019 ACM Symposium on User Interface Software and Technology (UIST) and the Hubs by Mozilla internal product team, we surveyed the motivations and experiences of remote attendees and discovered:

92% of all Social VR attendees would like to repeat the experience of attending a conference remotely using Hubs.

69% of all remote attendees rated the experience as very good or mostly good.

[...]

[Many] respondents reported difficulty hearing audio, poor visibility of the presentations, and lags in the presentation.

Only half of participants claimed they understood how to use the technology.

So, to be clear, of the attendees of this conference, half didn't know how to use the actual software, but almost all of them (92%) would like to repeat the experience anyways! How much of a burning need does a product need to fill that at least 42% of your users want to continue using your product, whilst claiming not to know how to actually use your product? To us, this signals amazing product-market fit, and I think we should jump on the bandwagon as early as possible.

What are its other selling points? In Mozilla Hubs you can visit the site in 3 dimensions. You'll also be able to see the other visitors, and engage in all of the social primate behaviours humans normally do at parties like talking, laughing, and continuously saying "is my microphone working?", "can anyone hear me?" and "how does walking work again?".

Have you ever wanted to be in a room with more than 25 people? No? Neither have we, so all rooms have a limit of 25 people who are allowed to enter. In addition, as we approach 25 people, the room will slow down on all devices and become significantly more laggy, causing the conversation to naturally slow down to prevent it from spiraling out of control.

If you've been part of LessWrong for any significant amount of time, you know how much effort we've spent thinking about how to avoid the problem of eternal september. Recently, after looking at our analytics for multiple minutes, we found out that a lot of users we don't want have much slower computers, or are using their phones to browse LessWrong.

So, by making LessWrong basically unusable on those devices, we are ensuring a continued high-quality discussion experience on the site, by filtering only for rational people who spend exorbitant amounts of money on their computer hardware. We've already had great success with this strategy when we drastically increased the processing power necessary to run LessWrong 2.0 by moving everything to a javascript based web-app architecture, so we consider this a natural next step for us to take.

(As a continuation of [Karma 2.0](#) we are working on a feature in which your avatar size can scale with your karma, such that users with the most karma can signal their superiority even better, and truly tower over their intellectual contemporaries.)

So it has come to this. For April 1st Day 2020, we give you the LessWrong 3.0 Homepage, in open beta only for today.

Experience the future

UPDATE

After fixing some performance issues with LessWrong 3.0, it now should run smoothly on most (many?) devices. To celebrate we are holding a surprise online meetup there tonight at 6:30 pacific time. We might or might not do a dramatic reading of HPMOR.

Link is here: [Expert Truthful Congregation v3.1](#)

Ethernet Is Worth It For Video Calls

I recently ran ethernet to my [desk](#) at [home](#), and I've been really impressed with how much it helps for video calls. Not only are garbled sections less common, I'm finding us talking over each other less.

This makes a lot of sense: wifi's going to struggle more with latency (lag) because the radio spectrum is a noisy shared space. I think I hadn't realized how much of an effect a wired connection would have on latency, or how much better a video call is when latency is lower. This is essentially the same reason serious multiplayer gamers use wired connections, but I hadn't been thinking of it that way.

The professional way to run ethernet is to use a roll of ethernet cable, jacks, and a cable crimper. This lets you make very small holes, just the size of the cable. I don't have these, it's not a good time for borrowing things, and I didn't expect to be doing enough ethernet for it to be worth it to buy one. So I got a long enough cable for the whole run, erring on the "much too long" side, and made holes big enough for the cable and jack. I was very careful of the jack as I went, since if the jack got damaged I'd have needed to start over.

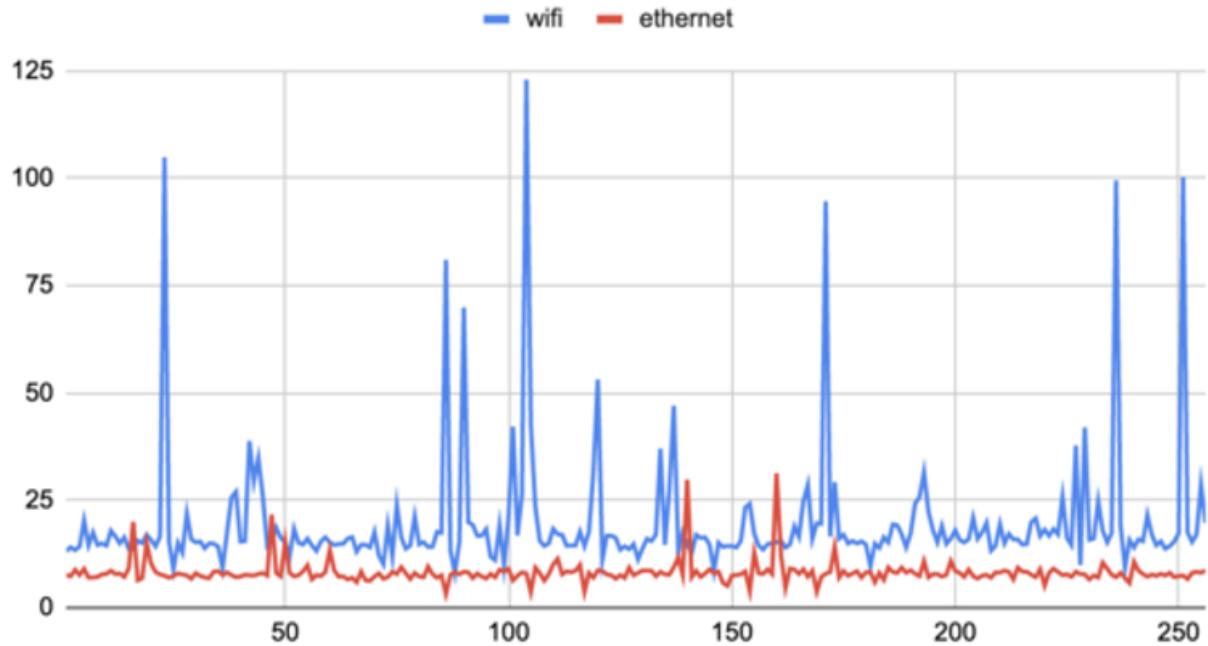
The router was three rooms away, and I could run the cable through the attic. I picked places at both ends that were relatively inconspicuous, poked nails up through the ceiling in each place, found the other ends of the nails in the attic, expanded the holes to be big enough for the cable, ran the cable, and cleaned up the dust. My laptop doesn't have an ethernet port, but I had an adapter. It took me about an hour from deciding to run the cable to having everything working.

While it's hard to quantify the effect on the calls, to check that I wasn't imagining things I ran some latency tests with ping. I picked an IP at my ISP and ran:

```
ping -c 256 -i 0.1 [IP]
```

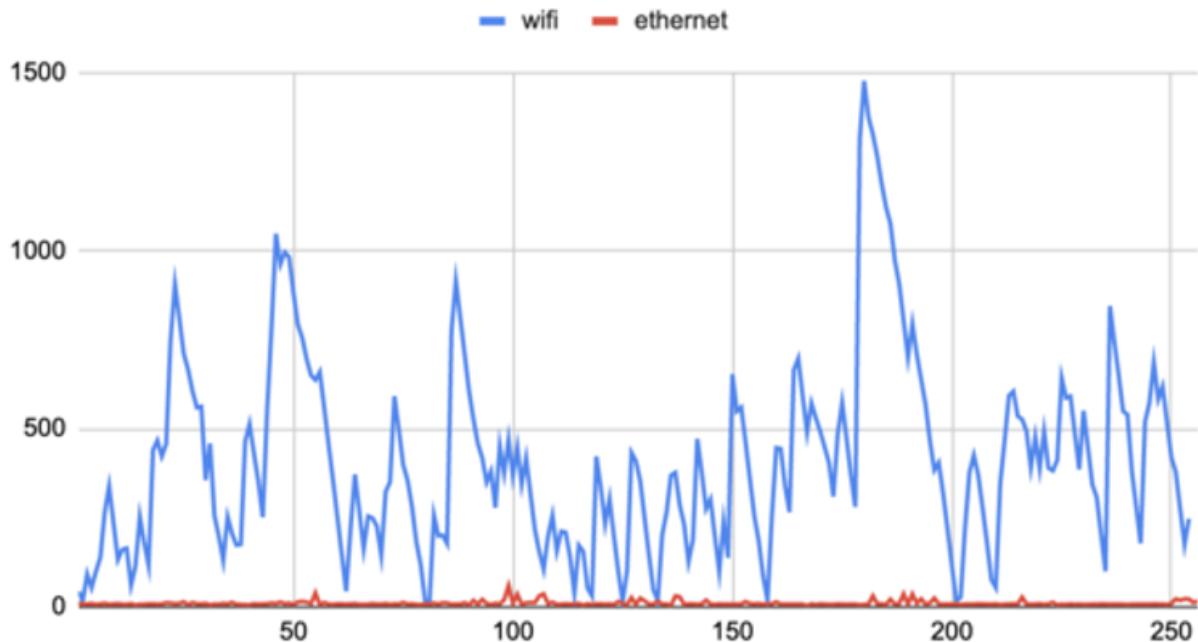
This tells my computer to send 256 packets to that IP, one every 100ms, and time how long they take to come back. First I tested on a relatively idle network, with my computer not doing anything and most of my housemates asleep:

ping time: idle connection



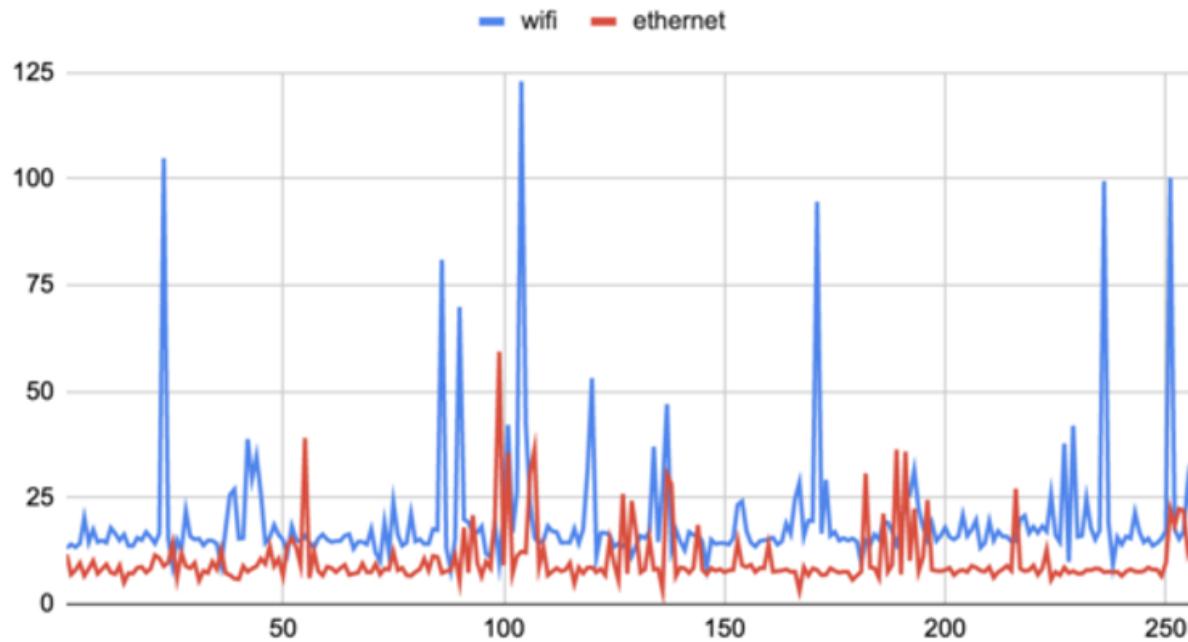
You can see that ethernet is generally a bit better than wifi, with occasional blips, though even the blips aren't that bad. Then I tested on a heavily loaded connection, by visiting a page with enough images that it wouldn't finish downloading them before the test stopped:

ping time: busy connection



Now you can see that wifi is really struggling, while ethernet is comparatively solid. In fact, heavily loaded ethernet does better than idle wifi:

ping time: idle wifi vs busy ethernet



Since you don't get the real benefits of a wired connection until both sides of the connection are wired, now I need to convince all my teammates to run ethernet.

Comment via: [facebook](#)

On “COVID-19 Superspread Events in 28 Countries: Critical Patterns and Lessons”

Analysis of / response to: [COVID-19 Superspread Events in 28 Countries: Critical Patterns and Lessons](#)

The article above, pointed out to me by many good sources, does one of the things we should be doing orders of magnitude more of than we are doing. It attempts to physically model Covid-19, and figure out the ways it spreads *and how relatively dangerous they are*. Then, based on that, it reasons out the wise policy responses and wise personal behaviors.

My analysis backs up the article’s conclusions. There are ways out, but they seem implausible.

There are three plausible vectors; that part of this article is matched by what I’ve seen everywhere.

Two of them are airborne:

According to the binary model established in the 1930s, droplets typically are classified as either (1) large globules of the Flüggian variety—arching through the air like a tennis ball until gravity brings them down to Earth; or (2) smaller particles, less than five to 10 micrometers in diameter (roughly a 10th the width of a human hair), which drift lazily through the air as fine aerosols.

...

And it is on this crucial scale that our knowledge is thinnest. Despite the passage of four months since the first known human cases of COVID-19, our public-health officials remain committed to policies that reflect no clear understanding as to whether it is one-off ballistic droplet payloads or clouds of fine aerosols that pose the greatest risk—or even how these two modes compare to the possibility of indirect infection through contaminated surfaces (known as “fomites”).

This seems super important because we are all choosing how paranoid to be about each of these three vectors, *and in what ways*.

Each of the three vectors has different implications. Here’s the article’s take on them.

1. If large droplets are found to be a dominant mode of transmission, then the expanded use of masks and social distancing is critical, because the threat will be understood as emerging from the *ballistic* droplet flight connected to sneezing, coughing, and laboured breathing. We would also be urged to speak softly, avoid “coughing, blowing and sneezing,” or exhibiting any kind of agitated respiratory state in public, and angle their mouths downward when speaking.
2. If lingering clouds of tiny aerosol droplets are found to be a dominant mode of transmission, on the other hand, then the focus on sneeze ballistics and

the precise geometric delineation of social distancing protocols become somewhat less important—since particles that remain indefinitely suspended in an airborne state can travel over large distances through the normal processes of natural convection and gas diffusion. In this case, we would need to prioritize the use of outdoor spaces (where aerosols are more quickly swept away) and improve the ventilation of indoor spaces.'

3. If contaminated surfaces are found to be a dominant mode of transmission, then we would need to continue, and even expand, our current practice of fastidiously washing hands following contact with store-bought items and other outside surfaces; as well as wiping down delivered items with bleach solution or other disinfectants.

Right now we're doing some mix of these three things, none of them especially consistently or well.

Large Droplets: Six Foot Rule is Understandable, But Also Obvious Nonsense

For large droplets, there is essentially zero messaging about angling downwards or avoiding physical actions that would expel more droplets, or avoiding being in the direct path of other people's potential droplets.

Instead, we have been told to keep a distance of six feet from other people. We've told them that six feet apart is safe, and five feet apart is unsafe. Because the virus can only travel six feet.

That's obvious nonsense. It is *very* clear that droplets can go *much* farther than six feet. Even more than that, the concept of a boolean risk function is insane. People expel virus at different velocities, from different heights, under different wind conditions and so on. The physics of each situation will differ. The closer you are, the more risk.

Intuitively it makes sense to think about something like an inverse square law until proven otherwise, so six feet away is about 3% of the risk of one foot away. That's definitely not right, but it's the guess I feel comfortable operating with.

Alas, that's not the message. The message is 72 inches safe, 71 inches unsafe.

[Unlike the previous case of obvious nonsense](#), there is a reasonable justification for this one. I am sympathetic. You get about five words. "Always stay six feet apart" is a pretty good five words. There might not be a better one. Six feet is a distance that you can plausibly mandate and still allow conversations and lines that are moderately sane, so it's a reasonable compromise.

It's a *lie*. It's *not real*. As a pragmatic choice, it's not bad.

The problem is *it is being treated as literally real*.

Joe Biden and Bernie Sanders met on a debate stage. The diagram plans had them *exactly six feet apart*.

In an article, someone invites the author, a reporter, to their house to chat. Says he's prepared two chairs, six feet apart. "I measured them myself," he says.

Lines have markings six feet apart everywhere.

The parking lot of a Las Vegas hotel marks off spaces six feet apart for homeless people to sleep in, while the hotel is closed. Then of course they sleep end to end within the spaces, so they're actually one foot apart or less, but then what did you expect.

And so on. People *really are* trying to make the distance *exactly six feet* as often as possible.

This isn't remotely a straw man situation. This is society sacrificing bandwidth to get a message across.

Again, *I get it*. The problem is we are *also* sacrificing *any ability to convey nuance*. We are incapable, after making this sacrifice, of telling people there is a physical world they might want to think about how to optimize. There is only a rule from on high, The Rule of Six Feet.

Thus, we may *never* be able to get people to talk softly into the ground rather than directly looking at each other and loudly and forcefully to 'make up for' the exact six foot distance, which happens to be the worst possible orientation that isn't closer than six feet.

In theory, we can go beyond this. You get infected because droplets from an infected person travel out of their face and touch your face.

Thus, a *line* is remarkably safe if everyone faces the same way, modulo any strong winds. The person behind you has no vector to get to your face. And we can extend that. We can have one sidewalk where people walk north, and another on the other side of the street where people walk south. If you see someone approaching from the other direction, *turn around and walk backwards* while they ensure the two of you don't collide. If necessary, stand in place for that reason. Either way, it should help - *if* this is the mechanism we are worried about.

It would be interesting to go through the SSEs (super spreader events) listed, and see which ones involved people facing only one way, if any.

Yes, it's annoying to not face other people, but you absolutely can have a conversation while facing away from each other. It's a small price to pay.

In similar fashion, it seems a small price to pay to *shut the hell up* whenever possible, while out in public. Talking *at all*, when around those outside your household, can be considered harmful and kept to a bare minimum outright (and also it should be done while facing no one).

Masks and goggles/glasses also obviously help block such infections, and we should all be forever furious at those who lied to us and pretended otherwise. At this point I'm assuming we've won that battle, at least among anyone reading this.

Aerosol Droplets: Embracing the Great Outdoors

The article covers this one well, as the implications are mostly straightforward.

In sufficiently dense places even being outdoors might not be good enough. At the height of the pandemic in New York City, it seemed likely that the air in recently crowded areas even outdoors was dangerous, even if you managed to stay six feet away slash behind other people. There's a limit where diffusing the air goes from 'if we diffuse then everything's fine' to 'if we diffuse then everything's not fine' and it's not obvious where that line might be. Also possible that all outdoor travel is at least tiny risk, at that point. Indoors we now care a lot about ventilation and about who has been in a place.

Surfaces: Are They Even Real?

I don't know that we know. We could find out easily enough, if we were willing to be a real civilization that understood that people can make choices and trade-offs, and that was capable of making choices and trade-offs and actually doing things. All you have to do is run physical experiments to find out whether people *actually get infected* from surfaces, at what rates and under what conditions. Then either we ramp up the sanitation messages, or we can stop worrying.

Alas, we are not a functional civilization in this sense. So we don't know, and articles like this one, written by non-professionals because someone had to and no one else would, are our best source of educated guesses. We are all doing our own research.

The precautionary principle here says that one continues to sanitize, so we do, *even though I think it's probably unnecessary*. A probably unnecessary action that might be very necessary remains necessary.

Identified Super Spreader Events are Primarily Large Droplet Transmission

The article makes a strong case that *in identified super spreader events* the primary mode of transmission is large droplets. And that large droplets are spread in close proximity, by people talking (basically everything) or singing (several choir/singing practices) frequently or loudly, or laughing (many parties) and crying (funerals), or otherwise exhaling rapidly (e.g. the curling match) and so on.

There is a highly noticeable absence of SSEs that would suggest other transmission mechanisms. Subways and other public transit aren't present, airplanes mostly aren't present. Performances and showings of all kinds also aren't present. Quiet work spaces aren't present, loud ones (where you have to yell in people's faces) do show up. University SSEs are not linked to classes (where essentially only the professor talks, mostly) but rather to socializing. And so on; see [the full text of the original](#), near the end, for full details.

Also strong is the concrete example of a restaurant where one individual infected many others and the direction of air flow seems convincingly to be the determining factor of who was at risk.

I buy the core thesis. Identified super spreader events, where lots of people get infected, are primarily fueled by large droplet transmission. The pattern is too consistent to be anything else.

The easiest way for that to be the case is for large droplets to be the primary means of transmission. But that doesn't *have* to be the case. What about the alternative

possibilities?

Are Unidentified Super Spreader Events Different?

The sample bias in identified events is not subtle. That doesn't mean we know its magnitude or direction. What would we expect to cause events to be identified?

Events where it is easy to track down participants are going to be included, whereas events where it is hard to track down participants are going to be excluded.

For an event to count, we'll need to track down participants to first confirm that an SSE took place, and then to figure out how many people likely were infected. If you can't do those things, it won't be counted, even if many were infected.

You'd also have to realize that you should try to do this in the first place.

Even if it is possible to track participants down, if you don't know to start doing so in the first place, you won't. So there needs to be an obvious pattern, *that is noticed and pointed out*, that allows the tracking down to even begin.

Thus, if a subway car was an SSE, would we ever know? It's not like you can send out a general call for car 5 of the 9:35 red line between stops 9 and 21. You *might* be able to figure out which subway car or which bus a given person was on, but often you won't be able to even if you're talking to them. I don't think people would even try to track these types of things down.

This explains *some* of the absent types of things, but not others.

Overall the pattern still holds.

Are Super Spreader Events Different from Regular Infections?

This is a much bigger issue. We'll take surfaces first, then small droplets.

Suppose a third of infections were via surfaces (which I don't believe).

Is it plausible that those infections could be distributed and diffuse, rather than creating super spreader events?

In theory, it's *possible*. Under this model, people are constantly touching things, and then other people touch those things, and *any one* interaction is low probability of infection, but there's a lot of them and they add up over time.

There's a power law on how often things get touched. A doorknob might be touched once per minute. Your package was touched maybe three times, period. Thus, one doorknob touch is two or three orders of magnitude more dangerous than one package if everyone touches the package in the same place. Since they don't, it's more like four or even five orders of magnitude.

The fall in exposure to surfaces with countermeasures would also seem very, very dramatic with intervention, because you have to actually touch your face before washing your hands in order for it to count. When you touch commonly touched surfaces you know you've done it. This has to fall almost entirely on the few people

not paying attention. But again, that's plausibly still a big deal, and doesn't answer the original question of whether this is something worth guarding against in the first place.

Still, there's an upper bound here. Surfaces don't cause SSEs. We'd probably know it from things like doorknobs and elevator buttons if they did. It's possible that each person touching an object ends up with a large part of its viral load somehow, which would in turn make subsequent people safer and prevent true SSEs that didn't have conflated potential causes. Maybe. Or perhaps it requires extensive touching of a surface on both ends, which again makes the infections more diffuse in location and time. But if that is required, it would be that much harder for this to be that big a vector.

Any one piece of missing evidence is easy to dismiss. But the absence of evidence keeps piling up for surfaces as a major vector.

Small droplets as a constant small risk is the other possibility. It *makes sense* that they don't cause SSEs while still perhaps causing a lot of infections. They're constantly there but not acute, so one is never at a super high risk *at any given time and place* from them, but they're there a lot because they linger for a long time. In the cases where people do linger in large groups for a while, such that the risk might compound from lots of different infected people continuing to put out small droplets over time that accumulate, there would almost always be a huge confounding with possible large droplets. So even if that happened, we would likely not have noticed it happening.

It still seems like a long shot. It's an especially long shot given that contact tracing has been shown to essentially work in multiple places. If small droplets are a major cause, and they linger for a while, contact tracing will break down. Thus, it's likely that this too is a minor factor in any situation where infection density is sufficiently low for contact tracing. Maybe that changes under mass social distancing plus mass infection, resulting in a meaningful risk from miasma in for example parts of New York City, at least for a while.

Focus Only On What Matters

So, yes. I think it's probably large droplets.

Focus on wearing a mask, on not facing anyone not in your household, on avoiding talking or anyone else who is talking. Aim down whenever possible. And so on.

That doesn't mean the other causes aren't worth avoiding. But unless I'm missing something big, we should be focusing the bulk of our efforts on large droplets, plus direct physical contacts, as the primary source of infection.

We shouldn't pay *zero* attention to packages and other surfaces. We shouldn't pay *zero* attention to small droplets. Better to be safe, even if all you get in most worlds is peace of mind. You feel safe, you know you did everything you could, and so forth.

As individuals trying to be responsible for ourselves and others, it makes sense to use 'an abundance of caution' in such spots. I approve.

But if I was *running an army that was fighting for survival*, and I had limited resources, I'd devote essentially no resources to those efforts.

Or if I was *trying to save a global economy*, and I had limited resources, I'd do likewise. I wouldn't *interfere* with efforts on other lines, but I also wouldn't sweat them.

The thing, from the beginning, is that *only the big exposures, and the big mistakes, matter*.

Within those big risks, *small changes matter*. They matter more than avoiding small risks entirely.

A single social event, like a funeral, birthday party or wedding, might well *by default* give any given person a *30%+ rate* to infect any given other person at that event if the event is small, and a reasonably big one even if large. You only need one. Keeping *slightly* more distance, speaking *slightly* less loudly, and so on, at one such event, is a big risk reduction.

Note that *within-household* transmission rates are not *that* much higher than that and [there are studies](#) saying it is lower! Simply being around a person is much less dangerous than the other methods being important would imply.

Whereas a 'close contact' that doesn't involve talking or close interaction probably gives more like (spitballing a guess, but based on various things) an *0.03% rate* of infection if the other person is positive, and likely with a lower resulting viral load. Certainly those contacts add up, but not that fast. Thus, a subway car full of "close contact" might give you 10 of them per day, most of whom are not, at any given time, infectious. If this model is correct.

That's not to minimize the risks one takes there. The big risk is *model error*. We might be wrong about what's happening. Thus, my best estimate of risk is different from the risk level I'm going to use when deciding what to do. That is as it should be. That's how we stay alive.

More thinking like this, please.

On R0

Epistemic Status: As with all of my Coronavirus posts, I am not any kind of expert. I am a person thinking out loud, who will doubtless make many mistakes. Treat accordingly. However, the concrete policy proposal contained herein seems right and I endorse it strongly.

Partly a response to (Overcoming Bias): [Beware R0 Variance](#)

Previously (not required): [Taking Initial Viral Load Seriously](#)

Also related to Covid-19: [Coronavirus is Here, An Open Letter To The Congregation Regarding The Upcoming Holiday, Let My People Stay Home,](#)

Ultimately, it's always all been about R0.

If you get and keep R0 substantially below one, infections fall off. Covid-19 is squashed.

If you can keep R0 below one and let normal life happen, life can return to normal. If you can't, life can't.

If R0 remains above one, infections continue to rise until something changes.

What will it take to do that?

Interventions that reduce contact and exposure, or reduce the danger of each contact or exposure, or change their dynamics in useful ways, reduce R0. Every person already infected is almost certainly immune (at least for now) which also reduces R0.

You need some combination of interventions and immunity from previous infection that sufficiently reduces the initial R0.

This is not an [o-ring production function](#). You are not as vulnerable as your weakest link. Not everything you do has to be correct. This isn't a designed puzzle where there are *exactly enough* interventions available to solve the problem and you need to do them all. Instead, we have a variety of possible actions and need to pick the cheapest basket that reliably accomplishes the mission.

What Was Initial R0?

As Robin notes, R0 does not start out and never is one number. It depends on the surrounding disease environment. In some places it will start out very high. I've seen plausible estimates for New York City as high as 8. In other places R0 might be lower than one to begin with, because people barely interact with each other in normal times. That hermit with fifty years experience social distancing? His local region's R0 is about zilch.

To get intuition pumping one still needs a good 'default' R0 before any substantial interventions take place, so we can adjust after that for interventions and anything about a particular place that changes R0 substantially versus the default.

[This document](#) collects, among other useful things, a bunch of estimates of R0. If you average out all numbers that don't involve an intervention it comes to about 3.36.

We need to reconcile that with serial interval and doubling times. Looking at the same document for serial interval, we get a minimum of 4 days and an average of 5 days, with an upper bound around a week.

It now seems clear that most of even those sounding early warnings predicted doubling times far longer than those we later observed. Community spread is fast. For a while a lot of places seemed to have doubling times of 2.5 days (and New York City may plausibly have been as low as 1.3 days). If the doubling time is 2.5 days and the serial interval is 5 days then R0 should be about 4, so each serial interval can let us double twice. I have been using that as a reasonable baseline overall guess.

It does seem Bucky and LessWrong [got it right on March 9](#), which *would have materially impacted my life plans* even that late in the game, but I didn't see that in time and only updated at least a week or so later. My usual sources all had 4-5 days or longer, causing me to move much too slowly.

Interaction Interactions

Exposures to Covid-19 outside the household and outside the medical system all seem to have sufficiently low individual probability of infection that we can add their infection probabilities together to get the expected infection count. This is a tiny error, but not enough to matter much.

If we presume that a typical place starts with R0=4, then we need roughly a 75% reduction in total quantity of exposures to get R0=1. It's more complicated than that, in a bit we will beware and also embrace R0 variance, but I want to start with the simple version first.

How much do various interventions cut exposures? What types of exposure are cut how much?

Here is a practical list of ways people might be exposed:

1. Inside of a household or living facility.
2. Social interactions.
3. Medical system.
4. Prison, jail, other detention.
5. Courts.
6. At work indoors.
7. At work outdoors.
8. Recreational indoors out of household.
9. Recreational outdoors.
10. Errands and tasks not otherwise included in the list.
11. Commuting, especially mass transit.
12. Travel other than commuting, especially air travel, trains and so on.
13. Crowd events now gone: Stadiums, theaters, concerts, rallies, etc.
14. Crowd events mostly gone: Religious services, holidays.
15. Grocery stores from other shoppers.
16. Restaurants, including takeout, excluding from the food.
17. Packages, excluding the food itself.
18. Food, including both groceries and restaurants.

19. Delivery contact points.
20. Schools.

What am I missing here?

It is easy to see that *most* of these have been squashed as vectors far more than 75%, but some of them have not been and may even have risen.

In particular, #1 (Inside of Household), #3 (Medical), #4 (Imprisonment), #9 (Recreational Outdoors) and #15 (Grocery Stores) are plausibly as bad or worse than their normal life baselines. I believe we have cut back on #5 (Courts) in most places but those that do remain still seem quite bad.

Inside of household exposure gets worse when you always stay home, which plausibly overrides any extra precautions taken. The good news is that it then gets very hard for many secondary in-household infections to spread further, as we'll discuss later.

The medical system frequently lacks proper protective equipment and is not giving health care workers proper time to quarantine. These are both very bad. The horrible consolation here is that each health care worker only gets sick once, they know to be careful when not on the job, and in heavily infected areas anyone who does not think they have Covid-19 or a true emergency situation is avoiding all health care workers entirely. There might not be that many others left to infect here after a few weeks.

The prison systems make even ordinary-life levels of social distancing or hygiene physically impossible, so basically everyone we don't release is probably going to get exposed. We should be releasing a lot of people but mostly aren't, although new arrests are slowing dramatically. Prisons are presumably headed quickly to herd immunity levels, and the guards will also mostly get infected at some point. Our prison system is pretty terrible even in normal times and this is worse. The 'good news' again is that this vector dies out relatively quickly due to herd immunity.

Grocery stores are often a madhouse now, and people need more stuff from them. Delivery services are maxed out and have failed to expand that much. Here in Warwick the local ShopRite is gigantic but impossible to safely use. The website is permanently overloaded and there are never available slots even for pick-up let alone delivery. Even the aisles are more dangerous than anything else we might do, and checking out is a disaster. Long lines in places others can't avoid. The counterargument is that before a lot of people went very often and would end up very, very close to each other, whereas now at least people are trying to not do that. Grocery delivery might both be the worst thing left for most people, and still better than it was before interventions.

Recreational outdoors could have gotten better or worse, since people take precautions but are also far more desperate to get outside and are crowding some areas.

#2 (Social Interactions), #6 (At Work Indoors), #7 (At Work Outdoors), #8 (Recreational Indoors), #10 (Errands), #12 (Travel), #13 (Stadiums), #14 (Worship), #16 (Restaurants), #19 (Delivery Contact Points) and #20 (Schools) all seem squashed reasonably well, at least 75% in most places. Some are effectively down almost 100%.

#11 (Commuting) is down a lot but perhaps not 75%. That could be a problem in places where commuting often involves mass transit, as riding in one's car is not a big

concern. One of New York City's biggest problems has been that the subway, even with ridership (and in a show of true bureaucratic insanity, number of trains) down by half or more, this vector is still a gigantic problem that most other places don't have.

#17 (Packages) and #18 (Food) depend on how many people are actually taking precautions on these, which is unclear. Presumably they haven't gotten net worse overall, but it is possible. I do not think they are large contributors to the initial R0. Those being careful are 90%+ safer. Those not being careful depend on the precautions on the other end, and are getting more packages.

Or another list, which would be infection vectors more generally:

1. Direct physical contact.
2. Droplets.
3. Surfaces.
4. Fecal/oral.
5. Miasma (to extent this is a thing).

Direct physical contact is way down. Even a simple 'do the things you'd normally do, other than those that involve a lot of direct physical contact, and try not to touch anyone while doing them' seems like it should be good for 75%+. We're going much farther than that.

Droplets are a function of social distancing and presumably follow roughly an inverse square law for each interaction because physics. People are being told to think of a Boolean at 6 feet, which is obviously wrong, but given people's defaults their actual reactions should result in large cuts. It's hard to think they aren't down 75%+ even before masks. Masks then help a lot as well, and the tide is turning on wearing them. If all of that isn't enough, it's presumably a large underestimation of the rise in grocery store exposures, which we can and should limit with extreme prejudice as I suggest in the next section.

Surfaces are something people previously mostly ignored and are now trying to dodge, plus we are mostly not going outside, plus a lot of wearing gloves and washing and not touching faces, so 75%+ reduction seems clear.

Miasma is a function of hanging out in packed places, which has to also be down 75%+ (and may or may not be a vector at all, or one worth worrying about).

The only vector that isn't obviously down at least 75% would be fecal/oral. Increased hand washing and reduced face touching seem like they'd be big games here, as would gloves worn during food preparation and doing most food preparation inside the household. Many are taking additional strong food precautions.

A third method is to look at interventions, and estimate how much reduction each one accomplishes.

How much do we get from hand washing and face avoiding? How much do we get from social distancing? Wearing masks? Working from home? Closing schools? Closing restaurants? Cancelling events? Closing stadiums and houses of worship? And so on.

Low Hanging Fruit: Safer Grocery Delivery

Doing that analysis made buying groceries stick out as the biggest remaining easily avoidable obstacle, and the one place things have gotten much worse for average people.

For most people able to work from home, getting groceries safely is the main barrier to maintaining an effective quarantine.

Delivery and pick-up services exist, but they are not scaling up fast enough and are at maximum capacity. Stores are packed. It is difficult to retain the work force, let alone expand it, as things get more dangerous and stressful. When we last investigated the local ShopRite, there were only three open checkout aisles. Getting even a pickup time has become impossible. Most people cannot afford Instacart's ~30% markup, which is the reason Instacart is still available at all, but that means it does not provide a general solution.

A single trip to many grocery stores, even while taking all realistic precautions and moving quickly, is likely *most* of the exposure for *most* people working from home or not working, even if you only count exposure to other shoppers. Store workers are at high risk from everyone coming in and out.

This is also the only major exposure that most people have that hasn't already been cut by 75% or more. If we could knock it out, it would be a huge blow to R₀. It would also be a huge safety boost to many of our most vulnerable.

The good news is that this problem is so big we can afford to throw massive amounts of money at it. I propose we do exactly that. There is no need to be subtle or careful, here.

I propose a straightforward \$20/hour direct wage subsidy to all grocery store and restaurant workers whose focus is a combination of check-out, pick-up and delivery, up to 20% of the revenue from goods sold (you need *some* cap if you're letting people hire at a good wage for free, these numbers don't seem obviously wrong, and going for the elegance of calling this "a 20/20 vision" seemed nice).

In exchange, we ask only that all take-out and delivery charges must be waved – delivery has to be same cost as in-store purchase. This encourages the use of the pick-up and delivery services without concern about price. As much as I love allocation by price in most situations, we want to force massive scaling here rather than efficient allocation before scaling is finished.

We include existing workers to not punish anyone who scaled up early, to encourage employee retention, and because they are heroes who deserve the hazard pay. As Henry Ford and many others have shown us, if you want extraordinary effort, pay well above the 'market wage' and you will get rewarded. Given what all this is costing us, the cost here is chump change.

And also because we already wanted to throw money at people, our existing methods are taking months to work, and this seems like as good a way as any to get things going.

The resulting lowering (or at least, not raising) of food prices across the board then acts as a progressive subsidy to everyone, taking the same role as sending out universal checks. Competition is a thing and everyone's gotta eat. We also create jobs.

That's version one. I'm sure it can be improved, but it seems like an obviously vastly better deal than anything else on the table, and seems shovel ready.

(A parallel wage subsidy to health care workers on the front lines likely also makes sense for many of the same reasons, especially as hospitals depend on elective procedures to keep the lights on and are in many places cutting doctor pay, and would presumably have broad support.)

With that out of the way, we now return to examining R0.

In-Household versus Out-of-Household

In versus out of household was a key distinction [when considering initial viral load](#). It's also a big consideration when modeling the spread of infections. This is especially true directly after instituting a lock down.

In-household transmissions go up rather than down when a lock down is instituted, since everyone in-household is now staying home more often, and thus interacting more often.

However, once you are locked down, are the resulting infections going to spread further?

The members of a household now have most of their interactions with each other, even more than they might have had before the lock down. Many, including most kids, will have a tiny amount of overall household exposure and an even smaller percentage of overall exposing of others to the household. If you stay home, you essentially cannot infect anyone outside the household, at all.

Thus, the bulk of in-household infections basically stop counting for the effective R0, because the resulting R0 for such infections is close to zero. We would be better off thinking about households as either 'infected' or 'not infected' the way we would previously have thought about individual people.

In turn, this means we will be doubly too pessimistic about lock down effects early on, since the ratio of in-household to out-of-household infections will spike before returning to normal levels and that is not only atypical but relatively harmless in terms of further infections.

Beware R0 Variance?

R0 as one number is a big simplification. R0 variance in some times and places is bad and we need to be beware. In other times and places it is helpful. Details matter a lot, as does the big picture.

R0 variance raises overall R0. If we have two groups are mostly distinct and only interact rarely, one with $R_0=0.5$ and one with $R_0=1.5$, we get most of our infections in the $R_0=1.5$ group, and R_0 ends up well above 1. Thus, if we say that $R_0=4$ in general, we're saying that the average across groups or regions is less than that. We are not doing good modeling if we say that $R_0=4$ in general but with variance so $R_0>4$, although there will be scenarios where it temporarily goes over 4.

If our goal is to squash entirely and then reopen everything, R_0 variance across places and groups naively looks quite bad. We need to squash in every group, or the high- R_0 groups will reinfect the low- R_0 groups where we squashed successfully. The nightmare is you get $R_0=0.5$ in general via strong suppression measures, but with a lot of variance, and in some places it still isn't enough or the measures weren't adapted, and you need to stay shut down or it all comes back.

In other situations, the variance works in our favor.

If we can properly target our interventions to particular groups, places and events, R_0 variance is great. R_0 is gigantic in sports stadiums so we shut down the sports stadiums. If R_0 was huge in New York City but under 1 in the rest of the country, we could close the bridges except to haul in supplies, and let 97% of our people carry on largely as normal. Wuhan being compact was great for China. National borders are a thing and can be enforced, and so on.

If we are willing and able to let things burn out to herd immunity in some places or among some groups, similar things happen. As an alternative to shutting New York City down indefinitely, we could simply declare defeat there, and allow New York City to lock down its most vulnerable as completely as possible while rapidly getting to herd immunity, while not following that strategy elsewhere. Then New York City is no longer a threat.

What is happening inside New York City likely looks a lot like that strategy, except for a subgroup rather than the whole area.

A Very Simple Model of New York City

Some New Yorkers were able and willing to self-isolate, or even to flee the city entirely. Others were not.

There is a continuum between my parents, who are literally only opening the door to take in packages that they then sanitize, and those who have no choice but to ride packed subways every day to make ends meet, or are so young and selfish as to be indifferent to what happens.

Let's simplify that a ton, and say that some people ride the subway and some don't.

The people in the subway are packed tight, because the MTA cut the number of trains in proportion to the cut in ridership. Those still in the subway have several times the exposure to people that others have, and those exposures are to other strangers who ride the subway.

Among those who ride the subway a lot, infection rates get out of control very quickly. We should expect by now that those still riding the subways daily have essentially all been exposed enough at least a week ago to get infected if they aren't somehow naturally immune - look at the admitted-to-be-undercounted death rates, assume those people were infected three weeks prior, look at the doubling rates during those times, and snowball the effects. There's no need to do any math here.

By the end of April and quite possibly the end of March, the subway group is mostly immune, and has blown past steady state herd immunity levels even with its high levels of exposure. The city then has R_0 levels that are even lower than if the subway was taken entirely out of play, since those people also interact elsewhere.

It's definitely not the solution we would want! It's a ton of infections. Also deaths. Still a lot less infections and deaths than doing the same to the whole city, which will now have a much easier time squashing than if everyone in the population was identical. It also means we can now use the subways to get essential workers where they need to get, without making the forward-looking problem worse.

Actual Exposure Inequality

I consider the above scenario a *relatively equal* distribution of exposures, compared to the real situation.

Even in normal times, different people have radically different levels of exposure to crowds, to being physically close to others, to interacting with different as opposed to the same people repeatedly. Different people have radically different levels of caution and hygiene.

People have speculated about 'super spreaders' who are much more infectious than others, and who might account for the bulk of out-of-household infections. That might or might not be a thing. What is definitely a thing is that some people are in position to create additional orders of magnitude more exposure in others, especially others that are socially disjoint from them, than most people are in position to create.

The people who are in position to infect lots of others are, mostly, also the people in position to be infected by many others in the first place. If you are the pastor and shake hands with your whole congregation each week under normal circumstances, you might both have 100 times as much exposure to catching the virus, and 100 times as many places to spread the virus.

Both directions follow a power law, and both directions are highly correlated.

Thus, the idea that to get R_0 from 4 to 1 via herd immunity requires 75% immunity rates *doesn't make even a little bit of sense* to me. People won't be infected at random. And yet this is the standard calculation used everywhere I look.

One follow-up question is, are those who have more exposure more commonly linked up with each other than with others?

In some cases, clearly yes. The subway is a clean example of those at risk mostly exposing each other. So are all the young men who continued not social distancing after they were ordered to distance. They mostly hung around each other. The reason Florida's spring break was likely to be bad was because those there were engaging in *unusual-for-them* behavior, and would later go back to other places and act normally. Under normal circumstances, if the beach or park is packed, then the group of people willing to go there will put each other at great risk but may not put others at such high risk. This gets us a lot of effective immunity at low cost.

Consider how we got to $R_0=4$. If one third of the people interact twice as much as everyone else, in both directions, and all interactions are random, than at the start they constitute half of all infections. Thus, we start out at $R_0=2.67$ for infections in the bigger group. If the third who interact more was immune, that's half of possible infections, so that goes to $R_0=1.33$, so you only need a quarter of the less interactive group immune to get $R_0=1$ with a 50% infection rate rather than 75%. You can't get that whole win since you wouldn't actually infect the whole more-interaction group that fast, but you could certainly get half this effect and get down around 62%.

The real ratios are *much bigger* than that. It doesn't take that much of this effect to get $R_0=1$ in the bulk of the population even before anyone is immune or does any social distancing.

If I had to guess what percentage of people we would need to *naturally* infect (rather than selecting them on purpose) to get herd immunity with no behavioral alterations? I would guess something like 35%, rather than 75%. And I'd expect that the first 3% infected, which I am guessing is about where the United States is right now, would get us *far more than 10%* of the effect of 35%.

That's without us doing deliberate infection or antibody testing in any intelligent way. We can do even better if we can create and/or identify immune citizens and put them in positions of high two-way exposure. Unless we completely drop the ball (which would fit our existing patterns so far) we should start to see substantial gains from antibody testing by the end of April.

Thus, I've become more of an optimist going forward. Variance makes the problem anti-inductive in the sense that the virus will thrive wherever conditions are best, but friendly longer term once containment fails anyway, because it separates out problems that solve themselves relatively quickly at lower cost.

I expect that the way we get out of this is a combination of herd immunity coming on stronger and faster than we expect, combined with the cumulative effects of various interventions, many of which can be maintained in a "Phase II" style world without that high a cost, mostly via voluntary action. We can continue to not shake hands and wear masks and avoid large gatherings and have a third of people work from home and so on at a relatively tiny cost. Then we combine that with test and trace, which helps even if it isn't ubiquitous, and we can have an acceptable situation while we await a vaccine.

Postscript on Recent Data

Since I started writing this several days ago, the official numbers have improved considerably in America and abroad. New infections are leveling off, test positive rates are down, and deaths are no longer growing exponentially. This is faster than even my at-the-time prediction that new infections (rather than new detected infections) peaked specifically in New York City on April 1, and the stock market has rallied accordingly.

