Best of LessWrong: March 2020

1. Credibility of the CDC on SARS-CoV-2
2. Interfaces as a Scarce Resource
3. LessWrong Coronavirus Agenda
4. Authorities and Amateurs
5. Cortés, Pizarro, and Afonso as Precedents for Takeover
6. A Significant Portion of COVID-19 Transmission Is Presymptomatic
7. Epistemic standards for “Why did it take so long to invent X?”
8. The Lens, Progerias and Polycausality
9. Ubiquitous Far-Ultraviolet Light Could Control the Spread of Covid-19 and Other Pandemics
10. My current framework for thinking about AGI timelines
11. Effectiveness of Fever-Screening Will Decline
12. "No evidence" as a Valley of Bad Rationality
13. Adding Up To Normality
14. Covid-19 Points of Leverage, Travel Bans and Eradication
15. Price Gouging and Speculative Costs
16. History's Biggest Natural Experiment
17. Can crimes be discussed literally?
18. Toby Ord’s ‘The Precipice’ is published!
19. What should we do once infected with COVID-19?
20. Thinking About Filtered Evidence Is (Very!) Hard
21. Crisis and opportunity during coronavirus
22. COVID-19's Household Secondary Attack Rate Is Unknown
23. Peter's COVID Consolidated Brief for 29 March
24. Near-term planning for secondary impacts of coronavirus lockdowns [assuming things don't blow up]
25. Simulacra and Subjectivity
26. Does the 14-month vaccine safety test make sense for COVID-19?
27. Coronavirus is Here
28. High Variance Productivity Advice
29. LessWrong Coronavirus Link Database
30. Is the coronavirus the most important thing to be focusing on right now?
31. [UPDATED] COVID-19 cabin secondary attack rates on Diamond Princess
32. ODE to Joy: Insights from 'A First Course in Ordinary Differential Equations'
33. LW Team Updates: Pandemic Edition (March 2020)
34. Blog Post Day II
35. The Case for Privacy Optimism
36. The Hammer and the Dance
37. Coronavirus Justified Practical Advice Summary
38. [Update: New URL] Today's Online Meetup: We're Using Mozilla Hubs
39. COVID-19 article translations site aims to bridge the knowledge gap
40. If I interact with someone with nCov for an hour, how likely am I to get nCov?
41. What is a School?
42. How Do You Convince Your Parents To Prep? To Quarantine?
43. How to have a happy quarantine
44. Zoom In: An Introduction to Circuits
45. Growth rate of COVID-19 outbreaks
46. What are the most plausible "AI Safety warning shot" scenarios?
47. Does SARS-CoV-2 utilize antibody-dependent enhancement?
48. Good News: the Containment Measures are Working
49. mind viruses about body viruses
50. Adaptive Immune System Aging

Best of LessWrong: March 2020

1. Credibility of the CDC on SARS-CoV-2
2. Interfaces as a Scarce Resource
3. LessWrong Coronavirus Agenda
4. Authorities and Amateurs
5. Cortés, Pizarro, and Afonso as Precedents for Takeover
6. A Significant Portion of COVID-19 Transmission Is Presymptomatic
7. Epistemic standards for “Why did it take so long to invent X?”
8. The Lens, Progerias and Polycausality
9. Ubiquitous Far-Ultraviolet Light Could Control the Spread of Covid-19 and Other Pandemics
10. My current framework for thinking about AGI timelines
11. Effectiveness of Fever-Screening Will Decline
12. "No evidence" as a Valley of Bad Rationality
13. Adding Up To Normality
14. Covid-19 Points of Leverage, Travel Bans and Eradication
15. Price Gouging and Speculative Costs
16. History's Biggest Natural Experiment
17. Can crimes be discussed literally?
18. Toby Ord’s ‘The Precipice’ is published!
19. What should we do once infected with COVID-19?
20. Thinking About Filtered Evidence Is (Very!) Hard
21. Crisis and opportunity during coronavirus
22. COVID-19's Household Secondary Attack Rate Is Unknown
23. Peter's COVID Consolidated Brief for 29 March
24. Near-term planning for secondary impacts of coronavirus lockdowns [assuming things don't blow up]
25. Simulacra and Subjectivity
26. Does the 14-month vaccine safety test make sense for COVID-19?
27. Coronavirus is Here
28. High Variance Productivity Advice
29. LessWrong Coronavirus Link Database
30. Is the coronavirus the most important thing to be focusing on right now?
31. [UPDATED] COVID-19 cabin secondary attack rates on Diamond Princess
32. ODE to Joy: Insights from 'A First Course in Ordinary Differential Equations'
33. LW Team Updates: Pandemic Edition (March 2020)
34. Blog Post Day II
35. The Case for Privacy Optimism
36. The Hammer and the Dance
37. Coronavirus Justified Practical Advice Summary
38. [Update: New URL] Today's Online Meetup: We're Using Mozilla Hubs
39. COVID-19 article translations site aims to bridge the knowledge gap
40. If I interact with someone with nCov for an hour, how likely am I to get nCov?
41. What is a School?
42. How Do You Convince Your Parents To Prep? To Quarantine?
43. How to have a happy quarantine
44. Zoom In: An Introduction to Circuits
45. Growth rate of COVID-19 outbreaks
46. What are the most plausible "AI Safety warning shot" scenarios?
47. Does SARS-CoV-2 utilize antibody-dependent enhancement?

48. Good News: the Containment Measures are Working
49. mind viruses about body viruses
50. Adaptive Immune System Aging

Credibility of the CDC on SARS-CoV-2

Introduction

One of the main places Americans look for information on coronavirus is the Center for Disease Control and Prevention (abbreviated CDC from the days before “and Prevention” was in the title). That’s natural; “handling contagious epidemics” is not their only job, but it is one of their primary ones, and they position themselves as the authority. At a time when so many things are uncertain, it saves a lot of anxiety (and time, and money) to have an expert source you can turn to and get solid advice.

Unfortunately, the CDC has repeatedly given advice with lots of evidence against it. Below is a list of actions from the CDC that we believe are misleading or otherwise indicative of an underlying problem. If you know of more examples or have information on any of these (for or against), please comment below and we will incorporate into this post.

Examples

Dismissed Risk of Infection Via Packages

On the CDC’s coronavirus FAQs pages on 2020-03-04, they say, under “Am I at risk for COVID-19 from a package or products shipping from China?”:

“In general, because of poor survivability of these coronaviruses on surfaces, there is likely very low risk of spread from products or packaging that are shipped over a period of days or weeks at ambient temperatures.”

However, this metareview found that various coronaviruses remained infectious for days at room temperature on certain surfaces (cardboard was not tested, alas) and potentially weeks at lower temperatures. The CDC’s answer is probably correct for packages *from China*, and it’s possible it’s even right for domestic packages with 2-day shipping, but it is incorrect to say that coronaviruses in general have low survivability, and to the best of my ability to determine, we don’t have the experiments that would prove deliveries are safe.

Blinded Itself to Community Spread

As late as 2020-02-29, the CDC was reporting that there had been no “community spread” of SARS-CoV-2. (Community spread means that the person hadn’t been traveling in an infected area or associating with someone who had). At this time, the CDC would only test a person for SARS-CoV-2 if they had been in China or in close contact with a confirmed COVID-19 case.

Testing Criteria as of 2020-02-11

Clinical Features	&	Epidemiologic Risk
Fever[1] **or** signs/symptoms of lower respiratory illness (e.g. cough or shortness of breath)	**AND**	Any person, including health care workers, who has had close contact[2] with a laboratory-confirmed[3,4] 2019-nCoV patient within 14 days of symptom onset
Fever[1] **and** signs/symptoms of a lower respiratory illness (e.g., cough or shortness of breath)	**AND**	A history of travel from **Hubei Province**, China[5] within 14 days of symptom onset
Fever[1] **and** signs/symptoms of a lower respiratory illness (e.g., cough or shortness of breath) requiring hospitalization[4]	**AND**	A history of travel from mainland **China**[5] within 14 days of symptom onset

This not only left them incapable of detecting community spread, it ignored potential cases who had travelled to other countries with known COVID-19 outbreaks.

By 2020-02-13, this had been amended to include

> The criteria are intended to serve as guidance for evaluation. Patients should be evaluated and discussed with public health departments on a case-by-case basis. For severely ill individuals, testing can be considered when exposure history is equivocal (e.g., uncertain travel or exposure, or no known exposure) and another etiology has not been identified.

(The CDC describes this change as happening on 2020-02-12, however the Wayback Machine did not capture the page that day).

Based on this announcement on 2020-02-14, when testing that could detect community exposure was happening it was in one of 5 major cities. However as of 2020-03-01 only 472 tests had been done, so no test could have been happening very often.

Between 2020-02-27 and 2020-02-28, the primary guidelines on this page were amended to

Clinical Features	&	Epidemiologic Risk
Fever[1] **or** signs/symptoms of lower respiratory illness (e.g. cough or shortness of breath)	**AND**	Any person, including healthcare workers[2], who has had close contact[3] with a laboratory-confirmed[4] COVID-19 patient within 14 days of symptom onset
Fever[1] **and** signs/symptoms of a lower respiratory illness (e.g., cough or shortness of breath) requiring hospitalization	**AND**	A history of travel from affected geographic areas[5] (see below) within 14 days of symptom onset
Fever[1] with severe acute lower respiratory illness (e.g., pneumonia, ARDS) requiring hospitalization and without alternative explanatory diagnosis (e.g., influenza)[6]	**AND**	No source of exposure has been identified

However guidance went out on the same day (the 28th) that only listed China as a risk (and even then, only medium risk unless they had been exposed to a confirmed case or travelled to Hubei specifically).

Testing Kits the CDC Sent to Local Labs were Unreliable

They generated too many false positives to be useful.

Hamstrung Detection by Banning 3rd Party Testing (HHS/FDA, not CDC)

One reason the CDC used such stringent criteria for determining who to test was that they had a very limited ability to test, hamstrung further by the faulty tests sent to local labs. Normally private testing would fill the gap, but the department of Health and Human Services invoked emergency measures that created a requirement for special approval of tests, and the FDA didn't grant it to anyone (source).

There are multiple harrowing stories of people with obvious symptoms and exposure to the virus being turned away from testing, often against a doctor's pleas:

- UC Davis patient
- NYC ER doctor complains about inability to test
- NYC man returning from Japan
- Northern California nurse who treated infected patient

There is also a rumor that the first case caught in Seattle, which has since turned into the US epicenter of the disease, was caught by a research lab using a loophole to perform unauthorized testing (raising the possibility that it's worse elsewhere and simply hasn't been caught).

Ceased to Report Number of Tests Run

Until 2020-03-02, the CDC reported how many tests SARS-CoV-2 tests it had run. On March 2nd, it stopped (before, after). There are many potential reasons for this, none of which inspire confidence. The official reason for this as told to reporter Kelsey Piper is that the number would no longer be representative now that states are running their own tests. So, best case scenario, the CDC can not coordinate enough to count tests performed by other labs.

Gave False Reassurances About Recovered Individuals

As of this writing (2020-03-05), the CDC's "Share Facts" page states that "Someone who has completed quarantine or has been released from isolation does not pose a risk of infection to other people."

While it is certainly true that being released from quarantine implies a significantly reduced risk, the quarantine that is typically performed is not stringent enough to say that people released pose no risk. The quarantine procedure performed by the CDC lasts 14 days, after which if symptoms have not appeared, they can be released.

There are case reports of individuals with incubation periods of 27 days and 19 days. There was a case in Texas where a person tested positive after being released from quarantine and visiting a mall.

While an epidemic is still contained, safely quarantining at-risk people means choosing a quarantine period long enough to be confident that, if they haven't shown symptoms, they don't have the disease. When a disease is still contained, this should be risk averse, since a single infected person could start an outbreak. The CDC's 14-day quarantine period was not long enough to catch the cases detailed above.

This was foreseeable. This paper, published Feb 6, estimated the distribution of incubation periods, including the incubation periods of outliers.

Percentiles	Incubation period distribution (days)					
	Weibull		Gamma		Lognormal	
	Estimate[a]	95% CI	Estimate[a]	95% CI	Estimate[a]	95% CI
2.5th	2.1	1.3–3.0	2.4	1.5–3.2	2.4	1.6–3.1
5th	2.7	1.8–3.5	2.9	2.0–3.6	2.8	2.0–3.5
50th	6.4	5.5–7.5	6.1	5.3–7.3	6.1	5.2–7.4
95th	10.3	8.6–14.1	11.3	9.1–15.7	13.3	9.9–20.5
97.5th	11.1	9.1–15.5	12.5	9.9–17.9	15.5	11.0–25.2
99th	11.9	9.7–17.2	14.1	10.9–20.6	18.5	12.6–32.2

The relevant row is the 99th percentile row, which estimates the longest incubation period per 100 people. If you quarantined 100 people, one of them would have an incubation period at least that long. The paper estimates this using three different methods; two of those estimates are greater than 14 days, and all three estimates put significant probability on incubation periods longer than 14 days.

There are also reports of the virus re-emerging in patients who were believed to have recovered.

Conflated Genetics and Environmental Exposure

This is a tough topic to write about.

Cruelty to people because they have or might have a disease is never okay. And the vast majority of people who were cruel to Asian-appearing people in the early days of an epidemic were doing it to healthy people out of knee jerk fear and antagonism, not a measured, well-informed cost-benefit analysis. When the CDC claimed on 2020-02-29 that "People of Asian descent, including Chinese Americans, are not more likely to get COVID-19 than any other American." they were surely trying to dampen attacks on people who had done nothing wrong and were hurting no one.

But the statement is false. Chinese-Americans are more likely to travel to China or associate with people who have, and thus were more likely to catch SARS-CoV-2. This doesn’t mean they are more likely to catch it *given exposure*, but they were more likely to be exposed.

The CDC admits this in the page specifically on stigma (2020-02-24), saying “People—including those of Asian descent—who have not recently traveled to China or been in contact with a person who is a confirmed or suspected case of COVID-19 are not at greater risk of acquiring and spreading COVID-19 than other Americans.”

However that same anti-stigma page goes on to say “Viruses cannot target people from specific populations, ethnicities, or racial backgrounds.” This is also false. About 10% of Europeans are immune to HIV, an immunity not found people originating from other areas. So we know it is technically possible for a virus to have differential effects based on race.

Does SARS-CoV-2 in particular have race-related effects? There are people claiming Asian men are more susceptible to SARS-CoV-2 than others due to a higher expression of a certain protein (example). Other people dispute this (example). Right now it is very much an open question.

We can see why the CDC prioritized calming racially-motivated violence over fully explaining their confusion over an unanswered question. It might have been the highest-utility thing to do. But it is important to know that “misrepresenting data in order to produce better actions from the public” is a thing the CDC does.

Discouraged Use of Masks

Which brings us to the CDC’s statement on masks:

CDC does not recommend that people who are well wear a facemask to protect themselves from respiratory diseases, including COVID-19.

The Surgeon General (who is not directly part of the CDC) takes a stronger tack:

While we can't hold the CDC responsible for the Surgeon General, they are being conflated in a lot of news articles saying or implying that masks are useless for healthy people. They're (probably) not.

Our best guess is that the CDC is trying to conserve masks for health care professionals and others with the highest need, in the face of a looming mask shortage. That could easily be the optimum mask allocation. I can't prove the lie wasn't justified for the greater good. But it is another example of the CDC placing "getting the outcome it wants" over "telling people the literal truth."

What Does This Mean?

These errors we've highlighted tend towards errors of omission: saying something is completely safe when it's not, saying something is unhelpful when it is, saying the current state is less dangerous than it is. You should include that bias when processing new information from the CDC. Notably **we're not saying any of the things they do recommend are bad**: to the best of our knowledge, you should be washing your hands and not touching your face. Vaccines are (mostly) great. But I would not take the CDC saying an activity is safe or unnecessary as the last word on the subject.

Interfaces as a Scarce Resource

Outline:

- The first three sections (Don Norman's Fridge, Interface Design, and When And Why Is It Hard?) cover what we mean by "interface", what it looks like for interfaces to be scarce, and the kinds of areas where they tend to be scarce.
- The next four sections apply these ideas to various topics:
    - Why AR is much more difficult than VR
    - AI alignment from an interface-design perspective
    - Good interfaces as a key bottleneck to creation of markets
    - Cross-department interfaces in organizations

Don Norman's Fridge

Don Norman (known for popularizing the term "affordance" in The Design of Everyday Things) offers a story about the temperature controls on his old fridge:

> I used to own an ordinary, two-compartment refrigerator - nothing very fancy about it. The problem was that I couldn't set the temperature properly. There were only two things to do: adjust the temperature of the freezer compartment and adjust the temperature of the fresh food compartment. And there were two controls, one labeled "freezer", the other "refrigerator". What's the problem?
>
> Oh, perhaps I'd better warn you. The two controls are not independent. The freezer control also affects the fresh food temperature, and the fresh food control also affects the freezer.

The natural human model of the refrigerator is: there's two compartments, and we want to control their temperatures independently. Yet the fridge, apparently, does not work like that. Why not? Norman:

> In fact, there is only one thermostat and only one cooling mechanism. One control adjusts the thermostat setting, the other the relative proportion of cold air sent to each of the two compartments of the refrigerator.

It's not hard to imagine why this would be a good design for a cheap fridge: it requires only one cooling mechanism and only one thermostat. Resources are saved by not duplicating components - at the cost of confused customers.

The root problem in this scenario is a mismatch between the structure of the machine (one thermostat, adjustable allocation of cooling power) and the structure of what-humans-want (independent temperature control of two compartments). In order to align the behavior of the fridge with the behavior humans want, somebody, at some point, needs to do the work of translating between the two structures. In Norman's fridge example, the translation is botched, and confusion results.

We'll call whatever method/tool is used for translating between structures an interface. Creating good methods/tools for translating between structures, then, is interface design.

Interface Design

In programming, the analogous problem is *API design*: taking whatever data structures are used by a software tool internally, and figuring out how to present them to external programmers in a useful, intelligible way. If there's a mismatch between the internal

structure of the system and the structure of what-users-want, then it's the API designer's job to translate. A "good" API is one which handles the translation well.

User interface design is a more general version of the same problem: take whatever structures are used by a tool internally, and figure out how to present them to external users in a useful, intelligible way. Conceptually, the only difference from API design is that we no longer assume our users are programmers interacting with the tool via code. We design the interface to fit however people use it - that could mean handles on doors, or buttons and icons in a mobile app, or the temperature knobs on a fridge.

Economically, interface design is a necessary input to make all sorts of things economically useful. How scarce is that input? How much are people willing to spend for good interface design?

My impression is: a lot. There's an entire category of tech companies whose business model is:

- Find a software tool or database which is very useful but has a bad interface
- Build a better interface to the same tool/database
- …
- Profit

This is an especially common pattern among small but profitable software companies; it's the sort of thing where a small team can build a tool and then lock in a small number of very loyal high-paying users. It's a good value prop - you go to people or businesses who need to use X, but find it a huge pain, and say "here, this will make it much easier to use X". Some examples:

- Companies which interface to government systems to provide tax services, travel visas, patenting, or business licensing
- Companies which set up websites, Salesforce, corporate structure, HR services, or shipping logistics for small business owners with little relevant expertise
- Companies which provide graphical interfaces for data, e.g. website traffic, sales funnels, government contracts, or market fundamentals

Even bigger examples can be found outside of tech, where humans themselves serve as an interface. Entire industries consist of people serving as interfaces.

What does this look like? It's the entire industry of tax accountants, or contract law, or lobbying. It's any industry where you could just do it yourself in principle, but the system is complicated and confusing, so it's useful to have an expert around to translate the things-you-want into the structure of the system.

In some sense, the entire field of software engineering is an example. A software engineer's primary job is to translate the things-humans-want into a language understandable by computers. People use software engineers because talking to the engineer (difficult though that may be) is an easier interface than an empty file in Jupyter.

These are not cheap industries. Lawyers, accountants, lobbyists, programmers… these are experts in complicated systems, and they get paid accordingly. The world spends large amounts of resources using people as interfaces - indicating that these kinds of interfaces are a very scarce resource.

When And Why Is It Hard?

Don Norman's work is full of interesting examples and general techniques for accurately communicating the internal structure of a tool to users - the classic example is "handle means pull, flat plate means push" on a door. At this point, I think (at least some) people

have a pretty good understanding of these techniques, and they're spreading over time. But accurate communication of a system's internal structure is only useful if the system's internal structure is itself pretty simple - like a door or a fridge. If I want to, say, write a contract, then I need to interface to the system of contract law; accurately communicating that structure would take a whole book, even just to summarize key pieces.

There are lots of systems which are simple enough that accurate communication is the bulk of the problem of interface design - this includes most everyday objects (like fridges), as well as most websites or mobile apps.

But the places where we see expensive industries providing interfaces - like law or software - are usually the cases where the underlying system is more complex. These are cases where the structure of what-humans-want is very different from the system's structure, and translating between the two requires study and practice. Accurate communication of the system's internal structure is not enough to make the problem easy.

In other words: interfaces to complex systems are especially scarce. This economic constraint is very taut, across a number of different areas. We see entire industries - large industries - whose main purpose is to provide non-expert humans with an interface to a particular complex system.

Given that interfaces to complex systems are a scarce resource in general, what other predictions would we make? What else would we expect to be hard/expensive, as a result of interfaces to complex systems being hard/expensive?

AR vs VR

By the standards of software engineering, pretty much anything in the real world is complex. Interfacing to the real world means we don't get to choose the ontology - we can make up a bunch of object types and data structures, but the real world will not consider itself obligated to follow them. The internal structure of computers or programming languages is rarely a perfect match to the structure of the real world.

Interfacing the real world to computers, then, is an area we'd expect to be difficult and expensive.

Augmented reality (AR) is one area where I expect this to be keenly felt, especially compared to VR. I expect AR applications to lag dramatically behind full virtual reality, in terms of both adoption and product quality. I expect AR will mostly be used in stable, controlled environments - e.g. factory floors or escape-room-style on-location games.

Why is interfacing software with the real world hard? Some standard answers:

- The real world is complicated. This is a cop-out answer which doesn't actually explain anything.
- The real world has lots of edge cases. This is also a cop-out, but more subtly; the real world will only seem to be full of edge cases if our program's ontologies don't line up well with reality. The real question: why is it hard to make our ontologies line up well with reality?

Some more interesting answers:

- The real world isn't implemented in Python. To the extent that the real world has a language, that language is math. As software needs to interface more with the real world, it's going to require more math - as we see in data science, for instance - and not all of that math will be easy to black-box and hide behind an API.
- The real world is only partially observable - even with ubiquitous sensors, we can't query anything anytime the way we can with e.g. a database. Explicitly modelling

things we can't directly observe will become more important over time, which means more reliance on probability and ML tools (though I don't think black-box methods or "programming by example" will expand beyond niche applications).
- We need enough compute to actually run all that math. In practice, I think this constraint is less taut than it first seems - we should generally be able to perform at least as well as a human without brute-forcing exponentially hard problems. That said, we do still need efficient algorithms.
- The real-world things we are interested in are abstract, high-level objects. At this point, we don't even have the mathematical tools to work with these kinds of fuzzy abstractions.
- We don't directly control the real world. Virtual worlds can be built to satisfy various assumptions by design; the real world can't.
- Combining the previous points: we don't have good ways to represent our models of the real world, or to describe what we want in the real world.
- Software engineers are mostly pretty bad at describing what they want and building ontologies which line up with the real world. These are hard skills to develop, and few programmers explicitly realize that they need to develop them.

Alignment

Continuing the discussion from the previous section, let's take the same problems in a different direction. We said that translating what-humans-want-in-the-real-world into a language usable by computers is hard/expensive. That's basically the AI alignment problem. Does the interfaces-as-scarce-resource view lend any interesting insight there?

First, this view immediately suggests some simple analogues for the AI alignment problem. The "Norman's fridge alignment problem" is one - it's surprisingly difficult to get a fridge to do what we want, when the internal structure of the fridge doesn't match the structure of what we want. Now consider the internal structure of, say, a neural network - how well does that match the structure of what we want? It's not hard to imagine that a neural network would run into a billion-times-more-difficult version of the fridge alignment problem.

Another analogue is the "Ethereum alignment problem": we can code up a smart contract to give monetary rewards for anything our code can recognize. Yet it's still difficult to specify a contract for exactly the things we actually want. This is essentially the AI alignment problem, except we use a market in place of an ML-based predictor/optimizer. One interesting corollary of the analogy: there are already economic incentives to find ways of aligning a generic predictor/optimizer. That's exactly the problem faced by smart contract writers, and by other kinds of contract writers/issuers in the economy. How strong are those incentives? What do the rewards for success look like - are smart contracts only a small part of the economy because the rewards are meager, or because the problems are hard? More discussion of the topic in the next section.

Moving away from analogues of alignment, what about alignment paths/strategies?

I think there's a plausible (though not very probable) path to general artificial intelligence in which:

- We figure out various core theoretical problems, e.g. abstraction, pointers to values, embedded decision theory, …
- The key theoretical insights are incorporated into new programming languages and frameworks
- Programmers can more easily translate what-they-want-in-the-real-world into code, and make/use models of the world which better line up with the structure of reality
- … and this creates a smooth-ish path of steadily-more-powerful declarative programming tools which eventually leads to full AGI

To be clear, I don't see a complete roadmap yet for this path; the list of theoretical problems is not complete, and a lot of progress would be needed in non-agenty mathematical modelling as well. But even if this path isn't smooth or doesn't run all the way to AGI, I definitely see a lot of economic pressure for this sort of thing. We are economically bottlenecked on our ability to describe what we want to computers, and anything which relaxes that bottleneck will be very valuable.

Markets and Contractability

The previous section mentioned the Ethereum alignment problem: we can code up a smart contract to give monetary rewards for anything our code can recognize, yet it's still difficult to specify a contract for exactly the things we actually want. More generally, it's hard to create contracts which specify what we want well enough that they can't be gamed.