I do think we should be deeply skeptical of these numbers. New York's numbers especially seem to mostly reflect hitting maximum hospital capacity and maximum testing capacity, both of which are mostly static. We only count deaths if they involve positive tests and in practice in New York that means hospitalization. So when we see deaths leveling off faster than would make sense given how deaths lag, we likely shouldn't interpret even deaths as meaning much more than 'the numbers are higher than this but we have no idea how much higher.' With even deaths clearly undercounted, calculations that start with official death counts, without adjustments, are going to be underestimates.

The percentage of tests that are positive is probably the best data we have, but even that seems to be strangely noisy. American positive test rates were in the 23%-26% range for four days, then dropped to 15% before increasing back to 19% for 4/5 and 4/6. This is even more stark if you exclude the epicenter in New York and New Jersey (where positive rates have been between 46% and 51% with 20k-31k daily tests each

day since 3/30), with the rest of the country having four days from 15.7%-18.3% positive rates, then three days from 9.5%-12.9% (all data from [the Covid tracking project](#)). That has to reflect strange distributions in reported tests and/or incorrect data. So it's all full of noise, and probably stays that way until we get antibody tests. Hopefully we get better data soon.

Coronavirus: Justified Key Insights Thread

This is a thread to list important insights and key open questions about the coronavirus and the coronavirus response. The inspiration for this thread is Eliezer's post below.

I'd like this thread to be a source of claims and ideas that are self-contained and well-explained. This is not a thread to drop one-liners that assume I've been following your particular news feed or know what's happening in your country or that I've read a bunch of studies on (say) viral load. There's a place for such high-context discussion, and it is not this thread.

Please include in your answers either a claim or an open question, along with an explanation or an explicit model under which it makes sense. I will be moving answers to the comments if they don't meet my subjective quality bar for justification - see [the last justified answers thread](#) for examples of what quality answers look like.

The purpose of giving models and data is to allow other people to build on your answer. Everyone can make arbitrary claims, but models and evidence allow for verification and dialogue.

The more concrete the explanation the better. Speculation is fine, uncertain models are fine; sources, explicit models and numbers for variables that other people can play with based on their own beliefs are excellent.

This thread is inspired by a [post](#) by Eliezer Yudkowsky which I'll reproduce below, in which Eliezer lists eight answers that this sort of post would come up with.

These are not justified to the standard of the thread, so you (you!) can get some easy karma by leaving an answer that justifies one of these with the sources/data/explanation needed to argue for it. It includes much of the discussion elsewhere on LW (e.g. by Wei Dai, Zvi, Robin, and others), so it shouldn't be hard to find the prior discussion.

Eliezer's post ([link](#)):

What do we early-warning cognoscenti now know about Covid-19 that others haven't currently figured out? What's the TOC of that blog post? [@WilliamAEden](#) [@robinhanson](#)

My stab at a TOC:

1: The Dose Hypothesis - the theory that C19 fatalities vary by how high the initial dose, and possibly how it's administered.

1a: So: Human trials of variolation are hugely urgent.

1b: So: Getting C19 from a roommate might be much worse than getting it on public transportation.

2: Challenge trials of vaccines save net lives.

3: Ventilators no longer look as important because they only save 15% of the patients on them.

4: There's huge apparent variation in CFR by country, and explaining this, or explaining it away, seems kinda important.

4a: CFRs may be underestimated by up to 3-fold, based on looking at excess death rates year-over-year.

4b: CFRs may be overestimated because of too little testing.

5: There was a huge EMH failure w/r/t C19, and it hasn't been explained away AFAIK.

6: Most of the economic damage from a real shock like this one is still due to the secondary demand shock, which can be prevented by decisive central bank action.

6a: We know the Fed isn't currently doing enough here because inflation expectations are dropping, showing the AD shock exceeds the AS shock.

6b: Stock prices take into account the next 15+ years of earnings. The real C19 shock only damages the next 2 years of earnings. A financial recession would damage many more years. Stock prices mainly reflect central bank policy, not C19.

7: Face masks do work, though others seem to have mostly figured this out.

8: The mainstream media's words on C19 may be best interpreted as not intended to mean things; like the way that MSNBC's talk about Bloomberg being able to give each American over \$1,000,000 can't have had a concrete model of reality behind it.

Any items I'm missing here?

The Unilateralist's "Curse" Is Mostly Good

People around here sometimes reference the "Unilateralist's Curse", especially when they want to keep someone else from doing something that might cause harm. Briefly, the idea is that "unilateral" actions taken without the consent of society at large are especially likely to be harmful, because people who underestimate the resulting harm will be the most likely to take the action in question.

The canonical formulation of the argument is in [a paper by Bostrom, Douglas, and Sandberg](#), most recently updated in 2016. I highly recommend reading the "Introduction" and "Discussion" sections of the paper. (I found the toy mathematical models in the middle sections less useful.)

While the paper mostly gives arguments that unilateralism could be harmful and ought to be stopped via a "principle of conformity", the authors concede that the historical record does not back this up: "[I]f we "backtest" the principle on historical experience, it is not at all clear that universal adoption of the principle of conformity would have had a net positive effect. It seems that, quite often, what is now widely recognized as important progress was instigated by the unilateral actions of mavericks, dissidents, and visionaries who undertook initiatives that most of their contemporaries would have viewed with hostility and that existing institutions sought to suppress."

For example, an [older 2013 version of Bostrom et al's paper](#) notes that "The principle of conformity could be seen to imply, for instance, that Galileo Galilei ought to have heeded the admonitions of the Catholic Church and ceased his efforts to investigate and promote the heliocentric theory." The 2016 update removes this and most of the discussion of unilateralism's importance to science, but retains the discussion of Daniel Ellsburg's decision to leak the Pentagon Papers to the media as an example of beneficial unilateralist action. A cursory glance through history finds many, many other examples of unilateralists having strong positive effects, from Stanislav Petrov to Oskar Schindler to Ignaz Semmelweis to Harriet Tubman.

Bostrom et al continue, "The claim that the unilateralist curse is an important phenomenon and that we have reason to lift it *is consistent with the claim that the curse has provided a net benefit to humanity*. [Italics mine.] The main effect of the curse is to produce a tendency towards unilateral initiatives, and if it has historically been the case that there have been other factors that have tended to strongly inhibit unilateral initiatives, then it could be the case that the curse has had the net effect of moving the overall amount of unilateralism closer to the optimal level."

Obviously there are any number of "factors that have tended to strongly inhibit unilateral initiatives", and not just historically. These continue today, and will last as long as the human condition persists. Examples include reflexive conformity a la [Asch's experiments](#); loyalty to friends, tribalism, and other [identity](#)-based sentiments; active punishment of dissenters via methods that range from execution to imprisonment and torture to civil lawsuits to "merely" withholding acclaim, funding, and opportunities for promotion; and most important of all, the sheer difficulty of figuring out better ways of doing things.

An undiscriminating “principle of conformity” would be just one more source of drag on progress across the board. In most areas, the benefits of unilateral action are far larger than the costs. In areas where particular caution is warranted, such as sharing weapons technology, the potential harms are widely understood and unilateral action by altruists is at most a minor factor. Advocates of inaction should rely on arguments about specific harms, not vague heuristics to buttress the natural human tendency towards deference.

College advice for people who are exactly like me

It's college decision season! To celebrate, I've been thinking about what I would have told myself in 2011 when I was deciding where to attend. Here's a stitching-together of several emails I sent to friends who asked for college advice, plus a few additional thoughts.

As a caveat: I really do mean that this advice is for people who are *exactly like me!* So please don't take my word for things—lots won't apply to you because we're different. As with all advice: think about it, see what resonates with you, consider [reversing it](#), and generally take it as one of many data points.

Keep the goal in mind

At the end of college you should probably have a good idea about what to do next.

Here's the default way to figure that plan out: (1) don't think about it for three years; (2) get really stressed during your senior year because you have no idea what you want to do; (3) notice that all your friends are interviewing for jobs in consulting, finance, and MicroFaceGoogAzon; (4) accept a job from consulting, finance, or MicroFaceGoogAzon.

This process... could stand to be improved. Plan ahead and don't make a "safe" decision based on social defaults or fear of not having a job. The most important question to answer is "what kind of jobs exist that (a) I would enjoy and (b) I have a reasonable shot at getting?"

Yes, it is scary and stressful to stare this question (especially b!) in the face, but it should be a *lot* scarier to waste years doing something safe instead of something awesome.

That isn't to say you need to know with any amount of certainty what you want to do. (That's unreasonable and it will change anyway.) Just that it's good if your choices during college are informed by *some* broad sense of what they're setting you up for later. "Plans are useless; planning is indispensable."

Social environment >> classes

The other useful thing to get out of college is a bunch of awesome, smart and competent friends. I'd recommend optimizing for that, probably more than educational quality or status.

Sub-points:

- On "density of people who are great at STEM," my impression is that research universities score a lot better than small liberal arts colleges. I didn't apply to Caltech, but by some (lagging) metrics it has an [even higher density of accomplished people](#) than other top schools.

- Control over who you live with/near is a huge selling point, since it mostly determines your social circle. For example, Harvard assigns dorms to groups of upperclassmen randomly—you can pick your roommates but not the other roommate groups that you live near. This makes it hard to find the most awesome people because they’re mixed into boring groups. Other schools like MIT will often have dorms with different cultures and types of people, which makes it easier to find clusters of awesome people. For that reason, if I were choosing colleges today I might have prioritized schools with better housing policies.
- If you go to a school with a strong subculture, make sure you retain your ability to interact well with “normal people.” The stereotype of, e.g., MIT alums that end up spending the rest of their life only hanging out with other MIT students, does happen to some people. It’s a reasonable personal choice (normal people can be annoying!), but if you’re optimizing for impact on the rest of the world, it will be a handicap.

Get better at deciding what to do

One thing I wish I’d had more of in college is something like a sense of taste: not aesthetic taste, but more “taste for what’s awesome” or “taste in good projects.” (My role model for good project-taste was my friend [Adrian](#), who coauthored articles in *Nature*, *Science* and *Cell* before graduating.)

Good taste is what will guide you to doing effective things, instead of pointless things, in college. But it’s also really hard to *develop* good taste while in college, because you have no models of good taste except for academics. Academics’ taste is often deeply weird and driven by what niche will get them tenure rather than what is exciting or useful, so it’s good to have other models of taste too.

My best guess for how to do that is to take a gap year, during which you can front-load learning how the real world works. Summer internships outside academia are also good.¹

Take a gap year

Yes, I just mentioned this above, but it’s concrete and important enough to merit its own section. I didn’t personally take a gap year, but everyone I know who took one found that it enormously improved how they used their time in college, and nobody regretted it.

If the point of a gap year is to improve how you spend your time in college, then you should spend it trying to learn as much as possible about how the world outside college works, and in particular, what that world considers valuable. Try to find projects that let you work with non-academics and produce something genuinely useful.

(It’s common for people to spend their gap years volunteering or traveling. A bit of that will expand your horizons, but spending the whole year on it, while fun, won’t maximize your learning.)

Bias towards action

One of the biggest new things about college is that you have way more independence. Because your parents no longer oversee your day-to-day life, it takes a lot less friction to do something random or unexpected.

Many college students take advantage of this by filling their schedule with (a) either parties or problem sets, plus (b) lunch dates where they complain about how busy they are. There are more exciting options! Here are some things I did that I think were more valuable on the margin than additional problem sets, even though a lot of them “didn’t work out:”

- interning at GiveWell (didn’t end up working there, but super valuable for exposure to various people/ideas)
- running [a student group](#) (I wasn’t very good at it but it turned out pretty well anyway)
- trying out being a math teaching assistant (turned out the grading:fun ratio was way too high)
- trying out research (couldn’t find projects I was excited about)
- joining a choir (was incredibly fun but I dropped after a year because the schedule was intense)
- writing a course catalog website (now unmaintained but was the first project I worked on from the beginning to shipping)
- writing blog posts

Of course, the problem sets are also important! But if you discover you can’t live without them, well, you can have six more years of grad school for that. If you discover you can’t live without parties, I’m not your guy for advice.

Get good advice

Getting good advice is one of the highest-impact things you can do! A one-hour conversation with a good advice-giver can totally change how you spend months of your time. Most people underrate good advice by a lot.

This is because most advice is bad.² For instance, if your college assigns you advisers they will probably be terrible.

My freshman year I was assigned a dental student as an adviser. I’m sure she was doing her best, but she had absolutely no idea how easy or hard any of the introductory courses were. (She could tell me things like “that course has a reputation for being hard,” but students come in with such different backgrounds that “hard” in the abstract is a completely meaningless description.) Her only contribution to my education was to convince me not to take [the hardest intro math course](#) during my first semester. I don’t think that made my life *that* much worse, but I do think I missed out on a great time!

My sophomore adviser would have been worse if I hadn’t learned my lesson by that time: when I handed her my study card she warned me that “introductory theoretical statistics” might ruin me, presumably because she hadn’t heard of algebraic topology. (There’s no reason she should have—she was studying education.) Fortunately, I successfully convinced her I’d be fine.

(Of course, the advice I really needed to hear was that “hard” and “important” aren’t [related](#), but that would have been way too much to ask for a 10-minute conversation slot!)

The problem is that by college, good advice is *incredibly* context-specific to you—what you enjoy, what frustrates you, what your background is like, what kind of goals you’re likely to have—that most formal advisers won’t ever have the time to understand.

Because of this, most of your best advice will probably come from bouncing things off friends—optimally friends who are a bit older than you, but people in your year are also probably better than random faculty. I think I under-utilized this a lot in college and would have benefited from reaching out more to people who I knew had a similar outlook.

(It’s possible to get good advice from authority figures at your college, but you’ll need to recognize their limitations and biases—no dental student will give you good advice about math courses, few professors will tell you that [grad school sucks](#), etc. If you treat them skeptically and do the work yourself to focus their advising on the areas where they’re well-informed, then you might have better luck than I did.)

Aggressively ignore bullshit

Coming in, you will probably think that your college experience has been carefully designed to provide you the best possible education. (HA HA HA.³) As a result, when your college asks you to do things, you will feel inclined to take them seriously.

This is a mistake. Your college will spew lots of bullshit at you. Ignore it when you can; when you can’t, try to limit the degree to which it seeps into your life.

Examples of things that, in my opinion, turned out to be bullshit (of course this will vary by school, field, worldview, etc.):

- Physically attending lectures (just listen to them at 2x speed)
- [“general education”](#)
- Paperwork (let me tell you the story of why I don’t have a minor in computer science!)
- Prerequisites for courses (note: not always!)
- Many humanities courses (but very much not always!)

Lest this sound too cynical, let me point out that (a) college is still awesome and (b) ignoring bullshit gives you *crazy superpowers* to focus on the actually awesome parts, like spending time with smart people, learning difficult things, trying out different jobs and activities, etc.

Don’t take this too seriously

I’ll close with some reasons you shouldn’t take this advice too seriously:

- This advice is very specific to my background—not only educational but also cultural/economic—as well as my general outlook, life goals, etc. If you’re different from me on any of these axes it will become less relevant to you.

- I was lucky to go to a very good high school, so I came into college with a very strong math/CS background already.⁴ As a result, I didn't learn any concretely useful skills in college except for some statistics and machine learning (though I think some other classes I took improved my general problem-solving ability a lot). If you're coming in with less background, the quality of instruction might matter more than it did for me.
 - I haven't been out of school for that long: someone with 10-20 years more life experience would have a different perspective on what was valuable and what wasn't. Unfortunately, I don't know anyone in that age bracket who was similar enough to me that I think their advice would apply well.
-

1. If you want to go to grad school, you might need to spend your summers cranking out research projects instead, but also, consider whether you really want to go to grad school... ↵
2. Including this advice! I know nothing about your situation or preferences! ↵
3. Actually, your college experience has been carefully designed to provide you with a piece of paper signalling to potential employers that you are intelligent and work hard, while in the process transferring as much money as possible from your family to administrators and construction firms. Also there are some people who will try to teach you stuff. The silver lining is that most of the brightest people in the world also want the same piece of paper so you can spend a lot of time hanging out with them. ↵
4. Linear and abstract algebra, real analysis, and a few years of post-AP CS. ↵

What are your favorite examples of distillation?

I'm a big fan of the [Distill](#) machine learning journal and the ideas of [Research Debt](#) and distillation. I consider Distill and LessWrong great repositories for distillations of ML / AI and some math topics. However, I've recently been hankering for distillations from other fields with which I'm somewhat familiar -- biology, algorithms, economics-- or even not that familiar. ([John Wentworth](#)'s recent series of posts on aging and constraints are good examples of one form posts like this could take.)

So, I figured I'd ask here: what are your favorite examples of distillation in different fields? I'm open to more ML / AI related posts but am especially excited about responses in the fields I mentioned above or other different fields (I would include math here too). Ideal answers would be posts that optimally trade off:

- Describing a non-trivial topic.
- Not "dumbing it down".
- Being accessible to non-experts.

Why anything that can be for-profit, should be

This is a linkpost for <https://rootsofprogress.org/organizational-metabolism-and-the-for-profit-advantage>

In a recent post I surveyed different [types of funding models](#), including nonprofit models such as universities or private foundations, and for-profit models such as startups. Although we need both models, I believe that for-profit models are underrated today. In what follows I will explain some fundamental advantages of the for-profit model.

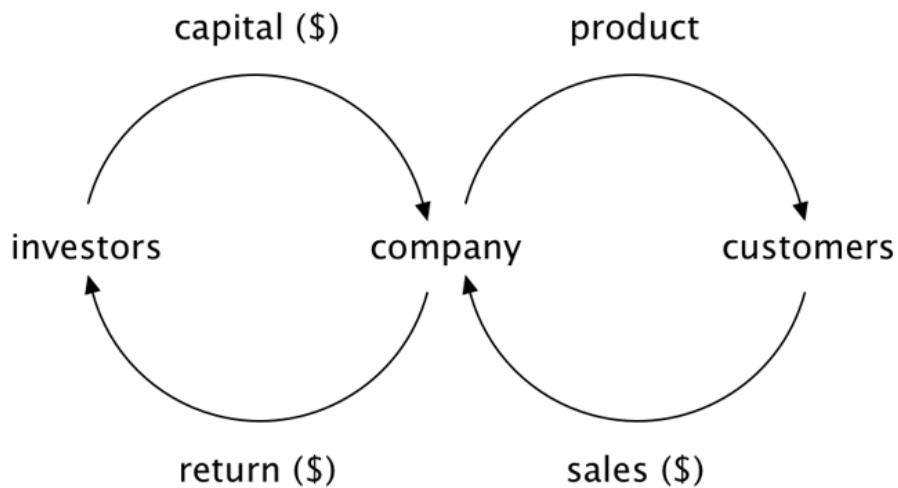
To understand these advantages and why they are essential, we have to understand the incentives and feedback loops that exist within the for-profit and nonprofit worlds. One framework I use to think about this is a concept I call *organizational metabolism*.

An organization, like a biological organism, has a metabolism: a core set of processes that keep it alive. These processes provide the resources it needs to act, to grow, and to sustain itself. For an organism, the basic resource is energy, whether from food, sunlight or chemicals. For an organization, the basic resource is money, for payroll and other expenses. Just as an animal must eat, an organization must collect revenue, or die.

Understanding an entity's metabolism is fundamental to understanding its role within an ecosystem of competing entities and the selective pressures it is under. An entity with a successful metabolism survives and grows; one that fails in its metabolism is eliminated. Over time, through natural selection, an ecosystem becomes dominated by the entities with successful metabolism. Different entities can have different designs and make different choices, but the laws of nature decide which of them thrive.

Metabolism can be represented as a set of feedback loops. Here is the metabolism of a for-profit business:

For-profit organizational metabolism



Jason Crawford / rootsofprogress.org

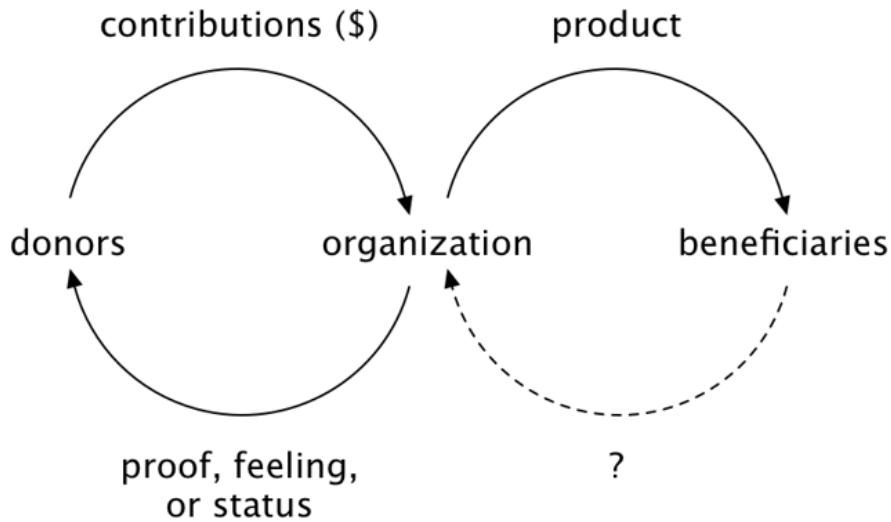
CC-BY-ND 4.0

The main loop is on the right: a business sells a product or service to its customers for a profit. To give birth to a business, investors can put in capital, which must produce a return; this is the loop on the left. Call the right-hand loop the “product loop” and the left-hand the “return loop”.

A business must be profitable over the long term to survive—but any business that makes a profit *can* survive. Similarly, a business must credibly promise a return to receive investment, and an investor must make a return over the long term to continue investing—but any promising investment can receive capital, and any successful investor can stay in business. And over time, resources accrue to the most scalable and profitable companies, and to the most successful investors.

In contrast, here is the metabolism of a nonprofit organization, such as a charity or foundation:

Non-profit organizational metabolism



Jason Crawford / rootsofprogress.org

CC-BY-ND 4.0

A nonprofit also typically provides a product to certain beneficiaries, whether clean water for poor rural areas, or a non-rival good such as research for the world at large. Note however that this product loop *is not its core metabolic loop*. As we will see, the nonprofit product loop is weak, and sometimes broken.

The loop that sustains a nonprofit is its return loop. Donors provide the organization with its revenue. What do they get in exchange? What does the organization have to provide to sustain or grow its revenue?

In theory, donors give because they support the organization's mission: they want to sustain its product loop. In the best case, the product loop is closed with evidence of value delivered, from data and statistics to testimonials and case studies. This is provided to the donors to close the return loop: objective demonstration of success at the organization's mission leads to more and larger donations.

Not all donors, however, are this discerning, thoughtful, or well-motivated. A charitable donation is often an emotional choice, fueled by moral sentiment. Donors may not demand, or even care for, evidence of success; they may be more persuaded by pictures and stories that tug at the heartstrings, or by the charisma of the executive director. Or they may be driven by motives less honorable than generosity, such as prestige, status, a public image. In short, donors may care less about *doing good* than about *feeling good* or *looking good*.

To the extent that donations are not driven by proof of effectiveness, the product loop is broken: it is no longer a loop, and it *does not matter* to the organization's metabolism. The organization may still provide a valuable service, but if so, this is incidental to its survival, an epiphenomenon. The organization is now in a loop of providing a feeling or a social image to its donors, and as long as it can do that, it can survive indefinitely whether or not it does any good for the world.

It's *possible* for a pathology like this to exist in a for-profit business—the nutritional supplements industry is probably an example—but it's restricted to a subset of products for which it is difficult to make rational buying decisions. And even a supplements company would quickly go out of business if it failed to physically deliver its product into the hands of customers—yet this kind of failure is *routine* in the nonprofit world, where disaster-relief donations are often discarded and international aid ends up in the pockets of warlords.

Nonprofits can of course be effective and efficient, and some are more so than many businesses—but this happens more by the grace of an executive with courage and integrity, or the rare strategic donor, than by inherent selective pressures.

A second thing to note about these diagrams is that every part of a business's loops can be measured. On the right, the business can measure costs, sales, and profit; on the left, capital, dividends, and return on investment. From fundamental measures such as these are derived a sophisticated system of financial metrics, and an entire discipline to manage them: accounting.

Financial metrics have a bad reputation; they are the implements of corporate "greed" and are blamed for everything from sweatshops to pollution to securities fraud. A myopic focus on financial metrics can indeed lead to long-term destructive behavior. But used properly, finance is a powerful tool to promote efficiency and effectiveness. It can identify waste, optimize portfolios, and justify long-term investment. Crucially, it can manage *risk*.

Precise, quantitative risk management is especially important in the allocation of resources to early-stage projects. Along the efficient frontier, risk is correlated with potential reward. A mechanism is needed to provide the investor with a higher rate of return in exchange for funding riskier projects. In debt, this is done with interest rates; in equity, by the mechanism of valuations: higher risk is reflected in lower stock prices. Particularly in the world of venture capital, this means there is a *disproportionate reward for being right early*—for being the first backer of a promising project *at its inception*, when risk is highest.

The result is that investors are actively incentivized to embrace *contrarian* theses: they win biggest when they bet on something that the rest of the world would not, once success is proven over time. I know of no such mechanism or incentive in the nonprofit world, where the incentive for donors is if anything the opposite: to seek out safe, consensus causes, especially to the extent that the donor's motivation is prestige or social approval.

Organizational metabolism helps explain the strengths and weaknesses of different [types of funding models](#), along the [lines I drew in a previous post](#).

The weaknesses of the money metabolism come from two key requirements that the nonprofit metabolism does not have. First, you must capture part of the value you create. Second, you must generate a return within the time horizon of investors (about a decade in venture capital). As a result, there is often pressure to fund specific, visible goals rather than completely undirected exploration.

Conversely, the strength of the nonprofit model is its lack of these limitations. It is probably a better way to support, for example, basic research, which tends to explore uncharted territory to generate open, often unpatentable knowledge that leads to economic value only on the timescale of a generation or more. (There are also reasons to believe that nonprofits are better suited for a variety of areas [from performing arts to nursing care](#).)

However, when it is possible, I see great advantages to the for-profit model:

- **Revenue is as scalable and reliable as demand itself.** When your market grows, your revenue grows; as long as your unit economics work out, you can scale supply to meet demand. In a nonprofit, if your market grows, *only your expenses grow*. Your revenue is at best loosely tied to demand. A charity that donates clean water to poor rural Africans may or may not find enough rich Westerners to pay the bill, but a business that figures out how to sell water to them at a profit will serve as many customers as are willing and able to pay. Again, this is not to say that such a charity should not exist, but simply to point out that it has inherent limitations to scale.
- **For-profits have the incentives and the metrics to drive effectiveness and efficiency.** A charitable donor may not think strategically about how many meals, mosquito nets, or doses of a vaccine can be delivered for their million-dollar donation. Indeed, to the extent they are motivated not by doing good but by feeling good or looking good, they are likely to measure themselves not on the results but on the *amount of money* they gave, or the portion of their wealth that represents—thus making the fundamental mistake of substituting metrics that measure input for those that measure output. In contrast, a business is constantly driven to serve more customers at lower cost, to expand into new markets and new product lines, to reduce waste, to cancel or overhaul failing programs. (Notable exceptions include many effective altruist organizations, [such as Open Philanthropy](#)—notable in part because the strategic thinking they evidence is so rare.)
- **For-profit investors are more likely to fund high-risk, high-reward experiments.** The incentive for investors to seek contrarian theses at the individual level leads to a *diversified global portfolio of bets* at the ecosystem level. The mechanism of equity, in particular, allows investors to be rewarded proportional to risk, which allows each investor to build a successful portfolio even when many of their bets fail. Given how many crazy-looking, risky experiments turned into breakthrough innovations that transformed the world, a mechanism that funds them when they are overlooked by others is extremely valuable.

These advantages are so great that I submit a bold principle:

Anything that can be for-profit, should be.

Corollaries:

- Nonprofit organizations should be formed only as a last resort to fill in the gaps where for-profits cannot work.
- If a given need cannot be served profitably, then finding ways to *make profit possible* in that area is even better than serving the need with nonprofit organizations.
- Institutions and mechanisms that create opportunities for profit—such as property rights, including intellectual property rights—create enormous value for a society.
- Conversely, when profit is prohibited in or drained from an industry, it represents an enormous *destruction* of value.

As an example, a challenging area of pharmaceutical development is repurposing drugs: discovering that a known drug (which may be off-patent) works for a disease it was not previously known to work for. Patents for this type of use can be [difficult to obtain and enforce](#); better market protection for repurposed drugs would make it more profitable to discover these uses, and pharmaceutical investment would follow.

Conventional thinking says that money is a corrupting influence. We hear calls to get money out of politics, out of health care, out of everything. I believe the opposite.

Money illuminates what is murky. It aligns interests. It keeps people honest. Money is, in fact, one of the greatest forces for social good. Instead of getting money out, we should find ways to bring money and profit *into* as many areas as possible.

Peter's COVID Consolidated Brief for 2 April

Happy April!

...All I can think of is that it's hard to imagine that just three weeks ago the world felt so different than it does today. I am still following COVID-19 a lot, so here's my second semi-regular installment of a public consolidated brief that tries to consolidate everything I read into one short, actionable list so other people don't have to re-create my work. This way I can save time and [fight research debt](#).

Maybe read this instead of spending a ton of your own time obsessing? (Though do be wary that I am not an expert by any means and may be off in my selection and interpretation.)

I have a lot more reporting that I wanted to put in here but didn't, due to lack of time. I will get them in the next issue and I will try to send out updates as fast as I can while maintaining a certain level of quality.

I do hope news will slow down at some point. ...I certainly will slow down at some point.

Previously:

- [29 Mar Brief](#)
- [My research questions](#) (27 Mar)

See also:

- [LessWrong links database](#)
- [EA Coronavirus Facebook Group](#)

Doing Your Part! How You Can Stay Safe and Help the Fight!

The Wikipedia page [“2019–2020 Coronavirus pandemic”](#) is currently the second most viewed Wikipedia page of all pages this week, with hundreds of thousands of page views. If you like reading links, it could be really helpful to add a few minutes to your day to join me in this fight and keep this and the associated pages up to date. Also update the [LessWrong links database](#) and the [Coronavirus Tech Handbook](#). And if you see parallel efforts, make sure they are aware of each other.

~

I don't want to wade too much into the Great Masks Debate, which feels too complicated for me to adequately analyze and summarize at the speed at which I am writing this. I am not an expert, but since some are asking, I will nonetheless briefly summarize my provisional opinion:

- There is some evidence that the public should wear masks - even DIY cloth ones. See [“The evidence for everyone wearing masks, explained”](#) and [“Face Masks: Much More Than You Wanted to Know”](#) and [“SSC Journal Club: Macintyre on Face Masks”](#). I am pro-mask and wear a mask when going to the grocery store.
- There is a lot of bad evidence and arguments out there. Do your best to only advocate for masks using *good* arguments, which I do think exist. You should be just as wary of DIY research as DIY masks.
- Let doctors and nurses get masks first. Avoid panic buying masks... if you even can at this point.
- [Consider making your own mask out of cloth or a T-shirt](#).
- Most importantly, still do the best you can to minimize the situations where you would need a mask. Stay inside. Do not let masks lure you into doing anything differently than you otherwise would do without a mask - this “risk compensation” could easily make masks net-negative.

Mask wearing is gaining steam. Austria already [requires residents to wear masks while grocery shopping](#). And now a [German city, just mandated for the first time](#) that people wear masks for shopping and using public transit.

~

Founders Pledge, in partnership with Silicon Valley Bank, have today launched a [Covid-19 Response Fund](#).

~

Are you an engineer? New York is putting together [the New York State COVID-19 Technology SWAT Team](#) and is accepting volunteers.

~

If you can do epidemiological modeling and are not already crazy busy with COVID stuff, maybe you could now heed [The Royal Society's urgent call for epidemiological modeling](#).

~

My previous [29 Mar Brief](#) has a bunch of advice on how to stay safe and contribute.

A Glance at The Latest Situation

Donald Trump [backed off his original plan to re-open the country by Easter](#) (insofar as he actually did have the power to do that) and is extending “social distancing guidelines” until April 30th. He also echoes claims made earlier by Dr. Fauci that around

200,000 people will die from COVID-19 by the end of the year.

While this is a very staggering and sad number, it is worth keeping in mind two things: (1) this number is only the modal part of the estimate... the actual death toll could be a fair bit higher or a fair bit lower and (2) this death estimate is still somewhat within the control of the United States government, should they be able to enact better policies to reduce the spread of COVID-19 and ensure hospitals are adequately resourced. [Now is a good time to act.](#)

~

There certainly are a lot of cases - with 952,171 worldwide cases as of the time of this writing, we look on pace to cross the 1M mark within a day.

Until recently, the most deadly infectious disease was tuberculosis, typically killing ~1.18M people per year, or 3231 people per day. Now COVID-19 has surpassed this daily death count and is the world's new most deadly infectious disease ([Source](#)).

A bit of good news though - while a bit too early to tell, it looks like we're finally starting to see declines in new cases, especially in Italy:



We should get significantly more information in about two weeks or less about the shape of these curves. Projecting the curve forward suggests that things could be relatively stable by June.

Note that due to the log scale, the US is still in for quite a lot of pain.

~

Again, actual numbers being reported can be hard to compare and are complicated by differences in measurement. [NYT reports a new source of discrepancies](#): "In Italy, authorities have conceded that their coronavirus death toll did not include those who had

died at home or in nursing homes. Similarly in France, officials have said that only those who died in hospitals had been recorded as pandemic-related — a practice they said would change in the coming days.”

~

Lastly, all this stuff is just hard! [The Atlantic reports:](#)

We rely on numbers to understand the size and scope of tragedy—to gauge what went wrong and put the damage in perspective. More Americans have now died from the coronavirus than were killed in the September 11 terrorist attacks, multiple news outlets announced yesterday.

But we likely won’t have an estimate of how many Americans have died as a result of the pandemic for a very long time—maybe months, maybe a year. We will almost certainly never know the exact number. “It sounds like it could be totally obvious—just count body bags,” John Mutter, an environmental-science professor at Columbia University who studies the role of natural disasters in human well-being, told me in an interview this week. “It’s not obvious at all.”

When Hurricane Maria flattened Puerto Rico in September 2017, the storm’s devastation was overwhelming. Yet the official death toll in December stood at 64 people—a number that almost no one believed, as my colleague Vann Newkirk II wrote at the time. Nearly a year after the storm, a team of researchers tried to develop their own estimate. They gathered months’ worth of mortality data from households across the island and published a study concluding that, in actuality, more than 4,600 deaths were potentially attributable to the hurricane—70 times the official number.

~

Of course, yet another reason numbers may be hard to compare is just *outright lying*. After a lot of public suspicion, [the US intelligence community is now reported as concluding](#) that “China has concealed the extent of the coronavirus outbreak in its country, under-reporting both total cases and deaths it’s suffered from the disease” and that “China’s public reporting on cases and deaths is intentionally incomplete”.

The US also [formally accuses Iran of lying about their numbers](#), saying, among other things that “[t]he regime is hiding a significant amount of information about the coronavirus outbreak. It is likely far worse that the regime is admitting. This lack of transparency poses a significant health risk to the Iranian people, as well as to Iran’s neighbors.”

If you’re confused by this - so are the US spies! Reuters reports [“U.S. spies find coronavirus spread in China, North Korea, Russia hard to chart”](#).

...So Just How Bad Could This All Get?

[Hungary descends into authoritarianism to fight the coronavirus:](#)

Hungary's parliament has voted to allow Prime Minister Viktor Orban to rule by decree indefinitely, in order to combat the coronavirus pandemic, giving the populist leader extra powers to unilaterally enact a series of sweeping measures.

The bill, which has been criticized by international human rights watchdogs, has no specified end date and allows Orban to bypass a number of democratic institutions in his response to the outbreak.

[This could be a cause for wider concern:](#)

In Hungary, the prime minister can now rule by decree. In Britain, ministers have what a critic called “eye-watering” power to detain people and close borders. Israel’s prime minister has shut down courts and begun an intrusive surveillance of citizens. Chile has sent the military to public squares once occupied by protesters. Bolivia has postponed elections.

As the coronavirus pandemic brings the world to a juddering halt and anxious citizens demand action, leaders across the globe are invoking executive powers and seizing virtually dictatorial authority with scant resistance.

Governments and rights groups agree that these extraordinary times call for extraordinary measures. States need new powers to shut their borders, enforce quarantines and track infected people. Many of these actions are protected under international rules, constitutional lawyers say.

But critics say some governments are using the public health crisis as cover to seize new powers that have little to do with the outbreak, with few safeguards to ensure that their new authority will not be abused.

Gaze into the Crystal - The Latest Modeling and Forecasting

According to the Institute for Health Metrics and Evaluation (you may know them as the folks who put together the Global Burden of Disease report) [the United States is potentially only two weeks away from peak hospital resource use](#) (though this will vary significantly by region) and still faces a shortage of tens of thousands of hospital beds and ventilators.

~

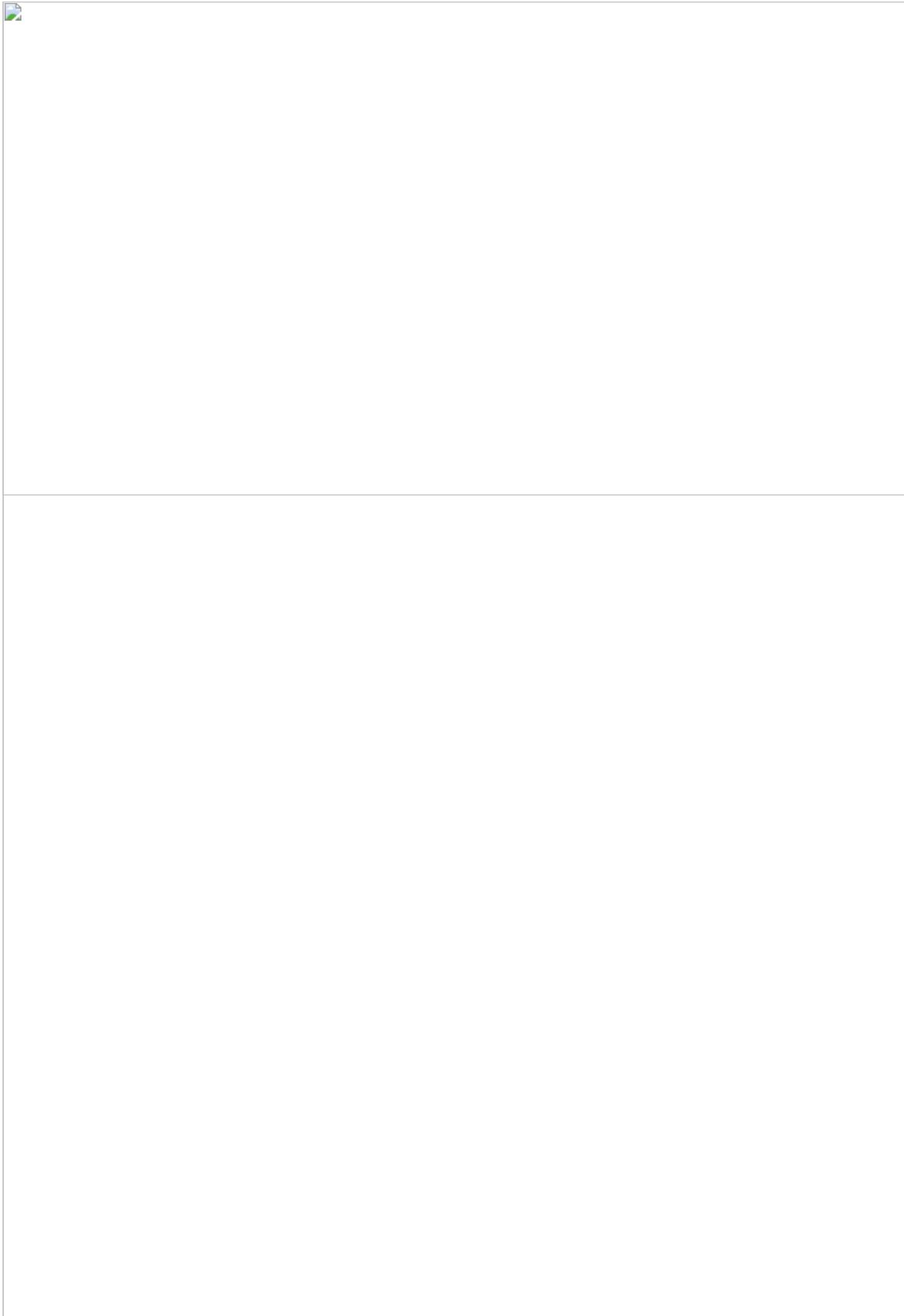
[COVID Act Now now provides county-level models for the US](#), showing time until projected hospital overload.

~

The crowdsourced prediction platform Metaculus has [fancy new graphs](#) displaying their predictions of cases, fatalities, and vaccine timelines:









~

So far experts have actually been underrating the spread of COVID even two weeks in the future... looks like they should've had wider error bars. Even a naive exponential trend extrapolation did much better than experts. Hopefully they will all learn their lesson and predict better next week:



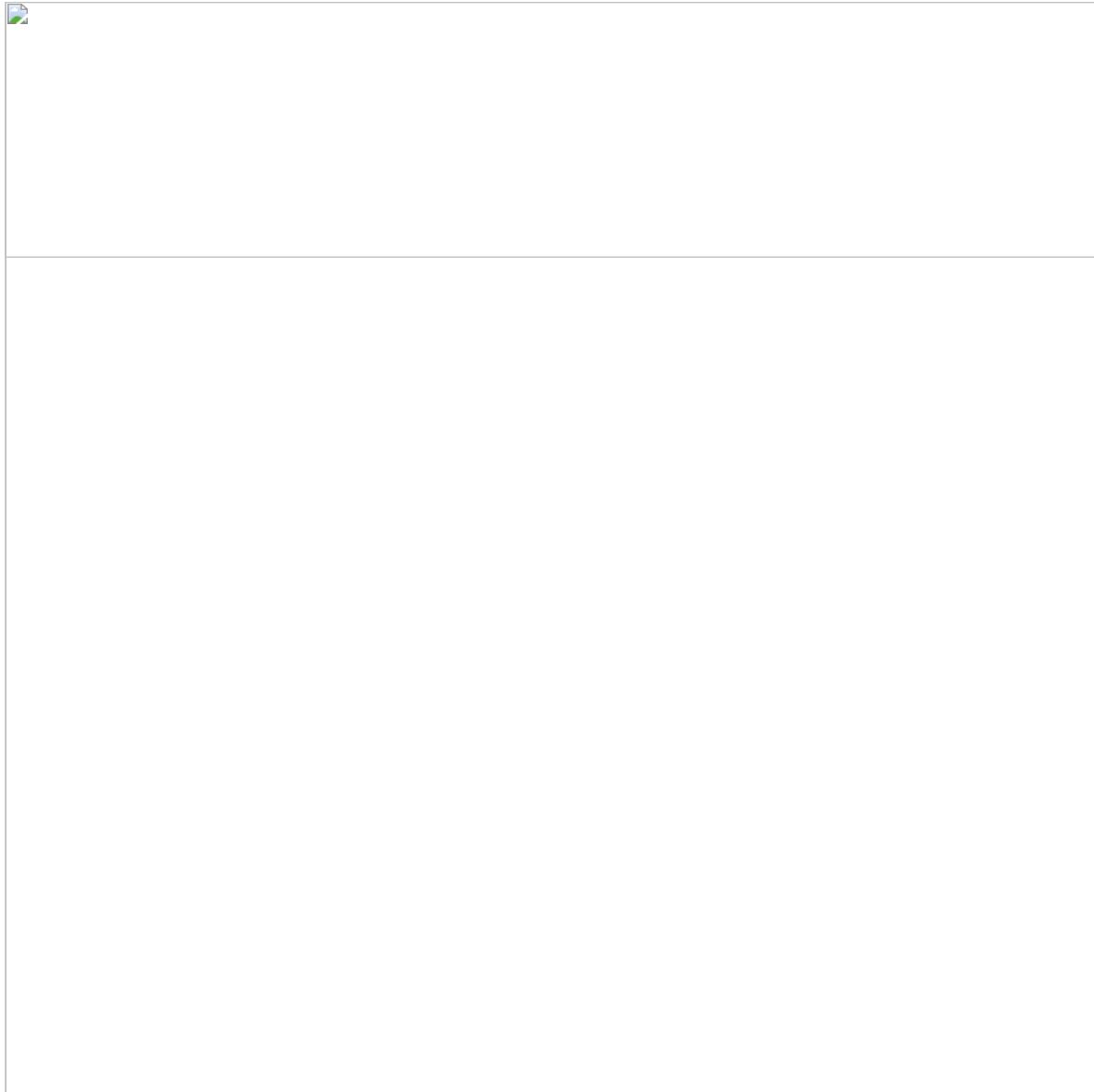
~

Why is it so hard to make a good COVID-19 model? [538 reports](#):

Consider something as basic as data entry. Different countries and regions collect data in different ways. There's no single spreadsheet everyone is filling out that can easily allow us to compare cases and deaths around the world. Even within the United States, doctors say we're underreporting the total number of deaths due to COVID-19.

The same inconsistencies apply to who gets tested. Some countries are giving tests to anyone who wants one. Others are ... not. That affects how much we can know about how many people have actually contracted COVID-19, versus how many people have tested positive.

And the virus itself is an unpredictable contagion, hurting some groups more than others — meaning that local demographics and health care access are going to be big determinants when it comes to the virus' impact on communities



Now Let's Talk Policy Response

Scott Gottlieb, the FDA commissioner from 2017-2019 and a current member of the Pence-led US government coronavirus taskforce, releases "[National Coronavirus Response: A Roadmap for Re-Opening](#)". This is apparently the closest thing we currently have to a national strategic plan for the United States.

The plan operates in three broad phases, plus one bonus phase:

Phase I: Slow the Spread. Current phase of the response, where the US has widespread school closures, work-from-home, close malls and gyms, and limit restaurants. This is intended to stay in place until transmission has measurably slowed down and health infrastructure has scaled up.

Phase II: State-by-State Reopening. Individual states are able to move to Phase II as they are identified to be able to safely diagnose, treat, and isolate COVID-19 cases and contacts. Testing must be scaled up rapidly. These states can gradually reopen schools and businesses, but will likely need to maintain some degree of physical distancing and limitations on larger gatherings. Older adults will also need to remain at home.

The trigger for issuing a stay-at-home advisory in a US state is when case counts are doubling every three to five days (based on the current New York experience) or when state and local officials recommend it based on the local context (for example, growth on track to overwhelm the health system's capacity). The trigger for issuing a recommendation to step down from a stay-at-home-advisory back to "slow the spread" is when the number of new cases reported in a state has declined steadily for 14 days (i.e., one incubation period) and the jurisdiction is able to test everyone seeking care for COVID-19 symptoms.

Phase III: Lifting restrictions. Phase III will also be gradually reached on a state-by-state basis "once a vaccine has been developed, has been tested for safety and efficacy, and receives FDA emergency use authorization" OR "there are other therapeutic options that can be used for preventive or treatment indications and that have a measurable impact on disease activity and can help rescue very sick patients".

Phase IV: Rebuild Readiness for the Next Pandemic. Phase III is basically returning to normal, so that makes Phase IV more of a wishlist for what to do after things are normal to not get in this mess again. The report asks for greater vaccine development capacity to reduce time-to-vaccine below one year, improve the hospital surge capacity to be able to rapidly scale ICUs if needed in the future, improve disease forecasting, and move policy away from decentralized response.

~

[Bill Gates: "Here's how to make up for lost time on COVID-19":](#)

1. consistent nationwide approach to shutting down
2. step up on testing
3. clear priorities for who is tested (prioritize health-care workers and first responders, followed by highly symptomatic people who are most at risk of becoming seriously ill and those who are likely to have been exposed)
4. allocate masks and ventilators nationally, not by having each state compete
5. don't speculate about treatments - run rapid trials involving various candidates and inform the public when the results are in
6. start building the infrastructure to make vaccines now

~

[Lawrence Summers summarizes the key policies:](#) "I think we need to be investing, in a way far beyond what we are, in developing an infrastructure for widespread testing, widespread contact tracing, and widespread separation of those who are sick and those who are most vulnerable"

~

A bit more on China, misreporting, the Principal-Agent Problem, and Goodhart's Law from the New York Times: ["China Created a Fail-Safe System to Track Contagions. It Failed."](#):

The alarm system was ready. Scarred by the SARS epidemic that erupted in 2002, China had created an infectious disease reporting system that officials said was world-class: fast, thorough and, just as important, immune from meddling.

Hospitals could input patients' details into a computer and instantly notify government health authorities in Beijing, where officers are trained to spot and smother contagious outbreaks before they spread.

It didn't work.

After doctors in Wuhan began treating clusters of patients stricken with a mysterious pneumonia in December, the reporting was supposed to have been automatic. Instead, hospitals deferred to local health officials who, over a political aversion to sharing bad news, withheld information about cases from the national reporting system — keeping Beijing in the dark and delaying the response.

China continues to have an issue of people not wanting to report bad news to higher-ups.

~

Another issue with creating good plans is when they aren't followed at all. Politico reports that [Trump team failed to follow the National Security Council's pandemic playbook](#):

The NSC devised the guide — officially called the Playbook for Early Response to High-Consequence Emerging Infectious Disease Threats and Biological Incidents, but known colloquially as "the pandemic playbook" — across 2016. The project was driven by career civil servants as well as political appointees, aware that global leaders had initially fumbled their response to the 2014-2015 spread of Ebola and wanting to be sure that the next response to an epidemic was better handled.

The Trump administration was briefed on the playbook's existence in 2017, said four former officials, but two cautioned that it never went through a full, National Security Council-led interagency process to be approved as Trump administration strategy. [...]

A health department spokesperson also said that the NSC playbook was not part of the current coronavirus strategy. [...]

Trump has claimed that his administration could not have foreseen the coronavirus pandemic, which has spread to all 50 states and more than 180 nations, sickening more than 460,000 people around the world. "Nobody ever expected a thing like this," Trump said in a Fox News interview on Tuesday.

But Trump's aides were told to expect a potential pandemic, ranging from a tabletop exercise that the outgoing Obama administration prepared for the president's incoming aides to a "Crimson Contagion" scenario that health officials undertook just last year and modeled out potential risks of a global infectious disease threat. Trump's deputies also have said that their coronavirus response relies on a federal playbook, specifically referring to a strategy laid out by the Centers for Disease Control.

It is not clear if the administration's failure to follow the NSC playbook was the result of an oversight or a deliberate decision to follow a different course."

Similarly, [the US should have learned important lessons from the Ebola response that it seems to have not](#):

In international crises, policymakers and politicians rarely have a dress rehearsal before their debut on the main stage. Yet in retrospect, the Ebola outbreak of 2013-15 amounts to exactly that—a real-life test of Washington's ability to detect and contain an infectious disease that threatens global security. [...]

It was clear to those who responded to the Ebola outbreak that the response system of the United States and the international response system would risk collapse if faced with a more dire scenario. It was equally clear that a more dire scenario taking place was a question of when, not if. [...]

Even before the Ebola epidemic ended, the U.S. government began pursuing a three-pronged strategy to contain a more dangerous outbreak. First, it doubled down on the Global Health Security Agenda, an initiative the Obama administration launched before the Ebola crisis to expand capabilities around the world to prevent, detect, and rapidly respond to infectious disease threats. [...] The strategy's second prong was to further build out the network of hospitals and testing centers in the United States designated to treat Ebola and to increase the size of the national medical stockpile with more of the personal protective equipment and materials needed to fight highly lethal pathogens. The third prong was to designate a health emergency response coordinator and create a new Directorate for Global Health Security and Biodefense within the National Security Council. [...]

As 2017 turned to 2018 and 2018 turned to 2019, each prong of this strategy fell away like wheels off a bus. When the money provided by the Ebola Response Supplemental ran out, the new administration continued to fund the Global Health Security Agenda. But the overall budget for the Centers for Disease Control was cut, and no robust, new investments were made in greater deployable capability in the United States or other countries. At home, the envisioned expansion of the original 35-hospital Ebola Treatment Network did not take place; the \$259 million appropriated for the network in 2014 was not followed by meaningful infusions of funds, setting it on track to expire in May 2020 and leading the Department of Health and Human Services to warn in November 2017 that "the current capacity of this system is not likely to be sufficient for many types of infectious disease outbreaks (e.g., pandemic influenza and other respiratory pathogens)." Nor was the national medical stockpile significantly bolstered. Congressional leaders passed budgets that had none of the vision or scale of the \$5.4 billion Ebola Response Supplemental.

The third prong of the strategy was the last to go. In his first month as National Security Adviser, John Bolton shuttered the new NSC Directorate for Global Health Security and Biodefense. Its leader departed the NSC staff just one day after the WHO declared a new outbreak of Ebola in the Democratic Republic of the Congo that to date has killed over 5,000 people."

We in the developed world should keep in mind that social distancing is a privilege that will be much harder to maintain in the developing world: "In neighboring Dharavi, Asia's largest slum with a population of 1 million in an area of less than a square mile, workers laughed when I asked them if they had been practicing social distancing. As I stood outside a hovel, trying to keep some distance from people the day after Modi's lockdown order was announced, even talking about social distancing felt obscene near a room of eight people crammed together with barely any space to breathe."

Project Syndicate also reports "As horrific as this sounds, the situation in the advanced economies is likely to be much more benign than what developing countries are facing, not only in terms of the disease burden, but also in terms of the economic devastation they will face. [...] Under these conditions, even if developing countries want to flatten the curve, they will lack the capacity to do so. If people must choose between a 10% chance of dying if they go to work and assured starvation if they stay at home, they are bound to choose work."

Notably, India and Pakistan have taken somewhat divergent approaches - India locked down with just four hours notice, whereas Pakistan has currently eschewed a countrywide lockdown, instead closing large malls and schools but letting individual states and provinces manage the rest.

COVID testing in India and Pakistan still remains too low to tell yet how each strategy is working.

~
IDInsight provides some advice for developing countries:

1. Lockdown as much as is possible
2. Use painted or marked "social distancing squares" to cue people to stay distant from each other for stores that remain open
3. Remind people to use frequent handwashing and cloth masks
4. Avoid outright closing public transit, but close especially risky forms of transit
5. Address issues of urban-rural migration
6. Make sure social distancing policies are localized and contextualized
7. Identify trusted leaders and communicate messages in many local languages



A Bit About Life Under Quarantine

You may have heard contestants on the "Big Brother" TV show were sequestered from news and [were some of the last to learn about the pandemic](#). But they're not the actual last. It turns out that [many submariners are still unaware of the pandemic](#):

Mariners aboard ballistic submarines are habitually spared bad news while underwater to avoid undermining their morale, say current and former officers who served aboard France's nuclear-armed subs. So any crews that left port before the virus spread around the globe are likely being kept in the dark about the extent of the rapidly unfurling crisis by their commanders until their return, they say. [...]

Speaking exclusively to The Associated Press, Salles said he believes submariners will likely only be told of the pandemic as they head back to port, in the final two days of their mission. [...]

"No matter how serious an event is, there is nothing a submariner can do about it. And since he cannot do anything, better that he know nothing," Salles said. "They know that they won't know and accept it. It's part of our deal."

~

[Buzzfeed reports](#): "The coronavirus outbreak has prompted an unprecedented surge in gun sales, exceeding a previous record set after the Sandy Hook mass shooting."

~

["It Felt Like a Black Mirror Episode": The Inside Account of How Bird Laid off 406 People in Two Minutes via a Zoom Webinar](#)

~
[Deals for businesses in the time of COVID-19](#)... now's a great time to get some free software!

If You Still Own Envelopes, Check Their Backs - Here's the Latest Cost-Benefit Analysis

[Another paper finds net benefits to lockdown](#):

We examine the net benefits of social distancing to slow the spread of COVID-19 in the United States. Social distancing saves lives but imposes large costs on society due to reduced economic activity. We use an SIR model to perform a benefit-cost analysis of controlling the COVID-19 outbreak. Assuming that social distancing measures can substantially reduce contacts among individuals, we find net benefits of roughly \$5 trillion in our benchmark scenario. We examine the magnitude of the critical parameters that would lead to negative net benefits. A key unknown factor is the time to economic recovery with and without social distancing measures in place. Our sensitivity analysis points to a need for effective economic stimulus when the outbreak has passed.

~
Also another paper suggests [locking down somewhere between 7 to 34 weeks](#):

We investigate the optimal duration of the COVID-19 suppression policy. We find that absent extensive suppression measures, the economic cost of the virus will total over \$9 trillion, which represents 43% of annual GDP. The optimal duration of the suppression policy crucially depends on the policy's effectiveness in reducing the rate of the virus transmission. We use three different assumptions for the suppression policy effectiveness, measured by the R0 that it can achieve (R0 indicates the number of people an infected person infects on average at the start of the outbreak). Using the assumption that the suppression policy can achieve $R_0 = 1$, we assess that it should be kept in place between 30 and 34 weeks. If suppression can achieve a lower $R_0 = 0.7$, the policy should be in place between 11 and 12 weeks. Finally, for the most optimistic assumption that the suppression policy can achieve an even lower R0 of 0.5, we estimate that it should last between seven and eight weeks. We further show that stopping the suppression policy before six weeks does not produce any meaningful improvements in the pandemic outcome.

Now Just What are the Tech Overlords up to?

[A new COVID-19 High Performance Computing Consortium](#) brings together the US government, IBM, Amazon, Google, Microsoft, Hewlett Packard, MIT, NASA, and others to "provide COVID-19 researchers worldwide with access to the world's most powerful high performance computing resources that can significantly advance the pace of scientific discovery in the fight to stop the virus".

~
[Palantir provides COVID-19 tracking software to CDC and NHS..pitches European health agencies](#)... but it will probably be ok?

Palantir, a secretive government-friendly big data operation that's able to ingest vast amounts of information to visualize trends and track individuals — useful tasks as the spread of COVID-19 threatens to overwhelm healthcare systems and ravage economies.

In mid-March, The Wall Street Journal reported that Palantir was working with the CDC to model the potential spread of the virus. Forbes reports that CDC staffers are now regularly using Palantir's web app to visualize the spread of the virus and to anticipate hospital needs. According to that report, Palantir is eschewing dealing with sensitive personally identifying information in its coronavirus efforts, instead providing analysis of anonymized hospital and healthcare data, lab results and equipment supplies through a platform called Palantir Foundry.

[...] Likely aware of its reputation as the shadowy tech giant that helps to power ICE's deportation machine, Palantir is apparently acknowledging the privacy implications of its new work. In a statement provided to The Wall Street Journal, Palantir's privacy lead Courtney Bowman asserted that privacy and civil liberty must be taken as "guiding concentrations" in any data-driven COVID-19 response, "not as afterthoughts."

While it appears to be taking on a new role with the U.S. COVID-19 response, Palantir has worked with the U.S. federal government on infectious health threats for years. In 2010, the CDC used Palantir to monitor an outbreak of cholera in Haiti."

~
[Amazon begins running temperature checks and will provide surgical masks at warehouses](#).

And How Do We Get Out of this Mess? Vaccines, Treatments, Testing, Tracing, etc.

[Johnson & Johnson announces](#) a vaccine that could be potentially available by early 2021:

J&J today announced the selection of a lead COVID-19 vaccine candidate from constructs it has been working on since January 2020; the significant expansion of the existing partnership between the Janssen Pharmaceutical Companies of Johnson & Johnson and the Biomedical Advanced Research and Development Authority (BARDA); and the rapid scaling of the Company's manufacturing capacity with the goal of providing global supply of more than one billion doses of a vaccine. The Company expects to initiate human clinical studies of its lead vaccine candidate at the latest by September 2020 and anticipates the first batches of a COVID-19 vaccine could be available for emergency use authorization in early 2021, a substantially accelerated timeframe in comparison to the typical vaccine development process.

J&J has also committed to produce more than a billion doses for global use, including production in both the United States and overseas.

~

[FDA Emergency Use Authorization is granted](#) to a new system for decontaminating N95 respirators for re-use. Innovations in this area could help with the mask shortage.

~

Does hydroxychloroquine help? [From Derek Lowe](#): there are now two blinded, randomized, and controlled trials from china. Unfortunately, "these two came out rather differently, with the Zhejiang study showing no detectable difference on treatment and the Wuhan one showing what looks like a real effect[...] which one (if either) reflects the real-world situation?"

[538 also urges some patience and caution:](#)

No drugs currently on the market have previously gone through control trials to treat COVID-19 specifically. Because of this, we can't be certain how effective or safe they would be. And even when a treatment seems promising, it may not end up being effective: Lopinavir-ritonavir, a combination of anti-HIV drugs, was considered a possible treatment for COVID-19, but a clinical study published March 18 showed the pair had no substantial benefit on patients.

Other clinical trials for existing medications, such as remdesivir, have already begun, including one sponsored by the National Institutes of Health and led by Kalil. It will take weeks, and possibly months, for the trials to be completed, and there's a chance that none of the drugs being investigated will effectively treat COVID-19. But it's the only hope we have of figuring out whether these drugs actually work and are safe to use.

The Non-Profit Impacts

Run a US 501c3 non-profit organization with less than 500 employees? Want grant money from the US government if you don't lay off employees? Reach out to your retail banking partner and tell them "I want to apply for the CARES ACT SBA LOAN 7(a)". [Details here](#). Act quickly!

Don't Forget About the Nonhumans!

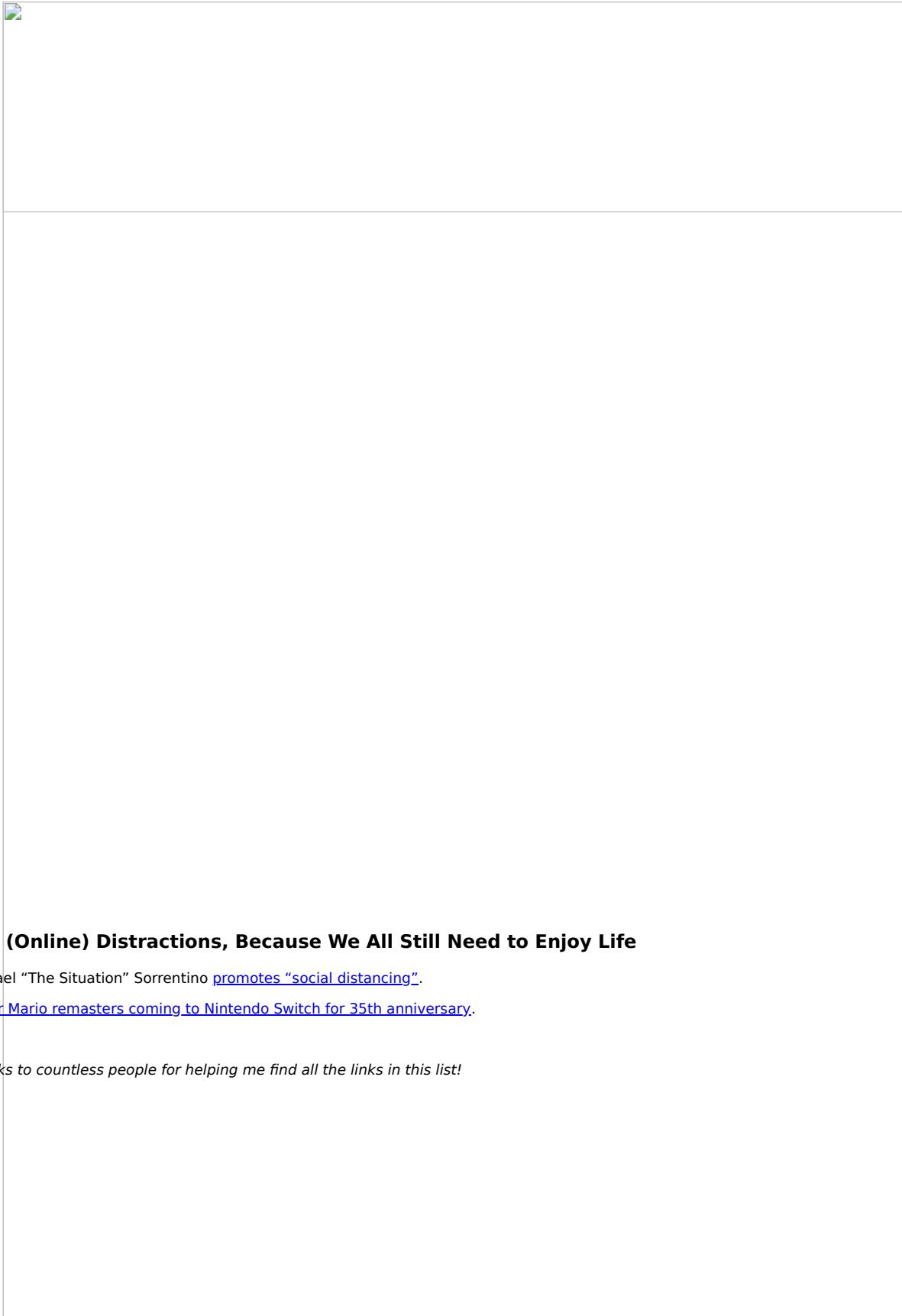
Cynthia Schuck and Wladimir Alonso, two public health experts who are also passionate about the plight of farm animals, have [prepared a short e-book to provide factual information on the connection between farm animals and pandemics](#).

Your Regular Dose of WTF

Guess who else is quarantined?



~



Fun (Online) Distractions, Because We All Still Need to Enjoy Life

Michael "The Situation" Sorrentino [promotes "social distancing"](#).

[Super Mario remasters coming to Nintendo Switch for 35th anniversary.](#)

-
Thanks to countless people for helping me find all the links in this list!

Treatments correlated with harm

There's a billion different ways studies can go wrong through failure of statistics to correctly capture reality.

One very specific one is: your successful treatments will often look like they're making things worse.

This is merely a special case of confounding, but it comes up often enough I want to highlight it. And it seems specifically important for the discussion of masks and respirators.

For example (not a doctor): every new leukemia drug has a survival rate of like 30%. Base rate survival is like 60%. Does every new leukemia drug work worse than normal? Obviously not, it's that the patients trying the new drug are the ones for whom everything else has failed, so they've been signed up for a clinical trial. These patients have very resistant cancers. And we can verify this, because when the drug does get adopted then survival rates go up.

Doctors have mostly figured this out with this subset of leukemia drugs. They do adjustments for how resistant the crazy-resistant cancers in the trial are, the adjustments aren't good enough and show the drug has no effect or is bad, they throw the adjustments out and do something else (like compare the results of this new drug to some other new drug that's almost like a control group), and they eventually seem to come to decent conclusions about whether to use the new drug in a setting that's more conducive to sane trial results.

However, in a lot of cases, they don't seem to figure this out, and it results in truly terrible treatment policies. The clearest example I've seen is in a few localized forms of cancer (like various sarcomas), where amputation shows vastly worse survival than local resection. As crazy as this seemed, it was minutely possible that side effects of amputation were in fact so bad that this was true, but after looking into it extensively I am very sure that in most extremity amputations it is not true. Doctors who see more aggressive tumors advocate amputation, these more aggressive tumors have worse overall survival regardless of how you treat, and amputation winds up looking statistically worse than local resection. And now the researchers who looked at this are advocating for fewer amputations, which will cost lives.

I've seen this in other cases too. The most recent: in patients with COPD, taking inhaled corticosteroids (ICS) [is associated with](#) higher pneumonia rates by a factor of 1.5 or something. This makes sense mechanistically, because corticosteroids do reduce immune function. But pneumonia rates are [~8x higher](#) in people with COPD! One of the first hypotheses that should jump out here is that people with mild COPD take ICS less than those with severe COPD, and ICS helps those with severe COPD have pneumonia less but not enough to make up the full difference! I don't know, maybe I'm the one missing something and people with severe COPD always use bronchodilators or something other than more ICS. But you at least have to address this in the study! And my hypothesis is supported by the fact that in asthma patients, ICS reduces pneumonia by a [factor of ~2](#).

So, the pattern is that disease A has treatment X. Those with worse disease A have worse outcomes by, say, a factor of 3. Treatment X improves outcomes by a factor of

2 and is given preferentially or in higher doses to those with worse disease A because of cost or side effects of unclear size. An observational study comes along and sees that treatment X is associated with 1.5x worse outcomes than not-X! Or maybe just that outcomes are about the same. They denounce X as showing "no appreciable benefit" or "signs of harm".

You won't always see this pattern. With randomization, you won't see it. If the treatment doesn't change in dose or type with increasing disease burden, you won't see it. Or if the beneficial effect of the treatment significantly outpaces the harmful effects of increased disease. The pattern is worst in cases where an observational study looks at a disease with quickly-scaling outcomes and a treatment that is partial or underpowered (cf cancer and COVID).

There's some old joke about how taking supplement X is associated with low levels of X. It's the same thing, just slightly more insidious.

Anyways, I'm concerned this pattern is showing up with the COVID discussions of masks and respirators. Both are things you'd common-sensically assume were helpful. Both show some weird nebulous signs of making things worse. (Certainly ventilators cause very serious side effects, but recently I saw some more serious claims about not boosting survival at all and thus being obviously net-costly. But of course, common sense says they have to be boosting survival some!) But if this is the news you'd expect to hear even in the world where there was no harm being done, the fact that you hear it is not much evidence that there's in fact harm being done.

I feel irresponsible for posting this without doing too much investigation of the masks and ventilators, because plausibly this is pointing in the wrong policy direction on them, but I don't have time for that at the moment. But in the meantime, I would like if every excited contrarian buzz about a treatment showing counterintuitive harm could be accompanied by SOME statement addressing the fact that you'd expect to see those results regardless of harm.

ETA: [Some evidence](#) SARS-CoV-2 attacks hemoglobin, which could inhibit gas exchange enough to cause oxygen poisoning in the lungs if oxygen concentrators or ventilators are used. On the other hand, [a response on chemrxiv](#) says the paper is terrible, and I haven't looked into it. None of this changes my feelings toward how people should be reacting to or talking about the observation that tons of people on ventilators die.

Peter's COVID Consolidated Brief - 29 Apr

It's been almost a month since my [last COVID Consolidated Brief](#) and I hope you are all doing ok. I've personally been settling into the new normal. On the other hand, I've witnessed first hand some of the risks that might be coming with COVID. I had to take shelter in a tornado warning for the first time in my life. While the tornado and the destruction were luckily quite minimal, there was a power failure for about a day and social distancing made it a lot harder to wait out the power failure in a nearby library or Starbucks. Overall, I'm lucky my life is so safe that this is the biggest problem, but I am worried about people who might be a lot less likely and face strong hurricanes or wildfires while also having to maintain social distancing. More on this in a bit.

If you're just joining us, I follow COVID-19 a lot and this is my third semi-regular installment of a public consolidated brief that tries to consolidate everything I read into one short, actionable list so other people can stay up to date without reading a ton on their own. For this issue, I spent over 25 additional hours trying to get to the bottom of everything so that you don't have to. This way I can save time and [fight research debt](#) and save you time from having to read all of this yourself. That being said, do keep in mind that I am not an expert and I have not been able to cover everything going on - I had to be fairly selective to make this brief actually somewhat brief.

I'm not sure how often I will do these, but I still intend to do them as I am able. Maybe it will be a monthly newsletter. Maybe I'll be able to do it every other week. We'll see!

Previously:

- [2 April Brief](#)
- [29 March Brief](#)
- [My research questions](#) (27 March)

See also:

- [LessWrong links database](#)
- [EA Coronavirus Facebook Group](#)

Doing Your Part! How You Can Stay Safe and Help the Fight!

Masks

Masks are a good idea - tell your friends! The opinion on masks has changed a lot since I last reported about a month ago.

[WHO is now onboard](#): "The World Health Organisation says it supports government initiatives that require or encourage the public wearing of masks, marking a major shift from previous advice amid the Covid-19 pandemic. [...] The WHO added that surgical masks should be reserved for medical professionals, while the public should use mainly cloth or home-made face coverings."

[The American CDC is now onboard](#): "CDC recommends wearing cloth face coverings in public settings where other social distancing measures are difficult to maintain (e.g., grocery stores and pharmacies), especially in areas of significant community-based transmission."

A lot more information on masks is now available in Thomas Pueyo's ["Coronavirus: The Basic Dance Steps Everybody Can Follow"](#).

Masks may even become mandatory as a part of the reopening plans - see more below.

Giving

Want to help give money to help people most affected by COVID-19 have a chance to get back on their feet? [GiveDirectly is now helping you give cash directly to those in most need.](#)

Want to give money to helping analyze and treat COVID? I would donate to the Center for Health Security at Johns Hopkins, which researches biosecurity and has been tracking COVID since early January. You can donate [here](#), or you can donate through [the Effective Altruism Funds](#).

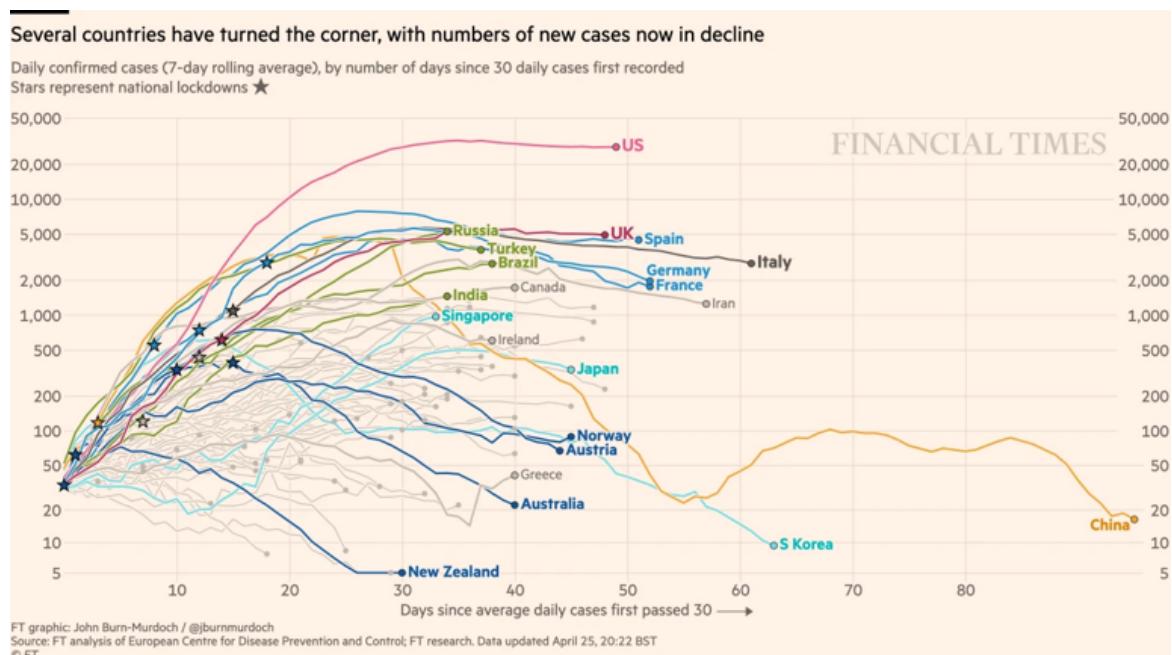
~

My previous [29 Mar Brief](#) has a bunch of advice on how to stay safe and contribute.

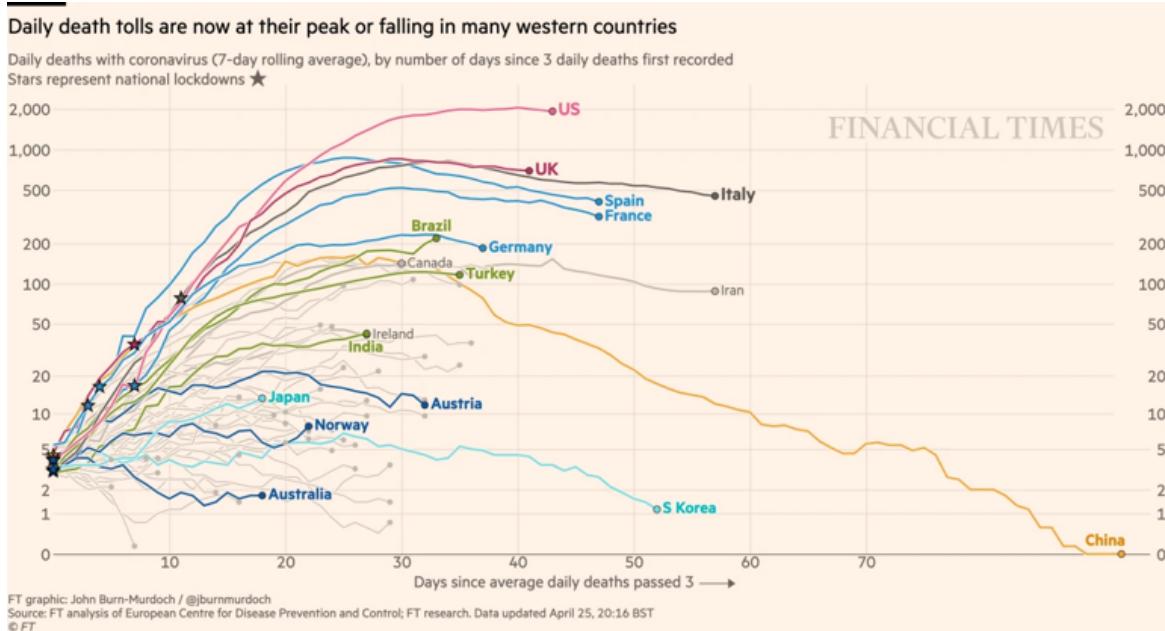
A Glance at The Latest Situation

The last time I wrote, we were just about reaching the one millionth case worldwide. Now we're above one million cases *just in the US* and we've exceeded three million cases worldwide.

Here's the latest based on cases...



and deaths...



It seems that the UK has now likely peaked and is coming down, whereas it is still too early to tell for the US. However, the descending part of the curve seems to be much slower than the sharp ascending part of the curve. It seems increasingly likely (and is even now acknowledged by the IHME) that we won't get a steep, bell-curve-type "mirror image" decline in cases / deaths as previously hoped for.

It looks like Austria, Australia, New Zealand, and Norway are joining Taiwan, South Korea, and Hong Kong as potential case studies in successful handling of COVID. On the other hand, previous darlings Japan and Singapore now look to each be facing a moderate outbreak.

~

Indeed, in good news, [New Zealand has already declared victory](#):

"There is no widespread undetected community transmission in New Zealand. We have won that battle," Ardern said Monday. "But we must remain vigilant if we are to keep it that way."

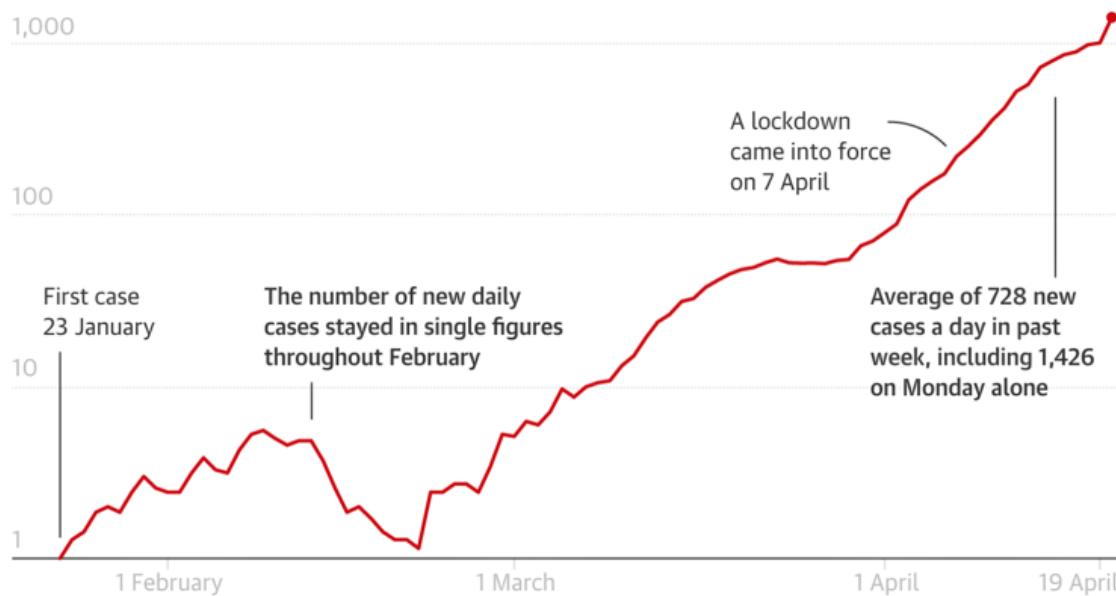
Asked whether New Zealand had eliminated COVID-19, Ardern replied: "currently."

~

Though Singapore's previous victory is now revoked and shows security may not be as absolute as it seems. They previously had ~20 cases per day in mid-March but jumped to ~1000/day in mid-April, declared a lockdown on 7 April and added additional measures on 21 April, and have seen cases drop to ~500/day.

Covid-19 cases have surged among Singapore's migrant workers

Rolling seven-day average of new cases since day of the first case, using a log scale



Guardian graphic. Source: Guardian analysis of Johns Hopkins CSSE data. Note: the CSSE states that its numbers rely on publicly available data from multiple sources, which do not always agree. Data to 20 April 2020

~

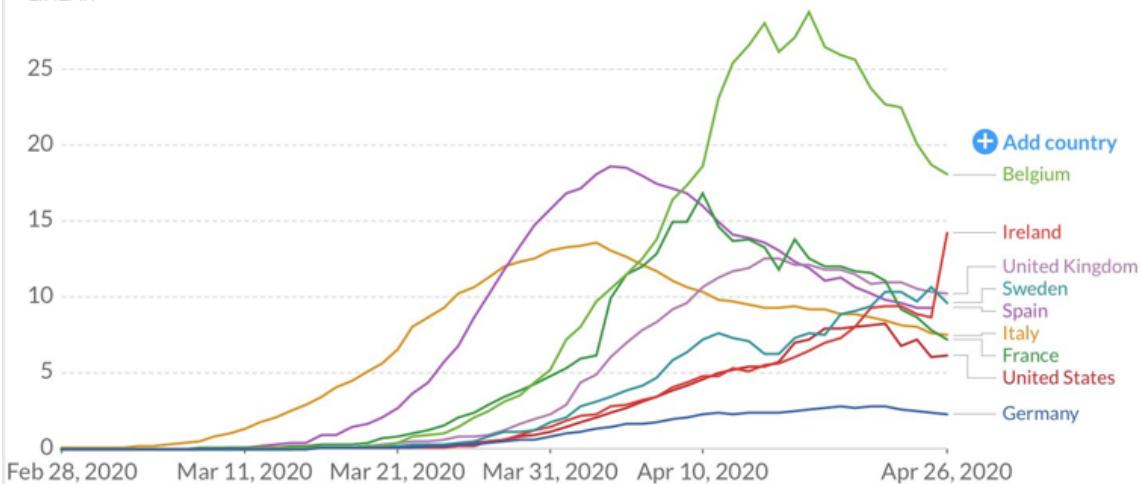
Here's another way to look at the numbers, which looks at new deaths per capita, averaged over the past week to smooth out some (but not all) reporting issues... makes you wonder what the heck is happening in Belgium and Ireland (if it is anything more than differences in how countries report death stats).

Daily confirmed COVID-19 deaths per million, 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.



LINEAR



Source: European CDC – Situation Update Worldwide – Last updated 26th April, 10:45 (London time)
OurWorldInData.org/coronavirus • CC BY

~

Reporting deaths is hard. Looking to “excess deaths”, [the death toll could be even worse than currently imagined](#):

The death toll from coronavirus may be almost 60 per cent higher than reported in official counts, according to an FT analysis of overall fatalities during the pandemic in 14 countries.

Mortality statistics show 122,000 deaths in excess of normal levels across these locations, considerably higher than the 77,000 official Covid-19 deaths reported for the same places and time periods.

If the same level of under-reporting observed in these countries was happening worldwide, the global Covid-19 death toll would rise from the current official total of 201,000 to as high as 318,000.

~

Also, [Business Insider reports](#) that latest forensics show that people died from COVID in the US earlier than we thought:

[A]utopsy results this week revealed that COVID-19 killed two people in Santa Clara County on February 6 and 17. That's at least three weeks earlier than the coronavirus death California officials previously considered the state's first. Indiana is also attempting to trace cases back to mid-February. The state reported its first death on March 16, but officials revised that date to March 10 earlier this month, according to the Indianapolis Star.

~

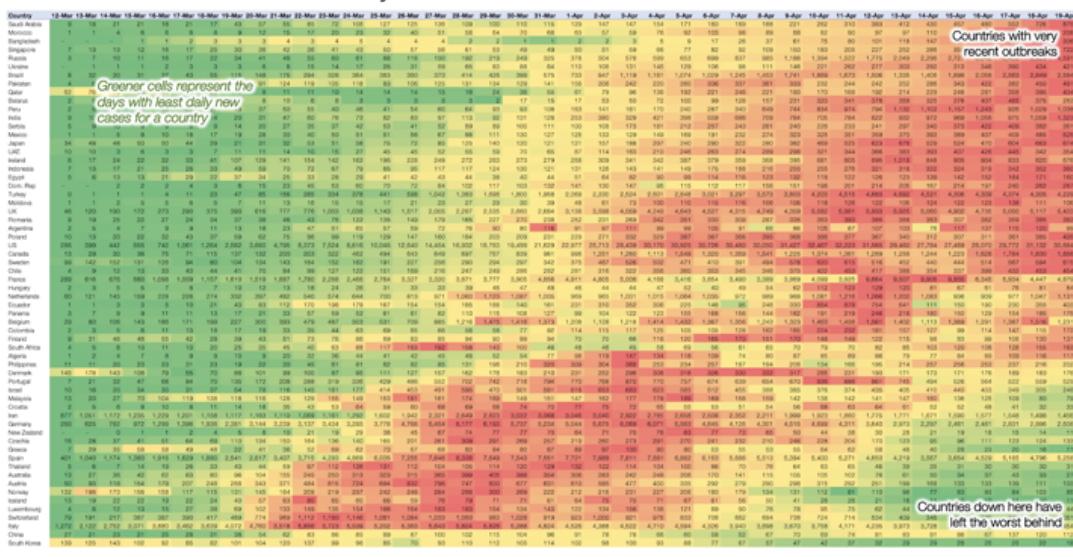
As a further sign of how difficult all this death-counting is - [“NYC death toll jumps by 3,700 after uncounted fatalities are added”](#):

Previously, the city had not counted people who died at home without getting tested for the coronavirus, or who died in nursing homes or at hospitals, but did not have a confirmed positive test result. Mayor Bill de Blasio admitted last week that the true number of deaths was far higher than the official tally, and said the city would start including presumed coronavirus cases in its data.

~

In “Coronavirus: Learning How to Dance”, Tomas Pueyo also breaks trends down on a country-by-country basis (numbers shaded based on relative size of new cases *within* that country):

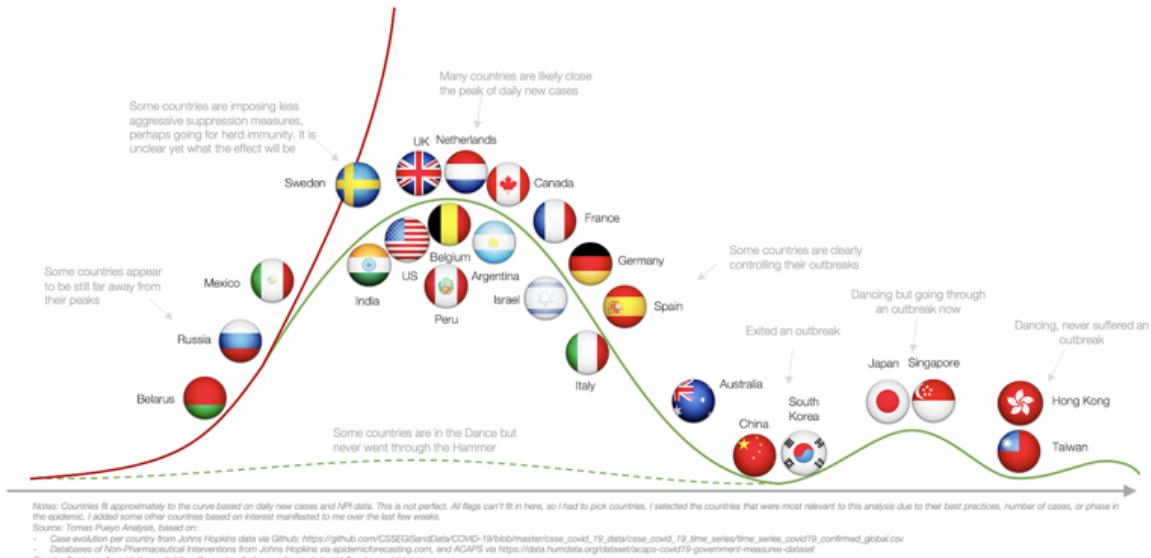
Chart 3: Daily New Coronavirus Cases Worldwide*



(see bigger)

It's worth taking a look at the bigger image, which shows that while outbreaks may be steady in more developed countries, they're just beginning to start taking hold in less developed countries. Here, we can see a lot of countries are just starting to see their outbreaks, while other countries have peaked and declined. From here we can visualize where all the countries are on a curve from handling the virus:

Chart 4: Approximation of Countries along the Hammer and the Dance Phases



([see bigger](#))

~

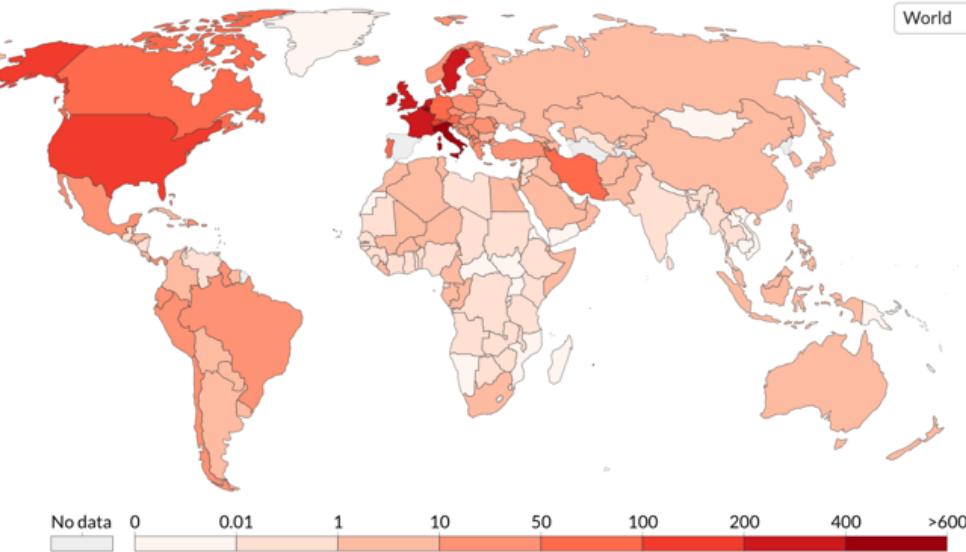
Here's a world map of current COVID deaths per person:

Total confirmed COVID-19 deaths per million people, Apr 29, 2020

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 total number of deaths from COVID-19.