(Definitional note: I'm using "contract" here in the broad sense, including pretty much any arrangement for economic transactions - e.g. by eating in a restaurant you implicitly agree to pay the bill later, or boxes in a store implicitly agree to contain what they say on the box. At least in the US, these kinds of contracts are legally binding, and we can sue if they're broken.)

A full discussion of contract specification goes way beyond interfaces - it's basically the whole field of contract theory and mechanism design, and encompasses things like adverse selection, signalling, moral hazard, incomplete contracts, and so forth. All of these are techniques and barriers to writing a contract when we *can't* specify exactly what we want. But why can't we specify exactly what we want in the first place? And what happens when we can?

Here's a good example where we can specify exactly what we want: buying gasoline. The product is very standardized, the measures (liters or gallons) are very standardized, so it's very easy to say "I'm buying X liters of type Y gas at time and place Z" - existing standards will fill in the remaining ambiguity. That's a case where the structure of the real world is not too far off from the structure of what-we-want - there's a nice clean interface. Not coincidentally, this product has a very liquid *market*: many buyers/sellers competing over price of a standardized good. Standard efficient-market economics mostly works.

On the other end of the spectrum, here's an example where it's very hard to specify exactly what we want: employing people for intellectual work. It's hard to outsource expertise - often, a non-expert doesn't even know how to tell a job well done from sloppy work. This is a natural consequence of using an expert as an interface to a complicated system. As a result, it's hard to standardize products, and there's not a very liquid market. Rather than efficient markets, we have to fall back on the tools of contract theory and mechanism design - we need ways of verifying that the job is done well without being able to just specify exactly what we want.

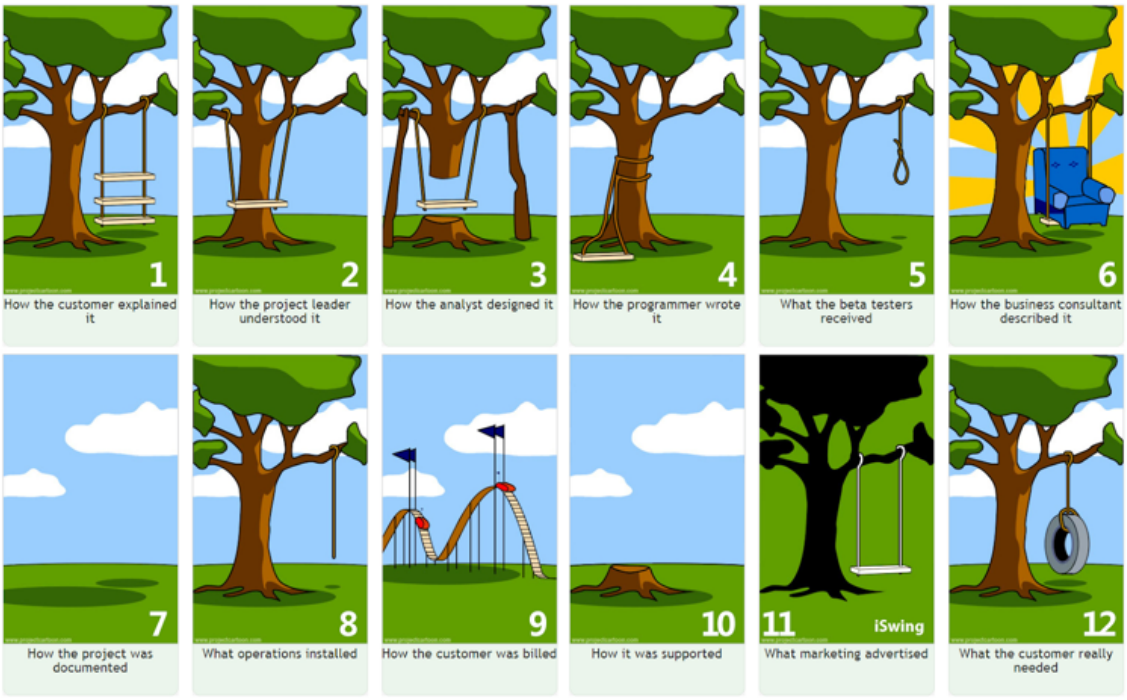
In the worst case, the tools of contract theory are insufficient, and we may not be able to form a contract at all. The lemon problem is an example: a seller may have a good used car, and a buyer may want to buy a good used car, but there's no (cheap) way for the seller to prove to the buyer that the car isn't a lemon - so there's no transaction. If we could fully specify everything the buyer wants from the car, and the seller could visibly verify that every box is checked, cheaply and efficiently, then this wouldn't be an issue.

The upshot of all this is that *good interfaces* - tools for translating the structure of the real world into the structure of what-we-want, and vice versa - enable efficient markets. They enable buying and selling with minimal overhead, and they avoid the expense and complexity of contract-theoretic tools.

Create a good interface for specifying what-people-want within some domain, and you're most of the way to creating a market.

Interfaces in Organizations

Accurately communicating what we want is hard. Programmers and product designers are especially familiar with this:



Incentives are a problem sometimes (obviously don't trust ads or salespeople), but even mostly-earnest communicators - customers, project managers, designers, engineers, etc - have a hard time explaining things. In general, people don't understand which aspects are most relevant to other specialists, or often even which aspects are most relevant to themselves. A designer will explain to a programmer the parts which seem most design-relevant; a programmer will pay attention to the parts which seem most programming-relevant.

It's not just that the structure of what-humans-want doesn't match the structure of the real world. It's that the structure of how-human-specialists-see-the-world varies between different specialists. Whenever two specialists in different areas need to convey what-they-want from one to the other, somebody/something has to do the work of translating between structures - in other words, we need an interface.

A particularly poignant example from several years ago: I overheard a designer and an engineer discuss a minor change to a web page. It went something like this:

Designer: "Ok, I want it just like it was before, but put this part at the top."

Engineer: "Like this?"

Designer: "No, I don't want everything else moved down. Just keep everything else where it was, and put this at the top."

Engineer: "But putting that at the top pushes everything else down."

Designer: "It doesn't need to. Look, just..."

... this went on for about 30 minutes, with steadily increasing frustration on both sides, and steadily increasing thumping noises from my head hitting the desk.

It turned out that the designer's tools built everything from the bottom of the page up, while the engineer's tools built everything from top down. So from the designer's perspective, "put this at the top" did not require moving anything else. But from the engineer's perspective, "put this at the top" meant everything else had to get pushed down.

Somebody/something has to do the translation work. It's a two-sided interface problem.

Handling these sorts of problems is a core function for managers and for anyone deciding how to structure an organization. It may seem silly to need to loop in, say, a project manager for every conversation between a designer and an engineer - but if the project manager's job is to translate, then it can be useful. Remember, the example above was frustrating, but at least both sides *realized* they weren't communicating successfully - if the double illusion of transparency kicks in, problems can crop up without anybody even realizing.

This is why, in large organizations, people who can operate across departments are worth their weight in gold. Interfaces are a scarce resource; people who operate across departments can act as human interfaces, translating model-structures between groups.

A great example of this is the 1986 Goldwater-Nichols Act. It was intended to fix a lack of communication/coordination between branches of the US military. The basic idea was simple: nobody could be promoted to lieutenant or higher without first completing a "joint mission", one in which they worked directly with members of other branches. People capable of serving as interfaces between branches were a scarce resource; Goldwater-Nichols introduced an incentive to create more such people. Before the bill's introduction, top commanders of all branches argued against it; they saw it as congressional meddling. But after the first Iraq war, every one of them testified that it was the best thing to ever happen to the US military.

Summary

The structure of things-humans-want does not always match the structure of the real world, or the structure of how-other-humans-see-the-world. When structures don't match, someone or something needs to serve as an interface, translating between the two.

In simple cases, this is just user interface design - accurately communicating how-the-thing-works to users. But when the system is more complicated - like a computer or a body of law - we usually need human specialists to serve as interfaces. Such people are expensive; interfaces to complicated systems are a scarce resource.

LessWrong Coronavirus Agenda

I've gone through a lot of introductions to this post but maybe this is the most honest one:

I am scared. Quite scared, actually. My chances of catching COVID-19 are actually quite low, and my chances of surviving it if I do are quite high, and I'm still scared. What if I get into a car accident and have to go to the ER? Will they have a bed for me? Will I leave with coronavirus? What are my pregnant friends going to do? What is anyone over 70 going to do?

My goal, and the goal of everyone on the LW staff, and I assume most everyone who's participated in all the coronavirus threads, has been to figure out what is happening and what we can do about it. We've already done a lot. Posts like Seeing the Smoke got coronavirus on people's radar faster than it otherwise would have been, aided by the numerous modeling threads backing it up. The Quarantine Preparations thread gave people a starting place to act from. The Justified Practical Advice (summary) thread let us share our expertise, in ways that led to concrete behavioral changes. More recently we examined asymptomatic transmission. I've had a legit, reasonably high ranking government official say they look at us to see where everyone else will be in weeks.

This is currently the LessWrong team's top priority, and they've done a number of things over the recent weeks to facilitate research and action on coronavirus, including hiring me to be a point person on it. To facilitate as much progress as possible over the coming weeks, habryka and I have compiled a list of what we consider the most important questions in fighting COVID, and are asking anyone with the skill to help us answer them.

That list is at the end of this post. But first, what is the overall plan here?

Who are we trying to help?

We have three broad categories of potential beneficiaries in mind:

1. **Individuals making choices for themselves and their loved ones**, who need accurate information about the current threat level and how to lower it with existing tech.

2. **Individuals creating the tools for the people above**, meaning anything from noticing that copper tape is anti-viral to creating plans for DIY non-invasive ventilators, who need accurate information about how COVID-19 operates and where the current gaps and bottlenecks are. We'd like to help people in this group get volunteers and money when appropriate.

3. **Organizations and institutions making decisions that affect many people**, who need all the information the previous two groups do, plus more to know what the effect of their decisions will be.

How Are We Doing That?

I am managing a Coronavirus Agenda, composed of what myself and habryka think are the most important coronavirus-related questions to answer (think we missed some? Please comment). But the full agenda is kind of overwhelming, and there are benefits to coordinating multiple people around the same question, so every so often I'll pull out Spotlight Questions to generate a critical mass of attention around the most critical questions. I want to say "every so often" will be once a week, but I feel like those kinds of commitments are for situations where I know within an order of magnitude how many people are going to die in that week. I will spotlight as often as seems merited by the situation at the time.

If your eye is caught by a question on the agenda that's not currently spotlighted, of course pursue your interest. That's the point of sharing the whole agenda. And if you think the agenda is missing something important, of course pursue that, and add a comment explaining it if you have time so I can add it.

Without further adieu, the spotlight questions...

Spotlight Questions

What is the impact of varying initial viral load of COVID-19?

The hypothesis that lower initial viral load leads to better outcomes, and might be worth pursuing deliberately, is a central assumption is Zvi's post Taking Initial Viral Load Seriously. Is it true?

Economics Questions

- What happens in a recession anyway?
- How will this recession differ from the last two?
- What will happen to supply chains in the era of COVID-19?

The Full Agenda

These are the questions about coronavirus I and habryka (and in the future, commenters on this post) most want answered. We'll be nudging LessWrong to pursue them over the coming weeks, but for clarity wanted to share the whole thing as a package.

Some of these someone has already answered, or attempted to answer, in which case I've linked to the (attempted) answers. I'll continue to update as more answers come in:

- How many people are infected?
    - Worldwide
    - In a location of your choosing

    - No one suggested a dashboard that met all of my or habryka's goals. PlaguePlus.com is the placeholder winner for at least attempting to do estimates of the true count instead of just reporting test results, and for showing any history instead of just cumulative cases, but I'd sure love for it to be replaced by something that can show history broken down by region.
- What projects need volunteers or donations?
    - We collected a number of suggestions and aggregations in the LessWrong Coronavirus Links DB (see Work & Donate tab), but ultimately didn't find any that were both widely applicable and exciting to us.
- What should I do if I get sick or am caring for someone sick?
    - My answer. This is 80/20ed, not completed.
- What is my prognosis if I get COVID-19?
    - Short Term
        - Attempt 1
    - Long term
        - Attempt 1
- What will the economic effects of covid be?
    - What will happen to supply chains in the era of COVID-19?
    - How will this recession differ from the last two?
    - What happens in a recession anyway?
- What is the basic science of coronavirus?
    - My favorite was this talk by a virology professor, it answered basically all of my questions, but requires too much background biology knowledge to be a perfect intro for everyone.
- What is the impact of varying initial viral load of COVID-19?**Q**
- What are the most predictable second order disasters?
- What problems are people running into when trying to work on all of this? Are there more things like the link database that we need?
- What skills should I be rapidly acquiring to be most useful to this whole situation?
- What mental health problems can we expect to spike hard in the next 1-6 months given people feeling shut in and helpless?
- What are the basic epidemiological parameters of C19, such as incubation rate, doubling times, probability of symptomatic infections, delay from disease onset to death, probability of death among symptomatics, etc?
    - Two documents collecting papers with estimates
- How much food do I need to have stored?
    - I've seen anywhere from 2 weeks to 9 months and given that neither the money nor the space is trivial to everyone, I'd really like to see model-backed estimates.
- What is actual hospital elasticity? Is there an existing gathering of data on this from previous disasters?
- How long should I be in isolation given the median assumptions about the world and the specifics of my area?
- Which physical objects have longer supply chains and thus can be expected to be less robust to disruption?
- What can we do to raise the standard of home care?
    - Potential answers are anything from electrolytes to DIY ventilators.
    - Could postural drainage help?
    - Justified Practical Advice Thread Summary
    - Are most NSAIDs dangerous?
- Is there an asymptomatic infectious phase?
    - Probably. Mean incubation period is 4-9 days, but the mean serial interval (period from when person A is infected to when they infect person B) is 4-6

(and estimates are closer to 4, although averaging different studies is not really appropriate)

- What are the risks of…
  - Accepting delivery food
  - Accepting packages
  - Using public transit
  - Going to work
    - For a variety of workplace types
  - Hosting a large gathering
  - Hosting a small gathering
  - Shaking hands with with an infected individual
  - Walking past an infected individual in a hallway
  - Standing or sitting 4 feet from an infected individual and having 5 minutes of conversation
  - Opening a piece of mail handed to you by an infected person
  - Opening a piece of mail left in your mailbox by an infected person 1 hour ago
  - Holding a grocery bag handed to you by an infected person
  - Picking up an item in the grocery store that was placed on the shelf by that person 1 hour ago
- How do I convince others to act?
  - Attempt 1: for parents
- What is the value of handwashing, when you are currently healthy? How much better is WHO-approved handwashing than what we do by default?
- What is the value of copper taping high-touch surfaces?
- What is the value of masks, when you are currently healthy?
- What is the value of goggles, when you are currently healthy?
- What is the value of contact tracing? How do you do it?
- What are the chances of vaccine development?
  - What are the costs, benefits, and logistics of opening up new vaccine facilities?
- What are the chances of treatment development?
- Do we actually have any chance of an approach that is not herd-immunity based? Is there still any chance at containment?

Authorities and Amateurs

People are writing a lot about the coronavirus, and I've seen a lot of pushback on how pieces often haven't been written by people with epidemiology or public health credentials. For example, Flatten the Curve of Armchair Epidemiology, Listen To Actual Experts On Coronavirus, and comments like this one. The argument that we should be listening to experts and not random people would make a lot of sense if the "armchair" folks didn't keep being right.

Let's look at the articles they're criticizing for having non-expert authors:

- Coronavirus: Act Today or People Will Die (3/10): Argues that we need to get everyone to stay home immediately.
- Cancel Everything (3/10): Argues essentially the same point.
- Flattening The Curve Is a Deadly Delusion (3/13): Argues that our medical capacity is so much lower than the likely peak that we need an immediate lockdown and a renewed focus on containment.

With two weeks of perspective, however, these articles were exactly right. They clearly laid out the case for decisive action, and if we had followed their prescriptions more closely we would be in much better shape right now.

This goes beyond a few articles, however. All the aspects of this crisis that have involved planning more than a couple weeks out have been very poorly handled:

- People weren't told to stock up on food so that they'd be able to reduce trips outdoors, and so that they'd have food in case they were quarantined for 2+ weeks. Stores weren't told to prepare for a rush. A government that was on top of things could have started advocating this in early February in an "if you can afford to" way. This would have spread people's buying over a longer period and avoided the empty shelves we see now. Instead, once restaurants were closing and people realized that they could be quarantined at any time, everyone simultaneously tried to buy weeks worth of ingredients and we had widespread shortages of basic goods starting in mid March.
- Hospitals weren't told (or allowed?) to ration personal protective equipment such as masks until they had shortages. The CDC didn't publish guidelines for sanitization and reuse, and start telling people to conserve. In weeks of handling initial cases, hospitals burned through amounts that would have lasted months with careful rationing.
- The federal government, state governments, or even hospitals could have placed emergency ventilator production orders in February, but didn't. Because we don't allow price gouging, ventilator companies can't ramp up production speculatively figuring that if there is really an epidemic then they'll make their money back. By mid March it was obvious that we were far short of where we needed to be and the companies started ramping up but we lost about a month of production increase.
- Masks were sitting on shelves across the country, and the government could have requisitioned them for emergency medical use, or even just gone and

bought them. Instead the Surgeon General tweeted a request that people not buy them.

- Testing has been completely messed up, though it's hard to tell how much was bad luck vs reasonable rationing of scarce tests. But we should have been quarantining people who seemed to have it based on symptoms, instead of saying "well, since we can't test you we have to assume you don't have it, so you're welcome to continue living your life".

- We could have planned and built out COVID-specific facilities to reduce the risk of others getting sick and make more efficient use of ventilators and personal protective equipment. We're just starting to do this now, much too late.

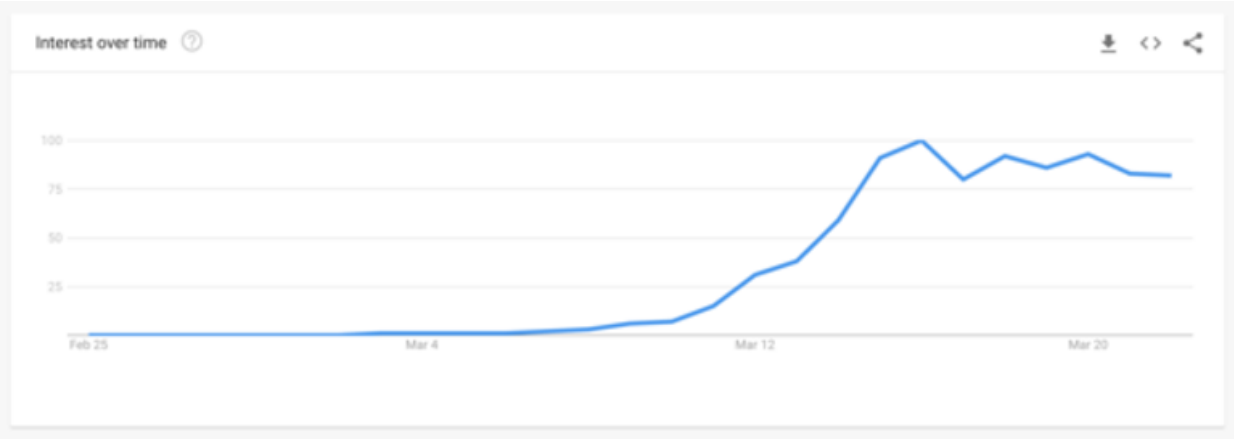
- The passengers of the Grand Princess were offered testing but told that if they tested positive they would be required to undergo quarantine. Of 858 passengers, 568 declined testing and were released without even advice to self-isolate.

- The CDC is still (3/24) not recommending the general public avoid contact with others unless you're sick, they're sick, or you know covid-19 is spreading in your community. They're only recommending people stay home if they're sick.

And I understand: this is moving very quickly, and authorities aren't used to needing to respond so rapidly. But there was a meme going around:

> **Neil Diamond**: touching hands
> **CDC**: no don't touch hands
> **Neil Diamond**: reaching out
> **CDC**: please avoid that
> **Neil Diamond**: TOUCHING YOU-
> **CDC**: everyone is Boston is doomed
> —@actioncookbook (2/27)

This joke and its many copycats feature the CDC we wish we had. A CDC that would have been pushing social distancing a month ago, when it would have helped so much more.



Google Trends: "social distancing"

If we had listened to the warnings and prepared better we would have the experts we need, with the influence to get policies changed, and we wouldn't need the advice of

the armchair epidemiologists. But that's not the world we've found ourselves in, and the amateurs have been doing critical work filling in for them in pushing policy.

A policy of "listen to random experts" is better than a policy of "listen to random amateurs". But rejecting the arguments of amateurs who were making clear arguments, solely on the grounds of their non-expert status, was harmful here.

Cortés, Pizarro, and Afonso as Precedents for Takeover

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Crossposted from AI Impacts.

Epistemic status: I am not a historian, nor have I investigated these case studies in detail. I admit I am still uncertain about how the conquistadors were able to colonize so much of the world so quickly. I think my ignorance is excusable because this is just a blog post; I welcome corrections from people who know more. If it generates sufficient interest I might do a deeper investigation. Even if I'm right, this is just one set of historical case-studies; it doesn't prove anything about AI, even if it is suggestive. Finally, in describing these conquistadors as "successful," I simply mean that they achieved their goals, not that what they achieved was good.

Summary

In the span of a few years, some minor European explorers (later known as the conquistadors) encountered, conquered, and enslaved several huge regions of the world. That they were able to do this is surprising; their technological advantage was not huge. (This was before the scientific and industrial revolutions.) From these cases, I think we learn that it is occasionally possible for a small force to quickly conquer large parts of the world, despite:

1. Having only a minuscule fraction of the world's resources and power
2. Having technology + diplomatic and strategic cunning that is better but not *that* much better
3. Having very little data about the world when the conquest begins
4. Being disunited

Which all suggests that it isn't as implausible that a small AI takes over the world in mildly favorable circumstances as is sometimes thought.

EDIT: In light of good pushback from people (e.g. Lucy.ea8 and e.g. Matthew Barnett) about the importance of disease, I think one should probably add a caveat to the above: "In times of chaos & disruption, at least."

NEW EDIT: After reading three giant history books on the subject, I take back my previous edit. My original claims were correct.

Three shocking true stories

I highly recommend you read the wiki pages yourself; otherwise, here are my summaries:

Cortés: [wiki] [wiki]

- April 1519: Hernán Cortés lands in Yucatan with ~500 men, 13 horses, and a few cannons. He destroys his ships so his men won't be able to retreat. His goal is to conquer the Aztec empire of several million people.
- He makes his way towards the imperial capital, Tenochtitlán. Along the way he encounters various local groups, fighting some and allying with some. He is constantly outnumbered but his technology gives him an advantage in fights. His force grows in size, because even though he loses Spaniards he gains local allies who resent Aztec rule.
- Tenochtitlán is an island fortress (like Venice) with a population of over 200,000, making it one of the largest and richest cities *in the world* at the time. Cortés arrives in the city asking for an audience with the Emperor, who receives him warily.
- Cortés takes the emperor hostage within his own palace, indirectly ruling Tenochtitlán through him.
- Cortés learns that the Spanish governor has landed in Mexico with a force twice his size, intent on arresting him. (Cortés' expedition was illegal!) Cortés leaves 200 men guarding the Emperor, marches to the coast with the rest, surprises and defeats the new Spaniards in battle, and incorporates the survivors into his army.
- July 1520: Back at the capital, the locals are starting to rebel against his men. Cortés marches back to the capital, uniting his forces just in time to be besieged in the imperial palace. They murder the emperor and fight their way out of the city overnight, taking heavy losses.
- They shelter in another city (Tlaxcala) that was thinking about rebelling against the Aztecs. Cortés allies with the Tlaxcalans and launches a general uprising against the Aztecs. Not everyone sides with him; many city-states remain loyal to Tenochtitlan. Some try to stay neutral. Some join him at first, and then abandon him later. Smallpox sweeps through the land, killing many on all sides and causing general chaos.
- May 1521: The final assault on Tenochtitlán. By this point, Cortés has about 1,000 Spanish troops and 80,000 - 200,000 allied native warriors. He had 16 cannons and 13 boats. The Aztecs have 80,000 - 300,000 warriors and 400 boats. Cortés and his allies win.
- Later, the Spanish would betray their native allies and assert hegemony over the entire region, in violation of the treaties they had signed.

Pizarro [wiki] [wiki]

- 1532: Francisco Pizarro arrives in Inca territory with 168 Spanish soldiers. His goal is to conquer the Inca empire, which was much bigger than the Aztec empire.
- The Inca empire is in the middle of a civil war and a devastating plague.
- Pizarro makes it to the Emperor right after the Emperor defeats his brother. Pizarro is allowed to approach because he promises that he comes in peace and will be able to provide useful information and gifts.
- At the meeting, Pizarro ambushes the Emperor, killing his retinue with a volley of gunfire and taking him hostage. The remainder of the Emperor's forces in the area back away, probably confused and scared by the novel weapons and hesitant to keep fighting for fear of risking the Emperor's life.
- Over the next months, Pizarro is able to leverage his control over the Emperor to stay alive and order the Incans around; eventually he murders the Emperor and makes an alliance with local forces (some of the Inca generals) to take over the capital city of Cuzco.

- The Spanish continue to rule via puppets, primarily Manco Inca, who is their puppet ruler while they crush various rebellions and consolidate their control over the empire. Manco Inca escapes and launches a rebellion of his own, which is partly successful: He utterly wipes out four columns of Spanish reinforcements, but is unable to retake the capital. With the morale and loyalty of his followers dwindling, Manco Inca eventually gives up and retreats, leaving the Spanish still in control.
- Then the Spanish ended up fighting *each other* for a while, while *also* putting down more local rebellions. After a few decades Spanish dominance of the region is complete. (1572).