Our World
in Data

World



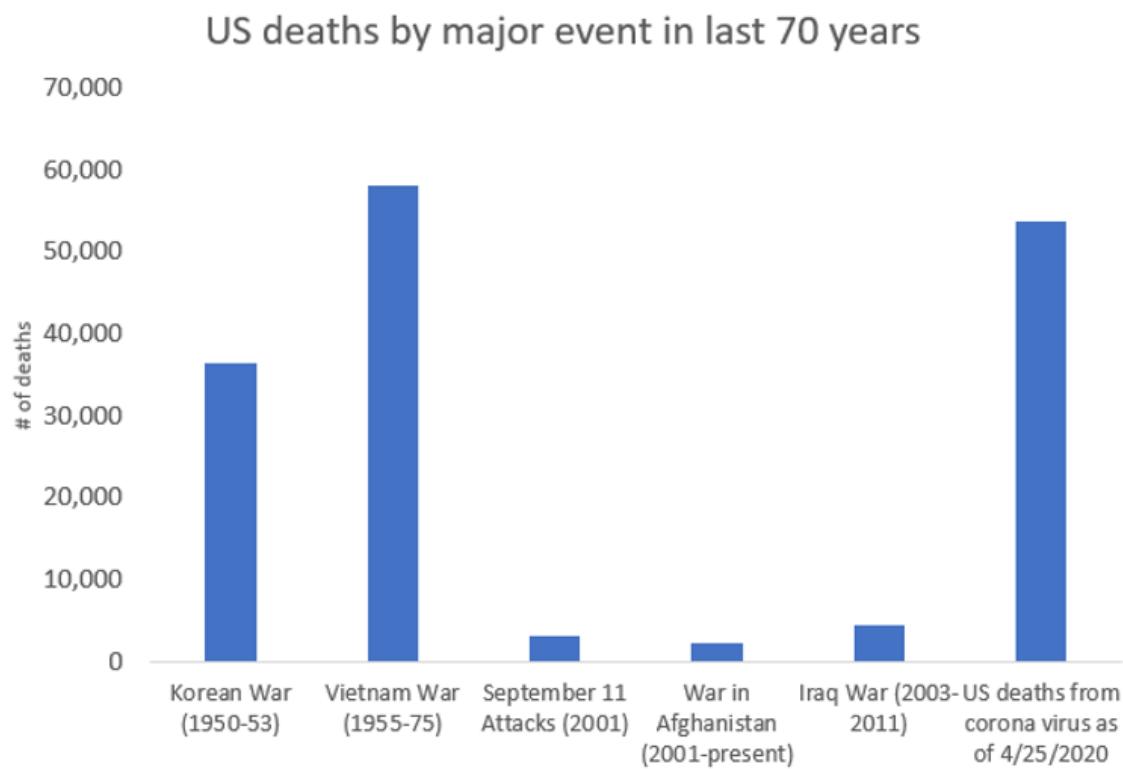
Source: European CDC – Situation Update Worldwide – Last updated 29th April, 11:30 (London time)

CC BY

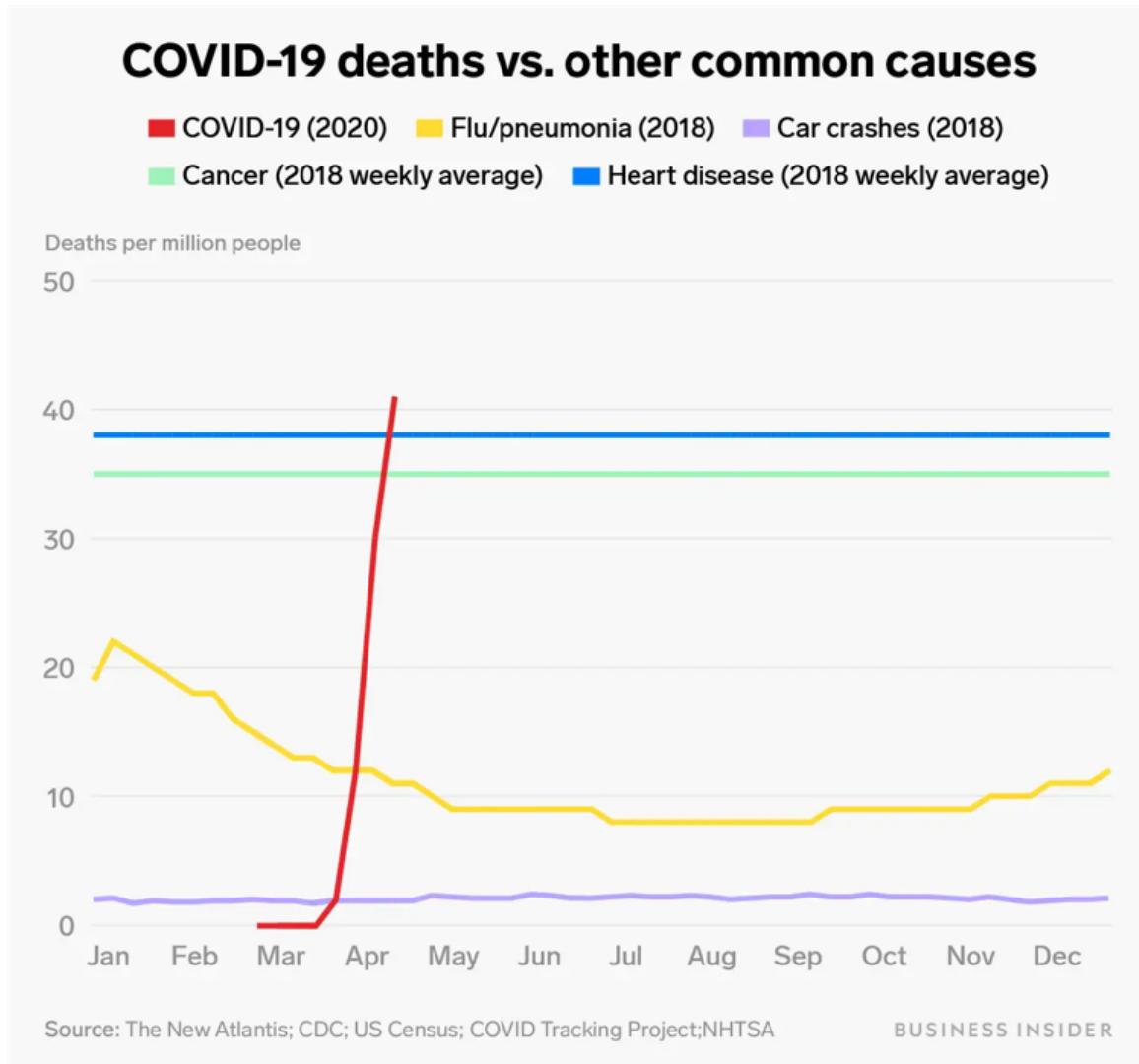
~

It's easy to get lost in the numbers, but the amount of death we've seen so far is a *lot* - the amount of deaths in the US over the past few months far exceeds American deaths in

September 11th and the Iraq War and recently also just exceeded the total number of American deaths in the twenty years of the Vietnam War ([source](#)):



And [COVID has become the leading cause of death in the US](#):



...So Just How Bad Could This All Get?

Just as we might start getting good news about coronavirus, some other disasters might make it worse. For example, Colorado State University [just issued a hurricane season forecast that doesn't look good](#):

As the world battles the coronavirus crisis, researchers are warning of a potentially active Atlantic Ocean hurricane season, which kicks off June 1 through the end of November.
[...]

Specifically, the team forecasts 16 named tropical systems; 12 is the average. Eight of those named systems are forecast to reach hurricane status, with winds greater than 74 mph; Six is the usual amount per year. CSU is also forecasting more major hurricanes than is typical per year: four as opposed to the average of 2.7.

[...] The forecast is also alarming in that it's calling for a nearly 70% chance of a major hurricane — which is at least a Category 3 storm with sustained wind speeds of 111 to 129 mph — makes landfall somewhere along the U.S. coast.

Needless to say, a hurricane like Katrina, Harvey, or Maria is devastating enough on its own, but would combine with already overloaded hospital systems in a very bad way. And as if hurricanes aren't bad enough, [there will almost certainly be large-scale wildfires again as there were last year.](#)

Also, needless to say, social distancing may be difficult during evacuations and this could lead to new outbreaks facing an even more overburdened hospital system.

~

Many conflicts could be potential time bombs amid the coronavirus. I'm still particularly worried China may take the opportunity to be more aggressive in the South China Sea, especially after [China rammed a Vietnamese fishing boat at the beginning of the month.](#) China has also been [bullying Taiwan with military flybys.](#) The [US has already called for \\$20B in new military spending to deter China.](#)

Meanwhile, [Politico reports](#) that the US State Department has continued to warn that Russia, China, and Iran are pushing a host of matching propaganda disinformation messages, including that the coronavirus was an American bioweapon.

Americans are increasingly negative on China. A [new Pew Research survey finds](#) 66 percent of Americans say they hold an unfavorable view of China, up 6 points from the previous year. And of course, recent ads from the [Trump camp referring to Biden as "Beijing Biden"](#) and [Biden retaliating by tying Trump to China](#) suggest this may only get worse.

While I am all for due criticism of the Chinese government and think there is a lot to rightfully criticize, I do worry that declining American sentiment toward China could risk us closer toward armed conflict.

~

To make matters worse, the [World Food Programme warns of "multiple famines of biblical proportions"](#):

David Beasley, head of the World Food Programme (WFP), said urgent action was needed to avoid a catastrophe. A report estimates that the number suffering from hunger could go from 135 million to more than 250 million.

Those most at risk are in 10 countries affected by conflict, economic crisis and climate change, the WFP says. The fourth annual Global Report on Food Crises highlights Yemen, the Democratic Republic of the Congo, Afghanistan, Venezuela, Ethiopia, South Sudan, Sudan, Syria, Nigeria and Haiti.

This is really bad.

~

Another question - less important but still of interest - will those famines happen in the developed world too?

On 27 March, [I wrote](#) that [Metaculus had put](#) the probability of a food shortage in a major US, UK, or EU metropolitan area before 6 June at 30%. I thought this was insanely high and offered my own prediction of 5%.

I see Metaculus has now come to their senses and now has a median prediction of 6%. However, now things might be starting to feel a bit different. It's hard to get more blaring than [Tyson Foods takes out full-page ad: 'The food supply chain is breaking'](#). This seems to mainly be limited to meat production in slaughterhouses, of which a lot have had to be closed or limited due to spread of COVID among slaughterhouse workers.

[The Wall Street Journal reports:](#)

U.S. grocers are struggling to secure meat, looking for new suppliers and selling different cuts, as the coronavirus pandemic cuts into domestic production and raises fears of shortages.

Covid-19 outbreaks among employees have closed about a dozen U.S. meatpacking facilities this month, including three Tyson Foods Inc. TSN -0.16% plants this week. Other plants have slowed production as workers stay home for various reasons.

[NBC reports:](#)

Tyson isn't the only company to shutter plants. Smithfield Foods, the world's largest pork processor, has temporarily closed plants across the country as some of its workers have become sickened. The company's CEO also warned of supply chain issues when it closed its Sioux Falls, South Dakota, plant after more than 400 employees tested positive for the coronavirus. Smithfield says the Sioux Falls plant is one of the largest pork processing facilities in the U.S.

Still, a meat shortage doesn't mean a food shortage writ large, and 6 June is now just a little over a month away and I expect we have more than a month of slack in food production, so overall I continue to put very low odds on this question (personally dropping now to 3%).

~

[Covid-19 May Worsen the Antibiotic Resistance Crisis](#): A bit of a double-whammy where patients ill with COVID are prescribed antibiotics to protect against other possible infections, thus increasing antibiotics use and thus increasing potential antibiotic resistance... at the same time antibiotics manufacturers are slowing the development of new antibiotics to focus on anti-COVID treatments instead.

And How Do We Get Out of this Mess? Vaccines, Treatments, Testing, Tracing, etc.

The Grand Reopening

Getting out of this mess is top of mind as the biggest news of the moment seems to be how we handle re-opening.

The United States

[NBC reports that according to US Vice President Pence](#) sixteen US states have unveiled "formal reopening plans" to lift coronavirus restrictions.

It's worth reiterating the ["National Coronavirus Response: A Roadmap for Re-Opening"](#) plan from the American Enterprise Institute that [I last covered in my previous brief](#) as it is now basically [the official plan endorsed by Donald Trump](#). As I mentioned before, this proceeds in three broad phases:

Phase I: Slow the Spread. Widespread school closures, work-from-home, close malls and gyms, and limit restaurants.

Phase II: State-by-State Reopening. Individual states are able to move to Phase II as they are identified to be able to safely diagnose, treat, and isolate COVID-19 cases and contacts. Testing must be scaled up rapidly. These states can gradually reopen schools and

businesses, but will likely need to maintain some degree of physical distancing and limitations on larger gatherings. Older adults will also need to remain at home.

The trigger for issuing a stay-at-home advisory in a US state is when case counts are doubling every three to five days (based on the current New York experience) or when state and local officials recommend it based on the local context (for example, growth on track to overwhelm the health system's capacity). The trigger for issuing a recommendation to step down from a stay-at-home-advisory back to "slow the spread" is when the number of new cases reported in a state has declined steadily for 14 days (i.e., one incubation period) and the jurisdiction is able to test everyone seeking care for COVID-19 symptoms.

Phase III: Lifting restrictions. Phase III will also be gradually reached on a state-by-state basis "once a vaccine has been developed, has been tested for safety and efficacy, and receives FDA emergency use authorization" OR "there are other therapeutic options that can be used for preventive or treatment indications and that have a measurable impact on disease activity and can help rescue very sick patients".

~

The point of these lockdowns has been to buy us time to (1) build up hospital capacity to handle a larger future wave and (2) build up testing/tracing/isolation capacity to be able to more finely quarantine just those with COVID. Insofar as we're accomplishing (1) and (2), we can start reopening the economy.

My current guess about re-opening is that there won't be a single binary "everything back to normal" event like it seems people are conceptualizing it. **People think of "reopening" as "everything goes back to how it was before the coronavirus", but that seems quite unlikely.**

Instead, **I expect the country to gradually reopen, largely in reverse of the way that it closed, except a lot slower.** That is, the things that were closed last (e.g., parks) will be reopened first - potentially really soon - whereas the things that we closed first (e.g., concerts, conferences) will be reopened last - and potentially not until after we have a widespread vaccine.

For a look at what the very first step might be, look to [the new "Safer at Home" policy of Colorado](#) that basically reopens private gatherings with less than ten people, one-to-one real estate home showings, curbside pickup, and not much else:

SAFER AT HOME

WHAT IT MEANS FOR YOU

WHAT IT IS:

- Continuing to stay at home as much as possible and if you leave, do it for very specific tasks
- For older adults and/or have a chronic condition, you **MUST** stay at home unless necessary
- Continuing to wear a facial covering and practice social distancing - 6 feet
- Recreating close to your home - no more than 10 miles
- Continuing to limit interactions to members of your household
- Gatherings no more than 10 people

WHAT IT'S NOT

- A free-for-all
- An opportunity to leave the house as much as possible and spread the virus to others
- An excuse to not wear a facial covering or hug or give a handshake
- Going to the mountains to spend the weekend
- Conducting unnecessary travel
- Having parties or get togethers
- Pick up soccer games or neighborhood BBQ's

#DoingMyPartCO



GOVERNOR
JARED POLIS

SAFER AT HOME

STARTING TODAY: APRIL 27



- **Retail:** Curbside pick-up and delivery can begin
- **Real Estate:** In-person showings can begin (no open houses)

*****UNLESS you live in a jurisdiction that is still under Stay at Home*****

#DoingMyPartCO



GOVERNOR
JARED POLIS

~

Some governors and mayors have been a bit more gung ho about reopening, however.

The semi-serious [Georgia is permitting a reopening](#) of “gyms, fitness centers, bowling alleys, body art studios, barbers, cosmetologists, hair designers, nail care artists, estheticians” with “screening workers for fever and respiratory illness, enhancing workplace sanitation, wearing masks & gloves if appropriate, separating workspaces by six feet, teleworking if possible & implementing staggered shifts”.

Even Trump seems to think this is too much, saying “I disagree strongly with [Governor Kemp’s] decision to open certain facilities which are in violation of the Phase One guidelines [...] I think spas, and beauty salons, and tattoo parlors, and barber shops, in Phase One ... is just too soon. ... They can wait a little bit longer.”

~

[Tennessee is also reopening restaurants and some other businesses](#):

The new restrictions include limiting capacity to 50% and ensuring tables are no less than 6 feet apart, with no more than half a dozen people per table. They also say businesses should screen both employees and customers for signs of illness.

Bar areas will remain closed, "live music should not be permitted" and employees must wear masks and gloves at all times, the governor's rules say. Self-serve buffets are also ruled out.

Retail businesses will follow similar guidelines on Wednesday, when they are slated to start reopening.

~

It gets worse - the widely mocked Las Vegas Mayor Carolyn Goodman just [wants to offer up Vegas as a "control group"](#) to measure the effects of lifting restrictions.

~

[Maryland has a proposal](#) that recommends lifting restrictions only after COVID deaths and new hospitalizations have seen a consistent two-week decline - more conservative than Trump's guidance of two weeks of decline *in new cases*.

~

Threading the needle between Georgia and Maryland, [Texas is going with a 25% plan](#):

Along with retail stores, restaurants and movie theaters, Abbott said that museums and libraries can also reopen on Friday at a 25 percent capacity. Sole proprietors of businesses can also open and doctors and dentists can resume normal operations as well. [...] Churches and places of worship, which were allowed to remain open during the state's stay-at-home orders, are also allowed to expand their capacity provided safe social distancing measures are still enacted. Barber shops, hair salons and bars will still remain closed.

~

[The Bay Area is extending their lockdown through the end of May](#), but now allowing drive-in religious services (stay in the car), one-on-one residential real estate viewings, golf courses, and driving ranges.

~

So when all is said and done, what might things begin to look like? CNN suggests that ["America's 'new normal' will be anything but ordinary"](#), potentially seeing some of the following changes:

- The world will be able to reopen somewhat
- People's temperature will be taken everywhere
- Face masks will be mandatory
- Sports and entertainment venues remain empty
- Lots of monitoring of cellphone locations
- Schools reopening to staggered classes ("You could have 9th and 10th grades come in the morning and 11th to 12th grades in the afternoon [...] Or half of the students could come Monday, Wednesday and Friday. The other half on Tuesday, Thursday and Saturday.") with smaller class sizes
- Elimination of school assemblies, physical education, and recess.
- Increased sanitization efforts and deep cleaning become standard
- Restaurants cut down the number of seats

- Mask-wearing diners greeted by servers in masks and gloves with disposable menus in their hands.
- Unessential airport travel will continue to be limited and airline passengers would be required to have the contact tracing app, confirm no proximity to a positive case, and have a temperature check or show documentation of immunity.

Another take from the Washington Post asks "[How much of our lives will coronavirus change?](#)" and speculates:

Deborah Birx, the White House's coronavirus task force coordinator, warns that social distancing will be in place through the end of the summer. [...]

Even if New York's plan and those put out by states such as Maryland come off without a hitch, they will take weeks, if not months, to ramp up, subject to any setbacks (e.g., a second wave). These states are coming to grips with the reality that much of life will not change to something approximating "normal" before we get a vaccine.

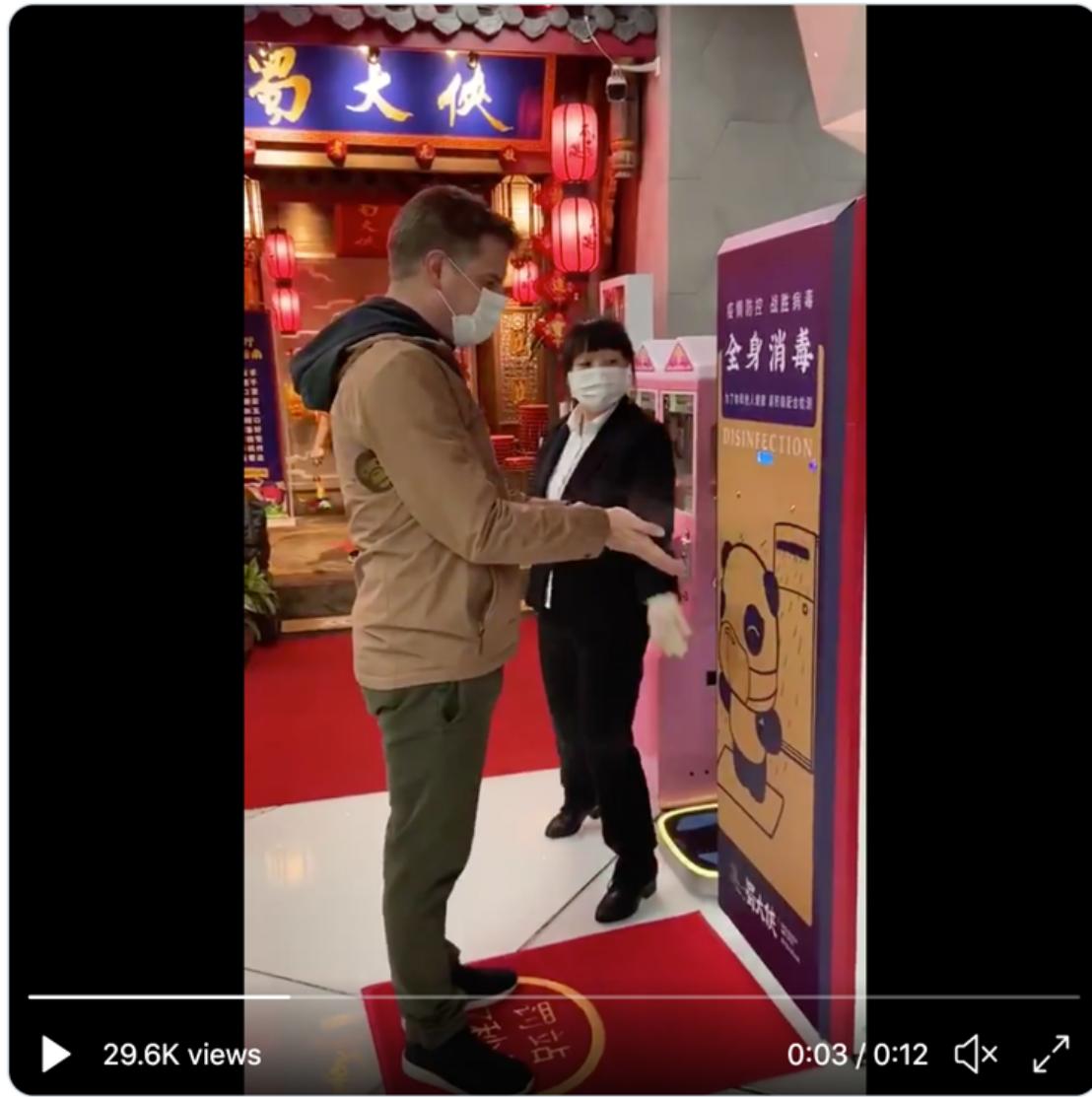
You likely will not enter a store without a mask, sit in a crowded movie theater or restaurant, or fly on a plane. Before there is a vaccine, you might not go to a gym, the beach or a mall — no matter what the social distancing. If you are working from home now, you may very well still be working from home six months or a year from now. Moreover, your employer may eventually decide the business can lease half the space it currently does and have you work from home permanently.

Students at K-12 schools and at colleges may go through a full year in which they never physically meet a teacher or attend a school play or an athletic event in person. Instead of live theater, concerts and sports, we might get our entertainment in a pay-per-view format. Movie theaters were dying off anyway with streaming services and big home televisions; most of the rest may vanish as well. Don't bank on watching a summer blockbuster movie in a theater.



A Starbucks in Hong Kong enforced social-distancing measures by taping off tables and chairs. Jerome Favre/EPA, via Shutterstock

Full body disinfectant spray machine before you enter this restaurant in #Shanghai. Staff in full mask and gloves.



~

It's hard to know when and how we *should* be reopening, and I imagine that many governors are willing to take on some additional death and some risk of disaster in order to reopen their states.

Furthermore, it's also unclear how much people will voluntarily take up these recently reopened businesses versus stay at home. While Georgia is allowing stores to reopen, one reporter reached out to twenty different small businesses and not one of them said they had plans to reopen this week.

...All of this could introduce a lot more uncertainty into the future of the effects of the coronavirus.

Again, [caution is warranted about making predictions about how things will unfold](#):

it's so seductively easy to double down on sweeping pronouncements: E-sports will replace football and basketball, movie theaters will never return, and telemedicine will become the new normal.

Anything is possible, but take a closer look at how often definitive predictions about permanent change are simply extrapolations of recently observable trends taken to some maximum extreme. [...]

Look back, for example, at pronouncements forged during our most recent financial crisis. In early 2009, in the depths of the Great Recession, Time magazine declared "The End of Excess." [...] All of this was perfectly plausible. But it is not what happened. For better or worse, unpredicted developments such as the fracking boom and other factors put an end to talk of "peak oil." [...] Some of this proved true but far less lasting than predicted. Consumer spending has risen by about a third since 2009, from about \$10 trillion a year to around \$13.5 trillion last year, and still accounts for more than two-thirds of the total economic activity. [...]

This is not an argument against predictions (and it is certainly not a critique of any specific prediction). Speculation about what might happen is useful; it can actually start interesting discussions about a future in which shareholder rights aren't so predominant or provoke us to imagine the implications of cities segregated by immunity status. Predictions of death rates or economic consequences can help shape or inspire responses that prevent those predictions from coming true.

~

We also have to wonder about the political will to roll this out. So far, we've done reasonably well at adhering to social distancing for long periods of time with only minimal protesting. I hope this can last for many more months, but I'm unsure. As Ezra Klein summarizes - ["I've read the plans to reopen the economy. They're scary. There is no plan to return to normal."](#):

~

How dangerous will coming out of lockdowns too early be? The real point is that we just don't really know, as we don't know how much voluntary social distancing will remain in place, among many other uncertainties. But previously we knew that it actually took an exceptionally strong lockdown to keep deaths down, so reversing that back to a not-all-that-strong lockdown could be disproportionately bad. Hence ["New Model Shows How Deadly Lifting Georgia's Lockdown May Be"](#):

As of Friday, by official counts in Georgia, at least 871 people statewide had lost their lives to COVID-19. If Georgia had maintained its pre-Friday lockdown policy, the Harvard/MIT team's simulation—which used data from the Johns Hopkins Coronavirus Resource Center and accounts for local demographics and health conditions based on Census and survey data—estimated the state would have logged a total of between 1,004 and 2,922 coronavirus fatalities by June 15. That fatality range, like all such ranges detailed in this article, includes deaths that had already been documented (in this case, 871).

By contrast, under Kemp's current plan to reopen, if approved businesses returned to just 50 percent of their pre-pandemic activity (or "contact") levels, that range could reach 1,604 to 4,236 deaths. At 100 percent of pre-shutdown activity, the projected final body count could soar to a range between 4,279 and 9,748.

One potential cautionary tale might be Japan's northern island of Hokkaido:

It acted quickly and contained an early outbreak of the coronavirus with a 3-week lockdown. But, when the governor lifted restrictions, a second wave of infections hit even harder. Twenty-six days later, the island was forced back into lockdown.

However, again, uncertainty remains a concern. If we make dire predictions that overestimate the return to normal activity or underestimate potential beneficial effects of mask wearing, weather, etc., it's possible things might not be as dire as they sound, which could lead to people distrusting these expert predictions more. ...On the other hand, things could end up worse than projected.

Colleges

For another question, when and how will Harvard reopen for the fall term? [Here's how they're thinking about it:](#)

Harvard will be open for the fall semester, but some or all instruction may continue to be online, the university's provost said Monday. [...] Several criteria must be met for the school to safely reopen its doors, Garber wrote, including models showing the disease is "mostly behind us," and that another outbreak is unlikely.

[Here's an alphabetical list of colleges that have either disclosed their plans, mentioned them in news reports, or set a deadline for deciding.](#) It looks like most of them are aiming to have in-person classes.

...I'm personally wary about how a large campus is supposed to reopen. After all, the risk of spreading COVID is proportional to the amount of people, their density, and the length of contact and college has a lot of all of that.

Indeed, [one study found](#) that "[t]he average student can "reach" only about 4% of other students by virtue of sharing a course together, but 87% of students can reach each other in two steps, via a shared classmate. By three steps, it's 98%."

Even on campus, presumably students will have to be aggressively pre-tested and quarantined before joining the campus population. Also, at least some steps to ensure isolation and some social distancing will have to take place on campus even after all of that.

Facebook

Another place to look to might be Facebook. Per [their latest announcement](#), they will be (1) requiring nearly universal work from home through at least the end of May, (2) cancelling all business travel through at least the end of June, and (3) and cancelling all large physical events with more than 50 people through June 2021.

The United Kingdom

[Buzzfeed News reports on the UK's reopening plan:](#)

A "best-case scenario" being worked on by the Scientific Advisory Group for Emergencies (SAGE) hopes to end lockdown restrictions for certain nonessential shops and industries in the short term, from early to mid-May.

Some social distancing measures could then gradually be relaxed in the medium term, in June and July, eventually leading to the reopening of pubs and restaurants towards the end of summer.

Long-term “shielding” for elderly and vulnerable people could mean limits on people seeing their parents or grandparents over 70 for as long as 12 to 18 months until a vaccine is found.

The timeline relies on SAGE scientists calculating how many new COVID-19 infections per day the UK's test and trace capabilities can manage and an “impossible” political decision for Downing Street on how many deaths per day they are willing to accept in order to be able to lift some restrictions before there is a vaccine.

South Korea

South Korea has remained relatively open, albeit with substantial restrictions in place.

One contact says people don't really notice the impact of COVID any more, though February and March felt pretty chaotic. Working from home is encouraged, esp. for employees that would have to use public transport, but many people come to offices. The next few weeks will be critical: will see if clubs and grocery stores can reopen in Seoul without causing a huge spike in infections.

[COVID-19: Testing, Isolation, Geolocation in Korea](#) gives a bit of a glimpse at how seriously South Korea is taking this:

The guy in the video has just returned to Korea from abroad. He is tested right away (results by next day), and asked to self-quarantine for 2 weeks. His location is monitored via phone (GPS) during this time. During quarantine the government supplies him with food free of charge.

Africa

[It looks like some African countries are coming out of lockdown:](#)

Some, like Ghana, are now easing these measures, concerned about their impact on the poor and because they've taken other steps against the virus.

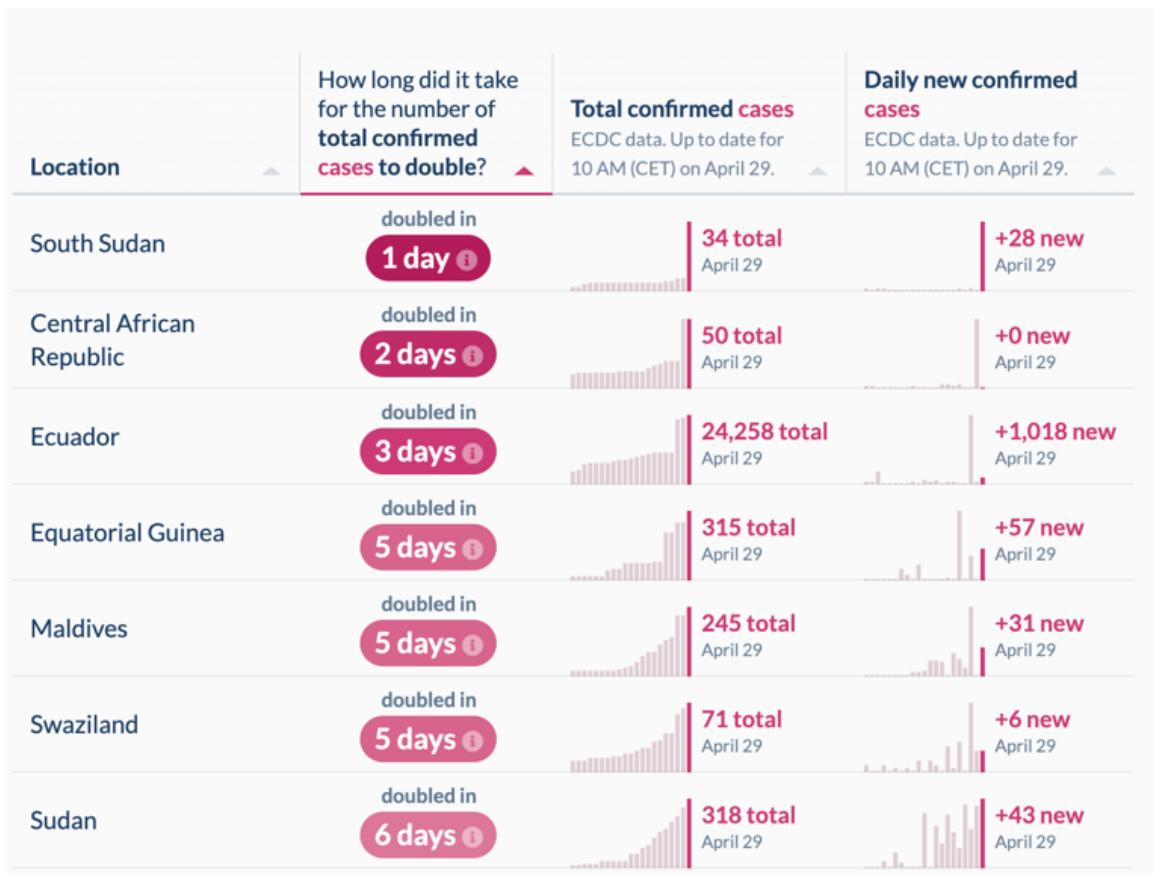
Ghana did place lockdown restrictions on its major cities - which it has now largely lifted. But social events and public gatherings are still banned, and school closures will remain in place for the time being.

However, [lockdowns in developing countries often force people to choose between starvation or disease.](#)

How can lockdowns cause more harm in poor countries than rich ones? When almost everyone works in an informal economy and needs to work every day to put food on the table — the situation in some of the poorest countries — calling a halt to economic activity can get rapidly disastrous. [...]

“If you're a day-wage laborer in a rural area of a developing country, and you don't have much of a buffer of savings, you may be reliant on your wage earnings in a given day or a given week in order to feed your family,” Mobarak argues. “If their family is going hungry that week, they're not going to follow all of your guidelines.”

Worse yet - for many African countries, this is still just the beginning:



Vaccines

What kind of treatments and vaccines might we expect? Derek Lowe outlines [“The Order of the Battle”](#) where first we try repurposing existing drugs (e.g., remdesivir, hydroxychloroquine, azithromycin, falapirivir, ivermectin), then monoclonal antibodies, then vaccines, and then potentially new treatments. It’s possible existing drugs or new treatments could greatly reduce the danger of COVID and allow for faster reopening, but it would take a vaccine to truly make it go away.

So let’s zoom in a bit on vaccines. Keep in mind that producing a vaccine in 1.5 years is by far the landspeed record for making vaccines. The ebola vaccine took five years and as I researched in 2017, it [historically takes on average thirty \(!! years\)](#). We expect unprecedented acceleration here, however, because we have unprecedented resource mobilization and an unprecedented willingness to not undergo all the same rigorous safety measures.

Derek Lowe provides updates on [“Coronavirus Vaccine Prospects”](#) - we need to make a vaccine (of which there are several possible types to try), it needs to be effective at treating COVID, it needs to be safe (not introduce any new diseases or symptoms), and it needs to be effectively manufactured and rolled out at scale. Any of these three prongs could slow down vaccine progress.

And Derek Lowe also provides [“A Close Look at the Frontrunning Coronavirus Vaccines As of April 28”](#), tracking eight major efforts:

So by my count, the biggest and most advanced programs include two inactivated virus vaccines, three different adenovirus vector vaccines, two mRNA possibilities, a DNA

vaccine, and a recombinant protein. That's a pretty good spread of mechanisms, and there are of course plenty more coming up right behind these. You cannot do the tiniest search for such information without being inundated with press releases about companies working on coronavirus vaccines

~

The aptly named VisualCapitalist [tracks the status of some ongoing treatments and vaccines](#):

TREATMENTS						Vasudev Bailey, PhD @vasudevbailey	Zoe Guttendorf @zoeguttendorf
Drug	Company	Target	Stage	Treatment Goal	Location		
1. Kaletra (lopinavir-ritonavir)	Abbvie	HIV protease inhibitor	Failed Trial	Anti-viral growth			
2. Arbidol	Pharmstandard	broad-spectrum antiviral	Failed Trial	Anti-viral growth			
3. Ganovo + Ritonavir	Asclexis	Hep C/HIV protease inhibitors	Phase IV	Treat pneumonia			
4. Actemra	Roche	IL-6 inhibitor	Phase III	Anti-inflammatory			
5. Lenzilumab	Humanigen	anti-GM-CSF	Phase III	Anti-inflammatory			
6. CD24Fc	Oncobiimmune	IL-6 inhibitor	Phase III	Anti-inflammatory			
7. Prezcobix	Shanghai Public Health Clinical Center*	HIV-1 protease inhibitor + CYP3A inhibitor	Phase III	Treat pneumonia			
8. Colchicine	Montreal Heart Institute*	tubulin disruption	Phase III	Anti-inflammatory			
9. Kevzara	Regeneron, Sanofi	IL-6 inhibitor	Phase II/III	Anti-inflammatory			
10. Chloroquine/ Hydroxychloroquine	Univ of Minnesota*	ACE-2 inhibitor	Phase II/III	Anti-viral growth			
11. Avigan	Fujifilm	RNA polymerase inhibitor	Phase II/III	Anti-viral growth			
12. Avastin	Roche	VEGF inhibitor	Phase II/III	Treat pneumonia			
13. Remdesivir	Gilead	adenosine analog	Phase II	Anti-viral growth			
14. Ivermectin (PRO 140)	CytoDyn	CCR5 antagonist	Phase II IND filed**	Anti-inflammatory			
15. Aviptadil	NeuroRx	IL-6 inhibitor	Phase II	Anti-inflammatory			
16. SNG001	Synairgen	IFN-beta-1a	Phase II	Treat respiratory illness			
17. Gilenya	Novartis	sphingosine 1-phosphate receptor modulator	Phase II	Anti-inflammatory			
18. AiRuiKa	Southeast Univ, China*	PD-1 inhibitor	Phase II	Treat pneumonia/sepsis			
19. Mesenchymal Stem Cells	VCANBIO Cell & Gene Engineering	Tissue regeneration	Phase I/II	Anti-inflammatory, Tissue regeneration			
20. Losartan	Univ of Minnesota	AT1R inhibitor	Phase I	Reduce organ failure			
21. Gimsilumab	Roviant	anti-GM-CSF	Phase I	Anti-inflammatory			
22. Sylvant	EUSA Pharma	IL-6 inhibitor	Observational	Anti-inflammatory			
23. Plasmapheresis	Mount Sinai	antibodies from recovered patients	Emergency use	Anti-viral growth, anti-inflammatory			

ARTIS VENTURES

*Trial sponsor **Emergency Use in Patients
Source: FDA, WHO, company websites, news. Available upon request.

VACCINES

Vasudev Bailey, PhD
@vasudevbailey

Zoe Guttendorf
@zoeguttendorf

Vaccine	Company	Platform	Stage	Description	Location
1. mRNA-1273	Moderna	RNA	Phase I-First Patient Dosed	First to dose a human in the US. Vaccine consists of a synthetic strand of mRNA designed to elicit an immune response to produce antibodies against SARS-CoV-2	🇺🇸
2. Ads-nCoV	CanSino Bio	Non-Replicating Viral Vector	Phase I	Benefits from previous success in the Ebola virus (time to market ~3 years). The vaccine being developed is based on viral vectors (adenoviruses) to deliver antigens to express the SARS-CoV-2 spike protein	🇨🇳
3. ChAdOxi nCoV-19	University of Oxford	Non-Replicating Viral Vector	Phase I/II	Enrolling 500+ individuals to test its vaccine candidate, which uses a non-replicating virus to deliver RNA into cells.	🇬🇧
4. LV-SMENP-DC	Shenzhen Geno-Immune Medical Institute	Lentiviral	Phase I/II	Began early testing of its vaccine candidate. The vaccine uses a lentiviral vector to deliver Covid-19 minigenes to modify dendritic cells and activate T cells.	🇨🇳
5. BCG Vaccine	Research Group, Netherlands	Live Attenuated Virus (LAV)	Phase II/III	Repurposing the BCG vaccine, originally for TB, to fight SARS-CoV-2 in healthcare workers at high risk of infection. 1,000 individuals will be enrolled across 8 hospitals to receive the vaccine or placebo.	🇳🇱
6. BCG Vaccine	Murdoch Children's Research Institute	Live Attenuated Virus (LAV)	Phase II/III	The BRACE trial will conduct a randomized, multi-center study of the TB vaccine in 4,000 healthcare workers across Australia.	🇦🇺

*Trial sponsor

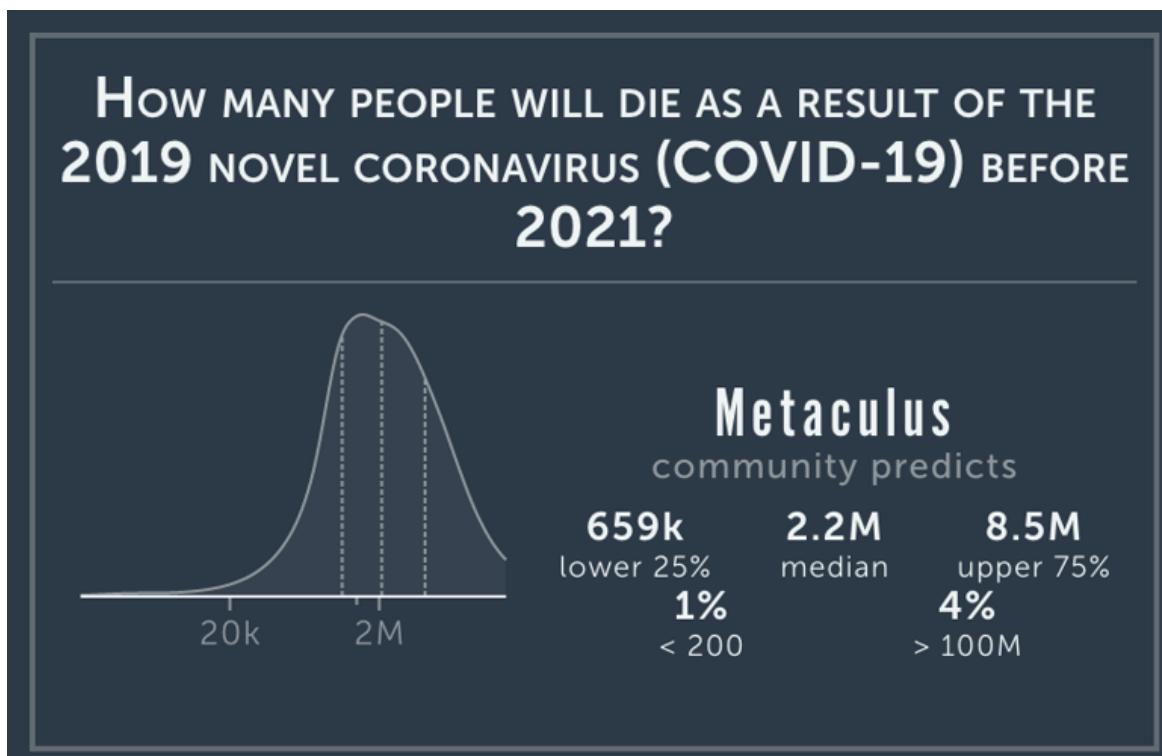
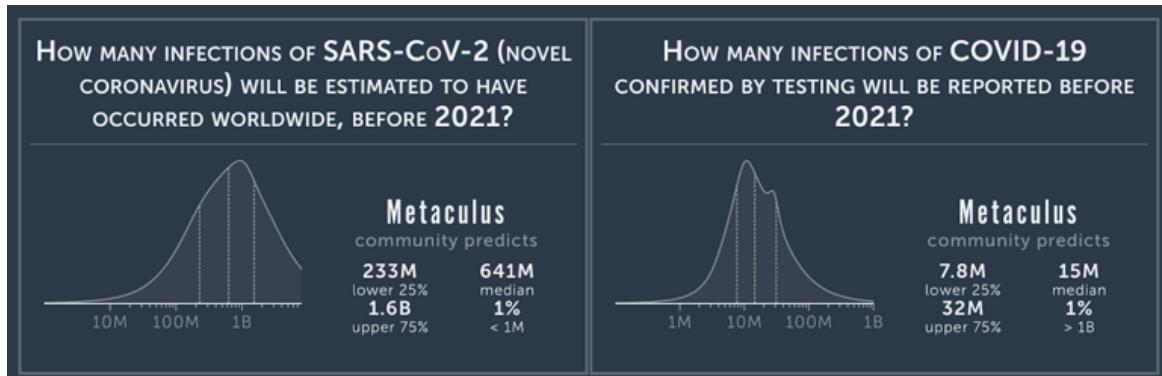
Source: FDA, WHO, company websites, news. Available upon request.