Afonso [wiki] [wiki] [wiki]

- 1506: Afonso helps the Portuguese king come up with a shockingly ambitious plan. *Eight years* prior, the first Europeans had rounded the coast of Africa and made it to the Indian Ocean. The Indian Ocean contained most of the world's trade at the time, since it linked up the world's biggest and wealthiest regions. See this map of world population (timestamp 3:45). Remember, this is prior to the Industrial and Scientific Revolutions; Europe is just coming out of the Middle Ages and does not have an obvious technological advantage over India or China or the Middle East, and has an obvious economic *disadvantage*. And Portugal is a just tiny state on the edge of the Iberian peninsula.
- The plan is: Not only will we go into the Indian Ocean and participate in the trading there -- cutting out all the middlemen who are currently involved in the trade between that region and Europe -- we will *conquer strategic ports around the region so that no one else can trade there!*
- Long story short, Afonso goes on to complete this plan by 1513. (!!!)

Some comparisons and contrasts:

- Afonso had more European soldiers at his disposal than Cortes or Pizarro, but not many more -- usually he had about a thousand or so. He did have more reinforcements and support from home.
- Like them, he was usually significantly outnumbered in battles. Like them, the empires he warred against were vastly wealthier and more populous than his forces.
- Like them, Afonso was often able to exploit local conflicts to gain local allies, which were crucial to his success.
- Unlike them, his goal wasn't to conquer the empires entirely, just to get and hold strategic ports.
- Unlike them, he was fighting empires that were technologically advanced; for example, in several battles his enemies had more cannons and gunpowder than he did.
- That said, it does seem that Portuguese technology was qualitatively better in some respects (ships, armor, and cannons, I'd say.) Not dramatically better, though.
- While Afonso's was a naval campaign, he did fight many land battles, usually marine assaults on port cities, or defenses of said cities against counterattacks. So superior European naval technology is not by itself enough to explain his victory, though it certainly was important.
- Plague and civil war were not involved in Afonso's success.

What explains these devastating conquests?

Wrong answer: I cherry-picked my case studies.

History is full of incredibly successful conquerors: Alexander the Great, Ghenghis Khan, etc. Perhaps some people are just really good at it, or really lucky, or both.

However: Three incredibly successful conquerors from the same tiny region and time period, conquering three separate empires? Followed up by dozens of less successful but still very successful conquerors from the same region and time period? Surely this is not a coincidence. Moreover, it's not like the conquistadors had many failed attempts and a few successes. The Aztec and Inca empires were the two biggest empires in the Americas, and there weren't any other Indian Oceans for the Portuguese to fail at conquering.

Fun fact: I had not heard of Afonso before I started writing this post this morning. Following the Rule of Three, I needed a third example and I predicted on the basis of Cortes and Pizarro that there would be other, similar stories happening in the world at around that time. That's how I found Afonso.

Right answer: Technology

However, I don't think this is the whole explanation. The technological advantage of the conquistadors was not overwhelming.

Whatever technological advantage the conquistadors had over the existing empires, it was the sort of technological advantage that one could acquire *before* the Scientific and Industrial revolutions. Technology didn't change very fast back then, yet Portugal managed to get a lead over the Ottomans, Egyptians, Mughals, etc. that was sufficient to bring them victory. On paper, the Aztecs and Spanish were pretty similar: Both were medieval, feudal civilizations. I don't know for sure, but I'd bet there were at least a few techniques and technologies the Aztecs had that the Spanish didn't. And of course the technological similarities between the Portuguese and their enemies were much stronger; the Ottomans even had access to European mercenaries! Even in cases in which the conquistadors had technology that was completely novel -- like steel armor, horses, and gunpowder were to the Aztecs and Incas -- it wasn't god-like. The armored soldiers were still killable; the gunpowder was more effective than arrows but limited in supply, etc.

(Contrary to popular legend, neither Cortés nor Pizarro were regarded as gods by the people they conquered. The Incas concluded pretty early on that the Spanish were mere men, and while the idea did float around the Aztecs for a bit the modern historical consensus is that most of them didn't take it seriously.)

Ask yourself: Suppose Cortés had found 500 local warriors, gave them all his equipment, trained them to use it expertly, and left. Would those local men have taken over all of Mexico? I doubt it. And this is despite the fact that they would have had much better local knowledge than Cortés did! Same goes for Pizarro and Afonso. Perhaps if he had found 500 local warriors *led by an exceptional commander* it would work. But the explanation for the conquistador's success can't just be that they were all exceptional commanders; that would be positing too much innate talent to occur in one small region of the globe at one time.

Right answer: Strategic and diplomatic cunning

This is my non-expert guess about the missing factor that joins with technology to explain this pattern of conquistador success.

They didn't just have technology; they had *effective strategy* and they had *effective diplomacy*. They made long-term plans that *worked* despite being breathtakingly ambitious. (And their short-term plans were usually pretty effective too, read the stories in detail to see this.) Despite not knowing the local culture or history, these conquistadors made surprisingly savvy diplomatic decisions. They knew when they could get away with breaking their word and when they couldn't; they knew which outrages the locals would tolerate and which they wouldn't; they knew how to convince locals to ally with them; they knew how to use words to escape militarily impossible situations... The locals, by contrast, often badly misjudged the conquistadors, e.g. not thinking Pizarro had the will (or the ability?) to kidnap the emperor, and thinking the emperor would be safe as long as they played along.

This raises the question, how did they get that advantage? My answer: they had *experience* with this sort of thing, whereas locals didn't. Presumably Pizarro learned from Cortés' experience; his strategy was pretty similar. (See also: the prior conquest of the Canary Islands by the Spanish). In Afonso's case, well, the Portuguese had been sailing around Africa, conquering ports and building forts for more than a hundred years.

Lessons I think we learn

I think we learn that:

It is occasionally possible for a small force to quickly conquer large parts of the world, despite:

1. Having only a minuscule fraction of the world's resources and power
2. Having technology + diplomatic and strategic cunning that is better but not *that* much better
3. Having very little data about the world when the conquest begins
4. Being disunited

Which all suggests that it isn't as implausible that a small AI takes over the world in mildly favorable circumstances as is sometimes thought.

EDIT: In light of good pushback from people (e.g. Lucy.ea8 and e.g. Matthew Barnett) about the importance of disease, I think one should probably add a caveat to the above: "In times of chaos & disruption, at least."

Having only a minuscule fraction of the world's resources and power

In all three examples, the conquest was more or less completed without support from home; while Spain/Portugal did send reinforcements, it wasn't even close to the entire nation of Spain/Portugal fighting the war. So these conquests are examples of non-state entities conquering states, so to speak. (That said, their *claim* to represent a large state may have been crucial for Cortes and Pizarro getting audiences and respect initially.) Cortés landed with about a thousandth the troops of Tenochtitlan, which controlled a still larger empire of vassal states. Of course, his troops were better equipped, but on the other hand they were also cut off from resupply, whereas the

Aztecs were in their home territory, able to draw on a large civilian population for new recruits and resupply.

The conquests succeeded in large part due to diplomacy. This has implications for AI takeover scenarios; rather than imagining a conflict of humans vs. robots, we could imagine humans vs. humans-with-AI-advisers, with the latter faction winning and somehow by the end of the conflict the AI advisers have managed to become *de facto* rulers, using the humans who obey them to put down rebellions by the humans who don't.

Having technology + diplomatic and strategic skill that is better but not *that* much better

As previously mentioned, the conquistadors didn't enjoy god-like technological superiority. In the case of Afonso the technology was pretty similar. Technology played an important role in their success, but it wasn't enough on its own. Meanwhile, the conquistadors may have had more diplomatic and strategic cunning (or experience) than the enemies they conquered. But not that much more--they are only human, after all. And their enemies were pretty smart.

In the AI context, we don't need to imagine god-like technology (e.g. swarms of self-replicating nanobots) to get an AI takeover. It might even be possible without any new physical technologies at all! Just superior software, e.g. piloting software for military drones, targeting software for anti-missile defenses, cyberwarfare capabilities, data analysis for military intelligence, and of course excellent propaganda and persuasion.

Nor do we need to imagine an AI so savvy and persuasive that it can persuade anyone of anything. We just need to imagine it about as cunning and experienced relative to its enemies as Cortés, Pizarro, and Afonso were relative to theirs. (Presumably no AI would be experienced with world takeover, but perhaps an intelligence advantage would give it the same benefits as an experience advantage.) And if I'm wrong about this explanation for the conquistador's success--if they had no such advantage in cunning/experience--then the conclusion is even stronger.

Additionally, in a rapidly-changing world that is undergoing slow takeoff, where there are lesser AIs and AI-created technologies all over the place, most of which are successfully controlled by humans, AI takeover might still happen if one AI is better, but not that much better, than the others.

Having very little data about the world when the conquest begins

Cortés invaded Mexico knowing very little about it. After all, the Spanish had only realized the Americas existed two decades prior. He heard rumors of a big wealthy empire and he set out to conquer it, knowing little of the technology and tactics he would face. Two years later, he ruled the place.

Pizarro and Afonzo were in better epistemic positions, but still, they had to learn a lot of important details (like what the local power centers, norms, and conflicts were, and exactly what technology the locals had) on the fly. But they were good at learning these things and making it up as they went along, apparently.

We can expect superhuman AI to be good at learning. Even if it starts off knowing very little about the world -- say, it figured out it was in a training environment and hacked

its way out, having inferred a few general facts about its creators but not much else -- if it is good at learning and reasoning, it might still be pretty dangerous.

Being disunited

Cortés invaded Mexico in defiance of his superiors and had to defeat the army they sent to arrest him. Pizarro ended up fighting a civil war against his fellow conquistadors in the middle of his conquest of Peru. Afonzo fought Greek mercenaries and some traitor Portuguese, conquered Malacca against the orders of a rival conquistador in the area, and was ultimately demoted due to political maneuvers by rivals back home.

This astonishes me. Somehow these conquests were completed by people who were at the same time busy infighting and backstabbing each other!

Why was it that the conquistadors were able to split the locals into factions, ally with some to defeat the others, and end up on top? Why didn't it happen the other way around: some ambitious local ruler talks to the conquistadors, exploits their internal divisions, allies with some to defeat the others, and ends up on top?

I think the answer is partly the "diplomatic and strategic cunning" mentioned earlier, but mostly other things. (The conquistadors were disunited, but presumably were united in the ways that mattered.) At any rate, I expect AIs to be pretty good at coordinating too; they should be able to conquer the world just fine even while competing fiercely with each other. For more on this idea, see this comment.

By Daniel Kokotajlo

Acknowledgements

Thanks to Katja Grace for feedback on a draft. All mistakes are my own, and should be pointed out in the comments. Edit: Also, when I wrote this post I had forgotten that the basic idea for it probably came from this comment by JoshuaFox.

A Significant Portion of COVID-19 Transmission Is Presymptomatic

Epistemic status: Not quite settled science, but preprints seem to agree.

Strong evidence points to presymptomatic sources as a major source of COVID-19 infections, possibly the majority. The exact proportion is environment-dependent; awareness and public health measures reduce symptomatic transmission more than they reduce presymptomatic transmission.

The main reasons for thinking presymptomatic transmission is significant are direct measurements of the serial interval and incubation period, and the outside view of what level of public health measures have and haven't succeeded at containment.

Before delving into papers, a quick aside. If COVID-19 were only transmissible when people were coughing or feverish, containing it would be pretty easy; just tell people to stay home if they have those symptoms. Some people might try to go out anyways, so you might also set up checkpoints where people have their temperature taken and have someone listen to whether they're coughing, but that would pretty much be sufficient. Empirically, however, COVID-19 is successfully spreading in countries which have taken these measures and other more extreme measures, which is what we would expect given presymptomatic transmission, but not what we would expect without it.

(Note: You might think this means that symptomatic people aren't contagious, but actually it just means that people who show symptoms are doing a good job of isolating themselves. People with COVID-19 symptoms are definitely contagious and need to isolate themselves and notify people they might have spread it to.)

(Note: Presymptomatic transmission is a separate issue from asymptomatic carriers. Presymptomatic transmission is when someone is contagious when they aren't symptomatic *yet*. An asymptomatic carrier is someone who is contagious but who never develops symptoms. Asymptomatic carriers seem to be rare, though not completely nonexistent.)

Serial Interval and Incubation Period

The serial interval is the average length of time between transmissions in a transmission chain; that is, given pairs of people A and B where A was infected on a known date and then infected B on a known date, the serial interval is the average amount of time between those dates. The incubation period is the amount of time between when someone is infected, and when they display symptoms.

If the serial interval is shorter than the incubation period, this implies that a large fraction of transmission must be presymptomatic. So, with that in mind, I went looking for studies which measure COVID-19's incubation period and serial interval. These are in two tables below.

One of these studies, Tapiwa Ganyani et al, estimated the proportion of transmission which was pre-symptomatic: 48% (95% CI 32-67%) for Singapore and 62% (95%CI 50-

76%) for Tianjin. No other studies estimated this quantitatively, but most stated that their results provided qualitative evidence that presymptomatic transmission is occurring.

Estimates of the Incubation Period

Study	Incubation Period (days)	Sample Size	Data source
Stephen A. Lauer et al	5.1 (Median)	181	Travellers "in areas with no known community transmission"
Wei-jie Guan et al	4 (Median)	1099	China outside Hubei
Qun Li et al	5.2 (Mean)	425	Wuhan
Jantien A Backer et al	6.4 (Mean)	88	Travellers from Wuhan
Sijia Tian	6.7 (Median)	262	Beijing
Lauren C. Tindale et al	7.1, 9 (Mean)	228	Singapore and Tianjin
Kaike Ping et al	8 (Mean)	162	Ghuizho, China

Estimates of Serial Interval

Study	Serial interval (d)	Sample Size	Data source
Shi Zhao et al	4.4 Mean 3.0 SD	21 chains, 12 pairs	Hong Kong public data
Nishiura H et al.	4.0 or 4.6 Median	28 or 18 pairs	Published case reports
Chong You et al	4.41 Mean 3.17 SD	71 chains	China, outside Hubei
Qun Li et al	7.5 Mean	5 clusters	Hubei case clusters
Zhanwei Du et al	3.96 Mean 4.75 SD	468 pairs	China, outside Hubei
Tapiwa Ganyani et al	5.21 Mean, 3.95 SD	226 cases	Singapore and Tianjin clusters
Lauren C. Tindale et al	4.56 Mean, 4.22 SD	228 cases	Singapore and Tianjin clusters
Shi Zhao et al	5.2 Mean	48 pairs	Hong Kong and Shenzhen
Kaike Ping et al	6.37 Mean	57 cases	Guizhou, China

Tapiwa Ganyani et al and Lauren C Tindale et al appear to have used overlapping public data sources. The sample size column for serial interval studies is unusually painful, as sample-size columns go, because many of the studies needed to account for uncertainty in who infected who; as such, sample sizes are reported varyingly in

units of (in order from most to least reliable per sample) pairs, chains, clusters, and cases.

The study with the longest estimated serial interval, Qun Li et al, looks at a small number of clusters and guesses which cases infected which other cases. While it estimates a mean serial interval of 7.5, Its data is also compatible with an interpretation in which the mean serial interval is shorter and some of the transmissions are indirect. This change in interpretation would bring it in line with other studies in this set, which estimate shorter intervals.

One of these studies, Zhanwe Du et al, estimated the serial interval using when people became symptomatic (rather than when they were exposed), and found that in 13% of cases, the infectee showed symptoms before the infector did. This would imply that either in those cases the infector transmitted presymptomatically, the infector had a relatively long incubation period, and the infectee had a relatively short incubation period; or that this data set had major issues identifying who affected who. The distribution of SIs fits a nice Gaussian, which is some evidence that it's the former.

Anecdotal Reports and Case Studies

To understand what presymptomatic transmission of COVID-19 would look like, I went looking for anecdotes and case studies of known COVID-19 transmission events. You can't use these to infer much about rates, but they're helpful for internalizing what presymptomatic transmission would look like.

> "I believe I caught it when attending a small house party at which no one was coughing, sneezing or otherwise displaying any symptoms of illness. It appears that 40% of the attendees of this party ended up sick."

https://www.facebook.com/EbethBerkeley/posts/10110434821081713

> (via Google Translate) "On January 24, Li and his grandfather, grandma, and father went to aunt's house for dinner, a total of 9 people. On January 28, Li developed fever. … all 9 people participating in the dinner were confirmed as confirmed cases."

http://hlj.people.com.cn/GB/n2/2020/0205/c220024-33767665.html

Practical Implications

The main practical implication is that contact tracing is really important.

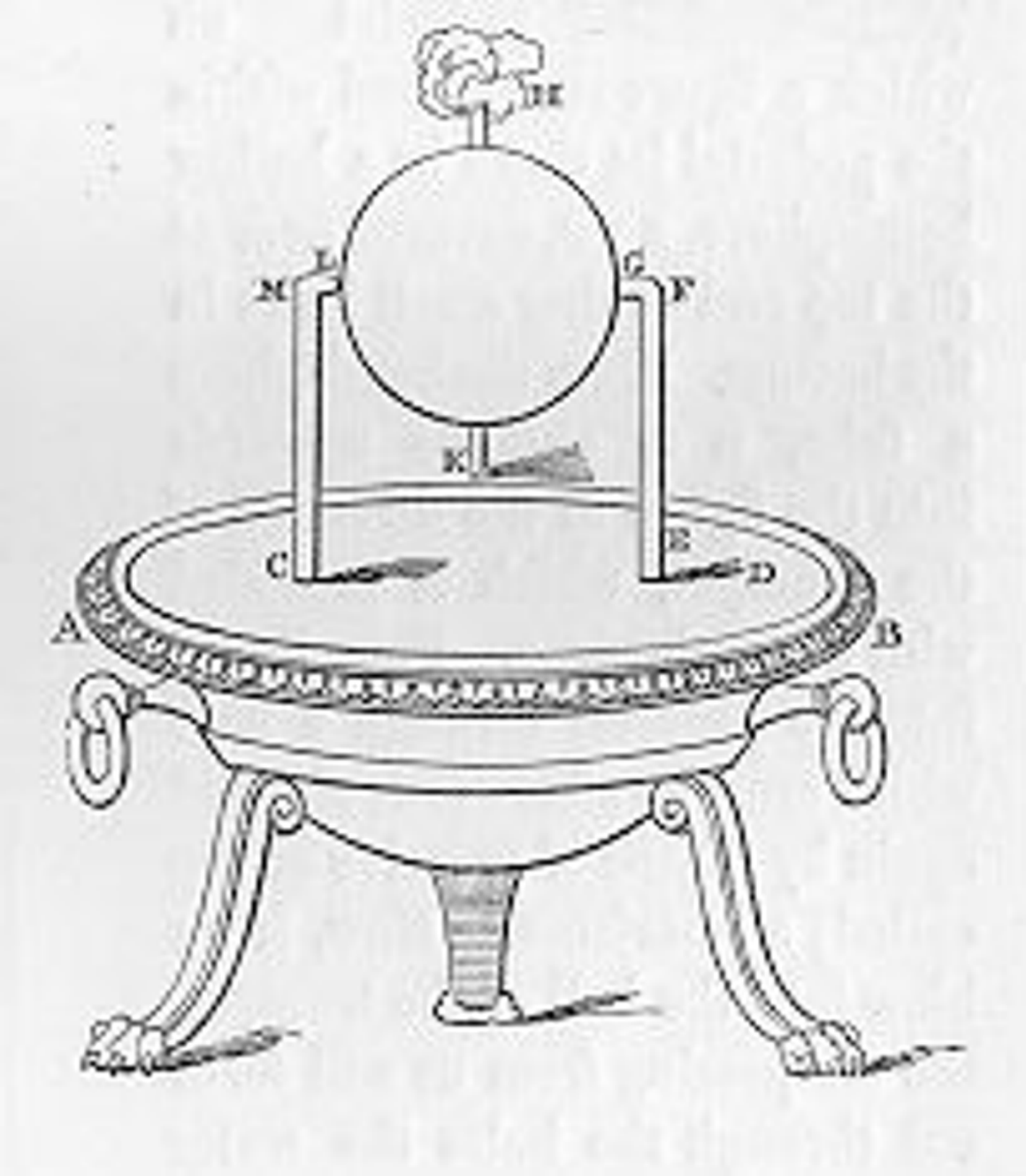
Contact tracing is where, when you find someone with COVID-19, you identify everyone they might have spread it to and warn them that they've been exposed. People who've been exposed are expected to quarantine themselves for 14 days, which is long enough that if they are in fact infected, there's only a 1% chance they are infected but not yet symptomatic . Back in January, this served two purposes: it ensured that if they had a cough, they wouldn't brush it off as something minor and keep going to work, and it ensured that if they *didn't* have a cough, they wouldn't transmit it while presymptomatic. The first issue is now less of a concern; everyone knows that if someone has a cough, they aren't supposed to go to work, even if it's definitely rhinovirus. The second issue is exactly as much a concern as it was before.

Epistemic standards for “Why did it take so long to invent X?”

This is a linkpost for https://rootsofprogress.org/epistemic-standards-for-why-it-took-so-long

In seeking to understand the history of progress, I keep running across intriguing cases of “ideas behind their time”—inventions that seem to have come along much later than they could have, such as the cotton gin or the bicycle. I’ve started collecting a list here, and will update that page with new analyses as I find them.

Debates on these questions sometimes oddly devolve into arguments, with people fruitlessly talking past each other (although if people on the Internet can argue over how many days are in a week, I guess we can argue about anything). So I want to comment on how we *think* about such cases and what the standards for evidence are.



Aeolipile, from Hero's Pneumatica
Wikimedia

To start, there is a need for precision. For example, take the steam engine. There was a thing in antiquity called an engine, that used steam: “Hero’s engine”, also known as the aeolipile. Some people see this and conclude that “steam engines existed in the 1st century” or that there’s no reason ancient Rome couldn’t have used this widely in industrial applications.

This is a mistake. The aeolipile is nothing like the steam engines of the 18th century and later: it’s a turbine, which means it is rotary, rather than using the reciprocating (back-and-forth) motion of a piston, as in Newcomen’s engine. Why does this matter? Because the aeolipile doesn’t generate enough torque for practical applications—one analysis says that Watt’s engine generated a *quarter of a million* times more torque.

Newcomen's steam engine

Even before finding that analysis, I had an hunch it would be the case—indeed, that's how I knew to search for "aeolipile torque", which led me to that link on the first page of Google results. My intuition was based on a few things. First, if a simple, primitive turbine like the aeolipile could be put to practical use, why didn't anyone reinvent it in the 18th or even 17th century? Why did Newcomen, Watt, and others focus on much more complicated piston engines? They were smart people and were obviously working hard on the problem, it seems impossible that such a simple solution would have escaped all of them. Second, the aeolipile is small—Hero's sketch above shows it sitting on a table—but Newcomen's engine was very large, to the point where a separate shed would be built to house one. Why did the engine have to be that large if a tiny one would do?

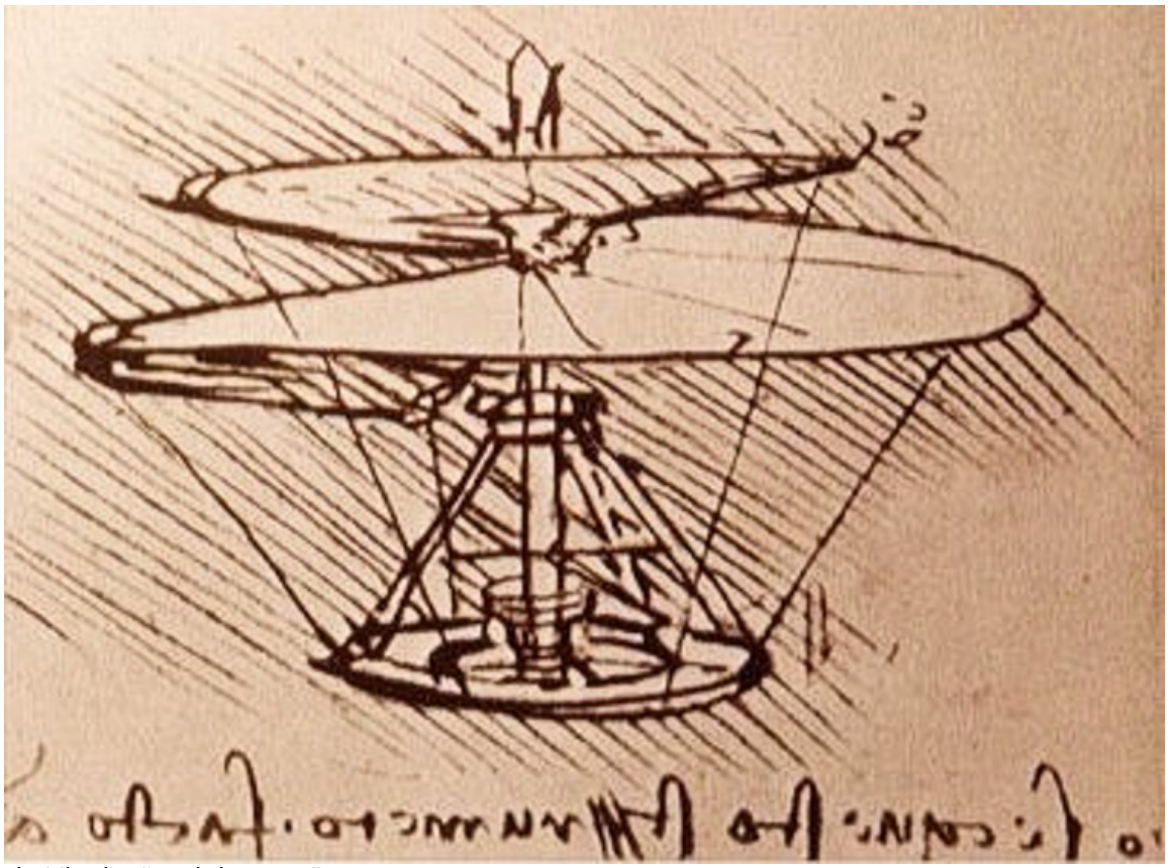
Again, a precise understanding of each invention will uncover relevant details like this. A *concept*, such as "an engine (of any type) that uses steam (in any way)", is not enough.