A much more detailed tracker of over 280 treatments and vaccines is available from the Milken Institute.

Gaze into the Crystal - The Latest Modeling and Forecasting

Metaculus has been doing some cool stuff, and I always like checking on [their latest dashboard](#):



It looks like compared to the numbers from 2 April, the lower 25% and median bounds of estimated infections and cases confirmed by testing have gone up. But some good news - the median estimated number of deaths has gone down slightly.

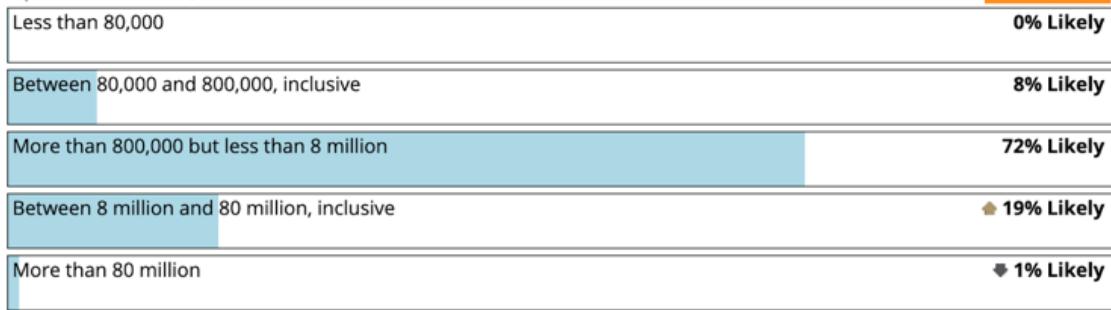
~

Another prediction dashboard, from [Good Judgement Inc.](#), also looks at cases and deaths, though over a slightly different timeframe, and seems to come up with fairly similar results:

How many deaths attributed to COVID-19 worldwide will be reported/estimated as of 31 March 2021?

Opened on 13 Mar 2020, scheduled to close on 31 Mar 2021.

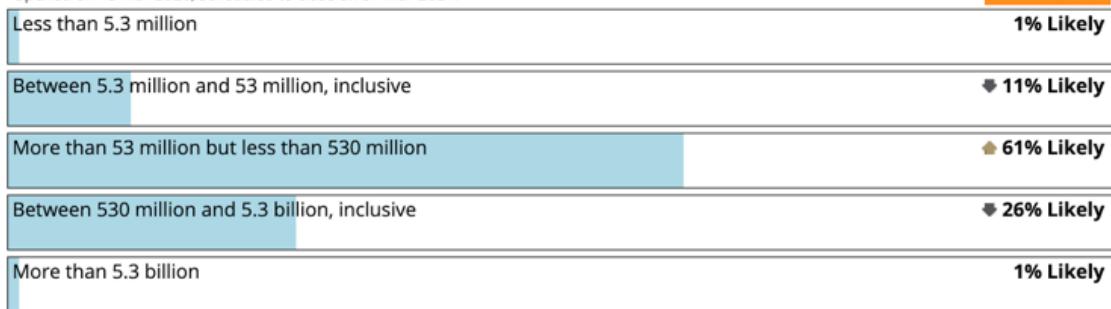
[SHOW MORE](#)



How many total cases of COVID-19 worldwide will be reported/estimated as of 31 March 2021?

Opened on 13 Mar 2020, scheduled to close on 31 Mar 2021.

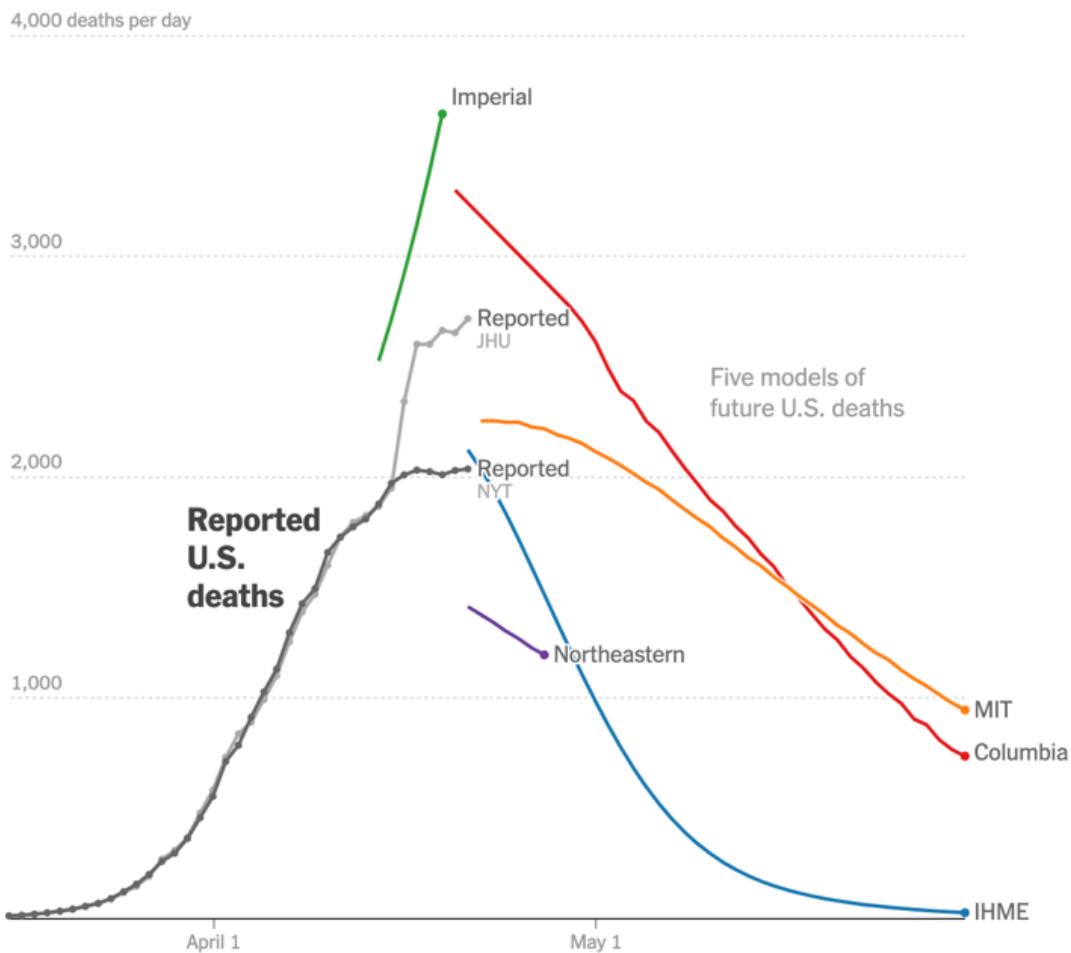
[SHOW MORE](#)



~

Looking at the actual models themselves used by the expert forecasts above, we can see they [actually struggle a bit to agree](#):

U.S. coronavirus deaths in five different forecasts

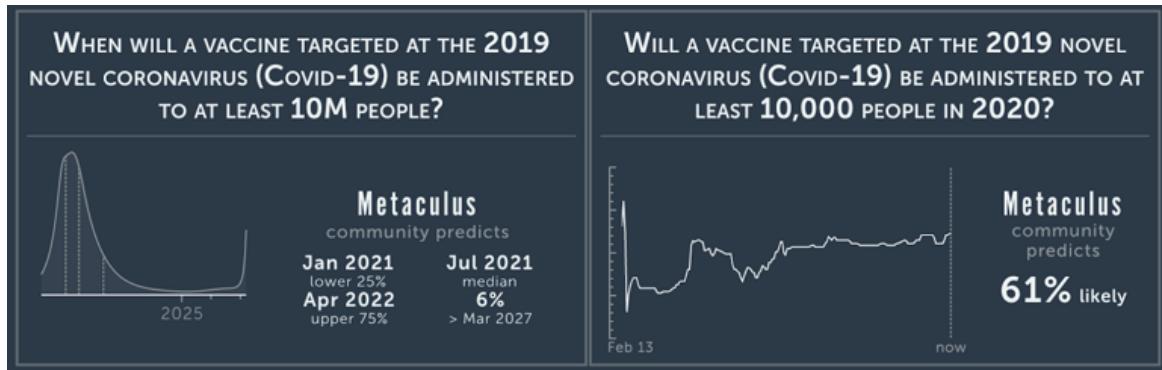


[Calibration of these models also continues to remain a concern:](#)

We have found that the predictions for daily number of deaths provided by the IHME model have been highly inaccurate.

The UW-IHME model has been found to perform poorly even when attempting to predict the number of next day deaths. In particular, the true number of next day deaths has been outside the IHME prediction intervals as much as 70% of the time. If the model has this much difficulty in predicting the next day, we are concerned how the model will perform over the longer horizon, and in international locations where the accuracy of the data and applicability of the model are in question.

~



More good news - the timeline for the vaccine has gotten much more optimistic over the past month, with the lower 25% and median both shifting six months earlier, the upper 75% shifting 1.5 years earlier, and the chance of a vaccine waiting until after March 2027 fell from 8% to 6%.

On the other hand, the Good Judgement forecasters appear a lot more pessimistic, putting almost double the odds (50% vs 25%) on a vaccine not appearing by April 2022. (Note that the scale of the distribution is different, but I'm doubtful this matters much for the timeline at this scale.)

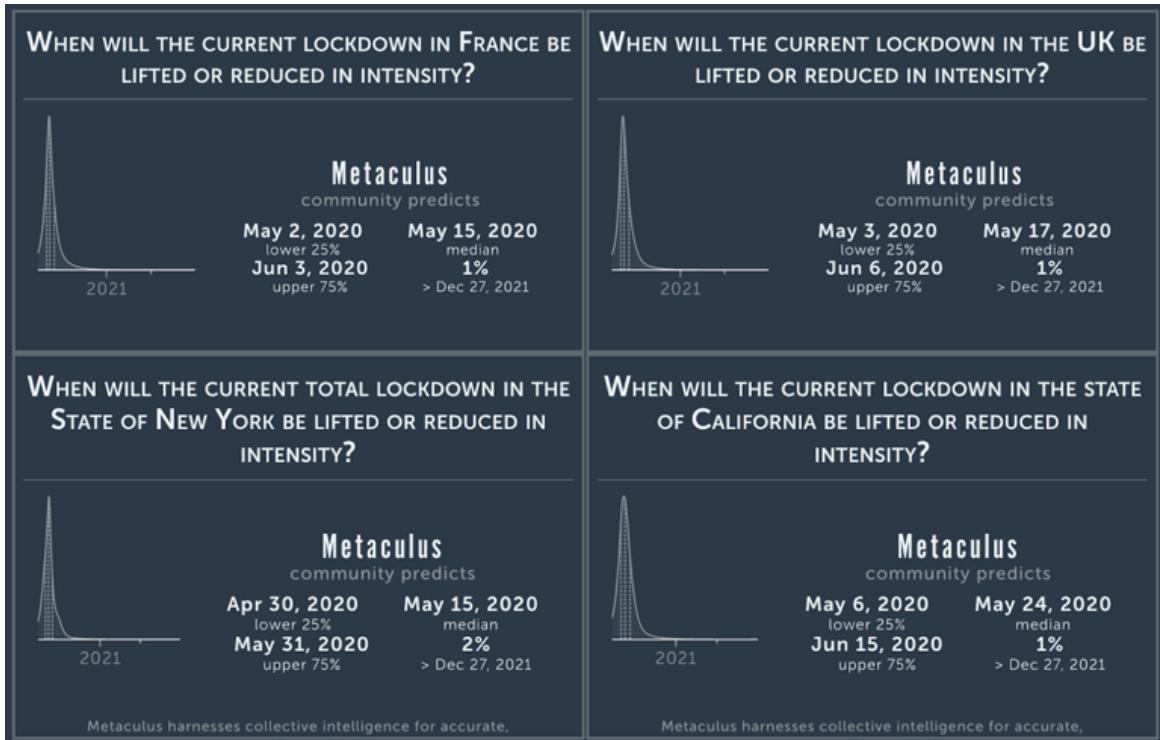
When will enough doses of FDA-approved COVID-19 vaccine(s) to inoculate 25 million people be distributed in the United States?

Opened on 17 Apr 2020, scheduled to close on 31 Mar 2022.

[SHOW MORE](#)

Before 1 October 2020	1% Likely
Between 1 October 2020 and 31 March 2021	4% Likely
Between 1 April 2021 and 30 September 2021	15% Likely
Between 1 October 2021 and 31 March 2022	30% Likely
Not before 1 April 2022	50% Likely

~



Metaculus has also been trying to forecast the easing of lockdowns. As seems confirmed by the reporting above, they're imminent. However, be careful to not interpret easing of lockdowns as anything but the first step in a gradual process - the way these questions are worded, *any* easement will count as a positive case, even if it is a lot less than would make it possible to enjoy the world to the extent we used to.

To get a bit more granular, Metaculus [currently has 47% odds](#) on “most of the classes for courses at Harvard College that would usually be scheduled to occur on September 2nd have in person instruction on September 2nd, 2020.”

Some other questions in a similar vain we might want to ask to get at the granularity of reopening... and my personal predictions:

- When will Chicago reopen Lincoln Park to the public (even with enforced social distancing)? **Median 15 May, 80% CI 1 May to 1 July**
- When will it be legally permitted in Chicago to get a haircut? **Median 1 June, 80% CI 7 May to 1 September**
- When will Chicago CTA subway cars reopen for non-essential travel? **Median 15 June, 80% CI 15 May to 1 September**
- When will it be legally permitted to eat dine-in in a Chicago restaurant? **Median 30 June, 80% CI 7 May to 1 September**
- When will the Chicago Symphony Orchestra have a public performance inside their concert hall (even with massively reduced and spaced-out attendance)? **Median 1 September 2020, 80% CI 1 June 2020 to 1 June 2022**

I made these about Chicago because that's the area where I live and I know well and because Chicago and Illinois have historically been more cautious about COVID. I added these questions to Metaculus to see what others think.

~

Laypeople can make a prediction too... it looks like most people are expecting a full “return to normal” sometime this year. I personally find it unlikely we would be able to see a “return

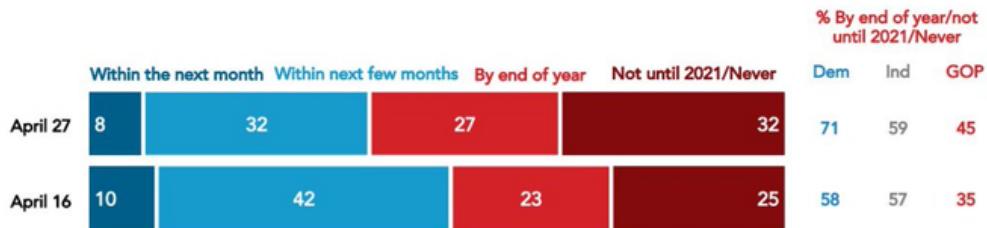
to normal" until after we have a vaccine, which does not seem likely to happen in 2020 at all.

More Say "Return To Normal" Will Take Longer

Fewer Americans now say that a "return to normal" is within a month or few months' reach, with more saying that it will take until the end of the year or beyond for things to go back to normal in the United States.

- Nearly half of Republicans (45%) now say that a return to normalcy won't happen until at least the end of the year.

When do you expect life in the United States to "return to normal," with businesses able to open and people able to go about their lives and interact as they did before the coronavirus pandemic?



Nationwide surveys of registered voters; Each wave represents approximately 1,000 interviews taken over the prior three days.
Latest wave conducted April 22-27, 2020. For more info, visit navigatorresearch.org.

navigator.

~

Even cooler, [Metaculus will be taking these forecasts head-to-head with experts](#):

We're excited to inform you that Metaculus will be participating in COVID-19 Expert Surveys, in collaboration with researchers at the University of Massachusetts Amherst. Your skills will be compared to the forecasting ability of infectious disease experts!

Each week we will be launching up to 8 questions in which the community will have 30 hours to lock in their predictions.

Researchers will concurrently distribute the same questions to some of the world's leading infectious disease experts.

The resulting Metaculus predictions (your predictions!) and expert forecasts will be gathered into a weekly report that is sent to the CDC.

~

If you want cool and useful plots about COVID cases, deaths, and testing broken down by US state, county, and city, check out covid19watcher.com!

Now Just What are the Tech Overlords up to?

[Google and Apple are joining forces on contact tracing](#). This is big news, as Google's Google's Android and Apple's iOS have approximately 100% of the mobile OS market share with [3 billion combined users](#). The plan is for Apple and Google to release interfaces next month for health officials to build apps and for contract tracing to be enabled over the next few months.

The app works by running Bluetooth in the background and broadcasting beacons that are logged by nearby phones. Once someone is diagnosed with COVID, they can consent to sharing the past two weeks of beacon logs with health professionals, who can then broadcast anonymous alerts to those who have been in contact.

Google and Apple emphasize that the system is built on a decentralized stack that does not broadcast any data without user's permission and does not use location data at all (only data of who you came in contact with, not where).

However, [the app needs about 60% of the adult population to use it for it to be fully effective](#) at controlling COVID on its own, but it could still be an important part of a larger contact tracing effort.

~

[Bill Gates is spending billions of dollars building manufacturing facilities for many vaccine candidates](#), even though many of these facilities will go unused:

Gates said he was picking the top seven vaccine candidates and building manufacturing capacity for them. "Even though we'll end up picking at most two of them, we're going to fund factories for all seven, just so that we don't waste time in serially saying, 'OK, which vaccine works?' and then building the factory," he said.

~

People sometimes criticize billionaires for not donating enough. Jack Dorsey gets it and is [donating a full billion](#) (~28% of his wealth) to fund global COVID-19 relief. [He's tracking his donations here](#) and it looks like he's already dispersed \$8M.

Now Let's Talk Policy Response

Polls consistently show that Americans still strongly support social distancing:

Few Americans Support Undoing Social Distancing

Despite coverage of protests against stay-at-home orders, the public continues to believe that we are currently doing the right thing or need more aggressive social distancing policies.

- Nearly all of the increase in relaxing distancing comes from Republicans, but even still, only one-in-four (23%) Republicans supports doing that.

When it comes to social distancing, what do you think we, as a country, need to be doing right now?



Nationwide surveys of registered voters; Each wave represents approximately 1,000 interviews taken over the prior three days.
Latest wave conducted April 22-27, 2020. For more info, visit [navigatorresearch.org](#).

navigator.

Majorities Across Partisans Say Shelter-In-Place Measures Are Worth It To Protect People

Share who say strict shelter-in-place measures...

- Are worth it in order to protect people and limit the spread of coronavirus
- Are placing unnecessary burdens on people and the economy and are causing more harm than good

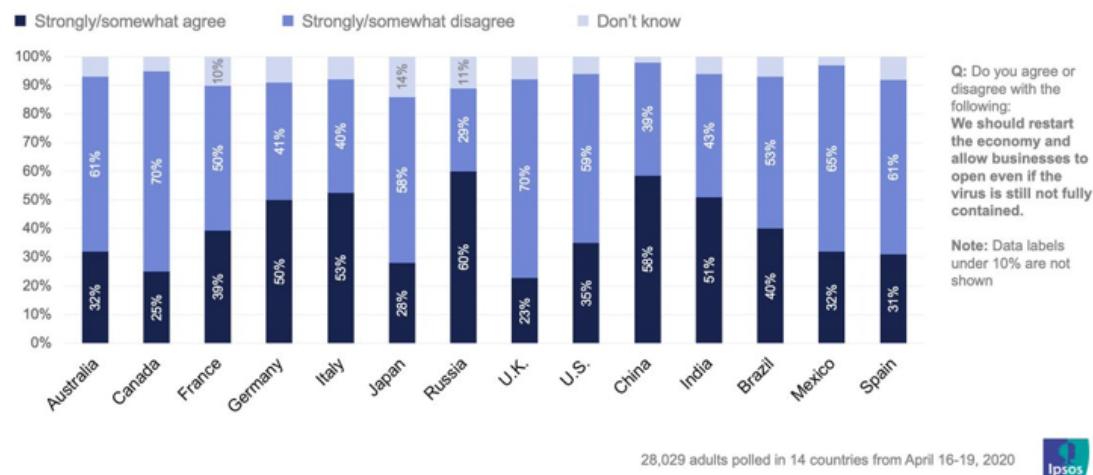


SOURCE: KFF Health Tracking Poll (conducted April 15-20, 2020). See topline for full question wording.



In fact, support among freedom-loving Americans seems to largely be in line with strong support in other countries:

SHOULD THE ECONOMY AND BUSINESSES OPEN EVEN IF THE VIRUS IS NOT FULLY CONTAINED?



28,029 adults polled in 14 countries from April 16-19, 2020



~

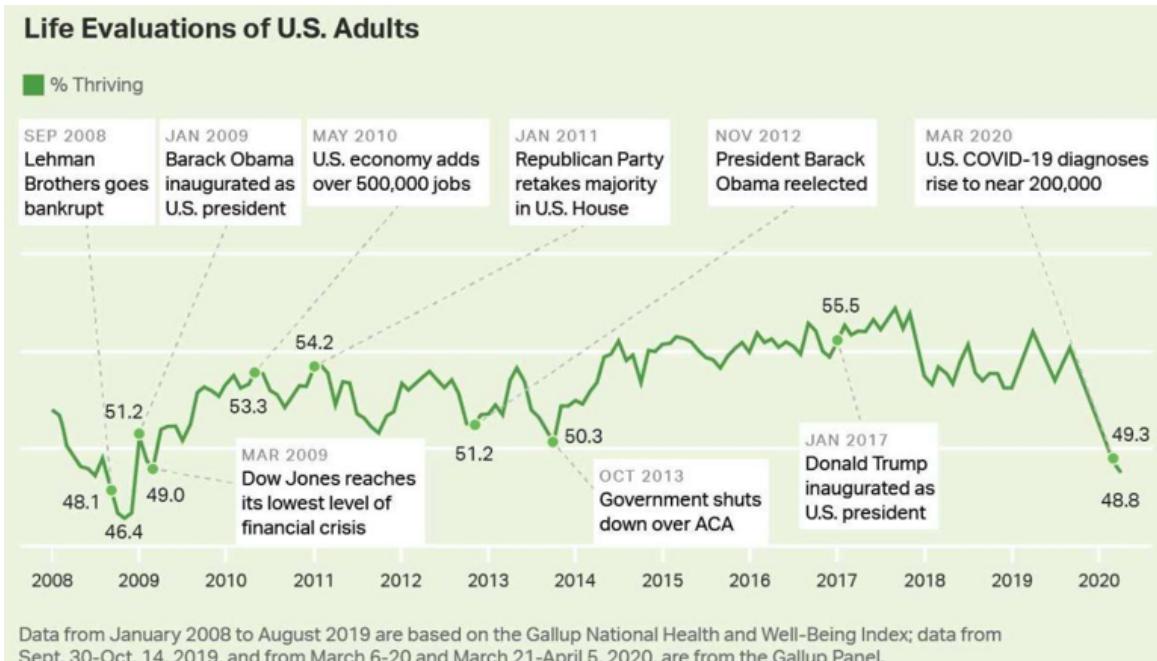
How can we scale up vaccine progress? There are many ways but one compelling way [might be option-based guarantees](#):

Effectively tackling COVID-19 will require rapidly scaling up the production of diagnostic tests, pharmaceutical treatments and vaccines. In each case, preparations for large-scale manufacturing, such as building factories, are typically delayed until the product is proven safe and effective. This makes sense from a commercial perspective, but incurs great costs in terms of lives lost and damage to the economy.

There are several potential solutions, but the most promising appears to be “option-based guarantees”. In essence, the government commits to paying a proportion of the manufacturer’s preparation costs should the product turn out not to be viable. (If the product is viable, it can be sold as normal.) This reduces the risk to the company while maintaining an incentive to produce a high-quality product quickly and at scale.

A Bit About Life Under Quarantine

COVID certainly is taking a toll on all of us. [Pete Davidson and Adam Sandler have teamed up to sing about it](#). Self-reported well-being in the US is at a 12 year low:



[What does all of the current modeling imply for your summer vacation?:](#)

Everyone is asking, “Can I travel this summer?” And the answer may be a cautious and optimistic “maybe,” at least for some of us to some destinations. Keep listening to government officials and research destinations that may make sense for a summer sojourn in our current more distanced realities. While a theme park visit may not be in the cards right away, a trip to a more secluded beach or mountains may be just what you need to recharge from your months spent at home once it is safe to do so.

Of course it depends on the month - travel in August seems a lot more likely than travel in June... and I don't expect you'll be going anywhere internationally.

[Baseball might be coming back this year:](#)

“Over the past two weeks, as states have begun to plan their reopenings, nearly everyone along the decision-making continuum -- league officials, players, union leaders, owners, doctors, politicians, TV power brokers, team executives -- has grown increasingly optimistic that there will be baseball this year. [...]

Finalize a plan in May. Hash out an agreement with the players by the end of the month or early June. Give players a week to arrive at designated spring training locations. Prepare for three weeks. Start the season in July. Play around an 80- to 100-game season in July, August, September and October. Hold an expanded playoff at warm-weather, neutral sites in November.

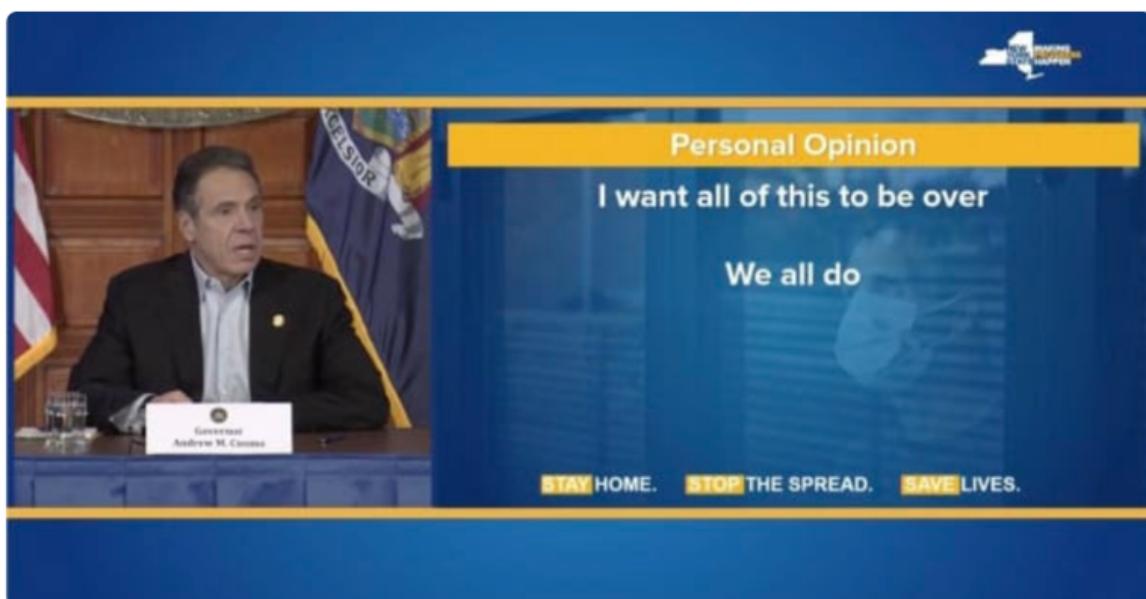
Now: This is not set in stone or anywhere close to it. But from the league to the players to the owners to TV executives, this, or some derivation of it, registers as the most realistic option at this point."

~

The real important question - how long until we run out of new TV? [Luckily it looks like Netflix can keep producing shows well into 2021.](#)

~

The latest victim of COVID-19? [The hotel mini-bar.](#)



If You Still Own Envelopes, Check Their Backs - Here's the Latest Cost-Benefit Analysis

[A new paper has come out](#) that may be the first proper attempt to evaluate COVID policy in terms of wellbeing and suggests that net benefits of releasing lockdown will be positive from June:

In choosing when to end the lockdown, policy-makers have to balance the impact of the decision upon incomes, unemployment, mental health, public confidence and many other factors, as well as (of course) upon the number of deaths from COVID-19. To facilitate the decision it is helpful to forecast each factor using a single metric. We use as our metric the number of Wellbeing-Years resulting from each date of ending the lockdown. This new metric makes it possible to compare the impact of each factor in a way that is relevant to all public policy decisions.

The paper has some shortcomings - among others, it would be good to have more probabilistic analysis given the highly uncertain inputs, the effect on mental illness remains guesswork, the GDP loss estimates may be overestimates, and there are some questionable assumptions. Therefore I wouldn't put much stock in the actual results, but I do like it as a better attempt at methods and much better than merely looking at death rates given an assigned statistical value of a life.

And Now a Word From the Lamestream Media

[The media is being impacted rather directly](#): “Roughly 36,000 workers at news companies in the U.S. have been laid off, been furloughed or had their pay reduced. Some publications that rely on ads have shut down.”

~

But assuming the media survives long enough for a retrospective, can we ask what went wrong with the media's coronavirus coverage? And can we do better? [Writing for Recode](#), [Peter Kafka argues that the issue](#) was about properly communicating uncertainty and risk, a question of which experts to trust, and how to properly communicate what they were saying. While the media was wrong, in many cases the experts were fairly wrong too. (This is why [the only place you can truly trust is LessWrong](#).)

Scott Alexander writes [“A Failure, but Not of Prediction”](#), arguing that predicting the spread of COVID was very difficult but we need to get better about making clear recommendations that are the best given the uncertainty:

Predicting the coronavirus was equally hard, and the best institutions we had missed it. On February 20th, Tetlock's superforecasters predicted only a 3% chance that there would be 200,000+ coronavirus cases a month later (there were). The stock market is a giant coordinated attempt to predict the economy, and it reached an all-time high on February 12, suggesting that analysts expected the economy to do great over the following few months. [...]

Their main excuse is that they were just relaying expert opinion – the sort of things the WHO and CDC and top epidemiologists were saying. I believe them. People on Twitter howl and gnash their teeth at this, asking why the press didn't fact-check or challenge those experts. But I'm not sure I want to institute a custom of journalists challenging experts. [...]

But I would ask this of any journalist who pleads that they were just relaying and providing context for expert opinions: what was the experts' percent confidence in their position?

I am so serious about this. What fact could possibly be more relevant? What context could it possibly be more important to give? I'm not saying you need to have put a number in your articles, maybe your readers don't go for that. But were you working off of one? Did this question even occur to you?

Nate Silver said there was a 29% chance Trump would win. Most people interpreted that as “Trump probably won't win” and got shocked when he did. What was the percent attached to your “coronavirus probably won't be a disaster” prediction? Was it also 29%? 20%? 10%? Are you sure you want to go lower than 10%? [...]

And if the risk was 10%, shouldn't that have been the headline. “TEN PERCENT CHANCE THAT THERE IS ABOUT TO BE A PANDEMIC THAT DEVASTATES THE GLOBAL ECONOMY, KILLS HUNDREDS OF THOUSANDS OF PEOPLE, AND PREVENTS YOU FROM LEAVING YOUR HOUSE FOR MONTHS”? Isn't that a better headline than Coronavirus panic sells as

alarmist information spreads on social media? But that's the headline you could have written if your odds were ten percent! [...]

The Vox article says the media needs to "say what it doesn't know". I agree with this up to a point. But they can't let this turn into a muddled message of "oh, who knows anything, whatever". Uncertainty about the world doesn't imply uncertainty about the best course of action! Within the range of uncertainty that we had about the coronavirus this February, an article that acknowledged that uncertainty wouldn't have looked like "We're not sure how this will develop, so we don't know whether you should stop large gatherings or not". It would have looked like "We're not sure how this will develop, so you should definitely stop large gatherings."

I worry that the people who refused to worry about coronavirus until too late thought they were "being careful" and "avoiding overconfidence". And I worry the lesson they'll take away from this is to be more careful, and avoid overconfidence even more strongly.

Vox (owner of Recode) Co-Founder Matt Yglesias [has a rejoinder](#) that maybe we can't have nice things:

This is a good take from @slatestarcodex on the media, but I hope people who read it will take seriously the question of what would happen to a media outlet that constantly featured blaring headlines warning about low-probability catastrophes.

Beyond the specific content of any one article any publication ran, the real "media failure" is the ratio of articles dedicated to early coverage of the 2020 election vs global public health issues is objectively indefensible.

But you've gotta do stories people want to read.

[Rob Wiblin counters that we actually could've predicted things just fine](#) (and Wiblin basically did):

In this post Scott Alexander says that in early February forecasting tournaments, financial markets, journalists, random public health people, and amateurs all gave a low likelihood to a serious SARS2 pandemic (<5%), and this shows it was legitimately hard to predict.

This is generous of Scott. But I want to argue it is wrong and we should not give up on better prediction, because our performance here was just surprisingly and unnecessarily bad.

These groups could have and should have given a much higher probability to what happened happening. Even, or especially, given the little we knew at the time.

[...] We suggest there's a ~20% chance of it being wiped out like SARS1. We don't understand how that could happen, but we also don't understand how SARS1 was eliminated, so uncertainty means there has to be a decent shot that the same will happen again. This leaves an 80% probability that it will spread widely, which means there's a decent chance it will go on to cause millions or tens of millions of deaths.

What great ability did we use to gain this insight which others missed? Nothing complicated. For myself, I just actually formed a super simple inside view of what was going on, and bothered to use it. [...]

As it turns out I was wrong about China not being able to control it internally. But it also turned out, i) SARS2 was pretty good at asymptomatic spread, ii) it was already in many countries by that point, and iii) most developed countries mounted no useful response early on. Any of these alone would likely have been enough for a serious global pandemic to result.

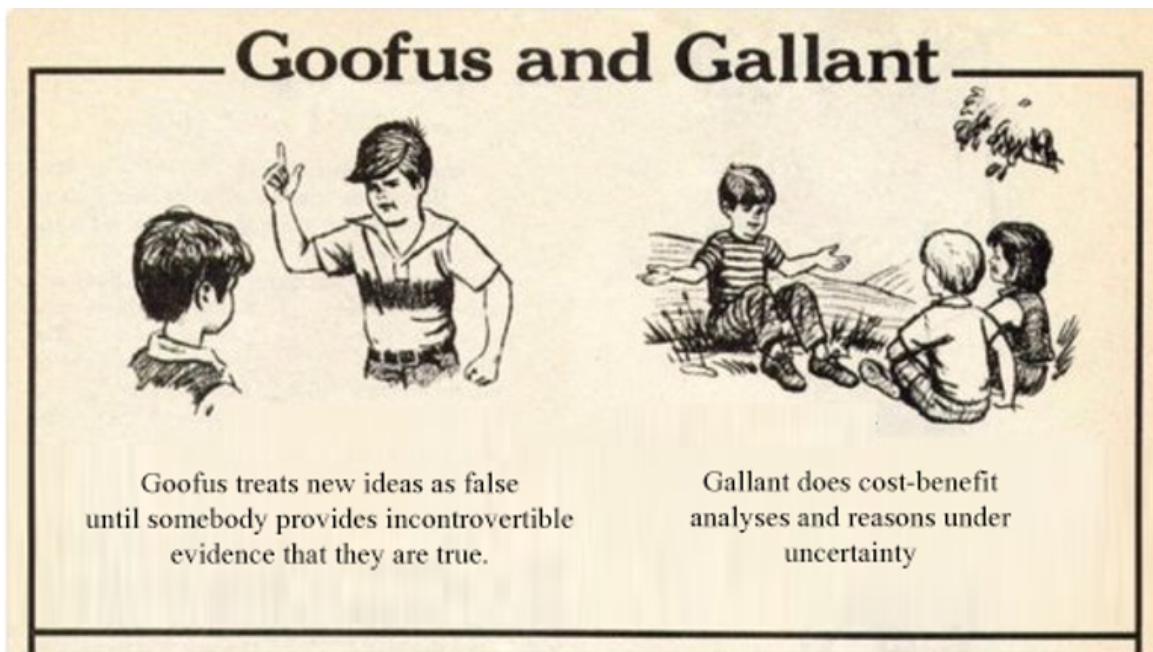
The mechanisms operating here were not so mysterious we needed to rely primarily on a general 'outside view' to see the future. But if we did, a single success with SARS1 should not have been that reassuring, relative to our failure to control almost all (maybe any?) human-transmissible respiratory viruses — something so difficult we've very rarely bothered to try. [...]

Saying the risk of a major global pandemic was 5% or lower was a strange contrarian bet against i) the really obvious, and ii) expert opinion, more narrowly defined. And as far as I could see it was a contrarian bet that nobody at the time was trying to justify.

What was everyone thinking? I don't know. Some possibilities are i) pinning too strongly to the recent examples of H1N1 and SARS1 not being so bad, ii) just not paying attention, as there's lots of things to worry about in the world, iii) general skepticism that any one thing going on can be such a big deal, especially something so random and meaningless as a virus jumping from cats to people.

I'm sympathetic, especially to the fact that there's always so many damn things going on that nobody has the capacity to form a sensible view about all of them.

But I maintain that our forecasters did worse than they realistically could have, people should learn from this and improve their thinking, and we should aspire to give the world more sensible credences and forewarning next time.



Don't Forget About the Nonhumans!

Rethink Priorities researcher Daniela R. Waldhorn [wrote a bunch about how nonhumans are being impacted](#):

The ongoing coronavirus pandemic is having several important consequences on the lives of non-human animals. In this report I overview its main direct effects on animals used for human consumption. Some main findings are:

Welfare concerns: Transport restrictions have prevented animal feed from getting delivered to farms. As a consequence, young chickens have been killed in massive numbers, in horrific ways. Moreover, animals who are transported alive are being subjected to extraordinary precarious conditions.

Fishing and aquaculture are the economic activities that have suffered the worst consequences of the pandemic.

Production and consumption of animal-based products: The disease is also affecting international meat and egg prices due to disruptions in supply and demand. While consumption of some animal-based products is temporarily rising (e.g., chicken meat, eggs), consumption of other types of meat (e.g., sea animals) is decreasing. Still, the future of the animal protein market looks complex and changing.

Plant-based meat: If the U.S. and Europe dip into a recession, the business of plant-based companies will probably be disrupted. In China, because of the coronavirus outbreak along with other health threats, consumers' willingness to try new plant-based proteins is likely to increase.

Finally, some practical implications of the pandemic are presented. One is that consumers are likely to be more concerned about obtaining safer and healthier food. Another is that organizations running corporate outreach campaigns should probably consider some strategic changes, focusing less on restaurant and hotel chains and more on grocery chains.

It should be noted that the situation is evolving rapidly and other factors not considered here may come into play in upcoming weeks. For the same reason, please note that we have waived our normal review and quality check standards to publish this report as soon as possible.

~

Previously I wrote about a short, scientific e-book about pandemics and animal farming. [That e-book is now available in a significantly prettier version.](#)

Your Regular Dose of WTF



New York Post

@nypost

Petition to name Dr. Anthony Fauci 'Sexiest Man Alive' gains momentum

trib.al/W266dus



April 2nd 2020

761 Retweets 4,374 Likes

A Second Dose of WTF



Joe Biden @JoeBiden · Apr 24

I can't believe I have to say this, but please don't drink bleach.

55.9K

319.8K

1.5M



Fun (Online) Distractions, Because We All Still Need to Enjoy Life

[If "Friends" were in quarantine.](#)

[Here's everything coming to Amazon Prime Video in May 2020.](#)

[Banksy's Working From Home and Says His Wife Hates It](#)

Looking for a little something to spice up your next meeting? [Invite a llama or goat to your next video call.](#) The pricing is definitely on the pricey side but could make it happen! “For \$65, you get a 20-minute virtual tour of the farm for up to six call participants. For a bigger meeting, you can pay \$100 for a 10-minute animal cameo or \$250 for a 25-minute virtual tour.”

[It's now possible to get a virtual haircut:](#) “How it works? Step 1: Get tools ready. Find or buy your best pair of haircutting scissors or razor for you men’s, women’s or kid’s haircut. Step 2: Book an appointment. Step 3: Video chat with your stylist who will coach you (or your friend) through your haircut session.”

[A Seaside Irish Village Adopts Matt Damon:](#) “Sightings of Mr. Damon have become common in recent weeks in Dalkey, a seaside resort town southeast of Dublin, where his presence has added yet another surreal layer to life under lockdown.”

[Prickles the Sheep Returns Home After Seven Years:](#) “After missing years of shears, the voluminous creature had ballooned to about five times the size of a typical sheep”

It's called quarantine coffee. It's just like normal coffee but it has a margarita in it and also no coffee.



Today's briefing was made possible with significant work by Peter Hurford, Derek Foster, Daniela Waldborn, and Neil Dullaghan. This brief greatly draws upon reporting by Johns Hopkins, The Dispatch, FiveThirtyEight, STAT News, Foreign Policy Magazine, Politico, and others.

How strong is the evidence for hydroxychloroquine?

There has been a lot of discussion of hydroxychloroquine (see the megathread on [Effective Altruism Coronavirus Discussion](#), note you need to answer two questions to gain access). Doctors treating COVID-19 have rated [hydroxychloroquine](#) the most effective drug based on their experience. But on the other hand, results have been mixed with a recent RCT showing [no effect](#).

At this stage how strong is the evidence for hydroxychloroquine and if it works, how effective does it appear to be as a treatment?

Disclaimer: Please seek medical advice before taking any substance, particularly those like hydroxychloroquine that have known side effects.

Hammer and Mask - Wide spread use of reusable particle filtering masks as a SARS-CoV-2 eradication strategy

by Marcel Müller (M.Sc. Biological Sciences)

Contact: marcel_mueller@mail.de

Epistemic Status: I am not a virologist and this is not medical advice. That said I am fairly sure about the core claims of this piece, though there may be unknown unknowns I overlooked. I think that among everything I have heard so far this is by far the most promising strategy to eradicate SARS-CoV-2 without crippling economic damage and / or millions of deaths.

You have permission to distribute this wherever you think it might do good and I actively encourage everyone to do so.

This is written from a European/German perspective since I live in Germany but should apply in most places with few changes necessary.

Abstract

SARS-CoV-2 emerged as a novel pandemic virus threatening to cause a world wide health and economic crisis. While measures have been put in place to temporarily slow the spread of the virus, a workable, economically feasible, long term strategy is needed to deal with the situation, to avoid both millions of deaths and severe economic disruption.

All strategies for dealing with the pandemic need to rely on controlling the spread of the epidemic by reducing R_{eff} either by generating immunity or by reducing the number of contacts within the population.

Therefore it is crucial to discuss the best ways of reducing R_{eff} . Economic lockdowns, school closures and social distancing are measures that were easy to implement but not necessarily optimal.

I propose the widespread use of reusable high quality (European P3 standard equivalent) particle filtering masks in both healthcare and community settings to reduce the risk of infection during contacts sufficiently to allow lifting of most other restrictions while still achieving continued exponential decay of new case numbers.

Introduction

In the last months of 2019, SARS-CoV-2 emerged as a novel zoonotic virus on a wet market in China. It quickly spread through the country and then through the rest of the world, threatening to overwhelm health care systems worldwide. Most nations

reacted by shutting down large parts of their economy, closing borders, and mandating different curfew variations. Some countries managed to significantly reduce the effective reproduction number R_{eff} , which caused exponential decay of their new case numbers, though at great cost to their economy. It is unclear how long these measures can be sustained before adequate supply of the population becomes questionable. [1] Unfortunately, lifting of these measures will return R_{eff} close to its original value again, leading to renewed exponential growth as long as there is no significant immunity in the population.

Several strategies have been proposed to deal with this problem:

Flattening the curve: Here the idea is to let the epidemic run its course at reduced speed until enough people have been infected to attain herd immunity. [2] Proponents of this idea hope to fine tune the rate of infection to the capacity of the health care system for critical care and especially ventilation. Since according to current estimates R_0 of SARS-CoV-2 is about 3, herd immunity is achieved when about 70% of the population have been infected.

This strategy has two problems:

1. Sufficient flattening would be extremely difficult to achieve, since even in Germany with its high density of critical care facilities and under very favourable assumptions (all critical care just for CoViD patients, 2% critical care patients, 10 treatment days) R_{eff} would have to be maintained between 1 and 1.25 for about a year to not overwhelm the healthcare system. [3,4]

Under less favourable assumptions and with less critical care density, even tighter control of R_{eff} would be necessary and the timeline would quickly expand to decades as discussed in [5].

2. Even if sufficient flattening is achieved, this approach would cause widespread death and disability among the affected population. With the best medical care possible CoViD 19 kills at the very least 0.4% of the infected population [6] and a few % require ventilation to survive [7]. In Germany alone this would result in 230000 dead and about 2 million on ventilation with all the morbidity and disability that follows. Applied to 5.25 billion infected (70% of the world population) this would mean 21 million dead and about 200 million on ventilation in the absolute best case. But since these numbers apply to a situation with critical care for everyone who needs it, which is, as demonstrated above, completely impossible to provide with that many cases most of the people who need ventilation would probably die.