This example also illustrates a second principle: *practicality matters*. A device that works in theory, but is too underpowered, inefficient, expensive, or unreliable, might as well not exist for practical purposes. It must work not only for a demonstration, but for real, human,

economic needs, in the context of consumers' lives, industrial processes, or business operations. Because of this, a difference in *degree* can become a difference in *kind*, when an invention crosses a threshold of practicality.

Edison's light bulb
Filip Mishevski / Flickr

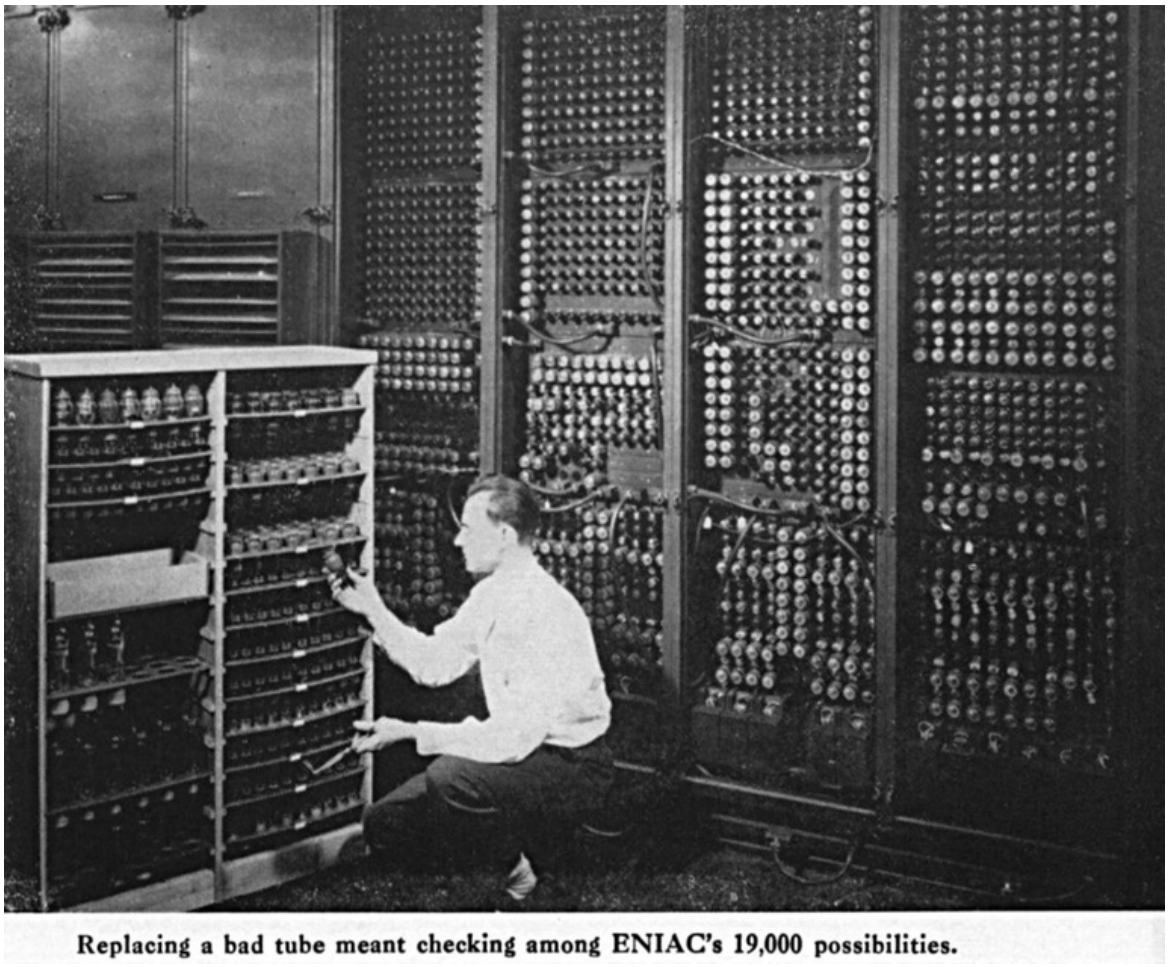
As a side note, this is why it's perfectly accurate to say that Edison's lab invented the light bulb, even though there were other light bulbs before it: they were too expensive (e.g., using platinum filaments), or they burned out quickly and thus needed to be replaced too often. In my opinion, it's redundant to say that someone invented "the first practical X"—this is the same as saying they invented X. To invent something *is* to invent a practical version of that thing. If your "invention" is impractical, it's just a demo or prototype. This can be useful to test ideas or to communicate possibilities, but it's the *practical* inventions—the ones that actually remove *all* the obstacles to widespread use—that move history.



da Vinci's "aerial screw"
Wikimedia

Another example of this is the computer. The computer was invented by J. Presper Eckert and John Mauchly at the University of Pennsylvania; their first model was the ENIAC, completed in 1945. It was a breakthrough because it was the first *fully electronic* computer, and this made it much *faster* than previous attempts, such as the IBM Automatic Sequence Controlled Calculator (ASSC, aka the Harvard Mark I), which was electromechanical, using magnetic relays. Based on the speeds for these machines given by Wikipedia, the ENIAC was about 600 times faster than the ASSC at division and over 2,000 times faster at multiplication. (The ENIAC was also over 2,000 times faster than a *human* using a mechanical calculator at calculating a ballistic trajectory, implying that the ASSC was probably not much faster than a human.) The ASSC was an interesting demo that got some press; the ENIAC was the machine that ignited the computer revolution. Again, a difference in degree becomes a difference in kind. (Going even further back, other predecessors such as the Atanasoff–Berry computer or Konrad Zuse's Z3 were also much slower than the ENIAC, and had other practical limitations. And Babbage's "computer" was only a concept with an

unfinished design that could never have been built with the technology of the day—which is why, despite my respect for his genius, I cannot regard Babbage as the inventor of the computer, any more than da Vinci was the inventor of the helicopter.)



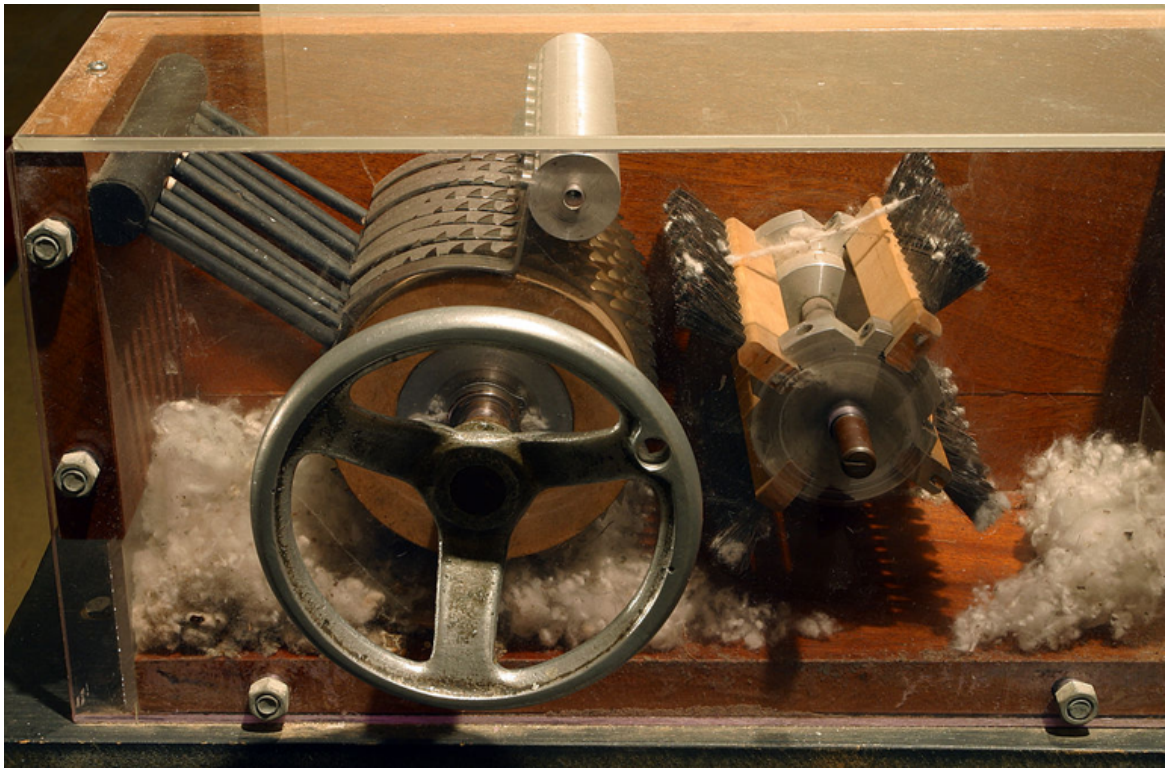
Changing a tube on the ENIAC
Wikimedia

If you want to argue that something could not have been invented before a particular “gate”, I regard this kind of history as the epistemic gold standard: strong economic motivation (and in the case of computers, military motivation in the context of WW2); multiple prior attempts, including some completed projects that were working and reasonably well-publicized; a measurable difference in a key practical dimension (in this case, speed); and an enabling technology that made a significant difference along that dimension (in this case, approximately three orders of magnitude). For this reason I confidently say that the computer—again, it’s redundant to say “practical computer”—could not have existed before the invention of the vacuum tube amplifier in 1907. (It’s plausible, but less certain to me, that it could not have existed until even later, when improved, more reliable vacuum tubes were invented. Plausible, because the ENIAC used over 17,000 tubes, and reliability was a concern among the engineers; less certain, because I don’t know of any failed attempt at building a fully electronic computer with less-reliable tubes, and because some statements from engineers indicate that reliability was less of a problem than feared, particularly if the tubes were operated continuously, to avoid thermal stress. So it might be that, after 1907, all that was holding back the engineering was a key insight.)

Karl Drais's Laufmaschine
Wikimedia

On the other hand, if you want to argue that something *could* have been invented much earlier than it was, you have to do better than glancing at its high-level concepts or components. You need to rigorously examine every part, material, and manufacturing process, and rule them *all* out as gating technologies. Any detail, even a minor one, can become crucial—especially when we remember that inventions need not only to work but to be practical, which includes performance, reliability and cost. As an example, in my analysis of the bicycle, I described the first proto-bicycle, known as the "draisine" or "Laufmaschine", as being made of wood with iron tires—both ancient technologies. However, Nick Szabo pointed out to me that it reportedly used brass ball bearings, a much newer technology, which might have been essential to reduce friction.

A related question: how *surprised* should we be that it took X years for invention Y after enabling technology Z? Inventions do not spring forth *immediately* upon becoming possible: ideas and information take time to spread, experiments are required, funding must be secured, laboratories organized, materials obtained; and at end of the day all this is performed not by automata or some clockwork mechanism, but by unpredictable individuals with their own vision, inspiration, hopes and fears, operating in complex network of teams, contracts, partnerships, and other social structures. Even in the best of circumstances, a gap of a decade or more from a key enabling technology to the commercial release of an invention is not surprising; if the enabler is a scientific discovery, two or three decades does not surprise me. And chance can intervene—the path to an invention can be derailed by a sudden disease, a financial panic, a war.



Eli Whitney's cotton gin
Wikimedia / Tom Murphy VII

In general, I think we should be *more* surprised at long gaps for inventions that have obvious, predictable impact on major industries. For this reason, the cotton gin and the flying shuttle are more compelling gaps to me than the wheeled suitcase, role-playing games, or the bicycle, which merely offer convenience or entertainment. I think we should also expect longer gaps in places and times that had lower population, less education, less economic surplus (to fund R&D), fewer or less effective financing mechanisms (such as venture capital), less political stability, etc.

My model for this is that innovation, at a societal level, is a stochastic process, with some parameters set by the environment and others by the particular invention in question. The more “pressure” there is to solve a problem (economic motivation), and the more opportunities there are for it to get solved (educated, inventive individuals or organizations, with time, space, materials, and funding, in a context of good legal institutions and political stability), the sooner you expect the leap to be made and the shorter a gap. In the limit, you get simultaneous invention, of which there are many stories (although some are overplayed, in part for the reasons discussed above regarding what counts as an “invention”). Some economics grad student could probably get a PhD thesis out of formalizing this model and fitting the parameters to data—both to quantify the “inventiveness” of a given place and time, and to identify outlier inventions that were truly, measurably, “behind their time.”

The Lens, Progerias and Polycausality

Fun fact: the lens of a human eye consists mostly of fiber deposits which are never broken down - they do not turn over. Furthermore, new fiber layers are constantly added throughout life, so the lens thickens linearly by about 25 microns per year. Starting at around 3.5mm in infancy, it reaches 5.5mm in old age.

The main clinical result of this is the practically-universal need for glasses for close-up vision in people over 55 years old.

(Source: Physiological Basis of Aging and Geriatrics; the section on the eye is one of the most detailed in the book.)

Besides being a simple, self-contained gear in its own right, the growth of the lens is a clear, knock-down example of an independent root cause of one symptom of aging. We know exactly what's accumulating in a nonequilibrium fashion: the fibers of the lens. It's wildly unlikely that the growth of the lens is a root cause for other symptoms of aging - like wrinkles, atherosclerosis, Alzheimer's, cancer, muscle degeneration, etc. So, we have a clear case for polycausality - at least for one symptom of aging.

That said, there's a fair bit of evidence that *most* symptoms of aging share a common root cause, or at least a common intermediate. Qualitatively, many/most symptoms of aging in a wide variety of tissues:

- Look similar at the cellular level - there's a loss of homeostasis, with cells dying off faster than they're replaced, high levels of misfolded protein aggregates (a.k.a. junk), and markers of chronic inflammation
- Follow a similar population-level onset/progression timetable: no noticeable problems from youth through mid-twenties, gradual onset/progression throughout middle age, then rapidly accelerating breakdown around 50-60 years of age and older. Some examples: cancer incidence, muscle loss, atherosclerosis. Google a performance metric which declines with age, and you'll probably see the pattern.
- Are correlated - someone who has one problem early is likely to have others early, and vice versa. See the literature on physiological/biological aging clocks for details.

The growth of the lens does *not* follow this pattern - it's just a straight-line linear growth starting from childhood, without any unusual role of chronic inflammation or misfolded proteins or other typical aging-associated characteristics. On the other hand, there are other contributing factors to old-age vision problems which *do* follow the usual pattern - for instance, the loss of pupil muscle mass.

Besides the growth of the lens, there are a handful of other possible root/intermediate causes of aging symptoms which don't follow the usual pattern. None of them are as conclusive an example as the lens, but they may be involved in nastier diseases. In particular: the thymus is an organ which trains adaptive immune cells to distinguish pathogens from healthy host cells. That organ begins to shrink (called "thymic involution") even in the first year of life, and steadily loses most of its mass by old age. I'll likely have a full post on that later.

Progerias

One interesting source of evidence about common root causes of aging symptoms is accelerated aging diseases, a.k.a. progerias. I'll talk about two: Werner Syndrome (WS) and Hutchinson-Gilford Progeria Syndrome (HGPS).

Werner syndrome is the progeria which most closely resembles true aging. People with WS develop normally through puberty, but then develop a laundry list of aging symptoms early:

- Gray hair
- Hair loss
- Wrinkles
- Skin hardening/tightening
- Loss of fat tissue
- Atrophy of gonads
- Cataracts
- Atherosclerosis
- Type 2 diabetes
- Muscle degeneration
- Bone loss
- Cancer

(you can find all this on the wikipedia page). Perhaps even more notable: changes in gene transcription associated with WS closely resemble the transcription changes associated with aging.

What causes this remarkably aging-like disease? Mutation of a gene called WRN (short for Werner), which is involved in repair of several types of DNA damage. The damage does still get repaired (otherwise people with WS wouldn't be alive at all), but it's slower, so presumably there's a higher steady-state level of DNA damage. This is consistent with other lines of evidence which I may talk about in future posts: high levels of DNA damage are associated with aging.

The other type of progeria we'll discuss is HGPS. HGPS also shows many aging-like symptoms:

- Hair loss
- Wrinkles
- Skin hardening/tightening
- Atherosclerosis
- Muscle degeneration
- Bone loss

But even more notable is the symptoms of aging which are *not* associated with HGPS, specifically:

- Cancer
- Arthritis

(Note: I didn't comprehensively check every symptom of WS against HGPS, so don't read too much into the differences between the two lists above.)

What would cause so many aging-like symptoms, but not cancer? HGPS is caused by mutation of a nuclear envelope protein; without it, the cell nucleus has a weird shape (striking picture here). The main result is that cells have trouble dividing - the folded-up nuclear envelope gets in the way of chromosome arrangement when the nucleus is

supposed to divide. The mutation limits cell division, which we'd expect to lower homeostatic counts of a broad variety of cell types.

Assuming that's the main mechanism, we'd expect HGPS to show the symptoms of aging associated with cell loss - e.g. hair loss, muscle degeneration - but not the symptoms associated with biological stressors like DNA damage - e.g. cancer and inflammatory diseases like arthritis. For some symptoms which aren't yet fully understood - e.g. wrinkles or atherosclerosis - HGPS is a hint that cell loss is probably a key mediating factor.

Ubiquitous Far-Ultraviolet Light Could Control the Spread of Covid-19 and Other Pandemics

Roko Mijic, Alexey Turchin

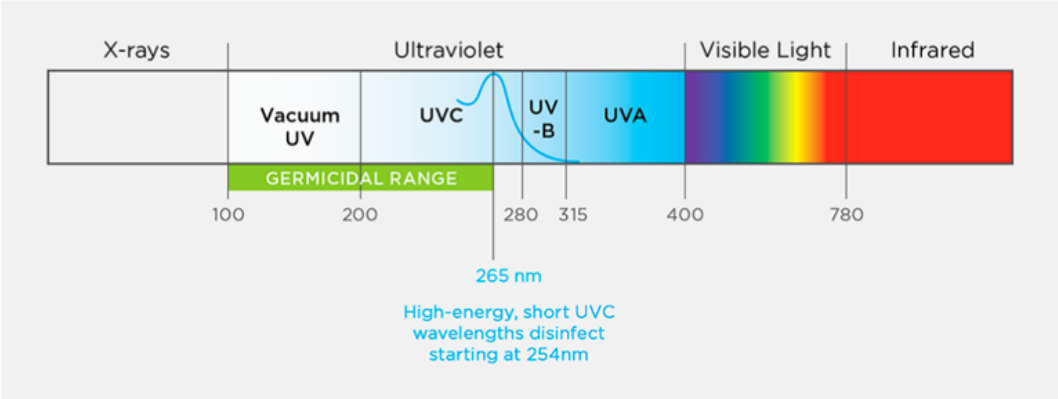
Epistemic status: Many different uncertainties here, but the idea has some good evidence in favor of it and a high potential payoff.

Tl;dr: We should urgently investigate putting special human-safe Far-UVC lamps all over our built environment to 'kill' virus particles whilst they are in the air, thereby vastly reducing covid-19 spread.

Inspired by: https://www.nature.com/articles/s41598-018-21058-w

One of the most promising and neglected ideas for combating the spread of covid-19 is the use of ubiquitous ultraviolet light in our built environment (trains, offices, hospitals, etc). Ultraviolet light is already being used as a disinfecting agent across the world; it goes by the acronym UVGI - "Ultraviolet germicidal irradiation". The energetic photons of UVC light break chemical bonds in DNA and kill/inactivate both viruses and bacteria.

Ultraviolet light on earth exists on a spectrum between 200nm and 400nm. Light above 400nm is **blue** visible light. Light below 200nm is called "vacuum UV" because it is strongly absorbed by the oxygen in ordinary air and therefore cannot exist except in a vacuum or some other non-air medium. Within the 200-400nm range we have UVA, UVB and UVC, and at the short-wave edge of the UVC band we have "Far-UVC", from roughly 200nm–220 nm.



Safety considerations

Human beings are also vulnerable to UV radiation. It causes skin cancer and serious eye damage.

However, recent research suggests that the Far-UVC band is actually safe for human skin because it cannot penetrate through the thin layer of dead skin cells on the surface of our skin.

This means that it might be possible to mount a long-term response to covid and other pathogens by constantly illuminating our built environment with light from specifically the Far-UVC band. If the Far-UVC light is indeed safe for humans, the Far-UVC could be on at all times and could destroy or deactivate viral particles before they can spread from person to person.

Why hasn't this already been considered by relevant authorities? Far-UVC appears in a literature review by WHO, but it is not currently being acted upon as the amount of evidence in favor of safety and efficacy is small.

There is some uncertainty about whether Ozone generation by this band (200nm-220nm) would be problematic. Ozone is not great for your health. However, it seems to be the case that the 200-220nm band is not a strong producer of Ozone. In addition, UV degradation of surfaces might result from chronic UV exposure.

Balancing harms of action and inaction

Even if Far-UVC is somewhat harmful it might still be a good idea to implement. Small harms from Far-UVC light might be much less bad than large harms from covid-19, or from the economic damage caused by the lockdown which one author estimates to be roughly $10 million per minute, plus much personal hardship which will be caused by the forthcoming recession.

Furthermore, UV light is easier to defend a person against than a virus. Sun-creams, clothing and eyewear that defend against UVC may be less bad than a semi-permanent lockdown or an exponentially growing covid-19 outbreak that results in millions or tens of millions of deaths. UV in the built environment could even be managed intelligently - computer vision could identify where the people were and turn on UV lights only in unoccupied areas, though such a project would at best be ready by the start of 2021 (and then only with wartime levels of effort and purpose).

If the safety claims of Far-UVC are partially true rather than fully true, a combination of using Far-UVC with physical protection like eyewear may still cause only acceptable losses to cancer and eye damage. In the longer term, such "almost safe" Far-UVC could be combined with intelligent management at various levels of granularity; imagine a lift that is bathed in Far-UVC every time people leave it, or "walls" of Far-UVC separating people that automatically turn off momentarily when a person walks through them. The ultimate system might even adjust the power of the Far-UVC using AI.

Epidemiological considerations

Even an ideal Far-UVC solution that was harmless to humans, 100% lethal to covid-19 particles and easy to deploy at scale might not be sufficient to reduce R to exactly 0. But the key question is whether it could reduce R below 1 whilst also allowing most economic activity. An easy preliminary experiment to run would be to put virus samples in mouse cages - perhaps in aerosolized form - treat some cages with Far-UVC, leave other cages alone, see if infection rates go down in the treated cages.

This is an important source of uncertainty and further research is needed.

Scaleup considerations

Even a perfect system is useless if it cannot be scaled up and implemented across the globe. Far-UVC can be produced from Krypton Chloride (Kr-Cl) Excimer Lamps, but modern Aluminium Nitride (AlN) Far-UVC LEDs are a better solution for the long term. In the even longer term, collimated Far-UVC could be produced by lasers. This is an important source of uncertainty and further research and expert input is needed.

Power considerations:

The amount of Far-UVC energy required to kill 99% of the viral particles is estimated to be around 20J/m^2. With a power of, say 5W/m^2, a system would need 4 seconds to mostly sterilize a viral aerosol that could travel from person to person. However a lower power system would still have some benefits - we know that people can be infected by air that was contaminated 30 minutes earlier. Higher power in these wavelengths could be difficult to achieve with Kr-Cl Excimer Lamps as the overall efficiency from electricity to Far-UVC is ~10%. AlN Far-UVC LEDs would likely have a much higher conversion efficiency.

Generality

One of the greatest benefits of Far-UVC is that it would be a very general weapon against pathogens. Far-UVC kills/deactivates bacteria, viruses and other pathogens. MRSA, C-DIFF, influenza, etc are all killed by UVC, as is the next problematic pathogen, whatever it is.

Summary

There are many different reasons that Ubiquitous Far-UVC might not work out, but if it did work out it could have huge benefits. For this reason the authors believe that it should get more attention at this critical time. Scaleup and safety and efficacy trials must all be carried out as quickly as possible, preferably in parallel. More importantly, the idea needs more attention from experts in the relevant fields - UV physics, epidemiology, and people who study the etiology of skin cancers. As of writing there are reports that the US government estimates the epidemic could last for 18 months, so a plan like Far-UVC that will take months to implement may still be a critical component of a response later this year.

Appendix. Other ways to use UV light to fight coronavirus

One of the explanations of the flu and other infections seasonality is that the Sun's UV kills viruses. However, people spend a lot of time indoors even during summer, and especially during self-isolation. Most of our infections are happening indoors: at home, in transport and in working places. UV from Sun could be part of the explanation of the lower instances of coronavirus in southern countries.

If we replace light bulbs everywhere with light sources which also emit UV light of some special wavelength, we will kill most of the airborne viruses and will clean fomites. Thus, we will create artificial summer everywhere and will lower R0 of coronavirus below 1.

The main obstacles are the duration of exposition and possible harm to people. Recently in Moscow 20 children had burns in their eyes after a school teacher forgot to turn off the UV cleaner in the classroom.

There are several other ideas, besides Far-UVC light, to prevent human eye and skin damage:

1) Intelligently *controlled UV lighting*. UV light source turns on the maximum level when there are no people in the room. We already have motion detectors for lighting, but here they will work in reverse. Light with motion detection could also direct light in directions, where there is no motion, so no people. On the video, one can see UV light sources on sale with motion detectors:

The power of light could be temporarily increased after the sound of sneezing. But it will make all the system more complex and its large-scale implementation will take longer time. If Krypton Chloride Excimer bulbs are used, their lifetime is not great, so they can't run constantly. But if we can get the Aluminium Nitride LEDS then lifetime and efficiency will be better.