Quick vaccination: Under this strategy, all measures are left in place until a vaccine has been developed and produced in sufficient quantity to vaccinate most of the population.

Here the problem is that even in the best case, development and safety testing of a vaccine will take at least 12 to 18 months [8] and it will likely prove to be economically infeasible to leave the current lockdowns in place until next year. So the only possibility would be to forgo all but the most rudimentary safety testing with the hope of having a vaccine available this fall, thereby risking severe adverse reactions in an appreciable fraction of the vaccinated population with somewhat unclear benefit.

While the outcome would probably be a lot better than with the „Flattening the curve“ strategy, at least currently it seems no state is willing to risk this.

Hammer and Dance: Here the „Hammer“ of the above described lockdowns is used to drive R_{eff} below 1 until the number of currently infected falls below a manageable number. This is followed by the „Dance“, where regulations are relaxed far enough to allow at least basic economic activity, while at the same time maintaining R_{eff} slightly below 1 so renewed exponential growth of new case numbers is prevented until a vaccine can safely be deployed to most of the population. [9]

The main problems with this approach are:

1. How many countries are able to impose sufficient measures to halt the spread of the epidemic in the first place?
2. Is it possible to guarantee continued economic functioning to supply the population with basic necessities while keeping R_{eff} below 1 for at least 18 months?

Contain and Eradicate: This strategy looks in many ways similar to Hammer and Dance. The difference is that after the “Hammer” phase, travel restrictions are left in place and every remaining infection chain is contained through extensive tracking and testing until the virus is eradicated from the population. While not relying on vaccination, this strategy shares every problem with “Hammer and Dance” but is even more demanding regarding the comprehensiveness of the testing and tracking regime and the extensiveness of travel bans needed [9]. Even a single uncontained case could lead to a full resurgence of the epidemic.

An approach missing from these discussions is the widespread use of high quality, reusable particle filtering masks which should, as demonstrated below, be a very promising strategy to drive R_{eff} below 1 and ultimately eradicate SARS-CoV-2 from the population. While this idea shares many similarities with both “Hammer and Dance” and “Contain and Eradicate”, the big advantage is that the demands on testing, tracking and continued reduced economic activity are much less severe than in the above scenarios.

Current discussion on mask usage

There has been much discussion about the use of masks to prevent SARS-CoV-2 infection [11]. Many experts, including the american CDC and the WHO, regard masks in community settings as not effective at all or only marginally effective against other aerosolized viruses such as influenza [12]. They also recommend against their use for self protection in the case of SARS-CoV-2 [13] though some contradictory studies exist [14]. This stands in stark contrast to both the expectation derived from principles of biology and physics and decades of experience in both health care [15] and lab settings [16]. To explore why this difference arises, a closer look at the different types of masks available is necessary.

Different kinds of masks

There are three different types of masks that could be used for infection control purposes:

1. Surgical Masks:

Cloth and surgical masks, which are not designed to form a seal on the face, and thus are unable to filter all the air inhaled by the wearer. While larger droplets on a direct path are stopped by these masks, aerosol particles that move primarily under the influence of air currents can gain access to the wearers airways by entering the gap between face and mask. Depending on the exact material used, small aerosol particles may even penetrate the mask itself [17].

2. Disposable “Filtering Facepiece Particle” (FFP) masks:

Disposable masks, built with a face piece made from a filtering material which is designed to form a seal on the face of the wearer. These are available in different grades, which differ on both seal quality and filter quality, with the highest quality generally deemed suitable for use against aerosolized and even airborne pathogens [18]. In Europe these grades are FFP1, FFP2 and FFP3 [18].

A big problem with these masks is that they need to be fitted to the user by bending wires in the mask and adjusting various straps, which many people, even trained medical personnel, often fail to do properly [19]. Also, not every mask can be fitted to every face and a filtering mask with a compromised seal is no better than a surgical mask since in both cases unfiltered air can enter from the sides. These are the masks almost universally used in health care for protection against infection with SARS-CoV-2.

Another downside of these masks is that the filtering element – the facepiece itself – is continuously exposed to the breath of the wearer, which soaks the filter with moisture. In the presence of liquid water it is possible for virus particles bound to filter fibers to become resuspended in this liquid. Since all particle filters used in masks are not membrane filters but adsorption filters with pore sizes much larger than the particles to be removed from the air [18], these resuspended virus particles can diffuse across the filter to the inside of the mask as soon as a continuous water column is formed across the filter material. Now they can either be re-aerosolized by the wearer's breath or taken up by lip contact to the inside of the mask. This necessitates frequent changing of damp masks. Since it is unclear if these masks can be cleaned and reused without significant drop in filter performance, the continued supply of the health care system with a sufficient number of masks is difficult and supply of the general population for everyday use is a logistical impossibility.

3. Reusable masks with replaceable filters

This category contains masks which consist of a gas impermeable mask body, typically made from silicone or rubber and one or two separate filter cartridges which fit into filter ports in the mask body. These filters are available in the similar qualities as the above disposable masks, called P1, P2 and P3 in Europe [18].

A properly fitting mask of this type is very easy to use compared to an FFP mask, since no delicate adjustments have to be made. Also fit testing with these masks is very easy since closing the filter ports while wearing the mask makes inhaling impossible if the mask seals properly.

Another big advantage is that these masks generally have a one way valve protecting the filter from the wearer's breath, which is discharged via a separate exhaust valve. This prevents the filter material from being soaked by the wearer's breath and thus prevents the diffusion of virus particles across the filter. Since these filters are generally designed to be worn for a couple of hours in environments with very high dust load (grinding wood or stone, spray painting etc.) these filters should easily last for days, weeks or even months worn in a relatively clean health care or community setting. While this is generally not recommended in the manufacturers guidelines, as long as the filter remains dry it should be quite safe [18]. A common misunderstanding is that full dust filters begin to leak. While this is true for gas filters when their absorption capacity is exhausted, dust filters do not leak, since the free space between internal surfaces available for absorption of particles does not increase but decrease, causing them to clog, i.e. resistance to air passage starts to rise which indicates that replacement is necessary.

Effectiveness of high quality particle filtering masks against SARS-CoV-2 infection

As mentioned above, on the one hand there has been some research on efficiency of different masks against community based dissemination of typical droplet and aerosolized infections like flu, rhino viruses and SARS-CoV-1 with mixed results.

On the other hand, there is ample precedent for successful use of well fitted P3 masks even against dry airborne threats like anthrax [20]. The known properties of the SARS-CoV-2 virus suggest that it should be even less likely to penetrate such masks.

Preliminary evidence suggests that SARS-CoV-2 becomes nonviable when dried out [21]. This also fits prior expectation since SARS-CoV-2 has a lipoprotein envelope, which should be denatured by desiccation. This means that very fine particles around 2.5 µm, which are most likely to penetrate a high quality particle filtering mask, should already be dried out at the point when they might be inhaled under most plausible circumstances. And even these particles are retained to 99,95% in a P3 filter with filtering efficiency steeply rising with larger (or smaller!) particle diameters [18].

This forms a stark contrast to the mixed results found for mask use against flu transmission in community settings.

This apparent contradiction is, in the case of FFP masks, easily explained by the general observation that they are often worn in a way that compromises their seal to the wearer's face even if worn by trained professionals [19]. In the case of surgical masks, their inability to form a seal in the first place makes them incapable to reliably protect against droplets small enough to move with air currents (single to double digit µm scale) [17]. These problems are further exacerbated by compliance issues, especially when people are asked to wear masks in their own household [22]. This makes the whole body of research conducted with surgical and FFP masks in community settings highly suspect if applied to easy to use, preformed, reusable masks with P3 filters.

Infection via intact or even damaged skin is highly implausible for SARS-CoV-2 and has never been observed. Food borne spread, while not completely implausible, has not been observed to play an appreciable epidemiological role in SARS-CoV-2 [23]. This leaves direct targeting of the respiratory system via droplet and aerosolized infection

and possibly smear infection to mouth and eyes as the only relevant transmission routes. Transmission via both of these should be severely curtailed by the use of a P3 standard equivalent mask. If the mask needs to be worn with spray tight eye protection or not is currently unclear, since it is currently unknown whether or not infection can occur through droplets entering the eye. A plausible mechanism for this exists, since liquids in the eye are evacuated into the nasal cavity via the nasolacrimal duct.

Therefore, until more research is conducted with well fitted P3 equivalent masks, we should go on to assume that at least P3 masks, worn properly, with appropriate eye protection while maintaining basic hand hygiene are efficient in preventing SARS-CoV-2 infection regardless of setting. The extraordinary evidence required to accept the quite extraordinary claim that a well fitted P3 mask does not prevent SARS-CoV-2 infections in community settings does not exist.

A policy proposal for the eradication of SARS-CoV2

This is a broad outline of a policy proposal that should be able to eradicate SARS-CoV-2 from a given population within a couple of months to a year while allowing nearly complete economic functioning over much of this period:

Over the next few months, infection numbers are maintained as envisioned under the conventional “Hammer and Dance” scenario, while at the same time public funding is used to build up a large scale production capability for 5 to 10 different mask body types to fit most faces, matching P3 equivalent filters and some form of eye protection with the aim to be able to supply most of the population with one properly fitted mask body and 3 to 5 matching P3 filters as well as spray tight eye protection.

While this is no trivial matter, it should be well within the capabilities of industrialized and most emerging economies. Prior to the Covid-19 pandemic, consumer prices for a half mask body were about € 30 to € 60, a P3 filter about € 5 to € 10 and protective goggles about € 20, forming an upper bound of € 100 per person with significant room for improvement through economies of scale. This suggests costs which are only a small fraction of the costs already incurred and still to be expected. Since mask bodies, valves and filter casings can be produced by simple injection molding, quick scale up of production capabilities should be possible by retooling of existing production lines.

Also the availability of raw materials should be no problem since a broad range of elastomeric polymers are suitable for the production of mask bodies, many of which are used in large quantities for other purposes. For example silicone (medical and industrial applications, gluing and sealing), Polyurethane (industrial and building applications, gluing and insulating) and butyl rubber (tyres and industrial applications).

As soon as masks are produced in large numbers they are issued to the population, beginning with health care workers and public facing employees with the directive to wear mask and eye protection whenever people not belonging to their household are (or have recently been) in the same room or within 5 meters.

This cuts down on wear time and thus reduces filter degradation and wearer exhaustion compared to wearing the mask at all times while outside the home without generating much risk of infection. To avoid possible infection via surfaces, everyone is instructed to carry a small bottle of 80% ethanol to disinfect hands

- before and after putting on the mask or taking it off
- touching the unmasked face for other reason
- eating

Ethanol is available in sufficient quantities, since large amounts could be diverted from E5 and E10 fuel production.

Even though this is not standard practice, for the reasons described above (filters remain dry, thus virus particles bound to the filters will stay there [18] and degrade over time) filters can be used for 2 to 4 weeks unless severe contamination with liquids or large amounts of dust has occurred. After use the mask body can be cleaned with soap and water and disinfected with the same alcoholic solution used as hand rub to maintain hygienic conditions. This greatly alleviates both production and logistic demands and allows to supply most of the population in a comparatively short amount of time.

Since people following the policies outlined above are much less likely to become infected, they are also much less likely to pass the infection on to other people much in the same way an immunised person no longer acts as a host to the virus. Thus as soon as about 70 % of the population follow this protocol, an “artificial herd immunity” will be attained and the number of active infections will start to decline even in the absence of additional measures or much earlier if combined with other measures. Thus it is not necessary to provide literally everyone with a mask for this to work. People medically incapable of wearing a mask would also be protected by this effect.

This regimen could be continued with little economic damage until SARS-CoV-2 is eradicated from the population or a safe vaccine has been developed.

Acknowledgements

Many thanks to Michael Albert, Simon Fischer, David Gretzschel, Stefan Heimersheim, Ursula Korten-Schmitz, Ulrike Mazalla and Dr. Gerd Schmitz for their valuable input.

References

1. Wikipedia “2019-2020 Coronavirus Pandemic”
https://en.wikipedia.org/wiki/2019–20_coronavirus_pandemic Retrieved 11.04.2020
2. Stevens H. “Why outbreaks like coronavirus spread exponentially, and how to “flatten the curve” Washington Post
<https://www.washingtonpost.com/graphics/2020/world/corona-simulator/> Retrieved 11.04.2020
3. Deutsche Gesellschaft für Epidemiologie (DGEpi) “Stellungnahme der Deutschen Gesellschaft für Epidemiologie (DGEpi) zur Verbreitung des neuen Coronavirus (SARS-CoV-2)”
https://www.dgepi.de/assets/Stellungnahmen/Stellungnahme2020Corona_DGEpi-21032020.pdf Retrieved 11.04.2020

4. Science Media Center Germany "Auslastung der Intensivstationen: Zahlen aus Deutschland und Europa" <https://www.sciencecenter.de/angebote/fact-sheet/details/news/auslastung-der-intensivstationen-zahlen-aus-deutschland-und-europa/> Retrieved 11.04.2020
5. Bach J. "Don't "Flatten the Curve," squash it!" Medium <https://medium.com/@joschabach/flattening-the-curve-is-a-deadly-delusion-eea324fe9727> Retrieved 11.04.2020
6. Streek, H. et al. "Vorläufiges Ergebnis und Schlussfolgerungen der COVID-19 Case-Cluster-Study (Gemeinde Gangelt)" https://www.land.nrw/sites/default/files/asset/document/zwischenergebnis_covid_19_case_study_gangelt_0.pdf Retrieved 11.04.2020
7. Phua, J. et al. "Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations" The Lancet [https://www.thelancet.com/journals/lanres/article/PIIS2213-2600\(20\)30161-2/fulltext](https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(20)30161-2/fulltext) Retrieved 11.04.2020
8. Kuznalla, R. "The timetable for a coronavirus vaccine is 18 months. Experts say that's risky" CNN <https://edition.cnn.com/2020/03/31/us/coronavirus-vaccine-timetable-concerns-experts-invs/index.html> Retrieved 11.04.2020
9. Pueyo, T "Coronavirus: The Hammer and the Dance" Medium <https://medium.com/@tomaspueyo/coronavirus-the-hammer-and-the-dance-be9337092b56> Retrieved 11.04.2020
10. Chen Shen, Nassim Nicholas Taleb and Yaneer Bar-Yam, Review of Ferguson et al "Impact of non-pharmaceutical interventions...", New England Complex Systems Institute (March 17, 2020) <https://necsi.edu/review-of-ferguson-et-al-impact-of-non-pharmaceutical-interventions> Retrieved 11.04.2020
11. Alexander, S "Face masks, much more than you wanted to know" Slate Star Codex <https://slatestarcodex.com/2020/03/23/face-masks-much-more-than-you-wanted-to-know/> Retrieved 11.04.2020
12. Centers for disease Control and Prevention "Interim Recommendations for Facemask and Respirator Use to Reduce 2009 Influenza A (H1N1) Virus Transmission" <https://www.cdc.gov/h1n1flu/masks.htm> Retrieved 11.04.2020
13. Centers for disease Control and Prevention <https://twitter.com/cdcgov/status/1233134710638825473> Retrieved 11.04.2020
14. Zhang, L. et al "Protection by Face Masks against Influenza A(H1N1)pdm09 Virus on Trans-Pacific Passenger Aircraft, 2009" Emerging infectious diseases <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810906/> Retrieved 11.04.2020
15. Centers for disease Control and Prevention "Respiratory Protection in Health-Care Settings" <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810906/> Retrieved 11.04.2020
16. Wikipedia "Biosafety Level" https://en.wikipedia.org/wiki/Biosafety_level#Biosafety_level_3 Retrieved 11.04.2020
17. Makison Booth, C. et al "Effectiveness of surgical masks against influenza bioaerosols" Journal of hospital infections Retrieved 11.04.2020 <https://www.sciencedirect.com/science/article/abs/pii/S0195670113000698> Retrieved 11.04.2020
18. M3 "Respiratory Protection for Airborne Exposures to Biohazards" Technical Data Bulletin <https://multimedia.3m.com/mws/media/4099030/respiratory-protection-against-biohazards.pdf> Retrieved 11.04.2020
19. Sutton PM et al "Tuberculosis isolation: comparison of written procedures and actual practices in three California hospitals." Infection Control and Hospital Epidemiology <https://www.ncbi.nlm.nih.gov/pubmed/10656351/> Retrieved 11.04.2020

20. OHSA "Anthrax, Control and Prevention"
<https://www.osha.gov/SLTC/emergencypreparedness/anthrax/controlandprevention.html> Retrieved 11.04.2020
21. Streeck, H. "Einzelne Übertragungen im Supermarkt sind nicht das Problem" Zeit Online <https://www.zeit.de/wissen/gesundheit/2020-04/hendrik-streeck-covid-19-heinsberg-symptome-infektionsschutz-massnahmen-studie/seite-2> Retrieved 11.04.2020
22. MacIntyre C. "Facemasks for the prevention of infection in healthcare and community settings" BMJ 2015;350:h694 <https://sci-hub.tw/10.1136/bmj.h694> Retrieved 11.04.2020
23. Bundesinstitut für Risikobewertung "Can the new type of coronavirus be transmitted via food and objects?"
https://www.bfr.bund.de/en/can_the_new_type_of_coronavirus_be_transmitted_via_food_and_objects_-244090.html Retrieved 11.04.2020

What Surprised Me About Entrepreneurship

When I was 24 I had a hard time getting a job as a software developer. As an self-taught engineer, I had no credentials. I was bad at writing resumes and cover letters. And I was bad at interviewing. Then I read [Hiring is Obsolete](#).

If you start a startup, you'll probably fail. Most startups fail. It's the nature of the business. But it's not necessarily a mistake to try something that has a 90% chance of failing, if you can afford the risk. Failing at 40, when you have a family to support, could be serious. But if you fail at 22, so what? If you try to start a startup right out of college and it tanks, you'll end up at 23 broke and a lot smarter. Which, if you think about it, is roughly what you hope to get from a graduate program.

Even if your startup does tank, you won't harm your prospects with employers. To make sure I asked some friends who work for big companies. I asked managers at Yahoo, Google, Amazon, Cisco and Microsoft how they'd feel about two candidates, both 24, with equal ability, one who'd tried to start a startup that tanked, and another who'd spent the two years since college working as a developer at a big company. Every one responded that they'd prefer the guy who'd tried to start his own company. Zod Nazem, who's in charge of engineering at Yahoo, said:

"I actually put more value on the guy with the failed startup. And you can quote me!"

So there you have it. Want to get hired by Yahoo? Start your own company.

"Hey," I thought, "I'm 24. I can game the system! If I start a startup with the deliberate intention to fail after a few months then I can get hired as a software developer."

I'll be 28 next week. On the one hand, things [are going well](#). On the other hand, I'm still waiting for Yahoo's recruitment email.

Here is a list of the biggest things that surprised me about starting a startup.

1. It's easier than I expected

Keeping a startup [alive](#) is harder than startup founders expect. Otherwise there wouldn't be a 90% failure rate.

When I was a teenager, "luxury vacation" meant sleeping in a campsite. "Typical vacation" meant sleeping with a machete for self-defense.

When I was 24, I flew to Shanghai. I rented the cheapest apartment I could find. To get there you take the subway as far west as it will go. Get off the subway and walk another half-mile west. On your left is an abandoned shopping mall with boarded-up stores and broken escalators. On the right is a large compound with a broken turnstile

at the entrance. Walk to the the house at the back of the compound. I slept in the kitchen cupboard.

If you grow up in Asia or Africa, or even in poverty in the USA, experiences like this are the norm. But they're unusual for Computer Science graduates. So when Paul Graham [says](#) "The best way to put it might be that starting a startup is fun the way a survivalist training course would be fun" followed by "When I look at the responses, the common theme is that starting a startup was like I said, but way more so", I have to reverse this advice.

2. Investors don't like hardware

Investors don't like hardware startups. [According to Paul Graham](#) "Out of 84 companies [in YC], 7 were making hardware. On the whole they've done better than the companies that weren't." [According to Paul Graham](#) this overperformance is evidence that YC is biased against hardware companies.

YC isn't alone here. I recently had a surreal conversation with an investor. He basically said "I love your team and you're making lots of money selling hardware, but you can't make billions of dollars selling hardware because Apple will crush you. Besides, Google just bought FitBit for billions of dollars. That means you can't make billions of dollars starting a hardware startup. Even if you could, venture capitalists would never fund you and I can't fund you because they won't."^[1]

The best thing about selling hardware is it provides immediate revenue. If investors won't fund us then we can bootstrap everything. If those same investors won't fund our competitors then we can take our time.

3. Hardware is easier and cheaper than I expected

Starting a hardware startup is easier and faster than I expected in every respect. It's not just us. Experienced investors recently estimated that such-and-such part would cost \$40,000 to make. My CEO got it done for \$12,000 in a rush order.

Maybe this is because my CEO and I speak Chinese and our family is from the Republic of China. Maybe we're unusually scrappy engineers. Or maybe hardware has gotten cheaper recently and the market [hasn't caught up yet](#).

4. Lisp is powerful

Machine learning is a core part of our product. We wrote a system to automate hyperparameter search. Originally this was written in Python, but as it got more and more meta, we ported most of it to [Lisp](#). Within a few months, we had a general-purpose system for hyperparameter search with layered caching for small data^[2]. While it is possible to write this sort of thing in Python, it would have been prohibitively expensive.

5. Younger is better

When I started this company I had 1.5 years of professional software development experience plus 1 year of working part-time at a physics lab where I programmed computers in-between cutting sheet metal and calibrating gamma ray detectors. I was afraid someone middle-aged with decades more experience and savings would crush us. I had it all backward.

It's true that the amount of money you have increases with age, but so do your expenses. Expenses are more important than savings in the startup game. A hypothetical 22-year-old with \$25k/year expenses and 1 year of runway has an advantage over a 40-year-old with \$150k/year expenses and 2 years of runway. The 22-year-old needs only 25k/year to hit [ramen profitability](#). The 40-year-old needs \$150k/year.

This applies even below ramen profitability. 20k/year gives the 40-year-old an extra 2 months of runway. The same revenue of 20k/year increases the 22-year-old's runway by 4 years.

What about experience?

When you're starting a startup, you have to do lots of different things. I've had to write full-stack web apps, native Android apps, native iOS apps, smartwatch apps, firmware for microcontrollers, machine learning systems and a compiler—and that's just my software development duties.

There is no way to know all these things in advance. You have to be adaptable. Older people are not more adaptable than younger people.

Lastly is the effect on one's career. A 22-year-old who flies a startup into the ground has just jumpstarted zir career in software development. A 40-year-old who flies a startup into the ground has postponed zir retirement. A 22-year-old who becomes a billionaire gets to enjoy it for the rest of zir life. A 40-year-old who becomes a billionaire should have done it before working a lower-paying job for the last 18 years.

6. Change

The longer I run a startup, the more I feel my personality drifting away from my engineering friends. They like software less and less every year. Meanwhile, I love software just as much as when I was a teenager, except now I have good [taste](#) and can write software instead of just admiring others'.

My friends also seem increasingly [docile](#). This one isn't them changing. It's me. This is disconcerting, even though 24-year-old me was just as docile as my friends are right now.

Maybe I was always destined to be an entrepreneur. I don't know. I never thought of myself as a "business person". Neither did my friends. But back when I was 24, a CEO asked me the following question before rejecting my job application.

Would you ever be happy working for someone else?

"Of course!" I replied, "That's why I'm applying to work for you!" I guess he noticed something back then. Maybe it had something to do with how my cover letter started "I built a competitor to your company last week...."

1. I appreciate how this MBA was direct about the irrationality of the situation. [←](#)
2. Small data is machine learning with strong Bayesian priors. The most most lucrative application of small data is alpha seeking in quantitative finance. We've considered open-sourcing our hyperparameter search tool. Please PM me if you're interested in discussing this, especially if you work in quantitative finance or a similar industry. [←](#)

Deminatalist Total Utilitarianism

TLDR: I propose a system of [population ethics](#) that arguably solves all major paradoxes, and formalizes some intuitions that prior systems have not. AFAIK this is original but I'm not an expert so maybe it's not.

This idea was inspired by a discussion in the "EA Corner" discord server, and specifically by a different proposal by discord user LeoTal.

Foreward: The Role of Utilitarnism

I don't believe [utilitarianism](#) is literally the correct system of ethics (I do endorse [consequentialism](#)). Human values are [complex](#) and no simple mathematical formula can capture them. Moreover, ethics is [subjective](#) and might different between cultures and between individuals.

However, I think it is useful to find simple approximate models of ethics, for two reasons.

First, my perspective is that ethics is just another name for someone's preferences, or a certain subset of someone's preferences. The source of preferences is ultimately intuition. However, intuition only applies to the familiar. You know that you prefer strawberries to lemons, just because. This preference is obvious enough to require no analysis. But, when you encounter the unfamiliar, intuition can fail. Is it better to cure a village of malaria or build a new school where there is none? Is it better to save one human or 1000 dogs? Can a computer simulation be worthy of moral consideration? What if it's [homomorphically encrypted](#)? Who knows?

In order to extrapolate your intuition from the familiar to the unfamiliar, you need *models*. You need to find an explicit, verbal representation that matches your intuition in the familiar cases, and that can be unambiguously applied to the unfamiliar case. And here you're justified to apply some Occam razor, despite the complexity of values, as long as you don't shave away too much.

Second, in order to cooperate and coordinate effectively we need to make our preferences explicit to each other and find a common ground we can agree on. I can make private choices based on intuition alone, but if I want to convince someone or we want to decide together [which charity to support](#), we need something that can be communicated, analyzed and debated.

This is why I think questions like population ethics are important: not as a quest to find the One True Formula of morality, but as a tool for decision making in situations that are unintuitive and/or require cooperation.

Motivation

The system I propose, deminatalist total utilitarianism (DNT) has the following attractive properties:

- It avoids the [repugnant conclusion](#) to which regular total utilitarianism falls prey, at least the way it is usually pictured.
- It avoids the many problems of average utilitarianism: the incentive to kill people of below-average happiness, the incentive to create people of negative happiness (that want to die) when the average happiness is negative, the sadistic conclusion and the non-locality (good and evil here depends on moral patients in the Andromeda galaxy).
- It avoids the problem with both totalism and averagism that killing a person and creating a different person with equal happiness is morally neutral.
- It captures the intuition many people have that the bar for when it's good to create a person is higher than the bar for when it's good not to kill one.
- It captures the intuition some people have that they don't want to die but they would rather not have been born.
- It captures the intuition some people have that *sometimes* living too long is bad (my dear transhumanist comrades, please wait before going for rotten tomatoes).

Formalism

I am going to ignore issues of [time discounting](#) and spatial horizons. In an infinite universe, you need some or your utilitarian formulas make no sense. However, this is, to first approximation, orthogonal to population ethics (i.e. the proper way to aggregate between individuals). If you wish, you can imagine everything constrained to your future light-cone with exponential time discount.

I will say "people" when I actually mean "moral patients". This can include animals (and does include some animals, in my opinion).

The total utility of a universe is a sum over all people that ever lived or will live, like in vanilla totalism. In vanilla totalism, the contribution of each person is

$$U_{\text{vanilla}} = \int_{t_{\text{birth}}}^{t_{\text{death}}} h(t) dt$$

where t_{birth} is the time of birth, t_{death} is the time of death, and $h(t)$ is happiness at time t (for now we think of it as hedonistic utilitarianism, but I propose a [preference utilitarianism](#) interpretation later).

On the other hand, in DNT the contribution of each person is

$$U_{\text{DNT}} = -u_0 + \int_{t_{\text{birth}}}^{t_{\text{death}}} (h(t) - h_0(1 - e^{-\frac{t-t_{\text{birth}}}{\tau_0}})) dt$$

- τ_0 is a constant with dimensions of time that should probably be around typical natural lifespan (at least in the case of humans).

- h_0 is a constant with dimensions of happiness, roughly corresponding to the minimal happiness of a person glad to have been born (presumably a higher bar than not wanting to die).
- u_0 is a constant with dimensions of utility that it's natural (but not obviously necessary) to let equal $h_0\tau_0$.

Of course the function $1 - e^{-\frac{t-t_{\text{birth}}}{\tau_0}}$ was chosen merely for the sake of simplicity, we can use a different function instead as long as it is monotonically increasing from 0 at $t = t_{\text{birth}}$ to 1 at $t = +\infty$ on a timescale of order τ_0 .

Analysis

For a person of constant happiness h and lifespan τ , we have

$$U_{\text{DNT}} = -u_0 + (h - h_0)\tau + h_0\tau_0(1 - e^{-\frac{\tau}{\tau_0}})$$

It is best to live forever when $h \geq h_0$, it is best to die immediately when $h < 0$ and in between it is best to live a lifespan of

$$\tau_{\text{opt}} = \tau_0 \ln \frac{1}{1 - \frac{u_0}{h_0}}$$

We can imagine the person in the intermediate case becoming "tired of life". Eir life is not good. It is not so bad as to warrant an earlier suicide, but there is only so much of it ey can take. One could argue that this should already be factored into "happiness", but well, it's not like I actually defined what happiness is. More seriously, perhaps rather than happiness it is better to think of h as the "quality of life". Under this interpretation, the meaning of the second correction in DNT is making explicit a relationship between quality of life and happiness.

Creating a new person is good if and only if $U_{\text{DNT}} > 0$, that is

$$(h - h_0)\tau + h_0\tau_0(1 - e^{-\frac{\tau}{\tau_0}}) > u_0$$

Creating a new immortal person is good when $h > h_0$ and bad when $h < h_0$. Assuming $u_0 \geq h_0\tau_0$, creating a person of happiness below h_0 is bad even if ey have optimal lifespan. Lower values of u_0 produce lower thresholds (there is no closed formula).

DNT is a form of total utilitarianism, so we also get a form of the repugnant conclusion. For vanilla utilitarianism the repugnant conclusion is: **for any given population, there is a better population in which every individual only barely prefers life over death**. On the other hand, for DNT, the "repugnant" conclusion take the form: **for any given population, there is a better population in which every individual is only barely glad to have been born** (but prefers life over death by a lot). This seems to me much more palatable.

Finally, again assuming $u_0 \geq h_0\tau_0$, killing a person and replacing em by a person of equal happiness is always bad, regardless of the person's happiness. If $u_0 = h_0\tau_0$ exactly, then the badness of it decreases to zero as the age of the victim during the switch goes to infinity. For larger u_0 it retains badness $u_0 - h_0\tau_0$ even in the limit.

From Happiness to Preferences

I believe that preference utilitarianism is often a better model than hedonistic utilitarianism, when it comes to adults and "adult-like" moral patients (i.e. moral patients that can understand and explain eir own preferences). What about DNT? We can take the perspective it corresponds to "vanilla" total preference utilitarianism, plus a particular *model* of human preferences.

Some Applications

So far, DNT made me somewhat more entrenched in my beliefs that

- [Astronomical waste](#) is indeed astronomically bad, because of the size of future supercivilization. Of course, in averagism the argument still has weight because of the high quality and long lifespan of future civilization.
- Factory farming is very bad. Although some may argue factory farmed animals have $h > 0$, it is much harder to argue they have $h > h_0$.

DNT made me somewhat update away from

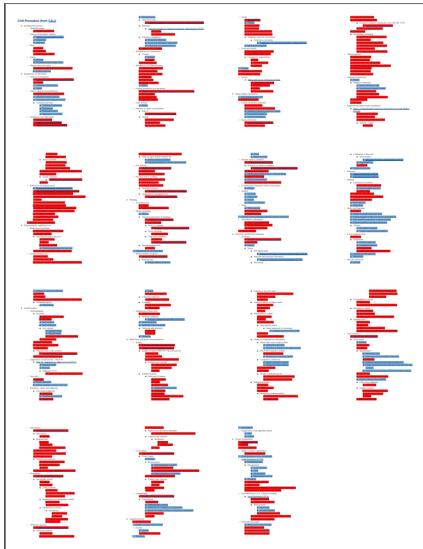
- The traditional transhumanist perspective that living forever is good unless life quality is extremely bad. Of course, I still believe living forever is good when life quality is genuinely good. (Forever, or at least very long: I don't think we can fully comprehend the consequences of immortality from our present perspective.)
- The belief that the value of humanity so far has been net positive in terms of terminal values. I think a random person in the past had a rather miserable life, and "but ey didn't commit suicide" is no longer so convincing. However, I'm still pretty sure it is *instrumentally* net positive because of the potential of future utopia.

DNT might also be useful for thinking about abortions, although it leaves open the thorny question of when a fetus becomes a moral patient. It does confirm that

abortions are usually good when performed before this moment.

Law school taught me nothing

[This is a list](#) of all the topics a standard Civil Procedure class in law school covers, with the concepts I remember highlighted in blue and those I don't in red.



There is a lot more red than blue, which is shocking because:

- I graduated from law school less than a year ago
- I paid slightly more than \$6,000 for my Civil Procedure class
- I received an A in the class

This post is an exploration of possible hypotheses for why this happened.

Epistemic status: spinning hypotheses from personal experience, so take lightly.

User error

In a world where even surgeons sometimes leave their tools inside their patient's body, the ghost of user error haunts us all. Maybe I am the problem, and my skills, situation, or environment are not suited for remembering basic facts that I paid \$6,000 to learn. There are a few ways in which I could be failing at this task.

First, maybe I am not good at school. I reject this hypothesis out-of-hand, mostly because I have a fragile ego, but also partly because it's not true. I scored in the 99th percentile on the LSAT, attended a top-3 law school in the US, and graduated with honors. All these points don't insulate me from being bad at school, but the *vast majority* of law school attendees are even worse at school (not learning) than me. You know that scene in *Taken* where Liam Neeson says "I have a very particular set of skills..."? That's me with school; I am *good* at school. If I can't retain this knowledge, the entire system might as well commit its tax-advantaged resources to rolling boulders up hills for all the good they're doing.

Second, maybe I have framed the problem incorrectly. It's true that retaining material from a Civil Procedure class is not a standard validated benchmark. However, I promise you that I'll receive similar results on any benchmark you devise for 90% of the law school classes I took; I only picked Civil Procedure for this exercise because it is alphabetically first in the list of required law school classes I took. My problem is not that I don't remember Civil Procedure; my problem is that I remember almost nothing.

Third, maybe I didn't study properly. This, unlike the previous two, is a plausible hypothesis. I did work my way through about ~70% of the study tactics the professor recommended, including attending class regularly, completing all the readings ahead of time, taking good notes, and reviewing all the material at the end of class. But if I had done more, like made flashcards, practiced with spaced repetition, joined a study group, or just spent more time with the material, I would remember more today. Some other students do remember much more than I do, partly because they used more intense study tactics.

School is just a signaling mechanism

As [Bryan Caplan](#) says:

There are TWO solid business reasons to pay extra for educated workers. One is that education teaches useful skills, *transforming* unskilled students into skilled workers. This is the standard "human capital" story. The other reason, though, is that education *certifies* useful skills, helping employers distinguish skilled workers from imitators. This is the "signaling" story. [...] My best estimate is that signaling explains 80% of the payoff.

In other words, students go to school because employers hire people with degrees, not because they want to learn. According to this hypothesis, it is perfectly rational for me to only remember odds-and-ends from my law school education; the only thing I need to retain is my degree, not any knowledge.

Legal employers largely behave in accordance with the signaling theory. It was common knowledge among the student body that (A) the highest paying or most prestigious jobs mostly go to graduates of top law schools and (B) as long as you are a graduate of one of the top law schools, you can get a high paying or prestigious job without having excellent grades. A prestigious degree essentially equals a good job. You might need a prestigious degree AND excellent grades to get a very rare job (e.g., a Supreme Court clerk, of which there are less than 40 each year), but most people were not aiming for those jobs.

Schools also behave in accordance with the signaling theory. Most blatantly, five of the top fourteen law schools in the US have adopted [idiosyncratic grading schemes](#) that make it hard for employers to understand exactly how competent a student is (inasmuch as grades ever indicate competence). These grading systems range from no fixed curves to pass/no-pass grades to UChicago's (hilarious) "numeric grade with median of 177." There is a discredited security strategy called [security through obscurity](#), where you secure your system by keeping the system secret, instead of making it resistant to attacks. This is analogous: grading through obscurity.

This privilege of using a grade scheme that obfuscates is exercised by market-making schools, whereas market-taking schools largely stick to A/B/C/D/Fs. But even these latter schools, while not prestigious enough to make up their own grading schemes,

have an array of schemes to choose from to make their students more attractive to employers. Employers judge schools by how many of their graduates find gainful employment. Schools want to maximize this number so that its graduates look like catches. In good times, they'll do so with good ol'-fashioned definition massaging, defining "long-term" employment as employment lasting a year and "J.D. advantage jobs" to include jobs such as FBI agent and paralegal, for which a J.D. is only marginally more advantageous than breathing oxygen. In bad times, such as the Great Recession, they'll do so by employing graduating students themselves in temporary positions.

Students, too, behave in accordance with the signaling theory to some extent. First year (1L, in the jargon) grades matter quite a lot for most student's first post-graduation job, so students do work hard in the first year. However, most 2L and 3L (3LOL, in the jargon) students skate by, trying to apply the bare minimum effort possible.

Poor teaching

I remember more from high school than I do from law school. I remember random facts that I have never used since high school, like the fact that the ideal gas law is $PV = nRT$ and the French verb *venir* requires special conjugation in the past tense. This is not as bizarre as it seems at first glance.

[A recent post](#) explained that the more people you need to coordinate, the less bandwidth you have for transmitting information to them. If you are talking to a handful of people, you can have a deep conversation, but if you are talking to thousands of people, you shout a short slogan at them. I think the same is true across time. If you are teaching a student in your class, you can give that student millions of words of information. By the time a student goes to his first job, that student remembers thousands of words of information. By the time a student reaches the peak of her career, that student remembers hundreds of words of information. By retirement, tens of words, if that.

This is not pure conjecture. [Middle school students can lose large chunks of their reading and math knowledge over summer break](#), even though those are basic skills continuously reinforced by real-world tasks! The information taught in law school bears only a passing resemblance to real-world tasks, and is almost never reinforced after graduation. Of course it is forgotten. Given that students will forget most things they learn, teachers should ideally teach in a way that mitigates the damage of the decline. The most important information should be impossible to forget.

Grade school teachers are better than professors at forcing unforgettable information into students' minds. (They're not perfect, but they're much better.) They'll use alliteration, repetition, rhymes, and chanting to get their students to memorize information *forever*. My high school chemistry teacher would say "piv nurt" every day for a few weeks when teaching us $PV = nRT$. That's not a good mnemonic; it's just two meaningless syllables! But the repetition worked. I will remember piv nurt on my deathbed. Same for DR MRS VANDERTRAMPP. And SOH CAH TOA. And "always add +c." And many other little factoids. My professors, on the other hand, when telling us important information, would preface it with "And if there is one thing you remember from this class, it is this..." That's it.

Law school professors will hate changing how they teach. They didn't get into their prestigious jobs to develop mnemonics for fully-grown adults. It's beneath their dignity. The responsibility for retaining information rests with their students. I get it. It does seem silly to suggest that famous credentialed lawyers spend their time treating their adult students like high-schoolers. But the current situation, where these very same famous credentialed lawyers spend months teaching students, only for the students to forget most of what they've learned within weeks, doesn't seem like its befitting their dignity either.

The point was to learn meta-skills

The canard that law school repeats often is "we teach you how to learn." It's fine that I remember almost nothing from my classes, because that was never the goal. The goal was to learn the skills needed to learn the law. Now the rest of my life can be an orgy of learning, limited only by my initiative.

Learning about anything, including meta-skills, requires looping through the below list many times:

- I do a task
- I receive feedback on the task, telling me what I did well and what I did poorly
- I update my mental model of how to do the task

Law school fails at this loop because it makes sure tenured professors rarely have to besmirch themselves by giving feedback.

Law school instructors, although all human, like the Eloi and the Morlocks can be split into two classes. There are tenured (or tenure-track) professors, and there are the rest. You can distinguish them by the size of their offices.

If you take a class with a tenured professor, it is extremely unlikely you will receive feedback on tasks you complete. It happened to me maybe twice in three years, and I'm including feedback given by the professors' TAs in that count. Even more annoyingly, you might not even have any tasks to complete! In almost all classes, the student's only job throughout the semester is to complete the reading, and if called upon in class, to answer reading-comprehension-type questions about the reading. (Law schools refer to this as the Socratic Method, but it's more like read-and-regurgitate.) Neither the reading nor the perfunctory questioning that follows is a meaningful task on which law students need feedback, given that law students are already good at reading and reading comprehension.

If you take a class with a non-tenured professor, on the other hand, you can receive lots of excellent feedback. However, these have their own problems. First, they require small student-to-teacher ratios so students can receive feedback, which means law schools need to hire lots of teachers to run them, which means they are expensive, which means law schools are wary of having too many of these. Second, the subject matter taught by non-tenured professors doesn't always overlap with that taught by tenured professors. Students end up experiencing a classic Catch-22: one can either properly learn an unwanted subject, or poorly learn a wanted subject. Third, almost all these classes are structured as practicums where students solve problems for real or fictional clients. Especially when the clients are real, these classes quickly devolve into base concerns of "How do we fix this client's problem?" rather than the lofty aim of

"Let's learn how to learn." On the whole though, these classes are much better than classes with tenured professors.

Summary

Below are my best guesses for how much each of these hypotheses explains why I don't remember anything.

- [25%] User error because I didn't study enough
- [30%] School is just a signaling mechanism, so remembering actual content would be pointless
- [40%] Poor teaching means I have and will continue to forget information at the whims of my subconscious
- [5%] The point of law school was to learn meta-skills, not information

Ways you can get sick without human contact

(TL;dr: food, animals, and yourself are all carriers of disease that *the average person gets ~.2x/year normally; if you include immune reaction despite non-illness, the rough base rate is probably anywhere from .02x/year to ∞ x/year and depends heavily on person; my own odds seem like ~10:1 non-COVID to COVID given a symptom, but yours are different; I'm not a doctor and don't have metis on this, maybe I say stupid things*)

Now that people have been shut away for weeks, they've (reasonably) expected to stop getting colds and things. But I've now seen at least 3 cases, one with an extremely high level of quarantine, where someone still got a cold-like illness. This affords a few hypotheses:

- Perhaps the base rate for sickness with no human contact is not as low as we thought
- Perhaps disease spread takes more advantage of tiny amounts of contact than one would expect from a model of p(illness) \propto amount of pathogen contacted
- Perhaps people are terrible at quarantining

I don't think it's the last, and the second one is interesting but hard to investigate, so this post will be about the first hypothesis: can people easily get sick while alone?

I'm going to sketch a catalogue of the options and roughly divide these apparent illnesses into two groups: those spread by significant external amounts of pathogen vs minor internal amounts of pathogen. (Base rate estimates at the end.)

Significant external amounts of pathogen

The three main sources of these illnesses are food, fauna, and fecal.

Food poisoning can be caused by lots of different types of bacteria. For example, Campylobacter, Clostridium, Staphylococcus, and E. Coli seem to be some of the main culprits, but there are many others (and [80% of cases](#) appear to be caused by agents we haven't identified!). These present like "stomach flu", often for 1-6 days after 1-10 days of incubation.

[Apparently](#) a lot of the diseases dogs spread overlap with food poisoning bacteria (and even norovirus, which also presents as stomach flu).

A lot of fecal-to-oral infections [appear](#) to be from similar bacteria. (Edited for clarity) For the purposes of this post, we're ignoring diseases from *other people's* fecal matter, and some of these are obviously spread person-to-person, like adenovirus. But the four listed above (can I call them the Big Four since they keep coming up?) appear to naturally occur in the lower digestive tract, so you can presumably self-contract

them unless there's some effect where you have serious immunity to the specific strains in yourself.

There are at least some others.

- [Fungal pneumonia](#) looks pretty nasty
- Streptococcus A is usually spread person-to-person but is also [found on the skin](#) and [regularly causes food poisoning](#) and can thus probably be acquired from oneself
- Mold and “stuff in dirt” also seem like prominent options, though maybe mostly minor symptoms
- Allergies are definitely minor but can mimic colds in some circumstances

In general these external infections seem mostly bacterial, since fungi are typically too weak to cause noticeable illness, and viruses mostly need other humans to reproduce unless they have animals. Not sure how many protozoans are in this group: I'd think lots, but haven't heard of any.

Minor internal amounts of pathogen

This is the category for a bunch of weird things that may compound on each other, which you don't really find much about in the medical literature.

The factors I see here are:

1. Fluctuating native pathogens
2. Variation in immune function
 1. Between people
 2. From health
 3. As calibration to ambient disease risk

As an example, you have a bunch of candida yeast on and in your body, not very deep because high temperature harms it. It is sometimes but very rarely invasive ([maybe because mammals are specifically warm-blooded to fight fungi!](#)). But in immunocompromised individuals, especially AIDS patients, suddenly invasive infections become vastly more common, and many of them have 50+% mortality. But they aren't *constant*, so this must be responding to various fluctuations in candida infectiousness and immune function.

Obviously variation in immune function also applies to the external pathogen sources from the above section, but the application to native pathogens is especially interesting. Is it plausible that immune function has a wide enough range of function that internal pathogen sources may cause a significant number of flare-ups in various circumstances much less extreme than AIDS?

Here's the pitch for plausibility.

Native pathogens are bacterial (especially the Big Four), fungal, and [viral](#); mostly they're on the surface of skin or mucosal membranes, but there is some penetration and humans may have a fair amount of internal viral load. We know highly-immunocompromised people regularly become infected by these native pathogens. But if normal people undergoing fluctuations in their viral load or immune function from health had these flare-ups, I don't think they'd be very distinguishable from colds

or stomach flu. To further complicate things, I think it's a reasonable hypothesis that your body may calibrate its immune response to what seems like the ambient disease load of your environment. Thus in times of high disease load you may have immunosensitivity and "sicknesses" from large immune response to small pathogen challenge (similar to immunosensitivity from allergies), and in times of low disease load you may be "immunodesensitized" and be at higher risk of infection. Evidence for a high rate of phantom colds, or partial viral colds being replaced by self-infection, seems extremely hard to come by, and I don't think the absence of evidence is much evidence of absence.

Lastly, there's some chance with such a large virome that chronic flare-ups may be more common than we think. Herpes is a classic. But other diseases may just be finicky: [supposedly](#) 2-10% of Campylobacter bacterial infections have sequelae or chronic presentation. [One highly-immunocompromised boy](#) was shedding influenza A from his stool for >2 months and from his respiratory secretions for >1 year, strongly hinting at some sort of chronic infection even though influenza A is supposed to be incapable of this. [At least a few people](#) now think chronic viral infections are a common cause of chronic fatigue syndrome, perhaps from mutation to [non-cytolytic form](#) or [abortive infection](#). Probably non-cytolytic ones are too weak to cause flare-ups, but this is the kind of weird edge case that, after realizing it's very difficult to obtain evidence against, makes me wonder if viruses do lots of weird stuff we just haven't classified yet. (Adenovirus-36 [may cause](#) a lot of obesity, h/t Adam Scholl—that should probably make us pause and consider the state of our knowledge.)

Base rates and takeaways

This is all somewhat useless without base rates that can help us infer how likely it is that our quarantine has failed vs we're just letting food sit out too long. Unfortunately, it's very difficult to get base rates on what we care about, so I'm going to have to resort to a lot of hand-waving.

A sampling of those that were reported: E. Coli infections are at [0.1%/yr](#) in the US. Supposedly food poisoning is [15%/year](#). Norovirus is about [5%/year](#), but probably most of that is person-to-person. Infections from pets are supposedly [1%/year](#), but most aren't serious. For comparison, flu is usually [3-15%/year](#).

Stomach-flu-like symptoms will be at least the food poisoning cases plus other cases. That's 15+%. If we relax from intensive cases to include the kind of minor confusing nausea I get once a year, I'd guess the rate goes up to about once/year based on personal experience. If there is weird immune modulation, I could see this being once per few months.

Common-cold-like symptoms don't seem to come up aside from immune modulation, allergies, and supposedly very-low-base-rate things like strep on skin. Healthy people probably have skin virus infection rates like .001-1%. Standard partially-immunocompromised people (that just get sick more than normal) may be more like .005-5%. I'd guess weird immune modulation could take that number to 1-200% if it exists (if it were more than 2x/year I'd think people would start noticing). But there's also allergies, mold, stuff I've missed, and all these exacerbated by overreaction to immune challenges. Empirically, I have ~thrice a year that I seem like I'm just starting to get sick and it turns out to be allergies or goes away mysteriously, and I think most people probably have this less but some have it substantially more. I could

imagine this rate being like 5x if my body was on high alert, and even a few times a year the symptoms mimicking an actual mild cold rather than a tease.

Of course, if your body can't pick up on external disease cues very well, or it only does so with respect to actual ambient pathogen load rather than leading mental indicators like disgust integration or anxiety tracking, then probably these base rates go down. If you're someone who gets sick a lot, they might be higher.

So I expect base rates of minor flu-like symptoms not contracted from another person to be about .1-10/year depending on person, and base rates of minor cold-like symptoms not contracted from another person to also be about .1-20/year depending on person (again, if more, we'd see that).

COVID ambient rates are (my guess) around .1-5% most places (more like 5% in the Bay). If you've been doing a good quarantine, your rate is probably about the rate it was when you started, e.g. I think I started around .1% and so that is probably about my current rate (I could easily be asymptomatic). **So I think if I were very healthy, my model didn't expect much mental immunomodulation, and I got cold or flu symptoms, I'd be about as likely to be seeing a COVID onset as seeing a false alarm. As it stands, being not super healthy and a little more sympathetic to various psychosomatic control theories (writ large, not just placebo), I think it's more like ~10:1 that I'm seeing a false alarm.**

Given the fact that these numbers are certainly going to be off by about an order of magnitude, obviously I would still monitor very closely and take thoughtful precautions that make sense both in worlds where I have COVID and don't have COVID (don't want to give housemates another sickness to make them more susceptible). If I hadn't been as careful with my lockdown I'd also consider getting sick as evidence that I should tighten things up a bit where possible. Also, how textbook the symptom match with COVID was would heavily effect this estimate. Etc etc. But I hope this helps you get a handle on how to grapple with these numbers and the implicated policies.

Research on repurposing filter products for masks?

A local organization is looking to produce masks for the hospital, and they reached out to me for guidance. Since this is outside of my expertise and I haven't been following all of the recent research, I'm passing the questions along here.

We are looking for some valid research on the use of commercial products for face masks. We have a couple of designs that can fit over or have a pocket to insert an additional filter. I have found multiple sources online that refer to furnace filters, air filtration system filters, even HEPA vacuum bags, as ideal products to use to either completely comprise a face mask or be inserted into a pocket between layers of fabric to provide extra protection.

My questions are:

- Which of these products are or are not safe to be breathing through? The only caution I have found is to be careful to check for fiberglass in the content.
- Which are most effective, or does it come down to the MERV/MRP rating? We have been looking at filters at MERV 13-16 and/or MRP 2500-2800 in range. Are these safe as long as they can be breathed through?
- Can these types of materials be layered to provide extra protection i.e. is layering two ply of MERV 11 like a MERV 22 or is there a loss ratio? Is there a formula which can be applied?
- Could other materials be used to enhance the filter's effectiveness, safety, smell, etc.? I have seen models that added an activated charcoal filter over the top, coffee filter, etc.? I have also heard that layers create a dispersion effect that increases the protective factor i.e. a thick bundled knitted scarf has a higher protective factor because of the density.

There is so much out there but it is mostly DIY crafters, and I am interested in a more empirical opinion!

Where should LessWrong go on COVID?

About 6 weeks ago the LW team chose to go really hard on COVID-19 (including contracting with me), for reasons [discussed here](#). The time for that seems to be winding down; people-besides-me are asking fewer covid questions and more non-covid questions, my posts get fewer comments every day, and there's not the same urgency adding to the links DB. Looking at people I know in the real world seems much the same: after a mad sprint to prep and then adjust to covid, people are settling in and catching up on the rest of their lives. So I and the LessWrong team are taking this moment to reevaluate what LessWrong's role in covid response should be.

My question for you is: where do we go from here? Should we slip back to normal, except with a covid section? Relegate covid work to personal posts since it's not timeless? Do you want practical posts like "how to disinfect a mask?", or big picture things like "what does the world look like in a year?"?

What would help you, specifically, right now? In a month? In a year?

Answers I've gotten from people I polled before posting:

- Answer questions with answers that can probably be synthesized from academic papers, e.g., "What is my risk from delivery food, given a certain prevalence?"
- Answer questions informing near-term behavior, e.g., "How do you tell when it's time in your specific region to start relaxing your protocols?" and "Which specific things do we think will be instant vectors the moment things open up?"
- Creates guides for questions about long-term behavior, e.g., "should plan on hibernating for a year and returning to my job, or retool for something entirely different?", "should I move out of [major metropolitan answer] for a year?", and "should I move to a more functional country forever?"
- Dataset on politicians/leaders/journalists etc evaluating how they handled covid
- A map of current efforts
- DB of coronavirus data sets

And remember that if you're entirely sick of this, you can use the new filter tool on the frontpage to make sure you never see covid content.

Market-shaping approaches to accelerate COVID-19 response: a role for option-based guarantees?

This is a policy brief directed at decision makers in the UK government, with a view to accelerating production of tests, drugs and vaccines for COVID-19; but it could be adapted for a wide variety of countries, products and crises. Critical feedback is very welcome.

Thanks to Sam Hilton, Tim Colbourn and anonymous others for input on previous drafts.

Summary

Effectively tackling COVID-19 will require rapidly scaling up the production of diagnostic tests, pharmaceutical treatments and vaccines. In each case, preparations for large-scale manufacturing, such as building factories, are typically delayed until the product is proven safe and effective. This makes sense from a commercial perspective, but incurs great costs in terms of lives lost and damage to the economy.

There are several potential solutions, but the most promising appears to be “option-based guarantees”. In essence, the government commits to paying a proportion of the manufacturer’s preparation costs should the product turn out not to be viable. (If the product is viable, it can be sold as normal.) This reduces the risk to the company while maintaining an incentive to produce a high-quality product quickly and at scale.

The problem

The UK, like most of the world, faces an urgent need for increased COVID-19 testing capacity. As more people become infected and recover, large-scale screening for past exposure (antibody tests) will be necessary, but the shortage is especially acute for tests of active infection (“swab” tests). The government recognises this need and has [set a target](#) of 100,000 tests per day – a combined figure for antibody and swab tests – by the end of April. While this increase is welcome, safely bringing the country out of lockdown could require far more widespread swab testing, potentially as many as 10 million tests a day. Post-lockdown policy is still being developed so this may become the strategy within a few weeks. The UK should prepare now by deciding how market forces can be leveraged to rapidly scale up testing.

A long-term solution to the crisis will involve effective pharmaceutical treatments or vaccines (ideally both). Promising candidates have been identified, but most will take at least several months to complete Phase 3 studies – perhaps less if [“human challenge trials”](#) are permitted. Since many products will not prove viable, companies have little incentive to invest in production facilities before the product achieves regulatory approval. Thus, scaling up production is likely to add a few more months to the overall timeline, costing thousands of lives and billions of pounds of lost GDP in the UK alone – far more than the cost of preparing to manufacture products that do not end up reaching the market.

Potential solutions

There are several ways to address this problem.

1. Prizes

The government could offer financial rewards for solutions to supply shortages. For example, companies could compete to offer the best idea for rapidly scaling up vaccines, and a contract to produce them could be part of the prize. By only paying out for the best solution (and perhaps not any, if certain criteria are unmet), this can be a fairly cheap option that incentivises innovation. However, there is necessarily a substantial delay between announcing and awarding the prize; and because the “losers” get nothing, there may not be adequate financial incentive to participate.

2. Public-private partnerships

PPPs can be an effective means of achieving social objectives, such as building infrastructure, by sharing the risk between government and private companies. However, they generally take a long time to negotiate and implement. Unless this process can be greatly accelerated, they are unlikely to be sufficient to ensure the most urgent needs are met in the current situation.

3. Direct purchase orders

The government could pre-order tests, pharmaceuticals and vaccines directly, well before efficacy or safety is established. This would legally guarantee that producers have a market and that the company will supply the product, thereby reducing risk to both parties.

However, this requires the government to first identify suppliers and producers, negotiate prices, and make orders. The process for governmental purchasing is complex, and purchasing something from a new vendor, or purchasing products not shown to be safe, will potentially be a violation of the UK’s public procurement policy. It is also likely to be wasteful as many final products will be unused, and it gives little incentive for producers to improve quality, speed, or cost-effectiveness through innovation.

4. Option-based guarantees

A new approach is for governments to enter into agreements with companies using [“put” options](#). A “put” (as in “put up for sale”) gives the holder the right, but not the obligation, to sell an asset at a specified price, by (or on) a specified date, to the provider of the put. So the government could supply a put option giving companies the right to sell certain items (e.g. vaccine factories, drugs, or diagnostic tests) to the government for a certain percentage of the cost of making them. Because there is no requirement to exercise the put, companies could sell viable products as normal, and would only use the option if their product turns out to be non-viable.

For example, suppose a manufacturer wishes to produce 100 million of a new type of test, but is delaying production because the product is currently being evaluated. They could approach the government, which could agree to provide a put option for, say, 90% of costs, capped at the company’s initial project cost estimate. If the test is found viable, the company would not exercise the option, the government would pay nothing, and the company would be able to sell the tests normally to the NHS and

others. If found non-viable, however, the company would have an incentive to stop production and exercise their option. At that point, a financial audit of costs would take place, and the government would accept delivery of any items purchased, built, and/or produced in exchange for 90% of costs. A further independent evaluation might be useful to resolve disputes about reasonableness of costs.

This approach has a number of advantages.

1. Commercial companies can continue to use traditional, non-governmental methods for financing and constructing a product without any government supervision.
2. Companies will be willing to take a larger risk in manufacturing not-yet-proven technologies, because the costs (to the company) of failure are reduced. This incentivises starting production earlier.
3. Both haste and high quality are still incentivised through normal market mechanisms: being the first and/or the best product on the market will increase sales and therefore profit.
4. Compared to some alternatives, it is relatively cheap.
5. It could potentially be implemented quickly – a very important consideration in the current circumstances, especially for diagnostics.

Recommendation

Overall, guarantees based on put options seem to show the most promise. When the viability of the final product is uncertain, they provide a relatively quick, low-cost way of incentivising the rapid production of new technologies. However, options 1-3 are also worth exploring further, and the optimal approach (or combination of approaches) may vary among products, time periods, and companies. For example:

- Prizes could work – perhaps alongside other incentives – when innovative solutions are likely to be needed, such as point-of-care tests for active infection, new vaccines, or new methods of scaling up production.
- PPP could be appropriate for less urgent and fairly low-risk products. Antibody tests, and drugs that are very likely to be used but will not run out soon, may fall into this category.
- A direct purchase order for a certain number of a certain diagnostic test could be effective if the safety, accuracy, cost and quantity required are known, the company is trusted, and the paperwork can be done quickly. This may be less risky than hoping a company will respond to financial incentives.
- A put option on production facilities (not final products) could be the best alternative for diagnostics, drugs and vaccines that are promising but whose viability, large-scale manufacturing methods and/or quantity required are substantially uncertain.

Over the coming weeks and months, making the right choices could save thousands of lives in the UK and millions around the world, while enabling economies and communities to reopen.

Annex 1: Potential variations on standard put options

Declining payout

The payout for the put options could be declining over time, so that the payment is, say, 95% at the outset, and declines by 1% per month. This will incentivise companies to exit as soon as possible if they think the project will fail.

Priced contracts

The government could decide to charge for the contracts, to dissuade unqualified or undercapitalised companies from taking huge immediate risks with small probabilities of paying off.

Early-ending bonus

Alternatively or additionally, there could be payments for ending the contract early. In this case, the government might refund a portion of the initial payment or pay some fixed amount if the company decides to end the option without exercising it. This would create another incentive for companies to move quickly, and reduce uncertainty and risk on the government's side.

Annex 2: Key questions about option-based guarantees

Q1: Isn't this a giveaway to corporations?

A: Yes, but in a sense it is a minimal giveaway. It does not subsidise companies to undertake projects that they expect cannot succeed, but does allow them to move forward schedules for production. Under the circumstances, it seems worthwhile.

Q2: Isn't this wasteful?

A: Yes, it is nearly certain that some items produced will not be viable, so the government will pay for unused products. However, the companies have an incentive not to spend more than needed, since they recover only part of the costs; and most importantly, the successful products will be available far faster. (The guaranteed percentage of costs can be adjusted to reach the desired trade-off between avoiding waste and hastening development of the needed product.)

The government may also be able to reduce the costs of the program by reselling some items: for example, a plant designed to manufacture an ineffective vaccine could eventually be adapted to produce an approved vaccine. There have been intermittent shortages of other vaccines, pharmaceuticals and tests, so excess capacity may not be entirely wasted.

Q3: Won't there be fraud or unnecessarily high costs?

A: There is a risk that some companies will try to take advantage of the programme. The put options, however, will pay less than the cost, so there is a reduced risk that companies will attempt to participate if they do not think the product has a reasonable chance of success. (Again, the right balance can be struck by adjusting the guaranteed percentage.)

For subcontracting, the structure of the payout means that the company, not the government, takes on the risk that costs will be considered excessive, so they have incentives to ensure they are not overpaying.

Q4: Won't safety and quality suffer?

A: With less "skin in the game", there may be a reduced incentive for companies to be successful. But if a product is viable, they will not want to exercise the put option, and they will have the same incentive to ensure quality and safety as they would otherwise.

Q5: How do you ensure the final product is affordable?

A: Put options do not ensure the final test, drug or vaccine is available at a reasonable price, but nor do they preclude price controls. This is an important but entirely separate issue that applies equally to products developed through other means. It is worth noting that, while the product must be cheap enough to roll out at massive scale, the price must also be high enough to reimburse the costs of constructing the production facilities.

The Chilling Effect of Confiscation

The federal government has been seizing mask imports:

- Three million, on arrival at the Port of New York and New Jersey on 2020-03-18 ([Boston Globe](#))
- Four hundred thousand, on arrival at JFK airport; 100k on 2020-04-06 and 300k on 2020-04-16 ([USA Today](#), [NY Daily News](#))
- One million, while in transit to South Florida firefighters last week ([The Miami Times](#), 2020-04-24)

Some states have responded by importing masks and other personal protective equipment (PPE) outside of standard channels:

- Massachusetts asked the owners of the New England Patriots to import masks on the team plane ([Wall Street Journal](#), [Time](#), 2020-04-02)
- Illinois chartered multiple FedEx flights (Chicago Sun-Times [2020-04-14](#), [2020-04-19](#))

And there's an article in the New England Journal of Medicine about a hospital administrator working to line up a shipment and [narrowly avoiding](#) confiscation.

Much more impactful than what these seizures are pushing people to do, however, is what they're pushing people not to do. A hospital administrator or governor who could be prioritizing sourcing PPE has many other things they could focus on, and if the PPE they import has a large chance of getting seized they're going to prioritize those other things on the margin. Independent importers likewise can't risk having their goods seized without payment, as Indutex reported ("[Let's not forget I paid \\$4 million for this product on March 18. I don't have any money and I don't have any product and there's people that are asking for it.](#)") Making deals with factories and prepaying so they can ramp up production is super valuable, and we'll see less of it the more risk there is of having your goods taken.

So much of this crisis has been [poorly handled](#) that any individual bit I pick to look at seems like it's not worth focusing on, but this policy is not just foolish, it's foolish in a way that you might have thought conservatives might have avoided? Like, conservatives saying we should avoid lockdowns and let people decide risk for themselves is the sort of policy disagreement I'd expect to have, but using the federal government to grab things, discouraging private initiative and investment, really?

Would 2009 H1N1 (Swine Flu) ring the alarm bell?

Introduction

In order to assist in anticipating and responding to the next pandemic, I created an ["alarm bell" questionnaire](#) to suggest whether or not one should behave as if the world is about to experience a stock market-crashing pandemic.

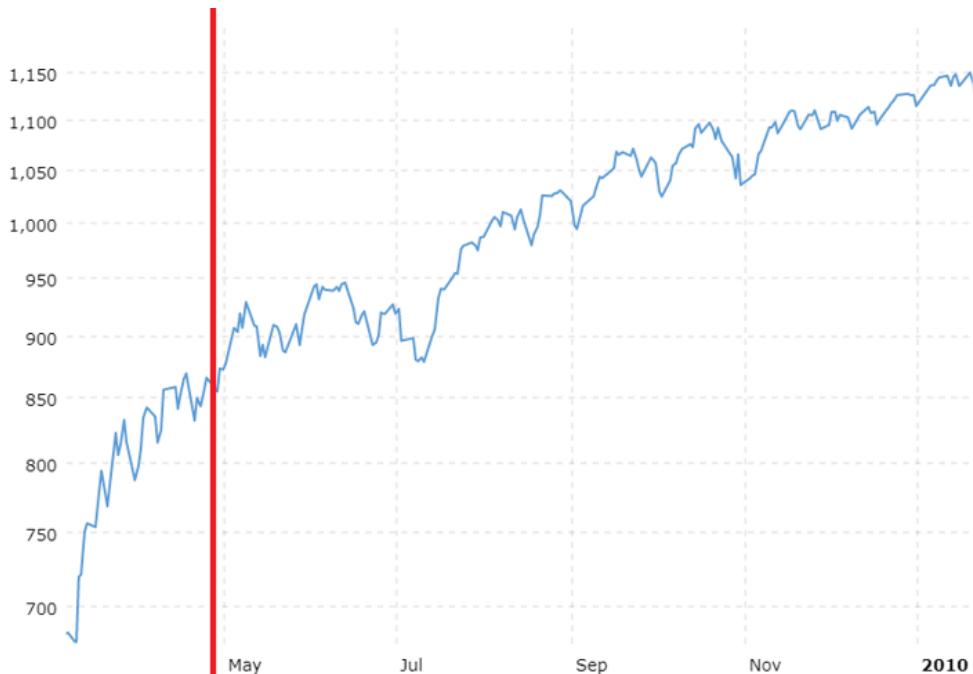
To evaluate whether it is correctly calibrated, my intention is to investigate whether the "alarm bell" would have rung during historical pandemics, and whether the stock market would have crashed before it rang, after it rang, or not at all. I have performed the first test, based on the 2009 H1N1/swine flu pandemic. The stock market did not crash, and in my analysis the alarm bell did not ring - but see the comments for caveats.

This is because 2009 H1N1 was much less deadly than COVID-19, and it also did not shut down the world economy. The "alarm bell" rings if 13 of the 16 below criteria are operative, but only 11 of the 16 criteria were ever met.

This is one point of evidence in favor of this alarm bell being correctly calibrated. I would also like to investigate this question for Ebola, MERS, SARS, HIV/AIDS, the 1968 Hong Kong flu, and (as far as it's relevant) the 1918 Spanish flu.

Stock Market Trends in 2009

The stock market was at a low point in 2009 around March 6th, but had begun a strong upward trend that would last the rest of the year by mid-March, when the first cases of swine flu were confirmed in Mexico. Although the New York Times reported [stock market tremors](#) related to swine flu on April 27th, 2009, looking at the data for the whole year, the daily fluctuations for the year look pretty much like a random walk (April 27th highlighted in red):



Factor 1: Transmissibility

- Does the disease appear to spread from human to human?

Yes, as with all H1N1 flu.

- Does the disease spread via indirect contact (coughing, sneezing)?

Yes, as with all H1N1 flu.

- Is there any evidence that the disease is transmissible before or just after the start of symptoms?

Yes. According to a [fact sheet](#) released by the state of New Hampshire, "People with H1N1 flu virus infection should be considered potentially contagious one day before the onset of symptoms and as long as they are symptomatic, and possibly up to 7 days following the onset of illness."

- Do any academic papers, especially in the Lancet or the New England Journal of Medicine, suggest the possibility of "risk of much wider spread," "exponential growth," or use similar phrases? Try searching "[DISEASE NAME] exponential growth" on Google Scholar.

A [paper published](#) on the 14th of May 2009 estimated that the "reproduction ratio was less than 2.2 – 3.1 in Mexico." For comparison, [R0 for COVID-19 is estimated at 1.5-3.5](#). I count this as yes as of on or before the 14th of May 2009.

Factor 2: Harm

- Is there evidence that the disease has a 3% or higher case-fatality rate among those 60 (or even younger), or 1%+ case-fatality among those 50 (or younger)?

This is one of the trickiest ones, because in early days when case rates are not known accurately, it's hard to predict. It's important to distinguish between confirmed case fatality rate and overall case fatality rate, and global vs. age-related CFR.

[One recent scholarly global CFR estimate for COVID-19 is 0.51%](#). An [estimate from the 7th of February](#) put it at 0.18%-2.8%. So I will consider adding a global, non-age-based CFR threshold to the criteria.

This project is also not about estimating the true value of CFR; it's about normalizing a heuristic for deciding when to behave as if a new illness could be as serious as COVID-19.

So we have to ask what our threshold for evidence is. I can think of two possibilities.

- Allow for extremely rough amateur division to meet this criterion. For example, in this April 27th 2009 New York Times piece, they wrote "Mexican health authorities have confirmed 149 deaths from that flu and are investigating the illnesses of 1,600 people." A reader who ignored the subsequent paragraph stating "... doctors have little information yet on the mortality rate, as there is no reliable data on the total number of people infected" might have divided 149 by 1600 to obtain a "case-fatality" rate of 9%, a wild overestimate of the true figure.
- Alternatively, we could require a scholarly or government/major NGO case-fatality rate estimate. [This scholarly editorial from September 2009](#) gave an Australian government estimate of 0.14% case-fatality. A [November 2009 estimate](#) was even lower, putting the rate at "0.026% (range 0.011-0.066%)." [The earliest CFR I found was from June 19th, 2009](#), estimating a 0.4% CFR (range 0.3%-1.8%).

Given that it was predictable that there were many more undetected cases that lowered the true CFR, I'm going to disqualify the "rough division" from meeting the criteria and require a scholarly CFR estimates.

That 0.4% June scholarly global CFR estimates for swine flu were within the 0.18%-2.8% range of the Feb. 7th global CFR range for COVID-19. However, that Feb. 7th paper was pretty skeptical of these figures. I'd like to see if there was another COVID-19 estimate prior to Feb. 20th that presented an estimate with more confidence.

I provisionally count this as a no, but may change this later.

- *Can the disease last for two weeks in more serious cases?*

From [an NCBI paper](#): "The acute symptoms of uncomplicated infections persist for three to seven days, and the disease is mostly self-limited in healthy individuals, but malaise and cough can persist for up to 2 weeks in some patients. Patients with more severe disease may require hospitalization, and this may increase the time of infection to around 9 to 10 days." I count this as a "no," because it is only malaise and cough that persisted for 2 weeks, not the need for hospitalization.

- *Do around 5% of patients seem to require hospitalization?*

[By CDC data](#), upper bound of US hospitalizations divided by number of cases is less than 1% of cases requiring hospitalization. I count this as a "no."

- *Are there no vaccines and no treatments that have been proven effective?*

Because flu was already known to be Tamiflu-resistant in 2008, by early January 2009 the CDC was already recommending Relenza (zanamivir) and other alternate treatments for H1N1. I count this as a "no."

Factor 3 - spread

- *Is the world death toll over 2,000?*

The [WHO's 61st pandemic update](#) put the world death toll at "at least 2185" on the 23rd of August 2009.

- *Has the disease been detected in at least 10 industrialized countries collectively containing a total of at least 1 billion people, found in people with no clear link to the original source?*

The first cases of what would later be confirmed as swine flu were diagnosed in Mexico in early March when 60% of the small town of La Gloria, in Veracruz, was sickened. It was confirmed in its 10th country, the Netherlands, in April 2009. It wasn't confirmed in a set of countries totaling a world population of at least 1 billion people until it was found in China on May 1st, 2009.

- *Is the disease present in major cities with strong international travel links?*

The first community outbreaks in the US were confirmed on April 25th, 2009.

- *Has the disease spread in advance of a lockdown, or escaped it?*

News of Mexico shutting down parts of its economy were [reported in Reuters](#) on April 29th, 2009. I'll count this as "lockdown."

Factor 4: Institutional response

- *Has a city-wide lockdown been attempted in a city of a population of at least 10 million, or have travel restrictions to or from major economies been implemented?*

I don't find reports of major city lockdowns. The most significant travel restriction I can find was of Mexican travel to Japan. I personally don't count 2009 Mexico as a "major economy" but that is a controversial and not pre-declared analysis decision. For future analyses I'll use the 2009 Mexican fraction of world GDP as my cutoff for "not a major economy." The WHO was recommending against travel restrictions early on. I count this as a no.

- *Has the WHO or a similar organization declared an "emergency of international concern" or "global emergency," or issued an even more severe warning?*

The WHO declared swine flu a "public health emergency of international concern" on April 25th, 2009.

- *Have there been several front-page news stories about the disease?*

[There](#) were [multiple](#) stories [on page](#) A1 in the New York Times by the last few days of April. There may have been earlier front-page swine flu news, but I'm not sure based on what's coming up on their digital search.

- *Are there reports or warnings of shortages of medical supplies from the most-affected regions?*

Yes, there were mask shortages.

Self-Experiment: Does Working More Hours Increase My Output?

[This piece is cross-posted on my blog here.](#)

After writing up [my research on limits to working](#), the sheer spread of possibilities amazed me. I genuinely wasn't sure if I would be able to tell the difference between a day with four hours of deep work and one with eight hours. Surely we could narrow down the hypothesis space from that!

So, I designed a simple experiment. I would do one hour of deep work each day for two days, then two days of four hours each, and finally two days of eight hours each.

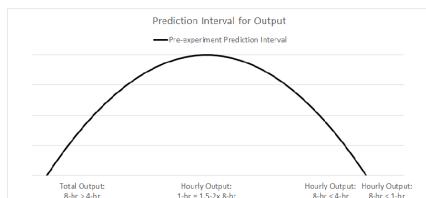
Why I chose this experiment

I optimized for quickly testing large effect sizes to narrow my uncertainty. I didn't expect this experiment to be rigorous or sensitive to small nuances -- n=2 per condition, and I would love to hear any suggestions for how to blind me to whether I was working one or eight hours that day.

But by doing such an extreme experiment, I would definitely see an effect.

My output would have to be uncorrelated with hours worked for me *not* to. If I couldn't easily tell a difference in the output between the conditions, it would indicate diminishing marginal returns massively influenced my output. Otherwise, I could get a rough guess at if and how much my output declined.

My guess was that I would get between 50% to 200% more done per hour on the one-hour day than the eight-hour day. I was less sure about the four-hour days, but I guessed my hourly output would fall in between that of the one-hour days and the eight-hour days. I would be quite surprised if I got more done per hour on the eight-hour day than the one-hour day. (Confession, I forgot to write these down before starting the experiment, so I'm writing them now after collecting data but before looking at the results.)



Methods

In order to have somewhat comparable results, I spent all twenty six hours writing and tracked how many words I wrote each hour. I scheduled coworking sessions on Focusmate.com to hold myself to a schedule. Since the Focusmate sessions are fifty minutes long, I operationalized "an hour" as a 50-minute pomodoro. (My Toggl tracked time ended up near 24 hours.)

When analyzing the data, I controlled for minutes worked to get words per minute. I tracked my output in a few categories (words drafted, words outlined, words edited,

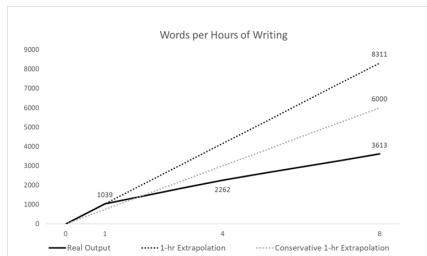
and words proofed), which I adjusted to reflect roughly similar amounts of effort. E.g., words proofed counted as 1/10 as much effort as words drafted. I also gave my word count “bonuses” for other particularly useful work that wasn’t reflected in word count, such as thinking hard on how to frame a tricky topic. I did these adjustments and bonuses before calculating totals to compare the days.

Here’s what I found

# of Hours	Average words per 60 minutes	Total words per day at that rate
1	1039*	1039
4	565	2262
8	453	3612

*I suspect my first day was more productive than normal (it was my most productive hour out of all 26). Given that, I’ve included a conservative estimate below based on the other 1-hr day (750 words).

UPDATE: I did another experiment where I wrote for 1-3 hours per day for five days in a row. During that time, I averaged 727 words per hour. So, my measured output was quite close to my conservative estimate below.



According to this, increasing the hours worked by 4x only increased the words written by 2.2x, and increasing hours 8x resulted in 3.5x more words. These equal 54% and 43% as many words per hour respectively as when I only worked for one hour per day.

So while increasing hours still led to increased output, I got much less done per hour. I’d be better off writing consistently for a few hours each day to efficiently maximize output.

A few other observations

Besides my main question of how working my hours affected my output, I noticed a few other observations that seemed noteworthy. However, these are post-hoc observations. Take them with an extra grain of salt - they might not replicate.

1. Go with the flow. My output varied significantly by hour, but it didn’t decrease linearly during the day. I had spikes throughout the eight-hour days.

However, the spikes were correlated with my subjective experience. (I drafted ≥ 750 words during 5 out of the 17 time blocks when feeling okay or good, but drafted that amount during 0 out of the 12 blocks when I felt blah.)

It worked best to do shallow work (such as proofing a transcription) or switch to a different piece when I felt stuck. My rule of thumb would be to stop if the writing isn’t flowing after ten minutes, though I could see this being easy to [Goodhart](#) in the future.

2. Switch up work. I scheduled my one-hour days for when I had the most calls. It ended up that I did a similar number of writing hours + coaching hours the first four days. Writing for one hour a day still felt easier than four hours even when the total hours spent working were similar. This weakly implies these limits may apply only to writing or maybe general deep work time.

3. Motivation matters, maybe. I was tired of writing, and my [RSI](#) was flaring up, by midway through the second eight-hour day. I wouldn't be surprised if some of the slump in output was related to motivation. So, if someone really wanted to write for twelve hours a day, they might be just fine.

Conclusion

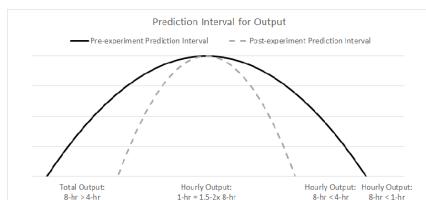
So, my data says I can do more than four hours even of deep work, but I'm probably better off consistently writing for a few hours each day.

And I also think it's possible that all of that could be wrong. [Did I notice a big effect because I expected a big effect?](#) Maybe. Was my sample size too small? Yeah. Were my methods susceptible to bias? Yes.

Do I know more than I did before this experiment? Definitely.

My prediction interval narrowed. Now, I would be quite surprised *not* to see more output from an eight-hour workday than from a four-hour workday even when I'm doing deep work. I would also be quite surprised if I got more done per hour from the eight-hour than from a four-hour workday.

Both of these seemed unlikely but plausible before running the experiment.



Even if I couldn't get to a final answer, I meaningfully reduced my uncertainty.

At least for myself. I don't expect these results to generalize widely. They might serve as a rough starting point, but much better to try it out yourself. You might have different types of work, different habits, and different levels of motivation - which could all lead to different results.

If you want to do a similar experiment, here are some questions to ask yourself:

- Did you crash the next day?
- Did your work quality go down? How motivated were you?
- How much of your top priority work did you complete in total? How much did you get done per hour?
- How happy did you feel? How stressed?
- How well did you prioritize?

If you do something similar, I'd love to hear the results!

The One Mistake Rule

Epistemic Status: [The Bed of Procrustes](#)

If a model gives a definitely wrong answer anywhere, it is useless everywhere.

This principle is doubtless ancient, and has doubtless gone by many names with many different formulations.

All models are wrong. That does not make them useless. What makes them useless is when they are giving answers that you know are definitely wrong. You need to fix that, if only by having the model more often spit out "I don't know."

As an example of saying "I don't know" that I'm taking from the comments, if you want to use Newtonian Physics, you need to be aware that it will give wrong answers in relativistic situations, and therefore slightly wrong answers in other places, and introduce the relevant error bars.

Of course, a wrong prediction of what is *probably* going to happen is not definitely wrong, in this sense. An *obviously wrong probability* is definitely wrong no matter the outcome.

The origin of this particular version of this principle was when me and a partner were, as part of an ongoing campaign of wagering, attempting to model the outcomes of sporting events.

He is the expert on sports. I am the expert on creating models and banging on databases and spreadsheets. My specialty was assuming the most liquid sports betting market odds were mostly accurate, and extrapolating what that implied elsewhere.

First we would talk and he would explain how things worked. Then I would look at the data lots of different ways and create a spreadsheet that modeled things. Then, he would vary the inputs to that spreadsheet until he got it to give him a wrong answer, or at least one that seemed wrong to him.

Then he'd point out the wrong answer and explain why it was definitely wrong. I could either argue that the answer was right and change his mind, or I could accept that it was wrong and go back and fix the model. Then the cycle repeated until he couldn't find a wrong answer.

Until this cycle stopped, we *did not use the new model for anything at all, anywhere, no matter what*. If a new wrong answer was found, we stopped using the model in question until we resolved the problem.

Two big reasons:

If we did use the model, even if it was only wrong in this one place, then the one place it was wrong would be the one place we would disagree with the market. Fools and their money would be soon parted.

Also, if the model was obviously wrong here, there's no reason to trust anything else the model says, either. Fix your model.

This included tail risk style events that were extremely unlikely. If you can't predict the probability of such events in a reasonable way, even if those outliers won't somehow bankrupt you directly, you're going to get the overall distributions wrong.

This also includes *the change in predictions between different states of the world*. If your model predictably doesn't agree with itself over time, or changes its answer based on things that can't plausibly matter much, then it's wrong. Period. Fix it.

You should be *deeply embarrassed* if your model outputs an obviously wrong or obviously time-inconsistent answer even in a hypothetical situation. You should be *even more embarrassed* if it gives such an answer to the actual situation.

The cycle isn't bad. It's good. It's an excellent way to improve your model: Build one, show it to someone, they point out a mistake, you figure out how it happened and fix it, repeat. And in the meantime, you can still use the model's answers to help supplement your intuitions, as a sanity check or very rough approximation, or as a jumping off point. But until the cycle is over, don't pretend you have anything more than that.

In my culture: the responsibilities of open source maintainers

If you maintain an open source project, what responsibilities do you have towards your users? Some recent drama (that I won't link to) reminded me that there are large differences in how people answer that question.

(In the drama in question, this wasn't the only thing at issue. But it was a relevant question.)

I thought I'd take a stab at describing my best guess as to how we answer it [in my culture](#): in the culture that exists only in my mind, but that I think (and hope) many copies of me would implement, if we had that opportunity. (That culture doesn't attempt to exclude people who *aren't* copies of me, so it does need to be robust to attack. In my culture, we do not just assume in defiance of all evidence that everyone is friendly and trustworthy.)

Some of this will probably seem obvious to many readers, like "in my culture, murder is considered bad". Probably different bits to different readers. I'm interested in discovering which bits seem obvious to almost everyone, and which bits are controversial.

A lot of it follows from how I think about responsibility in general. But if you start to think "extending this response to this other situation, you'd get this, and that's a terrible idea"... in my culture, we don't immediately assume from this that I'm endorsing a terrible idea. Instead we check. Maybe I [disagree](#) that that's how it extends. Maybe I hadn't thought about this, and you can change my mind about the initial response. Maybe I just straightforwardly endorse a terrible idea: in that case, it'll be much easier to push back once you've gotten me to admit to it.

I do not intend, in this essay, to discuss whether any particular person or group is living up to the standards I outline here. I may do that [in future](#). But how likely that is, and what that follow-up looks like, depends on whether the responses to this essay suggest a lot of people agree with my culture.

I think there are at least three important limitations to this essay. One is that I've never been a maintainer of an open source project that had users other than myself (that I knew of), though I've made small contributions to a few. As such, I don't really know what the experience is like or how my culture deals with its challenges, much like I don't really know how my culture deals with the challenges of living in Antarctica. I can make educated guesses, but that's all they are. I'm not going to explicitly flag them as such in the main body. (I'm already using the words "in my culture, we..." far too much. I'm not changing it to "in my culture, my educated guess is that we...") Ideally, I wouldn't write this essay because there was already a conversation taking place between people who knew what they were talking about. Unfortunately, as far as I've seen there mostly isn't, so I offer what I can.

Another is that I don't talk at all about the responsibilities of *users*, which is also an important part of the question. I'd like to, but... this essay has been knocking around in my head for at least a year and a half, I've made at least one previous attempt to write it that I gave up on, and I'm worried that if I don't publish it quickly I never will. I

hope and believe that even with this omission, it is better for this essay to be published than to not be published.

(I also omit the question "what is this responsibility thing anyway?", but that seems less important to me right now. I've probably also overlooked important parts of the things I do talk about, but that too may not be a big deal.)

And a third is that without specific examples, what I've written is less constrained than what's in my head. It may well be possible to read this and agree with what I've written, and then to discover that your culture has a much stricter or much looser conception of responsibility than mine does.

With all of that out of the way: there are three questions I'm going to be focusing on here. When is responsibility taken on; what does it entail; and how do we react if people fail at it?

When is responsibility taken on?

In my culture, taking on responsibility requires active involvement, but not explicit consent. It's possible to take on responsibility through negligence - to draw responsibility towards yourself without realizing that's what you're doing. In my culture, we're sympathetic towards people who do that, but we don't consider that this absolves their responsibility.

In my culture, merely making something available does not make you responsible for it. If you want to put something online and ignore anyone who tries to use it, you can do that. You are allowed to [shithub](#).

In my culture, you take on more responsibility for a project if...

- If you start *encouraging* people to use your project. If someone happens to stumble across a repository you never specifically shared, you have no responsibility for it. If you see them describe a problem they have, and you say "oh hey, I had that problem too, you might want to check out my repo" you have a little. If you create a website for your project where you describe it as "the best way to solve this problem", you have more.
- If you have *many* users. If you shirk your responsibilities, most of the harm done is to your users¹. With fewer users, the harm from not-acting-responsibly is lower, and so the responsibilities themselves are lower.
- If your users are *invested* in your project. If they can't easily stop using it, you have more responsibility to treat them well.

Your users' exit costs are probably low if you make a video game, or "libpng but 30% faster" or "find but with a nicer command line syntax". In that case your users can probably just play a different game, or easily replace your tool with libpng or find.

They're higher if you make a programming language, where migrating away requires rewriting a codebase. Or a document editor that can only save in its native binary format or pdf, so that there's no way to edit the documents without your tool.

- If you have the *ability* to accept responsibility. Handling responsibility takes time and it takes multiple skill sets. One person hacking on their side project simply cannot act in the same ways as a team of two full-time developers plus twenty volunteers. That team can't act in the same ways as the Mozilla Foundation.
- If you *act like* you're accepting responsibility. If you have a history of ignoring issues and pull requests, people should probably pick up on this. If you tell people you're going to break backwards compatibility, they shouldn't expect you to maintain it. Words like "production ready" increase your level of responsibility. Words like "alpha" decrease it, as do version numbers below 1.0.

A common thread here is that responsibility is related to *justified expectations*. "Expectations" in both the moral sense and the probabilistic sense. If someone can make a compelling argument "this person *morally ought to* accept responsibility here", or a compelling argument "I predict based on past behaviour that this person *will* act as though they've accepted responsibility here", then in my culture, that person has some degree of responsibility whether they like it or not.

Accordingly, in my culture you can disclaim responsibility in advance, simply by *saying that you're doing this*. Note that a pro forma warranty disclaimer in your LICENSE file isn't sufficient here. Instead you should say it at the entry points to your project - probably the README and the website (if there is one). Something like...

Note that while I think this project will be useful to many people, I am not devoting much attention to it, and I'm prioritizing my own use cases. Bugs may be fixed slowly if at all. Features may be removed without warning. If this isn't acceptable to you, you should probably not rely on the project; or at least you should be willing to fork it, if necessary.

Or even:

Note that while I think this project will be useful to many people, I have no interest in accepting responsibility for other people's use of it. Feel free to submit issues and pull requests; but I may ignore them at whim, and my whim may be arbitrary and capricious.

This won't usually be necessary. But if you feel in danger of taking on more responsibility than you'd like, you can do this.

If you have responsibility, what does that entail?

In my culture, if we've taken on responsibility and now we want to discharge it, we do so with care. We give advance notice, and we try to find replacement maintainers. If we can't find a replacement maintainer, we still get to quit. But we try.

In my culture, we acknowledge that different people have different needs and goals, and not every project will be suitable for all users. We try to help users figure out whether our projects will be suitable for them before they get invested. We're open about the limitations of our projects, and about the design goals that we explicitly reject.

In my culture, we don't need to reply to every critic. But we do try to notice common threads in criticism, and address those threads. ("Address" doesn't mean that we necessarily try to solve them in a way that will satisfy the critics. It simply means we

give our thoughts about them. We try to do that in such a way that even if the critics aren't happy with what we're doing, they at least feel like we've heard them and understood them.)

In my culture, we accept that some critics will be unkind, unfair and unproductive. This sucks, but it's a reality of life right now. We distinguish these critics from others. We are free to ignore them, and to ban them and delete their posts from the spaces we control. We do not use their actions to justify shirking our responsibilities to other critics.

In my culture, we also take care to check whether we're unfairly rounding people off as being unkind, unfair and unproductive. We don't require perfection from our critics. We try to hold ourselves and our supporters to the standards we expect of our critics. We moderate our spaces, but we try not to silence all criticism from them.

In my culture, we still introduce bugs, because we're still human. We try not to, of course. But we accept fallibility. When we fail, we do our best to fix it. We put more effort into avoiding more catastrophic bugs.

In my culture, we do not accept incompetence or indifference. These are signs that a person should not have taken on responsibility in the first place. We expect people to know the limits of their ability and of how much they care.

In my culture, we can set boundaries around how our future actions are constrained. We distinguish public APIs (where we mostly try to maintain backwards compatibility) from private APIs (where we mostly don't, and if people expect us to we point to the word "private"). We may reject bugfixes that we expect to cause too much ongoing future work, in hope of finding a better fix in future.

In my culture, sometimes we hurt people deliberately because the alternatives are worse. When we do, we own it. For example, sometimes we decide to remove a feature that people were relying on. We don't make that decision lightly. When we do it, we explain why we decided to do it, and we apologize to the people who were relying on it. We don't try to minimize their pain. We try to suggest alternatives, if we can; but we try not to pretend that those alternatives are any more suitable than they actually are. We give advance notice, if we can. We can be convinced to change our minds, if new information comes to light.