2) *Not "too strong" sources of UV*, which are producing Sun's intensity of UV and which act mostly on fomites. As we know, humans can survive at least 1 hour of sunlight UV exposure without strong damage (on beaches). We could use it as a reference point to calibrating UV sources.

3) *Strong UV lighting + gloves.* Everyone will wear gloves, masks and glasses outside. In that case, no skin will be exposed to the UV lighting (and to viruses). Wearing PPE will be effective anyway. Women in the East are wearing full cover clothes, and they are ok.

4) *Wearable headlight UV* will direct UV light in the opposite direction to the person's eyes but will cover everything he inhales or touches, as well as his hands. The light will be strongest near the human face (but not affecting the face), and will attack droplets which the person is about to inhale. However, the light will dissipate in the distance of 1- 2 meters to safer levels. UV headlamps already exist and on sale, but may be not strong enough for disinfection. It will be especially effective if wearables Far UVC light sources will be used.

5) *UV flashlight* - Torch that emits UV radiation in a wide beam. Runs off main power. Could be used by cleaners as an additional step when cleaning surfaces.

Pros:

1. Simpler, easier, cheaper and faster to build than other solutions
2. Less harm to people, as UV light can be directed, and is not always on
3. Proving ground (an MVP, in startup terms) for more advanced implementations
4. Mobile; could be used in multiple locations

Cons:

1. Less effective than always on UV lights and lamps

2. Requires additional time/effort on top of normal cleaning routines

Artificial light exists currently almost everywhere, where contemporary humans live: in homes, in any shop, in cars and even on the streets. All we need is to replace electric lamps. Large amounts of lamps could be manufactured in 0.5-1 year, and smaller amounts for critical places like elevators in the even shorter notice.

However, there is a problem of actual testing the technology until it will be approved as safe and effective by the FDA. It is technically difficult to make deep UV (220nm) light-emitting diodes.

A good start will be to put UV lights in the places of short use: elevators, shops, restrooms.

It is much more convenient to wear protection against light than protection against viruses, and after a few months of lockdown, the idea of returning to almost normal life but with sun cream and/or gloves will be quite nice.

References

Welch, D., Buonanno, M., Grilj, V. *et al.* Far-UVC light: A new tool to control the spread of airborne-mediated microbial diseases. *Sci Rep* 8, 2752 (2018). https://doi.org/10.1038/s41598-018-21058-w

Narita K, Asano K, Morimoto Y, Igarashi T, Nakane A (2018) Chronic irradiation with 222- nm UVC light induces neither DNA damage nor epidermal lesions in mouse skin, even at high doses. PLoS ONE 13(7): e0201259. https://doi.org/10.1371/journal.pone.0201259

Willie Taylor, Emily Camilleri, D. Levi Craft, George Korza, Maria Rocha Granados, Jaliyah Peterson, Renata Szczpaniak, Sandra K. Weller, Ralf Moeller, Thierry Douki, Wendy W.K. Mok, Peter Setlow DNA damage Kills Bacterial Spores and Cells Exposed to 222 nm UV Radiation Applied and Environmental Microbiology Feb 2020, AEM.03039-19; **DOI:** 10.1128/AEM.03039-19

Colorado company uses UV lighting technology to kill 99.9 percent of bacteria and viruses. Fox Denver, 7 Macrh 2020

My current framework for thinking about AGI timelines

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

At the beginning of 2017, someone I deeply trusted said they thought AGI would come in 10 years, with 50% probability.

I didn't take their opinion at face value, especially since so many experts seemed confident that AGI was decades away. But the possibility of imminent apocalypse seemed plausible enough and important enough that I decided to prioritize investigating AGI timelines over trying to strike gold. I left the VC-backed startup I'd cofounded, and went around talking to every smart and sensible person I could find who seemed to have opinions about when humanity would develop AGI.

My biggest takeaways after 3 years might be disappointing -- I don't think the considerations currently available to us point to any decisive conclusion one way or another, and I don't think anybody really knows when AGI is coming. At the very least, the fields of knowledge that I think bear on AGI forecasting (including deep learning, predictive coding, and comparative neuroanatomy) are disparate, and I don't know of any careful and measured thinkers with all the relevant expertise.

That being said, I did manage to identify a handful of background variables that consistently play significant roles in informing people's intuitive estimates of when we'll get to AGI. In other words, people would often tell me that their estimates of AGI timelines would significantly change if their views on one of these background variables changed.

I've put together a framework for understanding AGI timelines based on these background variables. Among all the frameworks for AGI timelines I've encountered, it's the framework that most comprehensively enumerates crucial considerations for AGI timelines, and it's the framework that best explains how smart and sensible people might arrive at vastly different views on AGI timelines.

Over the course of the next few weeks, I'll publish a series of posts about these background variables and some considerations that shed light on what their values are. I'll conclude by describing my framework for how they come together to explain various overall viewpoints on AGI timelines, depending on different prior assumptions on the values of these variables.

By trade, I'm a math competition junkie, an entrepreneur, and a hippie. I am not an expert on any of the topics I'll be writing about -- my analyses will not be comprehensive, and they might contain mistakes. I'm sharing them with you anyway in the hopes that you might contribute your own expertise, correct for my epistemic shortcomings, and perhaps find them interesting.

I'd like to thank Paul Christiano, Jessica Taylor, Carl Shulman, Anna Salamon, Katja Grace, Tegan McCaslin, Eric Drexler, Vlad Firiou, Janos Kramar, Victoria Krakovna, Jan Leike, Richard Ngo, Rohin Shah, Jacob Steinhardt, David Dalrymple, Catherine Olsson, Jelena Luketina, Alex Ray, Jack Gallagher, Ben Hoffman, Tsvi BT, Sam Eisenstat, Matthew Graves, Ryan Carey, Gary Basin, Eliana Lorch, Anand Srinivasan, Michael

Webb, Ashwin Sah, Yi Sun, Mark Sellke, Alex Gunning, Paul Kreiner, David Girardo, Danit Gal, Oliver Habryka, Sarah Constantin, Alex Flint, Stag Lynn, Andis Draguns, Tristan Hume, Holden Lee, David Dohan, and Daniel Kang for enlightening conversations about AGI timelines, and I'd like to apologize to anyone whose name I ought to have included, but forgot to include.

Table of contents

As I post over the coming weeks, I'll update this table of contents with links to the posts, and I might update some of the titles and descriptions.

Humans can design some things much better than evolution (like rockets), and evolution can design some things much better than humans (like immune systems). Where does general intelligence lie on this spectrum?

Tying it all together, part I

My framework for what these variables tell us about AGI timelines

Tying it all together, part II

My personal views on AGI timelines

---

1. https://en.wikipedia.org/wiki/Neocortex ↩
2. https://en.wikipedia.org/wiki/Cognitive_science ↩

Effectiveness of Fever-Screening Will Decline

COVID-19 Is Under Strong Selective Pressure

Fever screening is the practice of checking the temperature of everyone passing through a checkpoint, such as at an airport or public gathering, for fever. This can be done individually with regular thermometers, or on entire crowds with infrared cameras. Fever screening has been widely deployed against COVID-19, especially in China, since fever is its most common and often chronologically first symptom.

Fever screening applies an evolutionary pressure on pathogens to not cause fever, or to cause fever later relative to when it becomes transmissible. Pathogens mutate. In many diseases, this results in drug resistance; antibiotics, for example, lose effectiveness over time.

Drug resistance evolves slowly, because partial resistance is only a slight benefit to a pathogen which acquires it; most patients who receive antibiotics do not infect anyone after their treatment has started. Evasion of public-health screening procedures, on the other hand, is likely to evolve much faster, because any incremental improvement in a pathogen's ability to evade screening will give it a large advantage.

Evidence This Is Already Happening

I searched for papers recording the percentage of patients which had a fever in Google Scholar and in the citations on this page. For each paper, I extracted the percentage of patients reported to have a fever, and the date range of cases covered by the data. If the paper did not state the date range of cases, I wrote down the paper's publication date instead. I excluded papers which didn't report a percentage of patients with fever, or had n<20.

I found seven studies, listed below sorted by end date.

Dec 16-Jan 2: 98%
Jan 1-20: 83%
Jan 1-28: 98.6%
Jan 6-31: 58.3%
Pub. Feb 18: 78.2%
Jan 21-Feb 15: 85.7%
Before Feb 20: 87.9%

This is suggestive of a negative correlation between study date and percent of patients with fever. This is what you would expect if COVID-19 were evolving to evade fever screening; however, it could also be explained by later studies finding earlier and less-severe cases. This is weak evidence compared to the what could obtained by someone with access to a good dataset; these papers come from different sample populations and different points in the disease progression, and don't all specify what their cutoff temperature was for diagnosing fever. If someone has access to a suitable

dataset, I ask that they do the analysis and publicly state whether they see a declining rate of fever.

Other Properties Will Likely Also Change

It is well known that most diseases evolve to become less severe over time, because patients with severe cases are easier to detect and will take greater precautions. However, this takes place on very slow time scales. Evading fever screening, on the other hand, involves greater selective pressure and so may happen on a faster time scale, possibly fast enough to significantly influence the shape of the pandemic this year.

I don't know how fever-screening evasion relates to disease severity; I can see plausible mechanisms by which this would make it either increase or decrease.

"No evidence" as a Valley of Bad Rationality

Quick summary of Doctor, There are Two Kinds of “No Evidence”:

- Author has a relative with cancer. Relative is doing well after chemo and is going to a doctor to see if it's worth getting more chemo to kill the little extra bits of cancer that might be lingering.
- Doctor says that there is *no evidence* that getting more chemo does *any* good in these situations.
- Author says that this violates common sense.
- Doctor says that common sense doesn't matter, evidence does.
- Author asks whether "no evidence" means 1) a lot of studies showing that it doesn't do any good, or 2) not enough studies to conclusively say that it does good.
- Doctor didn't understand the difference.

Let me be clear about the mistake the doctor is making: he's focused on *conclusive* evidence. To him, if the evidence isn't conclusive, it doesn't count.

I think this doctor is stuck in a Valley of Bad Rationality. Here's what I mean:

- The average Joe doesn't know anything about t-tests and p-values, but the average Joe does know to update his beliefs incrementally. Lamar Jackson just had another 4 touchdown game? It's not conclusive, but it starts to point more in the direction of him winning the MVP.
- The average Joe doesn't know anything about formal statistical methods. He updates his beliefs in a hand-wavy, wishy-washy way.
- The doctor went to school to learn about these formal statistical methods. He learned that theorizing is error prone and that we need to base our beliefs on hard data. And he learned that if our p-value isn't less than 0.05, we can't reject the null hypothesis.
- You can argue that so far, the doctors education didn't move him forward. That it instead caused him to take a step *backwards*. Think about it: he's telling a patient with cancer to not even consider more chemo because there is "no evidence" that it will do "any" good. I think Average Joe could do better than that.
- But if the doctor continued his education and learned more about statistics, he'd learn that his intro class didn't paint a complete picture. He'd learn that you don't always have access to "conclusive" evidence, and that in these situations, sometimes you just have to work with what you have. He'd also learn that he was privileging the null hypothesis in a situation where it'd make sense to do the opposite. The null hypothesis of "more chemo has no effect" probably isn't true.
- Once the doctor receives this further education, it'd push him two steps forward.
- In the intro class, he took one step backwards. At that point he's in the Valley of Bad Rationality: education made him worse than where he started. But then when he received more education, he took two steps forward. It brought him out of this valley and further along than where he started.

I think that a lot of people are stuck in this same valley.

Adding Up To Normality

Related: Leave a Line of Retreat, Living In Many Worlds

> "It all adds up to normality." Greg Egan, *Quarantine*

You're on an airplane at 35,000 feet, and you strike up a conversation about aerodynamic lift with the passenger in your row. Things are going along just fine until they point out to you that your understanding of lift is wrong, and that planes couldn't fly from the effect you thought was responsible.

Should you immediately panic in fear that the plane will plummet out of the sky?

Obviously not; clearly the plane has been flying just fine up until now, and countless other planes have flown as well. There has to be *something* keeping the plane up, even if it's not what you thought, and even if you can't yet figure out what it actually is. Whatever is going on, it all adds up to normality.

Yet I claim that we often do this exact kind of panicked flailing when there's a challenge to our philosophical or psychological beliefs, and that this panic is entirely preventable.

I've experienced and/or seen this particular panic response when I, or others, encounter good arguments for propositions including

- My religion is not true. ("Oh no, then life and morality are meaningless and empty!")
- Many-worlds makes the most sense. ("Oh no, then there are always copies of me doing terrible things, and so none of my choices matter!")
- Many "altruistic" actions actually have hidden selfish motives. ("Oh no, then altruism doesn't exist and morality is pointless!")
- I don't have to be the best at something in order for it to be worth doing. ("Oh no, then others won't value me!") [Note: this one is from therapy; most people don't have the same core beliefs they're stuck on.]

(I promise these are not in fact strawmen. I'm sure you can think of your own examples. Also remember that panicking over an argument in this way is a mistake even if the proposition turns out to be false.)

To illustrate the way out, let's take the first example. It took me far too long to leave my religion, partly because I was so terrified about becoming a nihilist if I left that I kept flinching away from the evidence. (Of course, the religion proclaimed itself to be the origin of morality, and so it reinforced the notion that anyone else claiming to be moral was just too blind to see that their lack of faith implied nihilism.)

Eventually I did make myself face down, not just the object-level arguments, but the biases that had kept me from looking directly at them. And then I was an atheist, and still I was terrified of becoming a nihilist (especially about morality).

So I did one thing I still think was smart: I promised myself not to change all of my moral rules at once, but to change each one only when (under sober reflection) I decided it was wrong. And in the meantime, I read a lot of moral philosophy.

Over the next few months, I began relaxing the rules that were obviously pointless. And then I had a powerful insight: I was so cautious about changing my rules *because I wanted to help people and not slide into hurting them*. Regardless of what morality was, in fact, based on, the plane was still flying just fine. And that helped me sort out the good from the bad among the remaining rules, and to stop being so afraid of what arguments I might later encounter.

So in retrospect, the main thing I'd recommend is to **promise yourself to keep steering the plane mostly as normal while you think about lift** (to stretch the analogy). If you decide that something major is false, it doesn't mean that everything that follows from it has to be discarded immediately. (False things imply both true and false things!)

You'll generally find that many important things stand on their own without support from the old belief. (Doing this for the other examples I gave, as well as your own, is left to you.) Other things will collapse, and that's fine; that which can be destroyed by the truth should be. Just don't make all of these judgments in one fell swoop.

One last caution: **I recommend against changing meta-level rules as a result of changing object-level beliefs**. The meta level is how you correct bad decisions on the object level, and it should only be updated by very clear reasoning in a state of equilibrium. Changing your flight destination is perfectly fine, but don't take apart the wing mid-flight.

Good luck out there, and remember:

It all adds up to normality.

[EDIT 2020-03-25: khafra and Isnasene make good points about not applying this in cases where the plane shows signs of *actually dropping* and you're updating on *that*. (Maybe there's a new crisis in the external world that contradicts one of your beliefs, or maybe you update to believe that the thing you're about to do could actually cause a major catastrophe.)

In that case, you can try and land the plane safely- focus on getting to a safer state for yourself and the world, so that you have time to think things over. And if you can't do that, then you have no choice but to rethink your piloting on the fly, accepting the danger because you can't escape it. But these experiences will hopefully be very rare for you, current global crisis excepted.]

Covid-19 Points of Leverage, Travel Bans and Eradication

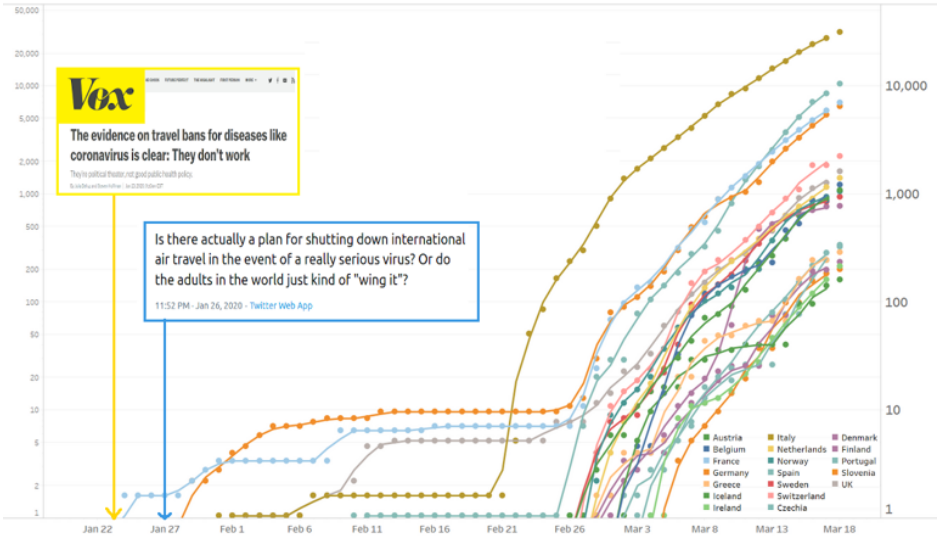
Covid-19 has become a major topic for discussion with talk about many different interventions and ideas that might help, from 3-D printing parts for respirators to drafting medical students into hospitals to thorough hand-washing procedures.

However as rationalists we should be asking which actions have the *highest* expected utility, not which actions have some positive utility. In an exponentially growing process, the actions with the highest expected utility are those actions which intervene early in the process, and actions like drafting medical students which intervene late in the process when the disease has already grown to a huge size are "nice to have" but by that point most of the damage has been done.

Proper and Prompt Travel Bans do Work

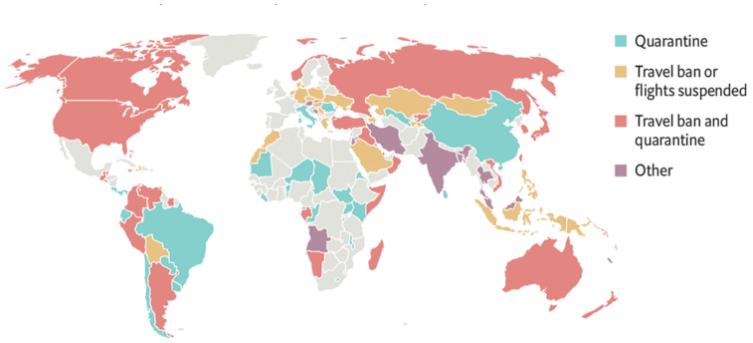
As early as January 26th, I called for cancellation of flights to limit the spread of covid-19; there was some pushback based on the idea that travel restrictions don't work which upon closer examination was actually the idea that late or half-hearted travel restrictions don't work:

> *During the height of the SARS outbreak in 2003, he had a colleague who wanted to return to the UK from Toronto, one of the cities most affected by the virus. So she caught a domestic flight from Toronto to Vancouver, then boarded a flight to London. "When she arrived at Heathrow [airport] and authorities asked her, 'Have you been to Toronto,' she said no and walked right through."*



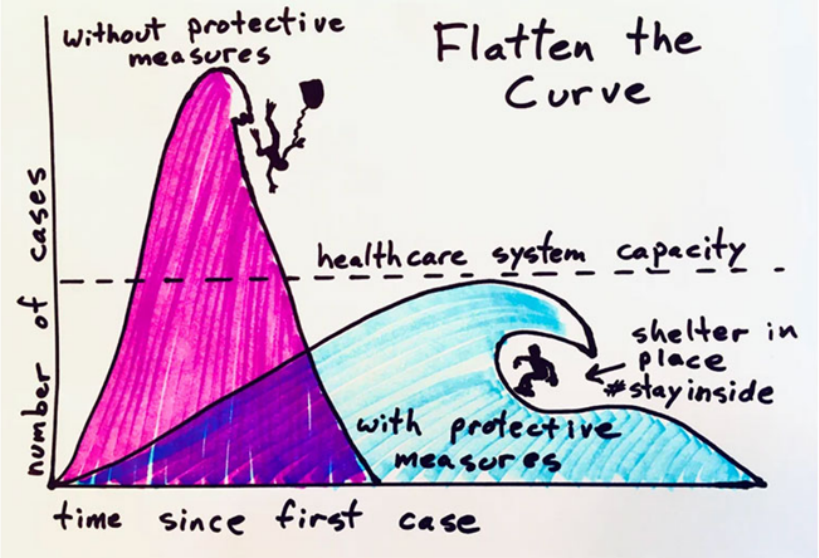
A policy that allows people to travel from an infected area to to an uninfected area is not a travel ban. It's containment theater. A real travel ban would be grounding *all* international flights and stopping passenger trains and boats until the disease had been eradicated or at least very well contained, as well as *aggressively* tracking down and contact tracing people who slipped through before the lockdown, for example using cellphone data from intelligence agencies. A key point here is that mopping up a small number of cases that slip through is in fact possible.

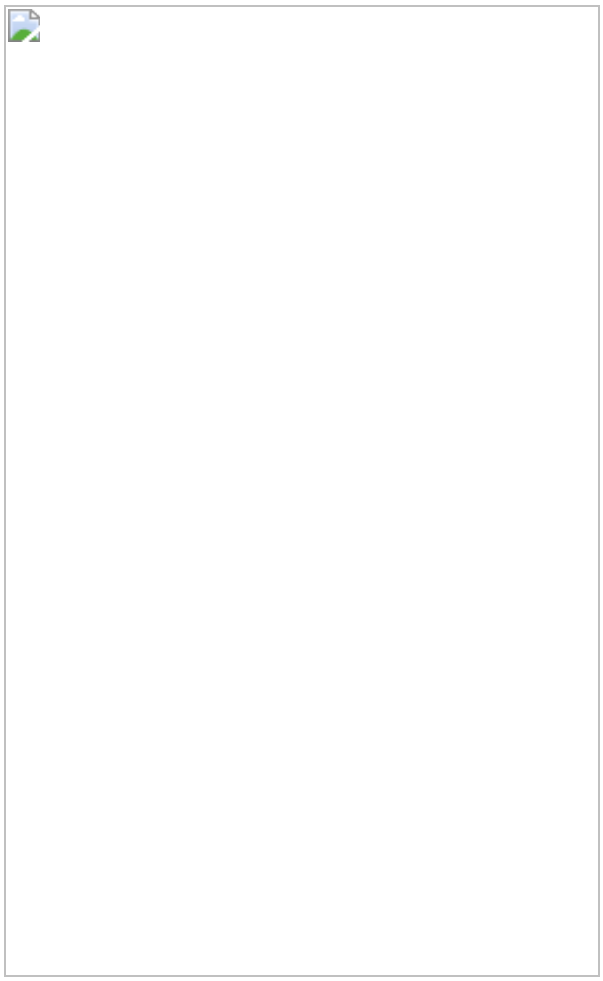
It would have been expensive to do all this, but the cost of not doing it is that the developed world is now on lockdown, the stock markets have fallen by around 33% and we have about 10,000 deaths at the time of writing. And we have ended up implementing the travel bans anyway!



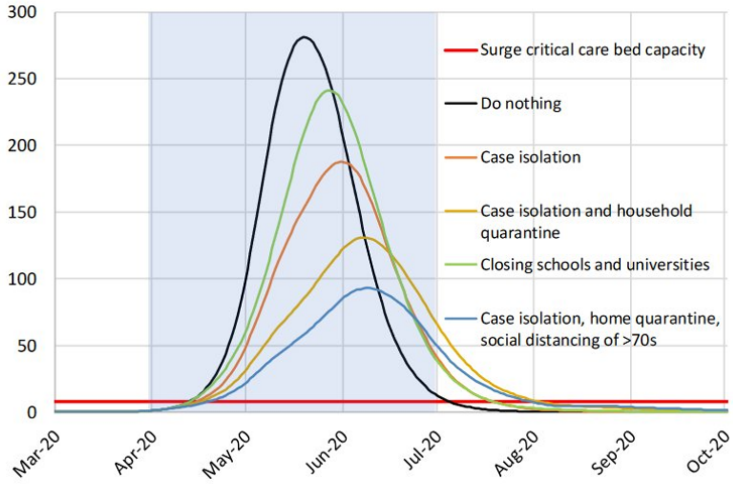
#Don'tFlattenTheCurve

The optimal strategy to defeat the disease is currently the subject of much debate. Several strategies have emerged, and a popular meme right now is #flattenthecurve. The idea of flattening the curve is that if we increase the *duration* of the pandemic, the number of people infected *at any one time* will be lower and our ability to treat people properly will be increased. People put a lot of time into creating convincing memes and diagrams showing how this works:


Flatten the Curve
without protective measures
number of cases
healthcare system capacity
shelter in place
#stayinside
with protective measures
time since first case



Unfortunately people didn't put much effort into getting the numbers right. Every single one of these diagrams is a steaming pile of nonsense because the line for "Healthcare System Capacity" is about 20-50 times too high, which was first pointed out by Joshua Bach. That tiny red line right next to the x-axis is our health system capacity:



(taken from The Imperial College COVID-19 Response Team's latest report).

The UK government's "herd immunity" strategy was another possible way forward, but the government reversed course on this when they realized it would involve at least a few hundred thousand deaths.

Contain and Eradicate

In my opinion, the correct strategy to beat covid-19 whilst minimizing losses from this point forward is a contain-and-eradicate strategy. The New England Complex Systems Institute's writeup on this, written by Nassim Nicholas Taleb of Black Swan fame outlines the strategy:

> *Since lockdowns result in exponentially decreasing numbers of cases, a comparatively short amount of time can be sufficient to achieve pathogen extinction, after which relaxing restrictions can be done without resurgence. ...*
>
> *Finally, the use of geographic boundaries and travel restrictions allows for effective and comparatively low cost imposition and relaxation of interventions. Such a multi-scale approach accelerates response efforts, reduces social impacts, allows for relaxing restrictions in areas earlier that are less affected, enables uninfected areas to assist in response in the ares that are infected, and is a much more practical and effective way to stop otherwise devastating outbreaks. ...*
>
> *A few other issues are of importance: They ignore the possibility of superspreader events in gatherings by not including the fat tail distribution of contagion in their model. This leads them to deny the importance of banning them, which has been shown to be incorrect, including in South Korea. Cutting the fat tail of the infection distribution is critical to reducing R0.*

Basically:

- Close borders and limit internal travel, lockdown and hygiene to drive R0 below 1

- Ban large events to cut off the long tail of the R0 distribution

- Use aggressive testing and contact tracing to clean up any remaining holdouts, and eradicate the virus on a region-by-region and country-by-country level.

- "Green" regions can return to mostly normal life, albeit without large events and travel. That means that people can go back to work and we can reverse the economic damage.

Contain-and-eradicate probably results in both less loss of life and less economic damage than any other strategy, and we can see this as a consequence of taking an exponential process and fighting it in the low orders of magnitude rather than the high ones. Flatten-The-Curve is bad because a flat curve that lasts for a long time is still, in log-terms, almost at the maximum power of the virus and therefore it can do huge amounts of damage. Herd-Immunity and Deliberate-Infection are bad for the same reason. The only other sensible plan I have seen is the idea of rushing a vaccine as quickly as possible, but that is beyond my expertise.

Travel bans and restrictions during a pandemic

Why do borders need to be closed now when the virus is already everywhere? Because there is still uncertainty about where the virus is and in what numbers. The virus wins when people with different amounts of virus mix, because areas of high virus can spread to areas of low virus whilst the reverse process doesn't work.

Similarly, if you were certain about who had the virus, this would almost be trivial because all the infected could be moved to containment facilities and everyone else could get on with running the economy.

The virus wants to maximize entropy (virus spread everywhere), humanity wants to minimize it (all virus in one place), for a given total amount of virus.

The need for borders is a result of this combination of uncertainty and mixing being bad. And as Taleb points out in the NECSI review, travel bans and restrictions during a pandemic should be multilevel.

As we approach the "endgame" where testing is ubiquitous and virus numbers get closer to 0, borders become more important, because adding 500 cases to an area with 1 case is *much* worse than adding 2000 cases to an area with 1000 cases (you have to think in logarithms).

Though even when numbers are high, closing borders is still useful and we should still do it; principally because the decision to close can lag behind rapidly developing facts on the ground, or worse the decision to close borders to a particular area could leak, at which point people start actively helping the virus to spread as they flee from the soon-to-be locked down area. Of course nobody would be stupid enough to leak that information, right?

What if we do Mitigation instead?

If we do go down the mitigation path - letting most people get the disease - there are some important "dice rolls" that will determine how it goes:

- The rate of long-term complications amongst covid-19 survivors,
- The rate at which young & otherwise healthy people die when hospital treatment is denied due to overcrowding
- Whether new pharmaceuticals like Chloroquine and Remdesivir are both effective and scalable, and how quickly covid-19 evolves resistance to them
- Whether summer weather substantially slows the spread
- Whether covid-19 picks up a mutation that makes it less lethal, or more lethal

A Test of Rationality

When competent Muggles make decisions, they're usually very empirical about it. They build a chair with one leg, it falls over, and then they don't do that again.

Covid-19's exponential dynamics, asymptomatic carriers and long lag time between infection and death punished the try-it-and-see approach very hard.

By the time it became absolutely obvious to people who build one-legged chairs that this was a *big deal* and needed attention, the virus had increased both its numbers and distribution most of the way to its goal of infecting every human being on the planet.

Covid-19 was a rationality test as well as a competence test. China failed on rationality but passed on competence. The West failed hard on rationality and is on course for a F+ on competence as well. Vox and the other mainstream media who either mocked those who took it seriously early, or got on a soapbox talking about racism (which is bad, but was not even remotely the most important thing at that time) should take a reputational hit. The various government agencies that dithered throughout February should be investigated, particularly in the USA.

Price Gouging and Speculative Costs

Let's say you see a potential pandemic coming, and you produce a product that could be critical. Maybe you make respirator masks, maybe you make ventilators, maybe you make PCR test reagents. You can see that if you and your competitors don't ramp up production and the pandemic happens, there will be a shortage. What do you do?

One option is to do nothing: keep producing at your regular rate. If the pandemic fears were overblown then you're fine. If the pandemic happens you quickly sell out, and start scrambling to ramp up production.

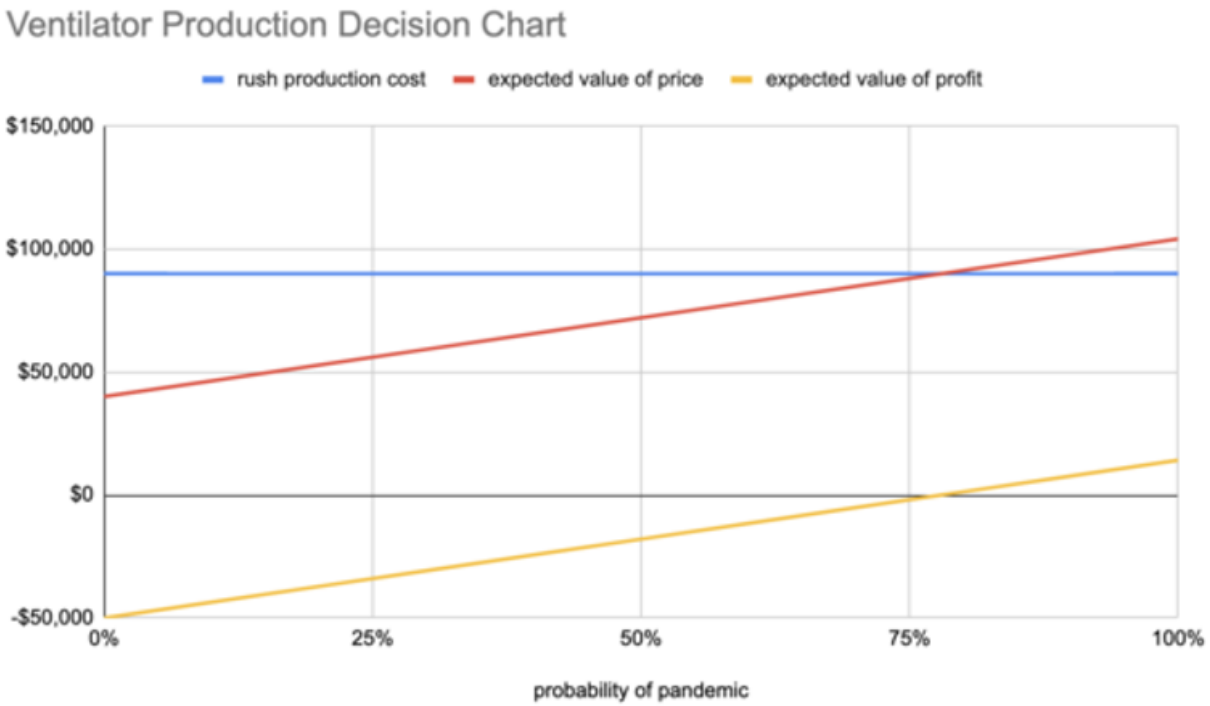
Another option is to ramp up production now, speculatively. Start paying workers extra to work longer shifts and run your assembly lines around the clock. Train extra workers. Find what you're bottlenecked on and figure out how to get that ramped up too. If the pandemic fears were overblown you lose a lot of money, but if the pandemic happens people need what you have so much that you can charge high prices. How much to ramp up production in advance depends on how likely you think the pandemic is, and how much you'd be able to increase prices if it does happen.

Except we have laws and customs against price gouging: if the pandemic does happen, you are going to have a lot of trouble raising your prices. The laws generally do allow passing along increased costs, but the problem here is that your costs were speculative. Let's work an example.

Imagine your ventilators normally cost $35k to make, and the market price is $40k each. If you push really hard to ramp up production you can make a lot more, but your cost goes up to $90k/each. In New Jersey, the state I looked at last time, you're allowed to pass along costs but can't increase your profit on "merchandise which is consumed or used as a direct result of an emergency or which is consumed or used to preserve, protect, or sustain the life, health, safety or comfort of persons or their property" by more than 10% (56:8-107, 108, 109). Your normal profit is 14%, so you would be allowed to sell them for $104k. Your possibilities are:

- No pandemic: you spent $90k each, but the market price is $40k. You lose $50k each.
- Yes pandemic: you spent $90k each, and the market price is way above that, but you can legally only charge $104k. You make $14k each.

If you think the pandemic is >78% likely to happen then you'll expect to make money by ramping up production, otherwise you'll lose money. So even if a pandemic looks, say, 75% likely, you don't ramp up.



Similarly, imagine you make respirator masks and you know that every so often there's an emergency where demand spikes. Could be a pandemic, but could also be widespread fires, or many other things. Since melt-blown fabric is a major bottleneck, you could decide to keep a large stockpile of it so you can easily ramp up production in an emergency. The same unfavorable math applies here: you're heavily limited in how much you can increase prices in an emergency, so keeping a large stockpile to be prepared for even a reasonably likely event is a money-loosing proposition.

In the current crisis people are likely to die because we don't have enough ventilators, and the marginal person probably needs one for about a week, and the peak lasts maybe two months, so the marginal ventilator saves about eight lives. At the US statistical value of life of ~$9M, that's $72M per ventilator. We're heading into a disaster where we don't have enough machines that we would value at ~$72M/each and normally cost ~$35k/each to make. This is really bad.

It's too late to fix this for the current situation, but I see three main ways out of this for the future:

- Allow price gouging: don't restrict what prices people can sell things at.
- Allow speculative production: require companies to disclose and document production plans, require them to share their probability estimates of how likely they think things are to be needed, keep them honest by allowing third parties to bet against them at their published probabilities.
- Have a government that will put in emergency ventilator orders at an early stage of the crisis even when they may not be needed, stockpile masks for potential pandemics, and generally stay on top of things.

These three approaches require increasing levels of government competence and foresight, and given recent performance I'm pretty skeptical. But the current approach

where the government doesn't handle the problem and also does not allow industry to make a profit handling the problem is a disaster.

Comment via: facebook

History's Biggest Natural Experiment

Faced with the threat of COVID-19, most countries in the world are suddenly, dramatically improving their hygiene and infectious disease control. This won't just slow the spread of COVID-19. As a byproduct, it will also suppress almost every other contagious disease. This creates a natural experiment, which answers the question:

What would happen if we fully suppressed influenza, the common cold, and all the other ignored, minor infectious diseases, all at once?

About 13% of cancers worldwide are attributable to known infectious causes. Many other health problems are suspected to be related to infectious disease, including chronic fatigue syndrome, type 1 diabetes, hypothyroidism, obesity, Alzheimers disease, bipolar disorder, and some mechanisms of aging. (And obesity again but this time it's a bacterium instead of a virus.)

In the coming year, we're going to see large reductions in many diseases we never realized were infectious-disease related. As bad as COVID-19 is, this will be a pretty substantial silver lining.

Most of the diseases won't stay gone; social distancing and citywide lockdowns are costly, and most diseases will retain enough reservoirs to bounce back after we return to normal. What we'll get is a period we can study retrospectively, different in different countries, in which almost no disease transmission occurred.

This will be confounded by COVID-19 itself, of course. People who themselves had COVID-19 will have a lot of health problems, which studies will need to account for. The isolation of social distancing, the anxiety of watching the disaster unfold, and the poverty of job-less and economic decline will leave their mark. In countries where COVID-19 took hold, during the height of the pandemic, cancers and autoimmune disorders and other conditions will have been left undiagnosed due to the shortage of hospitals. We won't easily be able to tell which diseases matched with which health conditions. With clever methodology and access to data in countries which had different outcomes, all of these issues will be possible to work around.

The world's attention is now very focused on the short term. Let's give a little thought to what next year will look like.

Can crimes be discussed literally?

This is a linkpost for http://benjaminrosshoffman.com/can-crimes-be-discussed-literally

Suppose I were to say that the American legal system is a criminal organization. The usual response would be that this is a crazy accusation.

Now, suppose I were to point out that it is standard practice for American lawyers to advise their clients to lie under oath in certain circumstances. I expect that this would still generally be perceived as a heterodox, emotionally overwrought, and perhaps hysterical conspiracy theory.

Then, suppose I were to further clarify that people accepting a plea bargain are expected to affirm under oath that no one made threats or promises to induce them to plead guilty, and that the American criminal justice system is heavily reliant on plea bargains. This might be conceded as literally true, but with the proviso that since everyone does it, I shouldn't use extreme language like "lie" and "fraud."

This isn't about lawyers - some cases in other fields:

In American medicine it is routine to officially certify that a standard of care was provided, that cannot possibly have been provided (e.g. some policies as to the timing of medication and tests can't be carried out given how many patients each nurse has to care for, but it's less trouble to fudge it as long as something vaguely resembling the officially desired outcome happened). The system relies on widespread willingness to falsify records, and would (temporarily) grind to a halt if people were to simply refuse to lie. But I expect that if I were to straightforwardly summarize this - that the American hospital system is built on lies - I mostly expect this to be evaluated as an attack, rather than a description. But of course if any one person refuses to lie, the proximate consequences may be bad.

Likewise for the psychiatric system.

In Simulacra and Subjectivity , the part that reads "while you cannot acquire a physician's privileges and social role simply by providing clear evidence of your ability to heal others" was, in an early draft, "physicians are actually nothing but a social class with specific privileges, social roles, and barriers to entry." These are expressions of the same thought, but the draft version is a direct, simple theoretical assertion, while the latter merely provides evidence *for* the assertion. I had to be coy on purpose in order to distract the reader from a potential fight.

The End User License Agreements we almost all falsely certify that we've read in order to use the updated version of any software we have are of course familiar. And when I worked in the corporate world, I routinely had to affirm in writing that I understood and was following policies that were nowhere in evidence. But of course if I'd *personally* refused to lie, the proximate consequences would have been counterproductive.

The Silicon Valley startup scene - as attested in Zvi's post, the show *Silicon Valley*, the New Yorker profile on Y Combinator (my analysis), and plenty of anecdotal evidence - uses business metrics as a theatrical prop to appeal to investors, not an accounting device to make profitable decisions on the object level.

The general argumentative pattern is:

> *A: X is a fraudulent enterprise.*
> *B: How can you say that?!*
> *A: X relies on asserting Y when we know Y to be false.*
> *B: But X produces benefit Z, and besides, everyone says Y and the system wouldn't work without it, so it's not reasonable to call it fraud.*

This wouldn't be as much of a problem if terms like "fraud", "lie," "deception" were unambiguously attack words, with a literal meaning of "ought to be scapegoated as deviant." The problem is that there's simultaneously the definition that the dictionaries claim the word has, with a literal meaning independent of any call to action.

There is a clear conflict between the use of language to punish offenders, and the use of language to describe problems, and there is great need for a language that can describe problems.

For instance, if I wanted to understand how to interpret statistics generated by the medical system, I would need a short, simple way to refer to any significant tendency to generate false reports. If the available simple terms were also attack words, the process would become much more complicated.

Related: Model-building and scapegoating, Blame games, Judgment, Punishment, and the Information-Suppression Field, AGAINST LIE INFLATION, Maybe Lying Doesn't Exist

Toby Ord’s ‘The Precipice’ is published!

[x-posted from EA Forum]

The Precipice: Existential Risk and the Future of Humanity is out today. I’ve been working on the book with Toby for the past 18 months, and I’m excited for everyone to read it. I think it has the potential to make a profound difference to the way the world thinks about existential risk.

How to get it

- It's out in the UK on March 5 and US March 24
- An audiobook, narrated by Toby himself, is out March 24
- You can buy it on Amazon now, or at theprecipice.com/purchase
- You can download the opening chapters for free by signing up to the newsletter at www.theprecipice.com

What you can do

- Read the book
- Talk about it with your friends and family, or share quotes you like on social media
- If you enjoy it, consider writing a review on Amazon or Goodreads

Summary of the book

Part One: The Stakes

Toby places our time within the broad sweep of human history: showing how far humanity has come in 2,000 centuries, and where we might go if we survive long enough. He outlines the major transitions in our past—the Agricultural, Scientific, and Industrial Revolutions. Each is characterised by dramatic increases in our power over the natural world, and together they have yielded massive improvements in living standards. During the twentieth century, with the detonation of the atomic bomb, humanity entered a new era. We gained the power to destroy itself, without the wisdom to ensure that we don’t. This is the Precipice, and how we navigate this period will determine whether humanity has a long and flourishing future, or no future at all. Toby introduces the concept of existential risk—risks that threaten to destroy humanity’s longterm potential. He shows how the case for safeguarding humanity from these risks draws support from a range of moral perspectives. Yet it remains grossly neglected—humanity spends more each year on ice cream than we do on protecting our future.

Part Two: The Risks

Toby explores the science behind the risks we face. In *Natural Risks*, he considers threats from asteroids & comets, supervolcanic eruptions, and stellar explosions. He shows how we can use humanity’s 200,000 year history to place strict bounds on how high the natural risk could be. In *Anthropogenic Risks*, he looks at risks we have imposed on ourselves in the last century, from nuclear war, extreme climate change,

and environmental damage. In *Future Risks*, he turns to threats that are on the horizon from emerging technologies, focusing in detail on engineered pandemics, unaligned artificial intelligence, and dystopian scenarios.

Part Three: The Path Forward

Toby surveys the risk landscape and gives his own estimates for each risk. He also provides tools for thinking about how they compare and combine, and for how to prioritise between risks. He estimates that nuclear war and climate change each pose more risk than all the natural risks combined, and that risks from emerging technologies are higher still. Altogether, Toby believes humanity faces a 1 in 6 change of existential catastrophe in the next century. He argues that it is in our power to end these risks today, and to reach a place of safety. He outlines a grand strategy for humanity, provides actionable policy and research recommendations, and shows what each of us can do. The book ends with an inspiring vision of humanity’s potential, and what we might hope to achieve if we navigate the risks of the next century.

What should we do once infected with COVID-19?

We’ve talked a lot about preparations and prevention, but statistically some of us, or people we care about, are going to actually get sick. What do we do once that happens?

Thinking About Filtered Evidence Is (Very!) Hard

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The content of this post would not exist if not for conversations with Zack Davis, and owes something to conversations with Sam Eisenstat.

There's been some talk about filtered evidence recently. I want to make a mathematical observation which causes some trouble for the Bayesian treatment of filtered evidence. [OK, when I *started* writing this post, it was "recently". It's been on the back burner for a while.]

This is also a continuation of the line of research about trolling mathematicians, and hence, relevant to logical uncertainty.

I'm going to be making a mathematical argument, but, I'm going to keep things rather informal. I think this increases the clarity of the argument for *most* readers. I'll make some comments on proper formalization at the end.

Alright, here's my argument.

According to the Bayesian treatment of filtered evidence, you need to update on *the fact that the fact was presented to you*, rather than the raw fact. This involves reasoning about the algorithm which decided which facts to show you. The point I want to make is that this can be incredibly computationally difficult, *even if the algorithm is so simple that you can predict what it will say next.* IE, I don't need to rely on anything like "humans are too complex for humans to really treat as well-specified evidence-filtering algorithms".

For my result, we imagine that a Bayesian reasoner (the "listener") is listening to a series of statements made by another agent (the "speaker").

First, I need to establish some terminology:

Assumption 1. *A listener will be said to have a* **rich hypothesis space** *if the listener assigns some probability to the speaker enumerating any computably enumerable set of statements.*

The intuition behind this assumption is supposed to be: due to computational limitations, the listener may need to restrict to some set H of easily computed hypotheses; for example, the hypotheses might be poly-time or even log-poly. This prevents hypotheses such as "the speaker is giving us the bits of a halting oracle in order", as well as "the speaker has a little more processing power than the listener". However, the hypothesis space is not *so* restricted as to limit the world to being a finite-state machine. The listener can imagine the speaker proving complicated theorems, so long as it is done *sufficiently slowly for the listener to keep up*. In such a model, the listener might imagine the speaker staying quiet for quite a long time

(observing the null string over and over, or some simple sentence such as 1=1) while a long computation completes; and only then making a complicated claim.

This is also not to say that I *assume* my listener considers *only* hypotheses in which it can 100% keep up with the speaker's reasoning. The listener can also have probabilistic hypotheses which recognize its inability to perfectly anticipate the speaker. I'm only pointing out that my result *does not rely on* a speaker which the listener can't keep up with.

What it *does* rely on is that there are not too many restrictions on what the speaker *eventually* says.

Assumption 2. *A listener* **believes a speaker to be honest** *if the listener distinguishes between "X" and "the speaker claims X at time t" (aka "*claims_t*-X"), and also has beliefs such that* $P(X|\ \text{claims}_t\text{-}X)=1$ *when* $P(\text{claims}_t\text{-}X) > 0$*.*

This assumption is, basically, saying that the agent trusts its observations; the speaker can filter evidence, but the speaker cannot falsify evidence.

Maybe this assumption seems quite strong. I'll talk about relaxing it after I sketch the central result.

Assumption 3. *A listener is said to have* **minimally consistent beliefs** *if each proposition* X *has a negation* X*, and* $P(X)+P(X^*)\leq 1$.

The idea behind minimally consistent beliefs is that the listener need not be logically omniscient, but does avoid outright contradictions. This is important, since assuming logical omniscience would throw out computability from the start, making any computational-difficulty result rather boring; but totally throwing out logic would make my result impossible. Minimal consistency keeps an extremely small amount of logic, but, it is enough to prove my result.

Theorem(/Conjecture). *It is not possible for a Bayesian reasoner, observing a sequence of remarks made by a speaker, to simultaneously:*

- *Have a rich hypothesis space.*
- *Believe the speaker to be honest.*
- *Have minimally consistent beliefs.*
- *Have computable beliefs.*

Proof sketch. Suppose assumptions 1-3. Thanks to the rich hypothesis space assumption, the listener will assign some probability to the speaker enumerating theorems of PA (Peano Arithmetic). Since this hypothesis makes distinct predictions, it is possible for the confidence to rise above 50% after finitely many observations. At that point, since the listener expects each theorem of PA to *eventually* be listed, with probability > 50%, and the listener believes the speaker, the listener *must* assign > 50% probability to each theorem of PA! But this implies that the listener's beliefs are not computable, since if we had access to them we could *separate theorems of PA from contradictions* by checking whether a sentence's probability is > 50%. □

So goes my argument.

What does the argument basically establish?

The argument is supposed to be surprising, because *minimally consistent beliefs* are compatible with computable beliefs; and *rich hypothesis space* is compatible with beliefs which are computable on observations alone; yet, when combined with a belief that the speaker is honest, we get an incomputability result.

My take-away from this result is that we cannot *simultaneously* use our unrestricted ability to predict sensory observations accurately *and* have completely coherent beliefs about the world which produces those sensory observations, at least if our "bridge" between the sensory observations and the world includes something like language (whereby sensory observations contain complex "claims" about the world).

This is because using the *full force* of our ability to predict sensory experiences includes some hypotheses which *eventually* make surprising claims about the world, by incrementally computing increasingly complicated information (like a theorem prover which slowly but inevitably produces all theorems of PA). In other words, a rich sensory model contains *implicit information about the world* which we cannot immediately compute the consequences of (in terms of probabilities about the hidden variables out there in the world). This "implicit" information can be *necessarily* implicit, in the same way that PA is *necessarily* incomplete.

To give a non-logical example: suppose that your moment-to-moment anticipations of your relationship with a friend are pretty accurate. It might be that if you roll those anticipations forward, you inevitably become closer and closer until the friendship becomes a romance. *However,* you can't necessarily predict that right now; even though the anticipation of each next moment is relatively easy, you face a halting-problem-like difficulty if you try to anticipate what the *eventual* behavior of your relationship is. Because our ability to look ahead is bounded, each new consequence can be predictable without the overall outcome being predictable.

Thus, in order for an agent to use the full force of its computational power on predicting sensory observations, it must have *partial* hypotheses -- similar to the way logical induction contains traders which focus only on special classes of sentences, or Vanessa's incomplete Bayesianism contains incomplete hypotheses which do not try to predict everything.

So, this is an argument against strict Bayesianism. In particular, it is an argument against strict Bayesianism *as a model of updating on filtered evidence!* I'll say more about this, but first, let's talk about possible holes in my argument.

Here are some concerns you might have with the argument.

One might possibly object that the perfect honesty requirement is unrealistic, and therefore conclude that the result does not apply to realistic agents.

- I would point out that the assumption is not so important, so long as the listener *can conceive of the possibility of perfect honesty,* and assigns it nonzero probability. In that case, we can consider P(X|honesty) rather than P(X). Establishing that some conditional beliefs are not computable seems similarly damning.
- Furthermore, because the "speaker" is serving the role of our *observations*, the perfect honesty assumption is just a version of P(X|observe-X)=1. IE, *observing*

X gives us X. This is true in typical filtered-evidence setups; IE, filtered evidence can be misleading, but it can't be false.
- However, one might further object that *agents need not be able to conceive of* "*perfect honesty*", because this assumption has an unrealistically aphysical, "perfectly logical" character. One might say that *all* observations are imperfect; none are perfect evidence of what is observed. In doing so, we can get around my result. This has some similarity to the assertion that zero is not a valid probability. I don't find this response particularly appealing, but I also don't have a strong argument against it.

Along similar lines, one might object that the result depends on an example ("the speaker is enumerating theorems") which comes from logic, as opposed to any realistic physical world-model. The example does have a "logical" character -- we're not explicitly reasoning about evidence-filtering algorithms interfacing with an empirical world and selectively telling us some things about it. However, I want to point out that I've assumed extremely little "logic" -- the only thing I use is that you don't expect a sentence and its negation to both be true. Observations corresponding to theorems of PA are just an example used to prove the result. The fact that P(X) can be very hard to compute even when we restrict to easily computed $P(claims_t\text{-}X)$ is very general; even if we do restrict attention to finite-state-machine hypotheses, we are in P-vs-NP territory.