Of course, the degree to which we feel bound by all of this depends on the degree to which we've actually taken on responsibility. And note that in all of this, there's very little requirement for positive action. If we don't want to include a feature, we don't need to write it, or even merge the PR if someone else writes it. If we don't want to go in the direction our users want us to, we don't have to go there. We just have to make it clear that that's what we're doing.

What if someone fails to act responsibly?

That is, if someone has taken on responsibility, whether intentionally or not, and then has failed to act like someone who's taken on responsibility... how do we respond?

In my culture, we mostly respond by acting as though this will happen in future.

We don't treat this as an indelible black mark against all aspects of the person. We'll still accept their contributions to other open source projects, for example. (Though

perhaps not technically complex contributions, that will likely need a lot of maintenance from the same person going forward.) We'll accept their conference talks. We'll stay friends with them.

But we won't rely on their projects, because we don't expect their projects to be reliable. We warn others about this, but if others decide to use their projects anyway, we consider that none of our business.

We accept that people have a right to run projects as they see fit. We acknowledge that our conception of "responsibility" is a mere social convention. But social conventions are important. We provide an opt-out for this particular social convention, but if you don't opt out *explicitly*, we'll make it explicit for you.

We give second chances. If they want to take on responsibility in future, and if they seem to understand how they failed in the past, and they say they'll try to do better, we give them a shot, because that's reason to think it won't happen in future. If they still fail, and want a third chance, we take their second failure into account too.

We may fork a project, even if the maintainer doesn't approve, but we'll rarely attempt a hostile takeover from the outside. We'll make sure our fork is clearly distinguished from the original project.

1. To the extent that this isn't true, I think most of the responsibility comes from something other than "being an open source maintainer". For example, if you're making a web scraper, I would suggest you have a responsibility to make it well-behaved by default - sleep between requests, stop after a certain number of failures, and so on. But that responsibility comes from the product, not the process. You can't avoid it with a disclaimer. And so it's not the focus of this essay. ↵

My stumble on COVID-19

A couple weeks ago, I started investigating the response, here and in the stock market, to COVID-19. [I found](#) that LessWrong's conversation took off about a week after the stock market started to crash. Given what we knew prior about COVID-19 prior to Feb. 20th, when the market first started to decline, I felt that the stock market's reaction was delayed. And of course, there'd been plenty of criticism of the response of experts and governments. But I was playing catch-up. I certainly was not screaming about COVID-19 until well after that time.

Today, I found [the most detailed timeline I've seen](#) of confirmed cases around the world. It goes day by day and country by country, from Jan. 13th to the end of March.

That timeline shows that Feb. 21st was the first date when at least 3 countries besides China had 10+ new confirmed cases in a single day (Japan, South Korea, Italy, and Iran).

That changes my interpretation of the stock market crash dramatically. Investors weren't failing to synthesize the early information or waiting for someone to yell "fire!" They were waiting to see confirmed international community spread, rather than just a few cases popping up here and there. Once they saw that early evidence, the sell-off began, and it continued in tandem, day by day, with the evidence of community spread in new countries and the exponential growth of COVID-19 cases in countries where it was already established.

[Scott Aaronson speculates](#) that he might have been able to tell what COVID-19 would turn into on Feb. 4th. I believed and agreed with him, hypothesizing that it was a lack of intellectual guideposts for synthesizing early information on a novel disease that made it hard to assess and act appropriately. To correct that lack, I created a model for what to look for and how to interpret information to judge the potential spread of a disease - [an alarm bell for the next pandemic](#).

Now that I've found day-by-day, country-by-country confirmed case data, I need to change a few of my points of view.

First, is lack of confirmed international community spread generally a good reason to not panic, even if in this case things turned out badly? COVID-19 had its first confirmed case in the USA on Jan. 21st, but never had more than 1-2 new cases on any given day until Feb. 29th - with several days up to 12 days in between new cases during that time. That's over a month of sporadic, occasional new case reports.

The coincidence of faster spread of COVID-19 internationally and the beginnings of the stock market crash suggest that hard evidence of community spread of a deadly disease is the world's evidential benchmark for an economic downturn.

And that seems perfectly reasonable to me.

To argue otherwise is to say that it should be obvious COVID-19 was highly contagious based on some other form of evidence. That's not impossible.

But in my case, I didn't know until weeks after I started researching that the world's largest encyclopedia had day-by-day, country-by-country confirmed case data.

So it's really not a stretch for an amateur such as myself to have missed or misinterpreted some critically important source of data. I still think it's possible for an avid, early COVID-19 amateur researcher to have felt reasonably confident in shorting the market prior to Feb. 20th. But they'd need to have searched *long and hard* for other forms of data to back up their argument. Not only evaluating COVID-19 on its own terms, but comparing it with other pandemics of the late 20th and early 21st century.

So I have to downgrade my confidence in the usefulness of the models I wrote for this. I need to call myself out on my willingness to assume with such confidence that the world had it wrong, rather than me. Maybe this experience has been necessary for my intellectual growth. Doing this research out of curiosity and being willing to accept my own conclusions, even when I ultimately felt I was mostly wrong, was perhaps the only way to get to this point of understanding, to get less wrong.

On a deeper level, I am learning and relearning a fundamental lesson lately. I don't anticipate that I'll have grasped it until it's bashed me over the head at least a few more times. Here it is, in free verse:

Intellectual knowledge is a slippery thing.

I can't readily grasp it.

I don't practice engaging with it every waking moment, the way I do with my five senses.

Human intellect is 2% of the evolutionary age of the nervous system.

This kind of intellectual work is extremely difficult; only a few people care about it enough to make the attempt, so I'm getting far less feedback than I would for other activities.

I am no savant.

Others have been doing this full-time for decades longer than I've spent as a part-time amateur.

A major league baseball player who hits a home run has far more in common with an MLB player who hits a foul ball than a child who hits a home run.

Reliable knowledge is a hard-won and precious thing.

What We Owe to Ourselves

[Cross-posted from [Grand, Unified, Crazy.](#)]

"You can never make the same mistake twice because the second time you make it, it's not a mistake, it's a choice."

Steven Denn

Something that has been kicking around my mind for the last little while is the relationship between responsibility and self-compassion. A couple of people recently made some very pointed observations about my lack of self-compassion, and it provoked a strange sadness in me. Sadness because while I know that their point is true – I am often very hard on myself, to the detriment of my happiness – those thought patterns seem so philosophically necessary that I have been unable to change them. This post is my attempt to unpack and understand that philosophical necessity.

Our ability to feel compassion is intimately tied to our judgement of responsibility in a situation. If you get unexpectedly laid off from your job, that's terrible luck and most people will express compassion. However, if you were an awful employee who showed up late and did your job poorly, then most people aren't going to be as sympathetic when you finally get fired. As the saying goes: you made your bed, you lie in it. More abstractly, we tend to feel less compassion for someone if we think that they're responsible for their own misfortune. This all tracks with my lack of self-compassion, as I have also been told that I have an overdeveloped sense of personal responsibility. If I feel responsible for something, I'm not going to be very compassionate towards myself when it goes wrong; I'm going to feel guilt or shame instead.

Of course, this raises the question of what we're fundamentally responsible for; the question of compassion is less relevant if I actually am responsible for the things that are causing me grief. People largely assume that we're responsible for our own actions, and this seems like a reasonable place to start. It makes sense, because our own actions are where we have the clearest sense of *control*. While we can control parts of the outside world, that process is less direct and less exact. Our control over ourselves is typically much clearer, though still not always perfect.

If we assume that we have control over ourselves and our actions, this means that we also have responsibility for ourselves and our actions. If we avoid or ignore that responsibility and we're not happy with the consequences, we don't deserve much, if any, compassion: it's our own damn fault. This all seems... normal and fairly standard, I think, but it's an important foundation to build on.

Freedom, and Responsibility over Time

Now let's explore what it means to be responsible for our actions, because that can be quite subtle. Sometimes our choices are limited, or we take an action under duress. Even ignoring those obvious edge cases, [our narrative dictates the majority of our day-to-day decisions](#). What responsibility do we bear in all of these cases? Ultimately I believe in a fairly Sartrean version of freedom, where we have a near-limitless range

of possible actions every day, and are responsible for which actions we take. Obviously some things are physically impossible (I can't walk to Vancouver in a day), but there are a lot of things that make no sense in the current framework of my life that are still theoretically options for me. If nothing else, I could start walking to Vancouver today.

Assuming that we're responsible for all of our actions in this fairly direct way, we also end up responsible for the consequences of actions *not* taken on a given day. There is a sense in which I am responsible for not walking to Vancouver today, because I chose to write this essay instead. I am responsible for my decision to write instead of walk, and thus for the consequence of *not being on my way to Vancouver*. This feels kind of weird and a bit irrelevant, so let's recast it into a more useful example.

A few hours from now when I finish this essay, I'll be hungry to the point of lightheadedness because I won't have eaten since breakfast. Am I responsible for my future hunger? There's a certain existential perspective in which I'm not, since it's a biological process that happens whether I will it or not. But it's equally true that I could have stopped writing several hours earlier, put the essay on hold, and had lunch at a reasonable time. I *am* definitely responsible for my decision to keep writing instead of eating lunch, and so there is a pretty concrete way in which I am at least partly responsible for my hunger.

This isn't to say that I'm necessarily going to be unhappy with that decision; even in hindsight I may believe that finishing the essay in one fell swoop was worth a little discomfort. But it does mean that I can't avoid taking some responsibility for that discomfort. And, since I'm responsible for it, it's not something I can feel much self-compassion over; if I decide in hindsight that it was a terrible decision, then it was still a decision that I freely made. If I experience a predictable consequence of that decision, it's my own damn fault.

This conclusion still feels pretty reasonable to me, so let's take a weirder concrete example, and imagine that in three months I get attacked on the street.

Predictability in Hindsight

I don't have a lot of experience with physical violence, so if I were to get attacked in my current state then I would likely lose, and be badly hurt, even imagining my attacker does not have a weapon. To what degree am I responsible for this pain? Intuitively, not at all, but again, the attack happens three months from now. I could very well decide to spend the next three months focused on an aggressive fitness and self-defence regimen (let's assume that this training would be effective, and that I would not get hurt in this case). Today, I made the decision to write this essay instead of embarking on such a regimen; in the moment today, this decision to write is clearly one that I am responsible for. Does this mean I'm responsible for the future pain that I experience? I'd much rather avoid that pain than finish this essay, so maybe I should stop writing and start training!

The flaw in this argument, of course, is that I don't know that I'm going to get attacked in three months. In fact, it seems like something that's very unlikely. In choosing not to train today, I can't accept responsibility for the full cost of that future pain. I should only take responsibility for the very small amount of pain that is left when that future is weighted by the small probability it will actually occur. If I did somehow know with perfect accuracy that I was going to be attacked, then the

situation seems somewhat different: I would feel responsible for not preparing appropriately in that case, in the same way I would feel responsible for not preparing for a boxing match that I'd registered for.

All of this seems to work pretty well when looking *forward* in time. We make predictions about the future, weight them by their probability, and take action based on the result. If an action leads to bad results with high probability and we do it anyway then we are responsible for that, and don't deserve much sympathy. We rarely go through the explicit predictions and calculations, but this seems to be the general way that our subconscious works.

But what about looking *backward* in time? Let's say I decide not to train because I think it is unlikely I will be attacked, and then I get attacked anyway. Was this purely bad luck, or was my prediction wrong? How can we tell the difference?

Depending on the perspective you take you can get pretty different answers. Random street violence is quite rare in my city, so from that perspective my prediction was correct. Hindsight is a funny thing though, because in hindsight I *know* that I was attacked; in hindsight probabilities tend to collapse to either 0 or 1. And knowing that I was attacked, I can start to look for clues that it *was* predictable. Maybe I realize that, while street violence is rare in the city overall, I live in a particularly bad neighbourhood. Or maybe I learn that while it's rare most of the time, it spikes on Tuesdays, the day when I normally go for a stroll. If I'd known these things initially, then I would have predicted a much higher probability of being attacked. Perhaps in that case I would have decided to train, or even take other mitigating steps like moving to a different neighbourhood. Who knows.

What this ultimately means, is that I can only possibly be responsible for being hurt in the attack if I'm somehow responsible for *failing to predict that attack*. I'm only responsible if for some reason I "should have known better".

Taking Responsibility for our Predictions

[I should take this opportunity to remind people that this is all hypothetical. I was not attacked. I'm just too lazy to keep filling the language with extra conditional clauses.]

At this point I've already diverged slightly from the orthodox position. A large number of people would argue that my attacker is solely responsible for attacking me, and that I should accept no blame in this scenario. This certainly seems true from the perspective of law, and justice. But in this essay I'm focused on outcomes, not on justice; ultimately I experienced significant pain, pain that I could have avoided had I made better predictions and thus taken better actions.

Let's return to the issue of whether or not I can be held responsible for failing to predict the attack. There is a continuum here. In some scenarios, the danger I was in should have been obvious, for example if my real-estate agent warned me explicitly about the neighbourhood when I moved in. In these scenarios, it seems reasonable to assign some blame to me for making a bad prediction. In other scenarios, there was really no signal, no warning; the attack was truly a stroke of random bad luck and even in hindsight I don't see a way that I could have done better. In these scenarios, I take no responsibility; my prediction was reasonable, and I would make the same prediction again.

As with most things, practical experience tells me that real-world situations tend to fall somewhere in the middle. Nobody's hitting you over the head with a warning, but neither is the danger utterly unpredictable; if you look closely enough, there are always signs. It is these scenarios where I think that my intuition fundamentally deviates from the norm. When something bad happens to me, then I *default* to taking responsibility for it, and I think that's the correct thing to do.

Of course, there are times when whatever it is *is* my fault in an uncontroversial way, but I'm not talking about those. I'm talking about things like getting attacked on the street, or being unable to finish a work project on time because somebody unexpectedly quit. These are the kind of things that I expect most people would say are "not my fault", and I do understand this position. However, I think that denying our responsibility for these failures is bad, because it causes us to *stop learning*. Every time we wave away a problem as "not our fault" we stop looking for the thing we could have done better. We stop growing our knowledge, our skills, our model of the world. We stagnate.

This sounds really negative, but we can frame it in a much more positive way: that there's always something to learn, and that we should always try to be better than we were. What I think gets lost for a lot of people is that this not a casual use of "always". Even in failures that are not *directly* our fault, there is still something to learn, and we should still use it as an opportunity to grow. Unless we perfectly predicted the failure and did everything we could to avoid or mitigate it, there is still something we could have done better. Denying our responsibility for our bad predictions is abdicating our ability to grow, change, or progress in life. Where does this leave us?

Any good Stoic will tell you that when something goes wrong, the only thing we have control over is our reaction, and this applies as much to how we assign responsibility as to anything else. We are responsible for the fate of our future self, and the only way to discharge that responsibility is to learn, grow, and constantly get better. The world is full of challenges. If we do not strive to meet them, we have no-one to blame but ourselves.

My experience with the "rationalist uncanny valley"

Epistemic status: **Very Uncertain.** Originally written with the belief rationality has harmed my everyday life; revised to reflect my current belief that it's been somewhat positive. The tone of this article, and some of the content, may constitute self-punishment behavior, signaling, or frustration rather than genuine belief.

Introduction

I'm currently a male second-year undergrad in computer science. I am not clinically depressed. My first exposure to the rationalist community was largely limited to reading HPMOR and ~40% of the original Sequences around 2013-2014; I've had rationalist/EA friends continuously since then; in mid-2019 I started following LW; in March 2020 I read almost every scrap of COVID-19 content. I'm not sure how to evaluate my strength as a rationalist, but I feel epistemically slightly weaker than the average LW commenter and guess that my applied skills are average for non-rationalists of my demographic.

Others described the phenomenon of the "rationalist uncanny valley" or "valley of bad rationality" as early as 2009:

[It has been observed that when someone is just starting to learn rationality, they sometimes appear to be worse off than they were before.](#)

I've known about the rationalist uncanny valley since 2013, and was willing to accept some temporary loss in effectiveness. Indeed, before March 2020, the damage to my everyday life was small and masked by self-improvement. However, with isolation, my life has ground to a halt, in part due to "rationalist uncanny valley" failure modes, and failures I'm predisposed to which rationality training has exacerbated. Looking back, one of these also happened during my exposure to the community in 2014. This post is an attempt to characterize the negative effects exposure to rationality has had on my life, and is not representative of the overall effect.

Examples

1. Bad form when reading LW material

I'm very competitive and my self-worth is mostly derived from social comparison, a trait which at worst can cause me to value winning over maintaining relationships, or cause me to avoid people who have higher status than me to avoid upward comparison. In reading LW and rationalist blogs, I think I've turned away from useful material that takes longer for me to grasp because it makes me feel inferior. I sometimes binge on low-quality material, sometimes even seeking out highly downvoted posts; I suspect I do this because it allows me to mentally jeer at people or ideas I know are incorrect. More commonly, I'll seek out material that is already

familiar to me. Worse, it's possible that all this reading has confirmed beliefs I was already predisposed to, and therefore been net-negative.

As a concrete example, Nate Soares has a post on the "dubious virtue" of [desperation](#). It's dubious because it must be applied carefully: one must be desperate enough to go all-out in pursuit of a goal, but not burn out or signal visible desperation towards people.

I am already strong at the positive aspects of desperation, but the idea of "dubious virtues" is appealing to me (maybe it's the idea that I can outdo others by mastering a powerful technique often considered too dangerous). I read the article several times, largely disregarding the warnings because they made me feel uncomfortable, with the result that I burned out and signaled desperation towards people.

Something similar but more severe happened in 2013-14, when I fell into the following pattern (not quite literally true): A friend links me [a LW article](#). Then my defense mechanisms of [epistemic learned helplessness](#) activate and I stop reading. (*didn't they make the basilisk thing? I should read all about that so I can identify suspect arguments*) Then I decide I should prove my defense mechanisms wrong by reading a quarter of HPMOR in one night and memorizing the Rationalist Virtues! Then I completely stop reading out of fear that rationality is a cult/mind-hacking attempt. I decide to wait several years to dampen the cycle before becoming a rationalist. It's possible I spent six years in the rationalist uncanny valley and I'm not sure there was a simple way out before approximately last year.

2a. Predictions and being a Straw Vulcan

Others have gone through a phase of [making all decisions by System 2 because they no longer trust System 1](#). I'm somewhat related. Over the last few months, I've worked on making [calibrated](#) predictions, including predicting my own future to inform career planning decisions. Perhaps due to the way I approach this exercise, I feel much less in touch with my emotions, and all predictions feel fuzzier. (It's also possible that my emotions are just unstable or suppressed due to the circumstances.) My feelings about the world vary with my mood, but now I try to correct for this and feel uncertain enough that I defer to a reference class or other people. Since I don't get to actually check on my own feelings, this is bad for practice.

2b. Predictions reify pessimism

Calibrating myself might be a good thing to do in ordinary times, but isolation has made me mildly depressed, causing reduced willpower. Consider a commitment I made recently to study with a peer over Zoom. In ordinary times, there's a 90% chance I keep this commitment. Taking into account my reduced agency, I predict an 80% chance I would do something that I normally do at 90%. However, there's a 65% chance that I actually do something I predict at 80%, so I continue until the fixed point, which is about 25%, which turns out to be accurate. Sometimes this fixed point is 0%.

In the past, I would mentally commit to actions I think I want to take (e.g. meditate regularly) then not actually follow through. Since I have realized how often this happens, I now make very few commitments. This has technically made me much

more trustworthy, but the number of commitments I keep (to myself and others) has decreased in absolute terms.

3. Not Actually Trying

Reading feels *much* better than trying, especially when it requires willpower and time and the outcome is uncertain.

- I've read most of the [Training Regime](#) sequence, but don't actually apply it to my life
- A friend pointed me to [Nonviolent Communication](#), which I have not practiced.
- Several interventions that could plausibly increase my everyday discipline, e.g. [The Real Rules Have No Exceptions](#)

I think this was out of reach in 2014 even if I had developed enough trust in LW to self-modify based on LW1 principles-- EY [notes](#) that the biggest mistake with the original Sequences was neglecting applications and practice opportunities.

4. Disorientation and miscellaneous disruptions

Anna Salamon says that a [particular type of disorientation](#) can result when a new rationalist discards common sense, and manifests as an inability to do the "ordinary" thing when it is correct. After LW discovered that the efficient market hypothesis [is sometimes false](#) relative to strong predictors, I updated strongly towards rationalist exceptionalism in general, which may be correct, but this also increased my disorientation. Some examples I can identify:

- I tried to convince my friend who's a good fit for climate policy to shift to AI policy.
- I noticed that I sometimes need to rationalize my curiosity about the world as something useful.
- [Rationalists beginning to see non-rationalists as normies, NPCs, or otherwise sub-human](#): I now find talking to non-rationalists much less interesting.
- I often sink into a debate mindset that *any proposition* can be true if I make the right argument for it, which I previously only liked to enter while playing devil's-advocate. When arguing for a point, it's slightly more common for me to be unsure whether I actually believe it than before. I have no idea what's going on here since I'm not much better at rhetoric than before. Is my unconscious rebelling against efforts to stop motivated reasoning? Am I trying to [play status games](#)? Should I have resolved my unwillingness to apologize?
- Several counterproductive, intrusive thoughts that haven't gone away for several months of discussions with friends and occasional therapy:
 - My self-worth is derived from my absolute impact on the world-- sometimes causes a vicious cycle where I feel worthless, make plans that take that into account, and feel more worthless.
 - I'm a bad person if I'm a dedicated EA but don't viscerally feel the impact I have.
 - I'm a bad person if I eat meat (despite that vegetarianism is infeasible in the short term due to circumstances, and is a long-term goal for me)

- After thinking about morality for a while, I'm 35% nihilist. This is supposed to not have an effect on my actions-- nihilism can just be subtracted out-- but everything feels approximately 35% hollow.

Conclusion

While I derived benefits from the content, I think it's plausible that COVID-19 was otherwise a bad time to dive headfirst into rationality. If I am to make guidelines for people **exactly like me**, they would be:

- Engage with material that interests you, but recognize discomfort and unhealthy reading patterns.
- Consume material when you can actually practice it (e.g. mentally stable, some minimum amount of slack).
- Practice it (still have to figure out how).

Curiosity: A Greedy Feeling

But what does the phrase “scientifically explicable” mean? It means that someone else knows how the light bulb works. - [Eliezar Yudkowsky](#)

In principle, I agree with the notion that it is unforgivable to not want to know, and not want to improve your map to match the territory. However, even the most curious person in the world cannot maintain equal curiosity about all things, and even if they could there are limits on time and energy. - [Elizabeth](#)

Rarely am I pulled by a specific desire to know. And most of my learning happens at those rare times. - [Anna Salamon](#)

Which question immediately sparks your *specific desire to know*?

- 1) Which enzymes are directly involved in the human transcription of lactase?
- 2) How do cells manufacture proteins?
- 3) Which theory is the best account of abiogenesis?
- 4) Is life inherently valuable?
- 5) If we could grow artificial human organs, would it be possible to cure aging by periodically replacing the worn-out parts of our bodies?

Without some research, I could not supply more than a rudimentary answer for any of these questions. But for me personally, #5 is the only one that sparks my curiosity.

Why? The first three questions represent different strata in the lake of systematized knowledge. If I ever needed an answer, I could look up fine-grained information on each of them. The fourth question is bottomless, and I know in advance that there will be no definitive answer. None of them suggest a pressing reason I'd want to know.

The last question is different. I can immediately see its practical and scholarly significance. Breaking it down into an array of tractable research questions, then determining what is known and unknown, would be an adventure.

In pursuing my inquiry to the limits of scientific knowledge, I would eventually arrive at strange questions, with a void where the answers should be. I'd never normally think to ask these strange questions, nor care particularly about the answers. Now they would be imbued with the significance of my original purpose.

Curiosity must have been a useful sensation for our ancestors, but a costly one. It motivates action, sometimes patterns of action that are tiring, dangerous, and require the efforts of the whole tribe. Learning is costly. So evolution shaped our curiosity into a getting feeling, a greedy feeling.

What is intrinsic motivation?

That makes curiosity sound like it's fueled by extrinsic motivations, which seems counterintuitive.

One [definition of intrinsic motivation](#) is:

Intrinsic work motivation is a cognitive state reflecting the extent to which the worker attributes the force of his or her task behaviors to outcomes derived from the task per se; that is, from outcomes which are not mediated by a source external to the task-person situation. Such a state of motivation can be characterized as a self-fulfilling experience.

OK, this needs some unpacking. Here are two different scenarios:

- John works for BizCorp on salary. He is put on a large team, tasked to build an e-commerce site to sell a new type of widget.
- John has invented a new type of widget. He builds an e-commerce site to sell it.

In the first case, whether or not BizCorp has an e-commerce website, John's getting paid. The outcome of his work, his reward, is "mediated by a source external to the task-person situation."

In the second case, John is not going to get his widget sold without an e-commerce site. His monetary reward, and the satisfaction of bringing a useful new widget to the world, depends directly on creating an e-commerce site. If he finds the creation of the site enjoyable, so much the better.

In the case of curiosity, the reward at the end, tangible or envisioned, will only be available when the curiosity has been satisfied, the quest completed. If you're trying to cure aging, you don't get the reward of immortality unless you succeed. And just think of how much more there would be to learn and experience with that achieved!

If you feel inspired by that objective, even knowing that you personally might die before the investment in research pays back society, then I believe the goal of immortality would constitute an intrinsic reward. So would the pleasure of discovery as you worked in the lab.

So curiosity can be a greedy feeling, a getting feeling, and still stem from intrinsic motivation.

Making yourself curious

Back in 2012, lukeprog wrote [Get Curious](#). It included visualization, meditation, and brainstorming activities for when you *should* be curious, but *aren't*. It follows the narrative in which some inner drive is weak, and we have to conjure up some mysterious inner force to overcome the deficiency.

I believe that lack of curiosity in a seemingly-important question does not stem from lack of drive or laziness.

I think it's that *effective curiosity* is the synthesis of playful, childlike imaginative interest in a topic we don't understand, and our skillful, adult capacities in an activity we're experienced with. If we don't practice synthesizing these two sides of ourselves, or worse, if we only practice one of these potentials and neglect the other, we'll lose our sense of curiosity.

We'll be left as either skillful but stuck, or imaginative but incapable: a Ben Wyatt or an Andy Dwyer, if you watch Parks & Rec. Part of the arc of both characters is a reconnection with the other half of themselves. Ben Wyatt comes to understand that even though his imagination got him into trouble at an early age, now that he's a responsible and focused adult, allowing his imagination back in will allow him to

accomplish amazing things and make his life far more full and rich than it could ever be as a small-town Indiana budget-slasher. Andy Dwyer discovers that when his natural playfulness find a structured outlet as a children's musician, it brings him not just money and respect, but a larger vision for his creativity and a chance to be a leader.

I think the reason why we subscribe to the flagging-drive model of curiosity is that life forces learning upon us. I'm in a calculus course right now, and boy, it would be a lot easier to master that body of knowledge if I was *hungry* to know it. If only there was a way to make myself feel charged up to learn calculus every single day.

I don't know how to do that.

But if the problem isn't forcing yourself to want to learn any arbitrary body of knowledge you choose, but stimulating your passion for investigation, then there is a way forward.

1) Imagine an ability or set of knowledge or invention or experience that would absolutely delight and amaze you *right now* if you could access it. I can think of lots of examples for myself:

- Completely and reversibly being able to change my sex and gender
- Being immortal, or at least very long-lived
- Being able to make new friends easily
- Knowing how to reliably find and observe large animals in the wild
- Be better than the stock market at predicting the economic impact of a pandemic.

2) Pick one, and write down some ideas for how we could potentially achieve that goal, based on what you know right now. For being immortal or very long-lived, I'd write:

- Learning how to grow new organs from my own DNA, and replacing my worn-out parts as they age
- Uploading my brain into a computer controlling a robot, and as my immortal robot-self figuring out how to manufacture a new human body for myself

3) Now start researching the efforts that are underway for #2 (almost certainly, there are some) and try to figure out which seems most practical. Keep a sense of all the sub-questions you might have about your original imagination. Do you need to investigate deeper into one of them, or transition to fleshing out your knowledge on a different sub-question? Always be motivated in your research by the desire to know how to do the impossible, by your desire to turn a fantasy into reality.

These have a strong relationship with lukeprog's three steps for getting curious:

- 1) Feel that you don't already know the answer.
- 2) Want to know the answer.
- 3) Sprint headlong into reality.

But lukeprog's writing is predicated on the idea that you feel as if you know the answer, don't really want to know, and feel no drive to discover it. And this is all because you've selected an arbitrary topic - perhaps a topic life forced on you, or one

your peers think is important, or one you feel you ought to be concerned with - and tried to make yourself get curious about it.

If instead, you start with an imaginative vision that you'd genuinely, immediately feel very excited about if it were realized, the basic circumstances are changed. I know right from the start that I've got no idea how to be immortal and have no idea whether or not progressive organ transplants might accomplish this. I do want to know the answer - that would be fascinating - and I can already think of the first questions I would look up. Googling "could organ transplants make us immortal" would be my first step. This is the sprinting headlong into reality.

Curiosity is not a cure-all: deliberate practice is still hard

What if I am trying to learn calculus, though, and want that hit of curiosity to drive me to study?

I already know that it will be helpful for my studies in grad school. But that's a fairly extrinsic motivation. Oh sure, I believe in an abstract way that calculus will eventually be directly useful for my intended course of study in computational biology. But I also have no ability to imagine how. I know I don't know enough to figure out how to apply it right now.

Games are generally a good way to get your brain activated for an arbitrary goal. So are stories. But part of what makes a game or a story good is that they make the challenges or the language interesting. Shoehorning in the solving of equations is a massive design challenge. You end up with a game that doesn't entertain and a homework assignment that's inefficient.

The best I've ever been able to do is to relate to my studies a little bit like it's a sport.

I get into the zone. I see how quickly, how effortlessly, how accurately I can answer problems. How fast can I read this chapter? What's the craziest organic molecule I can draw and then name?

I try to invent my own problems. I rewrite passages in my own words. I don't just try to answer the exercises and remember the concepts. I try to remember what the exercise *questions* were, or even to make up my own exercise questions. It's the next best thing to tutoring someone else.

I'm not being guided by curiosity. I'm being motivated by competitiveness and restlessness and the desire to test myself, to push the limits. When I'm at my best in my studies, I'm acting more like a fiercely competitive soccer player, or maybe like the stereotype of a hot-shot fighter pilot.

So what keeps a soccer player or a fighter pilot focused on their practice?

One of my skills is playing classical piano music. Some of the most challenging pieces I ever played are J'eux D'eau, by Ravel, and the Ballades of Chopin. I remember what motivated me there.

- The visceral beauty of the sounds they composed, which I was reproducing with my own hands.
- The physical coordination challenging of being able to play a complicated arpeggio flawlessly.

- The physical thrill of speeding up, releasing tension, and stretching my ears out to hear not just notes but music.
- The drive to complete a defined piece of music, to be able to play it from beginning to end, that motivated me to put my butt on the bench every day.
- The interest that comes with varying rubato and dynamics, listening to inner or bass lines, trying to hear melodies not as a sequence of sounds but as a voice.

There was a measure of curiosity here - interest in these pieces, the desire to make them my own. Also of competitiveness, the desire to play some of the most difficult music in the classical canon, or to impress and delight others with my ability

The second remains as I study calculus. It's partly the desire to master math that others find intimidating that motivates me to study calculus. Or to be able to help others with complex, important projects that need a "math guy."

That sense of curiosity is missing to some extent. When I start my third quarter of calculus, I'm not going to know in advance what it entails. There's no way to effortlessly understand as somebody else demonstrates it, the way I can listen to a recording of a Ravel piece I'm considering learning. The doing is identical to the observing. Math is not a spectator sport. That's probably why it's so hard to get people interested in learning it.

My guess is that doing a math assignment can be a thrill, just as much as scrimmage on a soccer team. And that this sense of thrill can be cultivated.

But if you approach a math assignment as though it's a task to complete in order to get a grade, or a tool you're using to force new neural connections in your brain, you won't be paying attention to that sense of thrill.

Where does it come from?

When you're little, playing dodge ball, the game's in motion whether you're ready or not. If the ball hits you, you're out. There's no re-do. It's not a big deal, but secretly, you want to be perfect at dodge ball. You want to hit the other kids with every throw, and never be hit yourself. And you want to do all that while time, and the game, run on.

Can your math assignment be the same way?

Can you connect with time as you work on calculus the same way you connect with it in a game of dodge ball? Can you approach organization on the page, accurate algebra, progress through the substitutions and graphing, with the same tenacity and joy you brought to targeting Johnny on the other team with your rubber ball of death?

Can you find a thrill in the right answer, the same way you felt a thrill when you heard the smack as Johnny got hit by your throw? Can you take delight in noticing you almost dropped a negative sign? Can you approach realizing you missed a step and have to redo the problem the same way you'd have felt getting hit by Johnny when your back was turned - not as frustrated at getting out as you were eager to get back in and seek vengeance?

Math is a lonely game, played against one.

And maybe the idea that a calculus assignment would be as exciting as a childhood game of dodgeball is a pipe dream.

But think about what actually goes into being a fighter pilot. Learning a lot of controls. Switches and dials and messing around with a joystick. You're not running around, you're sitting in a chair, seatbelt on. The wind is not in your hair. If it was the feeling of spinning and high speeds that you loved, a roller coaster would be a safer and easier way to get that sensation than joining the damned air force.

And yet people compete fiercely and work their asses off to be fighter pilots. They're all by themselves up in that plane. It's a lonely game, almost always played against one.

Likewise marksmanship. Likewise the piano. Likewise computer programming. Likewise SCUBA diving.

I think the people who stick with these activities must do a few things:

- They get their workspace set up so that they have everything they need to practice. They eliminate trivial inconveniences.
- They're in the habit of defining small but meaningful projects for themselves, breaking those projects down into steps, and practicing the steps until the whole project attains a smooth and satisfying flow.
- They're looking not just for goal-oriented success but for little, intuitive, hard-to-explain ways of finding a thrill in the practice itself. They make little games for themselves that would seem utterly neurotic and silly to other people. And those self-imposed, idiosyncratic games are actually an important part of the lived experience of practicing for these people.
- They look for opportunities to teach and show off their skills.
- They wear their heart on their sleeve. They want other people to know they're a marksman, a pianist, a programmer, a SCUBA diver, a fighter pilot. They need all those solitary hours of hard work to have social meaning. They want to make their isolated hours of toil visible and tangible to others and to themselves.

This quarter, I'm going to see if I can make my calculus and chemistry classes less like toil and more like soccer or piano practice used to be. I don't know what that means yet, or how in practice I can achieve that. But that impossible-seeming goal - what if calculus and chemistry could be as thrilling and joyful as sports or music - that's the first step in activating curiosity. I think I'm on the right track.

COVID-19: List of ideas to reduce the direct harm from the virus, with an emphasis on unusual ideas

This is a collection of ideas on what interventions could be investigated, funded, or implemented to potentially reduce the harms from COVID-19 by intervening directly on the causal chain leading to infection and death. In particular, it has a lot of “unusual ideas” that haven’t gotten much attention yet.

Caveats: Most of the “unusual ideas” (see later section) were generated or vetted only quickly. Before implementation, they will definitely need further investigation, and some are only meant to serve as inspiration for better interventions. But capacity to implement them if they *do* work can generally be built in parallel with investigating whether they work. Further, it’s also possible that some things here could have consequences worse than the disease (stuff like surveillance states, long term harm from bad vaccinations, or dual use dangerous biotechnology). Therefore, the suggestion here is to pour a lot of resources into investigating those particularly risky options, rather than to immediately and unilaterally implement them.

As the surrounding context to keep in mind, we want to:

- Stop this pandemic in its tracks
- Minimize lives lost, economic disruption, psychological harm, and political harm done directly by the disease or our responses to it
- Perhaps, if lucky, find a way to make humanity better off afterwards

For some related research topics check out [Coronavirus Research Ideas for EAs](#).

A general list of things for interfering with the causal chain of infection and death

In this section, I’ll list categories of interventions to reduce the harm from COVID-19. These should ideally be investigated and implemented in parallel at speed for something as disruptive as this virus. Some of these are being implemented by varying degrees in nations fighting the pandemic.

- Large scale prevention of exposure: Social distancing and managing packages and surfaces
- Small scale of prevention of exposure: Physically blocking the virus before it gets to the body
 - masks
 - sanitation
 - etc.
- Chemical interventions once it has entered the body: Vaccines, antivirals
- Physiological support once it has entered the body: Stuff like ventilator production and supplemental oxygen

- Meta: Investigate the alternative causal pathways to death and hypotheses about them; answering questions like: how much damage and death is due to lung damage, cytokine storms, damage to the nervous system, or interfering with oxygen transportation in the blood?
- Miscellaneous “unusual ideas” (see later section)

Policy for accelerating vaccine development (and similar for antivirals and possibly some other potential solutions)

- Many candidates can be tested in parallel for speed
- To speed up the testing of a specific vaccine we can skip directly to later stage trials - exposing humans to the virus
- We can accelerate testing for long-term negative effects by having a large and diverse set of volunteers take it at once - effects that are rare or slow to show themselves will be detected much faster this way
- Slow and rare effects can plausibly also be detected more easily by selecting people based on weakness to the plausible rare or slow effect, or alternatively exposing people to conditions that would cause a plausible rare effect or slow effect to show itself
- In parallel with testing, production capability and perhaps even stocks of plausible vaccine candidates can be built up and even distributed so they are ready to use on a moment's notice if the vaccine works. ([Bill Gates is now doing this.](#))
- See also the this [EA forum post](#)

Note: Some of these ideas will face ethical objections or may be hard politically to implement.

Unusual ideas

Minimizing environmental exposure

- Coppering surfaces to decrease the virus lingering time (this effect may not last much beyond when the copper surface oxidizes)
- Leaving soap residues on surfaces to decrease virus lingering time (plausible because covid-19 is surrounded by a lipid membrane, but this may not work at normal temperature and humidity)
- There may be other alternative better antiviral coatings that can be coated on surfaces
- [Having UVc lights of the ideal wavelength in people dense areas](#)
 - Possibly having them on when people are not around
 - Maybe you could even build something to shine in the lungs, unlikely though
- Assuming COVID-19 exhibits seasonality: changing the temperature, humidity, or wind patterns inside people-dense areas to induce artificial summer and

"outside"

Social distancing measures and management

- Applying even more sophisticated strategies for contact tracing and immunity
 - Remove the small-world network hubs or perhaps just wait for them to become immune (or replace them with immune folks) to reduce the [viral reproduction number](#)
 - Highly privacy preserving contact tracing: <https://covid-watch.org> and <https://www.novid.org>
 - [Take into account variance of R₀](#) when doing contact tracing and lockdowns
 - Do [pooled](#) and random sample testing to more efficiently use limited testing ability and search for community spread
 - As an idea for inspiration (not feasible as is), [change the network topology of interactions](#)
 - Another idea for inspiration (not feasible as is), move the recovered people around, integrate them into the social web, and have them act as firebreaks. You take recovered or otherwise immune people and move them to places where there is an outbreak in order to drop R₀ below 1
- Apply a more synchronous form of lockdown. In theory if everyone was in lockdown for a month at the same time we'd be able to identify essentially everyone who has it and have the mild cases non contagious by the end of that time period (we'd still have to be careful with contaminated surfaces (especially frozen or refrigerated items)). Something like this could very well be cheaper and easier than an extended partial lockdown.
- In order to spread out the timing of when people are sick, a staggered form of quarantine where you expose some people earlier than others could be implemented. This option is not a good option and more of a last resort to spread out when people are at the hospitals and give us more time to come up with something better. It is better than just letting the epidemic spread freely though. Perhaps implement something like this [controlled infection](#)
 - Special note: This strategy is risky because [the more generations it evolves for the more chances it has to evolve into another form](#), and flattening the curve increases the depth in generations
- Apply complicated forms of the "dance": things like varying the degree and intensity of tracing and lockdown over time and locations dynamically in response to contagion (perhaps using something like a [PID controller](#)). For some concreteness check out:
 - [Coronavirus: The Hammer and the Dance](#)
 - [Impact of non-pharmaceutical interventions \(NPIs\) to reduce COVID-19 mortality and healthcare demand](#)
 - [Health Before Wealth: The Economic Logic](#)
 - [Simulating Covid-19](#)
 - [COVID-19 modelling is wrong](#)
 - [Simulating an Epidemic](#)
 - [Views of economists' on relevant questions](#)
- Ubiquitous health monitoring (temperature sensors, cough sound detectors, ...) - this can be done using either attached or far away sensors
- Figure out how to make the symptoms more apparent to people (by education, customized biotracking, something that amplifies symptoms (stuff like the opposite of a cough suppressant or fever reducer?)), ...)

Decrease the initial viral load

- [Expose people to minimal viral loads in the best vector](#) - say an injection in the muscle, a scratch, or the gastrointestinal system

Chemical means of countering covid 19 infections

- Put the entire population on weak immune boosters (stuff like vitamins D and C, selenium, ...)
- Put the entire population on stronger forms of immune system boosters (I think there are some immune system boosters in certain vaccines). Alternatively maybe just hand them out to be used at the first sign of illness
- [Prime the immune system](#) before or during the earlier stages of the disease
- Put the entire population on other immune boosting strategies: intermittent fasting, cold showers, saunas, good sleep, getting to the [ideal weight for respiratory disease](#), oxytocin/social support/hugs, placebo pills, ...
- Avoid immune system suppressants like steroid chemicals, NSAIDs and Tylenol, and symptom suppressants (symptom suppressant use may also increase contagion by enabling people to go out without visible signs of illness (to themselves or others) and select for more dangerous versions of the virus)
- Lots of speculative chemicals like quercetin, rosemary, ...
- Transfer blood plasma/antibodies from the recovered to the ill (some people are implementing this) - possibly boost this by hooking up circulatory systems or boosting the recovered antibody production (perhaps by transferring some more viral load later)
- Transfer blood from the young to the old - probably need immune suppressors for this because you'd need to transfer immune cells and in any event this shouldn't work because you very likely need to generate new immune cells to handle the disease
- Disperse zinc lozenges to the population to use as soon as they even speculate they are coming down with something in order to kill the virus in their throat and perhaps upper respiratory system to decrease or perhaps eliminate their viral load. High percentage ethanol vapor can also produce this effect.
- Perhaps the high temperature and humidity of saunas can kill the virus in one's upper respiratory system and perhaps help clear the lower lungs
- Disperse a coat of antiviral and non-cytotoxic oil, docosanol, deep into the lungs - Danielle Fong had this idea but she worries about it gumming up the lungs and it hasn't gone anywhere yet
- See if chemicals useful for dealing with [high altitude](#) or [boosting oxygen](#) levels help
- Aerobic exercise to boost lung capacity and cardiovascular fitness before exposure (but not after)
- Decrease exposure to air pollution and smoke
- Look into what athletes use when doping to boost red blood cell count and hemoglobin levels
- Use chemicals that decrease cytokine storms (like possibly vitamin C, melatonin, [tocilizumab](#), and drugs that help with autoimmune disorders)
- Maybe some of the chemicals used in preventing autophagy, stroke damage, etc, can be used to prevent damage in this case
- Maybe boost antioxidants to inhibit damage to the lungs and inhibit cytokine storms

- Maybe inject lots of ACE2 receptors into the body and lungs to divert the virus away from its target
- Castration imitation chemicals, to regenerate the thyroid and possibly return it to a state similar to that of younger people where the death rate is lower. I recall reading a discussion on this somewhere but cannot find it again.
- Have people take estrogen or phytoestrogens, assuming that is relevant for the difference between the male and female fatality rates: [Estrogen receptor impairs interleukin-6 expression by preventing protein binding on the NF-kappaB site.](#)

Non-chemical ways of countering covid 19 infections

- Implement sophisticated forms of [liquid breathing](#)
- Maybe there is a way to ultrasonically shake the lungs to clear them, similar to what is used for kidney stones
- Maybe there is truth to the possibility that some forms of near infrared can boost healing. Maybe bathe the patient or their lungs (somehow) in it
- See if increasing partial air pressure can help for blood oxygenation
- Have people [implement prone positions](#) in the earlier stages of the disease or for self care if hospitals are overwhelmed

Conclusion

I again wish to emphasize that the ideas collected here range from strongly implied by science to be helpful (such as copper surfaces) to the very speculative (such as an antiviral "oil" for the lungs), and that some have the potential to cause more harm than the disease itself. Many of these ideas would need to be investigated further before implementation, though capacity to implement could be built up at the same time. And hopefully some of the "unusual ideas" can inspire more such ideas from others. In general, people need to move fast to combat COVID-19, but also need to be thoughtful, and to avoid [information hazards](#) and highly risky actions.

What do you think of these ideas; how probable, costly, and effective would they be?

Do you have additional ideas you think would be good to investigate and implement?

I hope that promising ideas from here will inspire additional progress and will be further investigated, forwarded to relevant parties, and acted upon if sufficiently vetted and developed.

Thanks to [David Kristoffersson](#), [Michael Aird](#), and [Elizabeth Van Nostrand](#) for editing help, to [Alexey Turchin](#) for sharing some ideas, and to the many people who have been coming up with great ideas for fighting the pandemic, many of which I've included here or have been inspired by.

DeepMind team on specification gaming

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

This is a linkpost for <https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity>

[Specification gaming: the flip side of AI ingenuity](#)