What does this result say about logical uncertainty?

Sam's untrollable prior beat the trollable-mathematician problem by the usual Bayesian trick of explicitly modeling the sequence of observations -- updating on I-observed-X-at-this-time rather than only X. (See also the illustrated explanation.)

However, it did so at a high cost: Sam's prior is *dumb.* It isn't able to perform rich Occam-style induction to divine the hidden rules of the universe. It doesn't *believe in* hidden rules; it believes "if there's a law of nature constraining everything to fit into a pattern, *I will eventually observe that law directly*." It shifts its probabilities when it makes observations, but, in some sense, it doesn't shift them *very much*; and indeed, that property seems key to the computability of that prior.

So, a natural question arises: is this an *essential* property of an untrollable prior? Or can we construct a "rich" prior which entertains hypotheses about the deep structure of the universe, learning about them in an Occam-like way, which is nonetheless still untrollable?

The present result is a first attempt at an answer: given my (admittedly a bit odd) notion of rich hypothesis space, it is indeed impossible to craft a computable prior over logic with some minimal good properties (like believing what's proved to it). I don't directly address a trollability-type property, unfortunately; but I do think I get close to the heart of the difficulty: a "deep" ability to adapt in order to predict data better stands in contradiction with computability of the latent probability-of-a-sentence.

So, how should we think about filtered evidence?

Orthodox Bayesian (OB): We can always resolve the problem by distinguishing between X and "I observe X", and conditioning on **all** the evidence available. Look

how nicely it works out in the Monty Hall problem and other simple examples we can write down.

Skeptical Critique (SC): You're ignoring the argument. You can't handle cases where running your model forward is easier than answering questions about what happens eventually; in those cases, many of your beliefs will either be uncomputable or incoherent.

OB: That's not a problem for me. Bayesian ideals of rationality apply to the logically omniscient case. What they give you is an idealized notion of rationality, which defines the **best** an agent **could** do.

SC: Really? Surely your Bayesian perspective is supposed to have some solid implications for finite beings who are not logically omniscient. I see you giving out all this advice to machine learning programmers, statisticians, doctors, and so on.

OB: Sure. We might not be able to achieve perfect Bayesian rationality, but whenever we see something less Bayesian than it could be, we can correct it. That's how we get closer to the Bayesian ideal!

SC: That sounds like cargo-cult Bayesianism to me. If you spot an inconsistency, **it matters how you correct it**; you don't want to go around correcting for the planning fallacy by trying to do everything faster, right? Similarly, if your rule-of-thumb for the frequency of primes is a little off, you don't want to add composite numbers to your list of primes to fudge the numbers.

OB: No one would make those mistakes.

SC: That's because there are, in fact, rationality principles which apply. You **don't** just cargo-cult Bayesianism by correcting inconsistencies any old way. A boundedly rational agent has rationality constraints which apply, guiding it to better approximate "ideal" rationality. And those rationality constraints don't actually need to refer to the "ideal" rationality. **The rationality constraints are about the update, not in the ideal which the update limits to.**

OB: Maybe we can imagine some sort of finite Bayesian reasoner, who treats logical uncertainty as a black box, and follows the evidence toward unbounded-Bayes-optimality in a bounded-Bayes-optimal way...

SC: Maybe, but I don't know of a good picture which looks like that. The picture we **do** have is given by logical induction: we learn to avoid Dutch books by noticing lots of Dutch books against ourselves, and gradually becoming less exploitable.

OB: That sounds a lot like the picture I gave.

SC: Sure, but it's more precise. And more importantly, it's **not** a Bayesian update -- there is a kind of family resemblance in the math, but it isn't learning through a Bayesian update in a strict sense.

OB: Ok, so what does all this have to do with filtered evidence? I still don't see why the way I handle that is wrong.

SC: Well, isn't the standard Bayesian answer a little suspicious? The numbers conditioning on X don't come out to what you want, so you introduce something new

to condition on, observe-X, which can have different conditional probabilities. Can't you get whatever answer you want, that way?

OB: I don't think so? The numbers are dictated by the scenario. The Monty Hall problem has a right answer, which determines how you should play the game if you want to win. You can't fudge it without changing the game.

SC: Fair enough. But I still feel funny about something. Isn't there an infinite regress? We jump to updating on observe-X when X is filtered. What if observe-X is filtered? Do we jump to observe-observe-X? What if we can construct a "meta Monty-Hall problem" where it isn't sufficient to condition on observe-X?

OB: If you observe, you observe that you observe. And if you observe that you observe, then you must observe. So there's no difference.

SC: If you're logically perfect, sure. But a boundedly rational agent need not realize immediately that it observed X. And certainly it need not realize and update on the entire sequence "X", "I observed X", "I observed that I observed X", and so on.

OB: Ok...

SC: To give a simple example: call a sensory impression "subliminal" when it is weak enough that only X is registered. A stronger impression also registers "observe-X", making the sensory impression more "consciously available". Then, we cannot properly track the effects of filtered evidence for subliminal impressions. Subliminal impressions would always register as if they were unfiltered evidence.

OB: ...no.

SC: What's wrong?

OB: An agent should come with a basic notion of sensory observation. If you're a human, that could be activation in the nerves running to sensory cortex. If you're a machine, it might be RBG pixel values coming from a camera. That's the only thing you ever have to condition on; all your evidence has that form. Observing a rabbit means *getting pixel values corresponding to a rabbit.* We don't start by conditioning on "rabbit" and then patch things by adding "observe-rabbit" as an additional fact. We condition on *the complicated observation corresponding to the rabbit*, which happens to, *by inference,* tell us that there is a rabbit.

SC: That's... a bit frustrating.

OB: How so?

SC: The core Bayesian doctrine is the Kolmogorov axioms, together with the rule that we update beliefs via Bayesian conditioning. A common extension of Bayesian doctrine *grafts on a distinction between observations and hypotheses,* naming some special events as observable, and others as non-observable hypotheses. I want you to *notice when you're using the extension rather than the core.*

OB: How is that even an extension? It just sounds like a special case, which happens to apply to just about any organism.

SC: But you're restricting the rule "update beliefs by Bayesian conditioning" -- you're saying that it only works for *observations*, not for other kinds of events.

OB: Sure, but you could never update on those other kinds of events anyway.

SC: Really, though? Can't you? Some information you update on comes from sensory observations, but other information comes from *reasoning*. Something like a feedforward neural network just computes one big function on sense-data, and can probably be modeled in the way you're suggesting. But something like a memory network has a nontrivial reasoning component. A Bayesian can't handle "updating" on internal calculations it's completed; at best they're treated as if they're black boxes whose outputs are "observations" again.

OB: Ok, I see you're backing me into a corner with logical uncertainty stuff again. I still feel like there should be a Bayesian way to handle it. But what does this have to do with filtered evidence?

SC: *The whole point of the argument we started out discussing* is that if you have this kind of observation/hypothesis divide, and have sufficiently rich ways of predicting sensory experiences, *and remain a classical Bayesian*, then your beliefs about the hidden information are not going to be computable, even if your hypotheses themselves are easy to compute. So we can't realistically reason about the hidden information just by Bayes-conditioning on the observables. The only way to maintain both computability and a rich hypothesis space under these conditions is to be less Bayesian, allowing for more inconsistencies in your beliefs. *Which means,* **reasoning about filtered evidence doesn't reduce to applying Bayes' Law**.

OB: That... seems wrong.

SC: Now we're getting somewhere!

All that being said, reasoning about filtered evidence via Bayes' Law in the orthodox way still seems quite practically compelling. The perspective SC puts forward in the above dialogue would be much more compelling if I had more practical/interesting "failure-cases" for Bayes' Law, and more to say about alternative ways of reasoning which work better for those cases. A real "meta Monty-Hall problem".

Arguably, logical induction *doesn't* use the "condition on the fact that X was observed" solution:

- Rather than the usual sequential prediction model, logical induction accommodates information coming in for any sentence, in any order. So, like the "core of Bayesianism" mentioned by SC, it maintains its good properties without special assumptions about what is being conditioned on. This is in contrast to, e.g., Solomonoff induction, which uses the sequential prediction model.
- In particular, in Monty Hall, although there is a distinction between the sentence "there is a goat behind door 3" and "the LI discovers, at time *t*, that there is a goat behind door 3" (or suitable arithmetizations of these sentences), we can condition on the first rather than the second. A logical inductor would learn to react to this in the appropriate way, since doing otherwise would leave it Dutch-bookable.

One might argue that the traders are implicitly using the standard Bayesian "condition on the fact that X was observed" solution in order to accomplish this. Or that the update an LI performs upon seeing X *is always* that it saw X. But to me, this feels like stretching things. The core of the Bayesian method for handling filtered evidence is to distinguish between X and the observation of X, and update on the latter. A logical

inductor doesn't explicitly follow this, and indeed appears to violate it. Part of the usual idea seems to be that a Bayesian needs to "update on all the evidence" -- but a logical inductor just gets a black-box report of X, without any information on how X was concluded or where it came from. So information can be arbitrarily excluded, and the logical inductor will still do its best (which, in the case of Monty Hall, appears to be sufficient to learn the correct result).

A notable thing about the standard sort of cases, where the Bayesian way of reasoning about filtered evidence is entirely adequate, is that you have a gears-level model of what is going on -- a causal model, which you can turn the crank on. If you run such a model "forward" -- in causal order -- you compute the hidden causes *before* you compute the filtered evidence about them. This makes it sound as if predicting the hidden variables should be *easier* than predicting the sensory observations; and, certainly makes it hard to visualize the situation where it is much much harder.

However, even in cases where we have a nice causal model like that, inferring the hidden variables from what is observed can be intractably computationally difficult, since it requires *reverse-engineering* the computation from its outputs. Forward-sampling causal models is always efficient; running them backwards, not so.

So even with causal models, there can be good reason to engage more directly with logical uncertainty rather than use pure Bayesian methods.

However, I suspect that one could construct a much more convincing example if one were to use partial models explicitly in the construction of the example. Perhaps something involving an "outside view" with strong empirical support, but lacking a known "inside view" (lacking a single consistent causal story).

Unfortunately, such an example escapes me at the moment.

Finally, some notes on further formalisation of my main argument.

The listener is supposed to have probabilistic beliefs of the standard variety -- an event space which is a sigma-algebra, and which has a P(event) obeying the Kolmogorov axioms. In particular, the beliefs are supposed to be perfectly logically consistent in the usual way.

However, in order to allow logical uncertainty, I'm assuming that there is *some embedding of arithmetic;* call it E[]. So, for each arithmetic sentence S, there is an event E[S]. Negation gets mapped to the "star" of an event: $E[\neg S] = (E[S])^*$. This need not be the compliment of the event E[S]. Similarly, the embedding $E[A \vee B]$ need not be $E[A] \cup E[B]$; $E[A \wedge B]$ need not be $E[A] \cap E[B]$; and so on. That's what allows for logical non-omniscience -- the probability distribution doesn't necessarily *know* that $E[A \wedge B]$ should act like $E[A] \cap E[B]$, and so on.

The more we impose requirements which force the embedding to *act like it should,* the more logical structure we are forcing onto the beliefs. If we impose very much consistency, however, then that would already imply uncomputability and the central result would not be interesting. So, the "minimal consistency" assumption requires

very little of our embedding. Still, it is enough for the embedding of PA to cause trouble in connection with the other assumptions.

In addition to all this, we have a distinguished set of events which count as observations. A first pass on this is that for any event A, there is an associated event obs(A) which is the observation of A. But I do worry that this includes more observation events than we want to require. Some events A do not correspond to sentences; sigma-algebras are closed under countable unions. If we think of the observation events as *claims made by the speaker*, it doesn't make sense to imagine the speaker claiming a countable union of sentences (particularly not the union of an uncomputable collection).

So, more conservatively, we might say that for events E[S], that is *events in the image of the embedding*, we also have an event obs(E[S]). In any case, this is closer to the minimal thing we need to establish the result.

I don't know if the argument works out exactly as I sketched; it's possible that the rich hypothesis assumption needs to be "and also positive weight on a particular enumeration". Given that, we can argue: take one such enumeration; as we continue getting observations consistent with that observation, the hypothesis which predicts it loses no weight, and hypotheses which (eventually) predict other things must (eventually) lose weight; so, the updated probability eventually believes that particular enumeration will continue with probability > 1/2.

On the other hand, that patched definition is certainly less nice. Perhaps there is a better route.

Crisis and opportunity during coronavirus

Note: please put on your own oxygen mask first. Don't engage with this post if you haven't taken appropriate measures to prepare yourself and your family; and plausibly don't engage if you haven't taken measures to ensure you can do so while staying stable and grounded.

We face a time of global crisis. But in spite of the unfolding tragedy – or perhaps *because of it* – this will also be a time of great opportunity. If you have the skills, slack, and willingness to act, it might make sense to start looking for ways to contribute (regardless of whether you're seeking personal gain or altruistic benefit).

Why does this seem like a good opportunity?

There's a question of whether the current situation should change your overall cause-prioritisation, in determining what's most useful to work on over a >1 year time-scale.

This depends on where you started in your beliefs. To many readers, this likely does not provide any high-level updates, as we already believed that pandemics were a major risk for which the world was underprepared, and that the major institutions in charge were dysfunctional. (Nonetheless, I am learning a *massive* amount by living through a time of global crisis when I have the epistemic ability and agency to understand what's happening and take action.)

Beyond that, there's the question of whether this is a *window of opportunity*. Even if your long-term goals remain after this pandemic, are there actions which will have an extraordinarily high leverage now, compared to other times?

I think there are a few reasons for thinking so.

- **Underpreparation.** The world wasn't prepared. Everyone is scrambling to figure things out and there's too much for anyone to do. Hundreds of millions of people are suddenly changing their lives. The same goes for hundreds of thousands of companies and hundreds of governments. Most of them have no routines or experience in handling situations like these, which means they'll be facing *problems they have never faced before*.
- **Exponential growth.** Each infected person can be responsible for thousands of downstream infections, so the impact of behaviour change has a large multiplier. (Though this is modulo some uncertainty about counterfactual infections which I'm unsure how to think about).
- **Scale.** The pandemic might grow to directly affect hundreds of millions of people, and it will indirectly affect billions. It is also a *memetic* pandemic (it probably consumes >90% of my FB and Twitter feeds, and >70% of my conversations). People are actively trying to find information, products, and similar.
- **Direct exposure and quick feedback loops.** Most startups die because no one wants what they're building. A common warning sign is that the founders

aren’t themselves users of the product. But in the coming months, you’ll have to solve lots of problems for yourself, and chances are high that others might benefit from your solution (e.g. many spreadsheets and documents that went viral were initially just a single person trying to figure out how they should prepare, when their company should work remotely, etc.) Even if you're solving problems for others, you'll quick learn if there's any demand.

How can you contribute?

I want to distinguish two kinds of windows of opportunities. For lack of a better term, I’ll call them “social” and “causal”.

Social windows of opportunity. Suddenly people are willing to listen to new advice, and consider different actions than they previously would. Hence there are attempts to get people to sign social pledges to self-quarantine and epidemiologists are signing open letters to tech giants. More nefariously, lawmakers are smuggling their pet policy proposal into things that look like corona response measures. When all is said and done, it seems plausible the overton window for biorisk policy will have shifted massively, and that might bring with it other surprising opportunities as well.

Causal windows of opportunity. There are also many new problems to be solved, where you can build a tool or other solution that actually changes the world in a mechanistic way, and which isn't primarily about convincing other people of things.

For example:

- **Problem:** stop touching your face. **Example solution:** build an app that uses your webcam and machine learning to warn when you’re about to touch your face, or reconfigure your existing hardware startup to produce bracelets that vibrate when you’re about to touch your face
- **Problem:** figure out who you might have infected after you realise you’re ill. **Example solution:** build an app that uses Google maps location tracker to figure out who you’ve been in contact with after you get ill.
- **Problem:** we'll run out of ventilators and other medical equipment. **Example solution:** coordinate to design and manufacture open-source emergency medical equipment
- **Problem:** Many tens of millions of people will switch to full-time remote work for the first time. What new problems will they encounter? **Example solution:** I don’t know! (Not a solution: buying stock in the wrong company because it had the same name as a video-calling company)

The Coronavirus Tech Handbook is an excellent resource summarising what people are building to fight the outbreak. Many projects are urgently looking for collaborators.

Even if you don't have tech skills, there are other ways of finding great opportunity in times of crisis.

Some of history’s most successful trades (1, 2) occurred during crises. There will likely be many financial opportunities during this crisis as well. (LessWronger user Wei Dai posted about how he successfully shorted the S&P500 a few weeks back, and saw 700% returns already *before* the crashes of the recent week.) (This is not financial advice and if you have no trading experience now might be a particularly bad time to start.)

We have also seen a massive failure of responsible institutions to respond appropriately and provide reliable information. This means there's a shortage and need for reliable research and advice. This situation requires thinking for ourselves.

Due to the existence of niche online communities doing this, I started seriously thinking and preparing when there were 2 cases in my home country. A week and a half later I went home to my family and helped them prepare, as the only mask-wearing person at an empty row in the back of an otherwise full plane. There were 20 confirmed cases. My mom initially yelled at me and felt embarrassed when none of her friends were taking action, and asked why the authorities didn't say much. I left 5 days later. The case count had grown exponentially to >400. A friend in med school told me to "wash my hands and don't panic". I left from an airport where staff wore neither gloves nor masks, and shorted the local stock market. As I'm writing this two days later the case count is almost 700.

The jury is still out, but sadly this seems to be a time where it's critically important to be able to take your beliefs seriously even when they go much further than official advice and mainstream behaviour . There will likely be many more opportunities over the coming months were good judgement and independent research can make an important difference. The LessWrong page of posts tagged coronavirus is one place to find and contribute to open questions.

Addendum on profiting from outbreaks

I strongly believe that that traders and entrepreneurs who try to gain profit during this crisis are *not immoral.* Rather, they are *incredibly important*. All this “flatten the curve” business is about smoothing out demand peaks over time. And financial markets (futures markets in particular) are one of the key technologies our society has for coordinating to efficiently allocate resources across time. For example, it might have been hugely beneficial if someone with foresight would have stockpiled massive amounts of medical ventilators months back and thereby caused suppliers to increase production (it seems plausible this might have been worth it *even if* they wouldn’t have sold those stockpiled ventilators at a massive markup). The actual stockpiling of masks that happened might also have been beneficial for this reason (but I am highly uncertain about this claim and wouldn't bet highly on it).

More generally, the coming year will present a massive wealth transfer, in various ways, to the prepared from the unprepared (or those *unable* to prepare, due to lack of money, knowledge, or some other key prerequisite). I don’t know what the implications of this will be. But once again, it might be worth a few hours of your time thinking about what opportunities it might generate.

tags: Coronavirus

COVID-19's Household Secondary Attack Rate Is Unknown

For group houses, one of the most important factors when deciding how to relate to COVID-19 is the question: If one of my housemates gets infected, how likely am I to also get infected? This is known as the household secondary attack rate, and it determines how much you need to worry about your housemates' level of precaution (as compared to your own), and how much within-house social distancing is necessary.

Household secondary attack rate is context dependent; it could mean one of several things, which I will illustrate with two scenarios:

- Scenario 1: A member of your household is returning from a trip to Wuhan, which you have heard is high risk. You react… however you react to that. You might find a way to be out of the house for awhile, or prepare an isolated room they can lock themself in. If they're your spouse, you might decide that isolation isn't worth it. Some time not long after they become symptomatic, they check into a hospital, where they remain until after their infectious period is over.
- Scenario 2: You live in a group house. You all have separate rooms, but share a kitchen and living room. One of your housemates is infected during a trip to the grocery store. Some time later they become symptomatic, and you react… however you react to that. You might or might not have somewhere to move to, or somewhere to move them to, but they can't check into a hospital because they're all full. You might or might not have been exposed to presymptomatic transmission. If neither of you has the ability to move, you'll be in the same house until after they've recovered.

Right now there are two studies which purport to measure the household secondary attack rate:

- Active Monitoring of Persons Exposed to Patients with Confirmed COVID-19 — United States, January–February 2020
- Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts

I'll refer to these as the CDC report and the Shenzhen study, respectively. These are the only studies I have been able to find which make any quantitative claim about COVID-19's secondary attack rate, and all other claims I've found have traced back to either of these two. The CDC study finds a household secondary attack rate of 10%; the Shenzhen study finds a household secondary attack rate of 15%.

When I started writing this post, I thought I'd be focusing on the difference between scenarios 1 and 2. Unfortunately, as I dug into the studies in detail, I found evidence of severe problems which make me think that these two studies provide almost no evidence whatsoever about COVID-19's household secondary attack rate, even in scenario 1.

CDC Report

On March 3, CDC published a report on the results of a contact-tracing program started on January 20. The report statistics on contacts of the first 10 patients with travel-related confirmed COVID-19 reported in the US; presumably, all of these travellers came from Hubei on or after January 20. They trace 445 contacts total, of which 54 developed concerning symptoms, became "persons under investigation", and were tested. It doesn't sound like anyone besides those 54 were tested.

The 445 contacts break down as follows:

- 222 were health care personnel
- 100 were "community members who were exposed to a patient in a health care setting"
- 104 were "community members who spent at least 10 minutes within 6 feet of a patient with confirmed disease"
- 19 were members of a patient's household

Out of the 54 people who were tested, two were positive; of those two, both were members of a patient's household. The CDC report does not provide any further information about those two positive cases, but they can be pretty easily matched to public news coverage. The first case was in Illinois and is described in detail in this Lancet paper; the second was in San Benito and is described in this announcement from a local public health agency. Both transmissions were to the spouses of travellers who returned from Wuhan.

The Lancet paper describes the first instance of person-to-person spread in detail, with a complete timeline of travel, symptoms, and tests. A woman who returned from Wuhan to Illinois on January 13 tested positive on January 20; her husband tested positive on January 24. In the Lancet paper, a few things are striking.

The first striking thing is that the husband was not tested until he developed a fever, at which point his wife had been hospitalized with a positive test for 4 days. So, testing was very much not proactive.

The second striking thing is that they ran many different tests in parallel, and appear to have been grappling with false negatives.

The third striking thing about the Lancet paper involved monitoring of 372 contacts, of which 44 became PUIs and were tested. Of these 44, one was her husband and was positive, and this was the only household contact. The CDC report had 445 contacts and 54 people tested. So after subtracting out the Lancet study and the San Benito case, we're left with 17 household members, 56 miscellaneous contacts, and... only 9 people tested. There is no information on how those 9 tests were allocated, except that they were negative.

Of the 19 household members in the CDC study, five stayed in the house with an infected person after they were diagnosed.

So to summarize: Two family members of index cases were tested and were positive. Nine more tests were allocated between 17 household members and 56 miscellaneous other contacts; none of those nine tests were positive. From this, the CDC report concludes that the household secondary attack rate is 2/19 (~10%).

I would say that this is laughable, but unfortunately it isn't funny. The practical upshot of all this is that the CDC report provides *almost no information whatsoever* about the household secondary attack rate.

Shenzhen Study

Shenzhen is a Chinese city in Guangdong province. The Shenzhen study looks at 391 cases and 1286 close contacts between Jan 14 and Feb 12, and estimates a household secondary attack rate of 15%.

298 (76%) of the index cases were travelers. Sick people were isolated an average of 2.57 days after symptom onset (if they were being monitored for symptoms because they had been labelled as at-risk by contact tracing) or 4.64 days after symptom onset (if they weren't). The study estimates R during the observation period to have been 0.4, implying successful containment.

I have a few concerns with this study.

My first concern is that the household secondary attack rate is an important factor in peoples' decision whether to stay put when a household member is sick, which might create political pressure to find a low number. If people tried to move out when their housemates got sick, they wouldn't lower their own risk much, but they would spread it wherever they moved to.

My second concern is that 9 days before the Shenzhen study was published as a preprint, the Report of the WHO-China Joint Mission on COVID-19 stated that

> Household transmission studies are currently underway, but preliminary studies ongoing in Guangdong estimate the secondary attack rate in households ranges from 3-10%.

I believe the Shenzhen study is the preliminary study referred to (the geographic location matches, and I can find no other studies in that geographic region which attempt to measure the rate). This seems like evidence of political pressure to report a low attack rate. (10% was the household secondary attack rate for SARS, and was used in some preliminary modeling of COVID-19 transmission dynamics before data was available.)

My third concern is that the paper contains three different household secondary attack rates: 15% in the Findings section, 14.9% in the Transmission Characteristics section, and 12.9% in Table 3. I cannot reconcile these numbers, and my attempts to cross-check numbers between different sections and tables within the paper all ended in mismatches and muddle.

My fourth concern is that in table 3, adding up the numbers within the category labels implies a substantial amount of data is missing, in ways that make no sense. 19% of contacts are missing a gender, 17% are missing an age, 10% are missing the annotation of whether they're a household-member or not, and 14% are missing the annotation for whether they interacted with the contact rarely, moderately often, or often. I am having a hard time imagining what sort of data collection process could do this, without being such a mess that serious errors are likely.

My fifth concern is that during the period studied, China was having significant issues with false negatives. Feb 12, the last day covered in the Shenzhen study, is the day before China changed its diagnostic criteria and reported a 34% one-day increase in cases. The study itself states that it changed its definition of a confirmed case changed on Feb 7, to require symptoms, "but sensitivity analyses show that truncating the data at this point does not qualitatively impact results". The paper reports results

for many variables, and does not state which variables had sensitivity analysis performed.

These issues add up to extremely low confidence in the paper. I might change my mind if the authors release data that someone else can analyze, or someone manages to make sense of the seeming inconsistencies within it. Either of these things would surprise me.

Conclusion

The unfortunate practical upshot is that there's no good quantitative estimate of the household secondary attack rate (or attack rates in general). My belief, based on priors and on the observed large values for R0, is that it's probably quite high, and I will be acting accordingly; but even a small amount of non-terrible evidence could shift this belief greatly.

Peter's COVID Consolidated Brief for 29 March

...See the latest in my 2 Apr Brief!

COVID-19 is a rapidly changing situation and it is hard to keep up to date. However, right now I am following COVID-19 full time and I read *widely* and I read *a lot*. I'm going to experiment with providing 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.

This brief assumes you are up to date on most things that have happened since around the 25th of March and will aim to keep you up to date on the latest over the past three days or so.

Note that I am not a domain expert and I urge some caution in over-relying on my selection and interpretation of these links.

This brief follows my research agenda. I am going to keep that up to date as well. I will also keep eyes on the LessWrong Coronavirus Agenda and submit to the LessWrong links database. Further discussion will be in the EA Coronavirus Facebook Group.

Doing Your Part! How You Can Stay Safe and Help the Fight!

Rob Besinger offers some advice for staying safe that I have not had the time or expertise to verify, but will reprint uncritically:

1. definitely, definitely self-quarantine
2. Avoid people
3. If you do need to be around people, wear something over your mouth and nose
4. Don't touch your face (duh)
5. Wash your hands (duh)
6. Eat well, sleep well, get exercise
7. Consider stockpiling a month of food (...if you still can at this point, IMO good to have a at least a week or so above your usual amount of food)
8. Consider printing out copies of your health records
9. Regularly disinfect commonly touched surfaces like door handles and light switches
10. Consider covering commonly touched surfaces with copper tape
11. Probably stop taking NSAIDs like ibuprofen
12. Probably even-better-advice-than-normal to consume 2000-6000 IU of Vitamin D daily, in the morning
13. Consider running an air purifier
14. Understand how COVID-19 usually presents and progresses, so you can make an informed guess about how likely you are to have it
15. Take zinc immediately if you start feeling any cold-, flu-, or COVID-19-like symptoms
16. Start monitoring your oxygen immediately if you develop a fever or experience significant chest tightness or difficulty breathing

~

80,000 Hours puts out a list of things people can do to help with COVID:

1. Research to understand the disease and to develop new treatments and a vaccine.
2. Determine the right policies, both for public health and the economic response.
3. Increase healthcare capacity, especially for testing, ventilators, personal protective equipment, and critical care.
4. Slow the spread through testing and isolating cases, as well as mass advocacy to promote social distancing and other key behaviours, buying us more time to do the above.
5. We also need to keep society functioning through the progression of the pandemic.

80,000 Hours thinks it is people should switch to working on COVID-related projects if they're roughly in the top 4% of people best suited to work on it - typically people who:

1. have highly relevant skills and/or useful connections - especially those who have medical training, can help with urgent hardware or software engineering efforts, or have knowledge of vaccines, public health, and government institutions
2. are not otherwise doing really important work
3. are highly motivated and informed on COVID
4. can switch into COVID-related work and switch back after without derailing one's long-term career

~

Here's my personal list of things you can do:

- Find your equilibrium, prepare yourself, and make sure you are okay first before trying to help. Make sure you have what you need to continue to be healthy and successful. Find a way to have a happy quarantine. Here's a bunch more ideas. There are also a million articles on this topic (these are my favorite four out of the 40+ I've seen).
- If you are already working in an essential industry, are a valiant healthcare worker, etc., definitely keep doing that.
- If you have the skills to contribute to vaccines, antivirals, etc... obviously do that.
- Rest a lot if you feel sick. Do what you need to do to self-care and look after your mental health.
- Contact your government decision-makers and let them know you support the shutdown and value public health. Now is an unusually important time to make your voices heard and convince others to do the same!
- If you are a publicist, social media influencer, or have celebrity contacts, consider getting them onboard with maintaining public support.
- If you have social media experience and/or online advertising experience, consider helping out with some social media campaigns.
- Research one of my research ideas for coronavirus and publish your findings.
- If you have expertise in data science or forecasting, besides trying to work on these research questions, it seems worth throwing significant time to various forecasting efforts like Metaculus's Pandemic Questions, the Good Judgment Open, and/or Kaggle. This could potentially scale to consume a significant amount of EA talent, though it may not be that neglected.
- If you have deep learning and image recognition experience, you could try to join https://www.covid19challenge.eu/
- Find a project on "Help with Covid", which also lets you filter by skill. Read through LessWrong and the "Effective Altruism Coronavirus Discussion" FB group. Look through this list of EA approaches. However, be wary of low neglectedness and widespread duplication of work.
- Spend time helping aggregate and organize information, maybe by making the Coronavirus Tech Handbook nicer and updating Wikipedia.
- Reach out to your local community, friends, family, and neighbors and make sure they feel supported and are doing okay in this trying time.
- With the pause in normal work now could be a great time for some personal and organizational reflection. Self-evaluation can pay big dividends in the long-term.

> Perhaps now you have time to re-evaluate long-term strategy, evaluate hiring practices, management style, employee morale, team culture, etc.

~

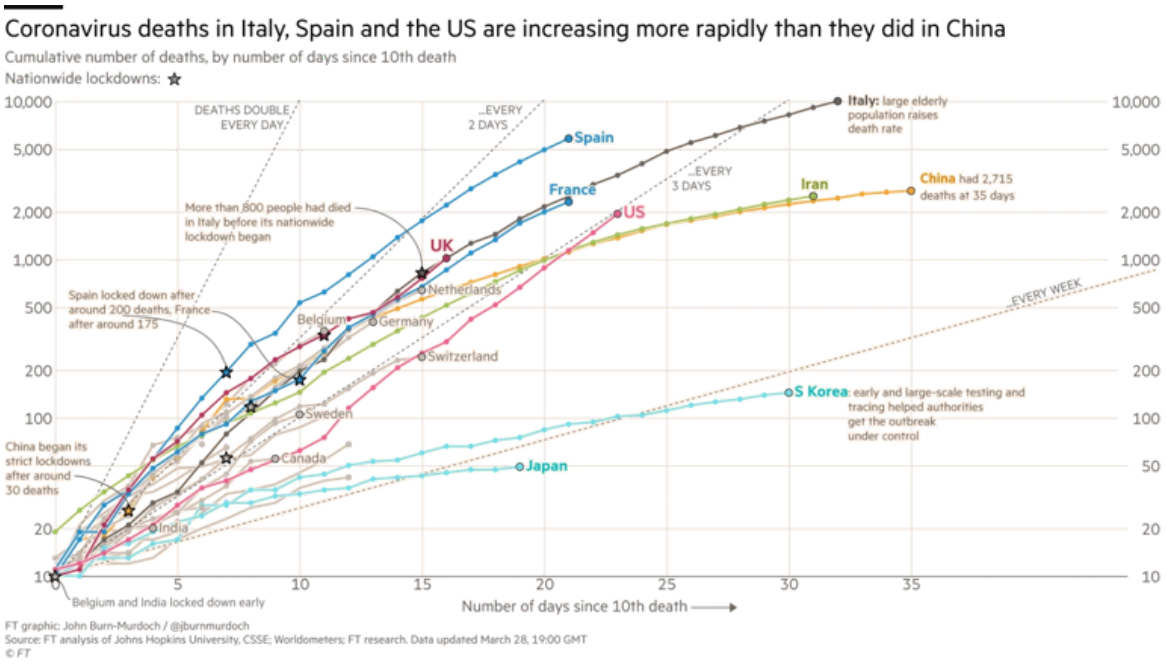
**[80,000 Hours also suggests donations for COVD-19 relief](): ** Center for Health Security at John Hopkins, Gates Foundation COVID-19 Funds, and the Center for Global Development.

~

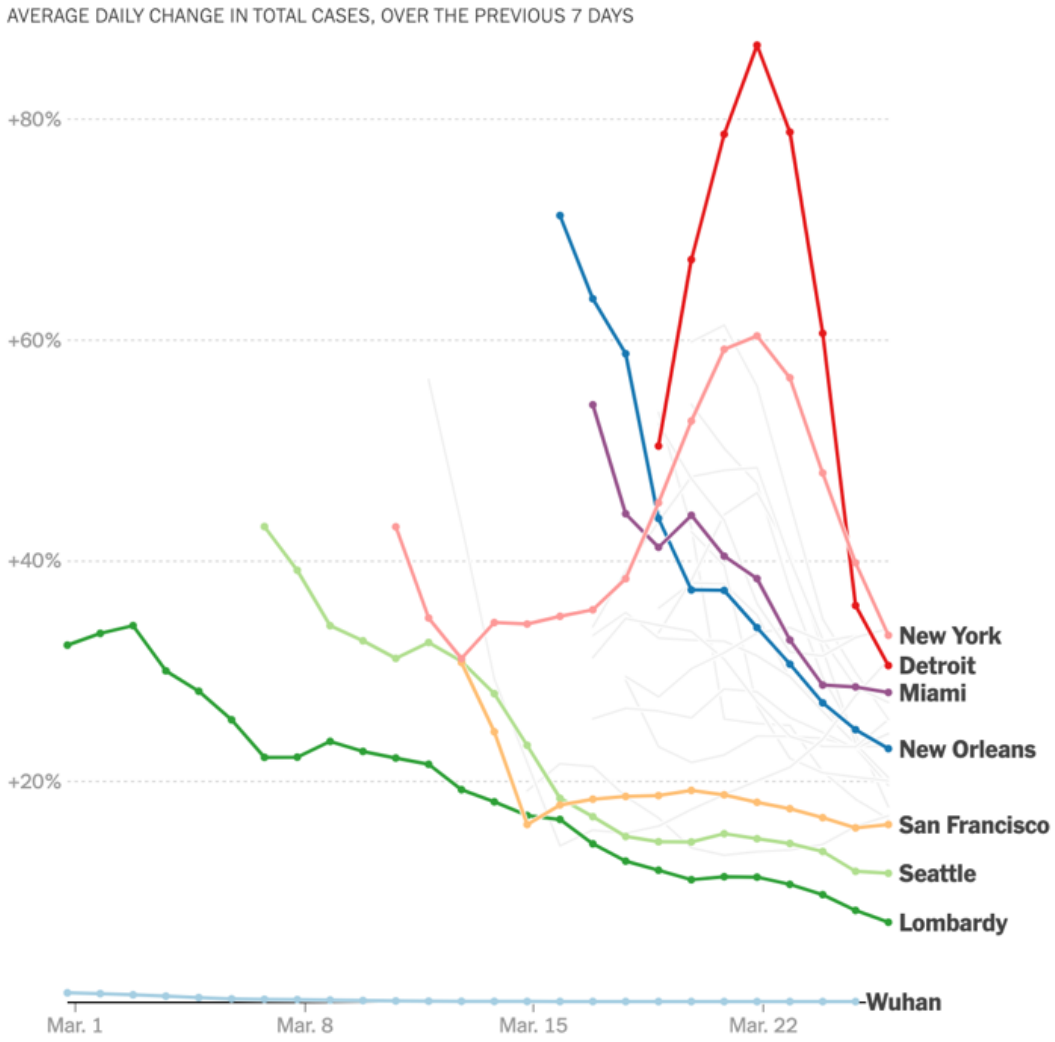
You can now help fight COVID using your laptop's spare cycles via Folding@Home.

A Glance at The Latest Situation

Things are still getting bad quickly. See FT's latest graph reprinted below. ...As Justin Wolfers puts it: "Project the U.S. line forward just 7 days, and we'll be at 10,000 deaths in total. Project it forward a week after that, and we'll be at 10,000 per day." (Hopefully we'll have flattened the curve a bit since then.)



The case numbers look bad, but at least the second derivative (growth in growth) shows some good news. The New York Times reports:



However, the situation still is dire: “The rate of increase in cases [in New York City] is far higher for the number of cases than it was in Wuhan or Lombardy, once they had reached similar numbers of cases. Other metropolitan areas, like Detroit and New Orleans, stand out as places where a coronavirus outbreak might escalate quickly without preventive measures. The Seattle and San Francisco areas, in contrast, seem to have made serious progress in flattening the curve.”

~

Here’s a very good visualization of percent of people infected by region in the world.

~

Note that case numbers are related to testing numbers, differences in who gets tested, and how cases, tests, and deaths are reported - and these all can differ country-to-country. This might be why Germany has such a low death-per-case rate. Italian data may also be underreported. Same with Spanish data. We should be prepared for the data to be a bit wonky and for comparisons to not be entirely apples-to-apples.

For example, I’ve been watching China-related COVID-19 reporting with a lot of anticipation… they seem to be doing very well at containing the virus and could become a model for the

rest of the world. However, there's also a lot of disinformation and misinformation around China. There's good reason to distrust their numbers. There's also good reason to distrust the distrust of their numbers. Some Chinese doctors who tested negative for coronavirus have later tested positive. ...One thing that is easy to verify: China re-closes all cinemas over fear of a second coronavirus outbreak.

Similarly, I share Scott Alexander's deep confusion over what the hell is going on in Japan, Iran, Nigeria, and Mexico:

Japan should be having a terrible time right now. They were one of the first countries to get coronavirus cases, around the same time as South Korea and Italy. And their response has been somewhere between terrible and nonexistent. A friend living in Japan says that "Japan has the worst coronavirus response in the world (the USA is second worst)", and gets backup from commenters, including a photo of still-packed rush hour trains. [...]

But actually their case number has barely budged over the past month. It was 200 a month ago. Now it's 1300. This is the most successful coronavirus containment by any major country's, much better than even South Korea's, and it was all done with zero effort.

The obvious conclusion is that Japan just isn't testing anyone. This turns out to be true – they were hoping that if they made themselves look virus-free, the world would still let them hold the Tokyo Olympics this summer.

But at this point, it should be beyond their ability to cover up. We should be getting the same horrifying stories of overflowing hospitals and convoys of coffins that we hear out of Italy. Japanese cities should be defying the national government's orders and going into total lockdowns. Since none of this is happening, it looks like Japan really is almost virus-free. The Japan Times is as confused about this as I am. [...]

Also, what about Iran? The reports sounded basically apocalyptic a few weeks ago. They stubbornly refused to institute any lockdowns or stop kissing their sacred shrines. Now they have fewer cases than Spain, Germany, or the US. A quick look at the data confirms that their doubling time is now 11 days, compared to six days in Italy and four in the US. Again, I have no explanation. [...]

The third world ...is in really deep trouble, isn't it?

The numbers say it isn't. Less developed countries are doing fine. Nigeria only has 65 cases. Ethiopia, 12 cases. Sudan only has three!

But they probably just aren't testing enough. San Diego has 337 diagnosed cases right now. The equally-sized Mexican city of Tijuana, so close by that San Diegans and Tijuanans play volleyball over the border fence, has 10. If we assume that the real numbers are more similar (can we assume this?), then Mexico is undercounting by a factor of 30 relative to the US, which is itself undercounting by a factor of 10 or so. This would suggest Mexico has the same number of cases as eg Britain, which doesn't seem so far off to me (Mexico has twice as many people). [...]

Nigeria and Mexico and so on make me confused in the same way as Japan – why aren't they already so bad that they can't hide it? If the very poorest countries in sub-Saharan Africa were suffering a full-scale coronavirus epidemic, would we definitely know? In Liberia, only 3% of people are aged above 65 (in the US, it's 16%). It only has one doctor per 100,000 people (in the US, it's one per 400) – what does "hospital overcrowding" even mean in a situation like that? I don't think a full-scale epidemic could stay completely hidden forever, but maybe it could be harder to notice we would naively expect.

Tyler Cowen also asks "Where does all the heterogeneity come from?": *"Can anyone shed light on why the death rate is not higher in Iceland? Is it that the death rate is about to burst a week from now? [...] Similarly, Sweden hasn't restricted public life very much and they do*

not seem to be falling apart? [...] It is possible that Cambodia, Thailand, and Vietnam still will be hit hard, but so far the signs do not indicate as such. Warm weather may play a positive role, though that remains speculative. The latest weather paper appears credible and indicates some modestly positive results. Of course weather won't explain the relative Icelandic and Swedish success, if indeed those are truly successes."

Similarly, why aren't a ton of Swedes on their way to being dead?

Maybe it's still just a matter of time? Vox argues that Mexico's coronavirus-skeptical president is setting up his country for a health crisis and Japan's coronavirus crisis may be just beginning.

...Also, beware Goodhart's Law in these metrics - originally US states were strongly incentivized to underreport, but now that relief is tied to caseload, they are strongly incentivized to overreport.

~

There is now data on ICU beds by US county.

...So Just How Bad Could This All Get?

"From Spain to Germany, Farmers Warn of Fresh Food Shortages" warns a Bloomberg article, mainly due to fewer workers being available to pick fruit. I'm pretty skeptical of COVID-related food shortages, but I still think it is important to monitor and be on the lookout. The NYTimes reports that there is still plenty of food in storage and supply chains are currently getting replenished just fine.

Gaze into the Crystal - The Latest Modeling and Forecasting

A Stanford team produces a stochastic model to forecast a "lightswitch approach" to lockdowns, where we alternatively lockdown and un-lockdown to continually keep the case load manageable enough to not overwhelm hospitals.

Vipul Naik estimates when we will get out of lockdown. He thinks the strict "shelter in place / go out for emergencies only" might get relaxed back to "most things closed"-level lockdown by mid-June and to the "most things open except large groups still banned"-level lockdown by summer 2021 and back to business as usual by summer 2022.

A survey of 18 epidemiologists say COVID-19 will cause approximately 195,000 deaths in the US and that a "second wave" of the virus is likely to occur between August and December.

Dr. Fauci predicts over 100,000 Covid-19 deaths in the United States.

A hospital-specific model tries to use more specific information about shortages. And another model, not tied to specific hospitals, but still modeling the impact of PPE shortages.

Now Let's Talk Policy Response

Harvard Business Review outlines some key lessons from Italy:

1. *“Recognize your cognitive biases. In its early stages, the Covid-19 crisis in Italy looked nothing like a crisis. [...] Threats such as pandemics that evolve in a nonlinear fashion (i.e., they start small but exponentially intensify) are especially tricky to confront because of the challenges of rapidly interpreting what is happening in real time.”*
2. *“The most effective time to take strong action is extremely early, when the threat appears to be small — or even before there are any cases. But if the intervention actually works, it will appear in retrospect as if the strong actions were an overreaction. This is a game many politicians don’t want to play.”*
3. *“The systematic inability to listen to experts highlights the trouble that leaders — and people in general — have figuring out how to act in dire, highly complex situations where there’s no easy solution.”*
4. *“Avoid partial solutions. A second lesson that can be drawn from the Italian experience is the importance of systematic approaches and the perils of partial solutions. The Italian government dealt with the Covid-19 pandemic by issuing a series of decrees that gradually increased restrictions within lockdown areas[... I]t backfired for two reasons. First, it was inconsistent with the rapid exponential spread of the virus. [...] Second, the selective approach might have inadvertently facilitated the spread of the virus.”*
5. *“Consider the decision to initially lock down some regions but not others. When the decree announcing the closing of northern Italy became public, it touched off a massive exodus to southern Italy, undoubtedly spreading the virus to regions where it had not been present.”*
6. *“An effective response to the virus needs to be orchestrated as a coherent system of actions taken simultaneously.”*
7. *“These rules also apply to the organization of the health care system itself. Wholesale reorganizations are needed within hospitals (for example, the creation of Covid-19 and non Covid-19 streams of care).”*
8. *“Finding the right implementation approach requires the ability to quickly learn from both successes and failures and the willingness to change actions accordingly. Certainly, there are valuable lessons to be learned from the approaches of China, South Korea, Taiwan, and Singapore, which were able to contain the contagion fairly early. But sometimes the best practices can be found just next door. Because the Italian health care system is highly decentralized, different regions tried different policy responses. The most notable example is the contrast between the approaches taken by Lombardy and Veneto, two neighboring regions with similar socioeconomic profiles.”*
9. *Good policies: “Extensive testing of symptomatic and asymptomatic cases early on. Proactive tracing of potential positives. If someone tested positive, everyone in that patient’s home as well as their neighbors were tested. If testing kits were unavailable, they were self-quarantined. A strong emphasis on home diagnosis and care. Whenever possible, samples were collected directly from a patient’s home and then processed in regional and local university labs. Specific efforts to monitor and protect health care and other essential workers.”*
10. *“It is especially important to understand what does not work. While successes easily surface thanks to leaders eager to publicize progress, problems often are hidden due to fear of retribution, or, when they do emerge, they are interpreted as individual — rather than systemic — failures.”*
11. *“Collecting and disseminating data is important. Italy seems to have suffered from two data-related problems. In the early onset of the pandemic, the problem was data paucity. More specifically, it has been suggested that the widespread and unnoticed diffusion of the virus in the early months of 2020 may have been facilitated by the lack of epidemiological capabilities and the inability to systematically record anomalous infection peaks in some hospitals.”*
12. *“More recently, the problem appears to be one of data precision. In particular, in spite of the remarkable effort that the Italian government has shown in regularly updating statistics relative to the pandemic on a publicly available website, some commentators have advanced the hypothesis that the striking discrepancy in mortality rates between Italy and other countries and within Italian regions may (at least in part) be driven by different testing approaches.”*

13. *“In an ideal scenario, data documenting the spread and effects of the virus should be as standardized as possible across regions and countries and follow the progression of the virus and its containment at both a macro (state) and micro (hospital) level.”*

~

Rhode Island essentially declares war on New York City:

Rhode Island police began stopping cars with New York plates Friday. On Saturday, the National Guard will help them conduct house-to-house searches to find people who traveled from New York and demand 14 days of self-quarantine.

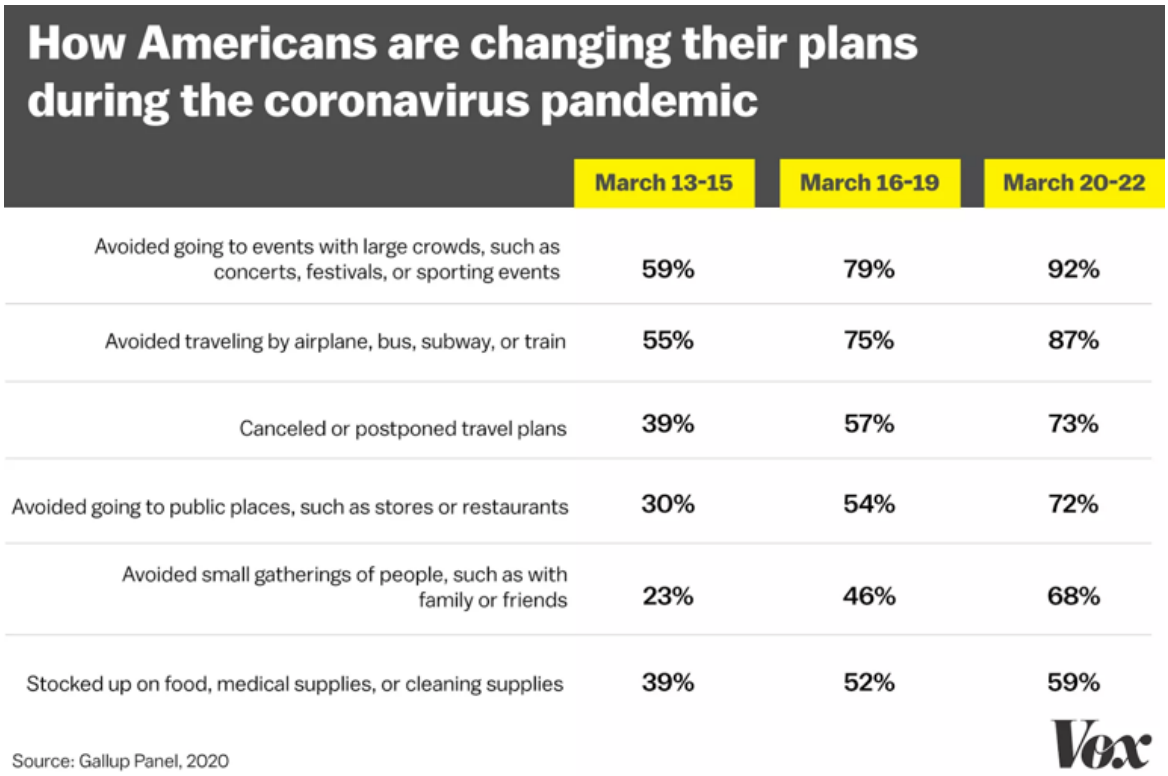
“Right now we have a pinpointed risk,” Governor Gina Raimondo said. “That risk is called New York City.”

~

WHO is very cagey / weird about Taiwan. WHO Director General, Bruce Aylward is asked about Taiwan’s membership in WHO… he responds by hanging up and then pretending to have not heard the question. Taiwan is not a member of WHO due to China’s insistence that Taiwan is a part of China.

A Bit About Life Under Quarantine

Americans have been changing their plans a lot:



~

Wild… “Our estimate is that about 12-13% [of churches in the US] were still open to in-person worship as of this week.”

The Psychologists Also Have Something to Say About This

For those studying the social psychology of the pandemic, there is now a research tracker.

~

The EA Australia Research Collaboration is running surveys to help policymakers with decision making about how to allocate resources to tackle COVID-19:

We doing something quite different from other ongoing surveys. Most of these are about country level comparisons and not about understanding behavioural drivers. For example, we can report things by location and demographic, but also why people are not doing the behaviours, e.g., 75% of males between 30 and 40 are social distancing but only 50% of males between 20-30. At only 55% adherence, the inner west region reports the lowest amount of social distancing. The main capability barriers are commuting and desire to see friends. 97% of those surveyed are always washing their hands, suggesting that this need no longer be a key communication target. [...]

Please consider joining us in collaboration. You can contribute by:

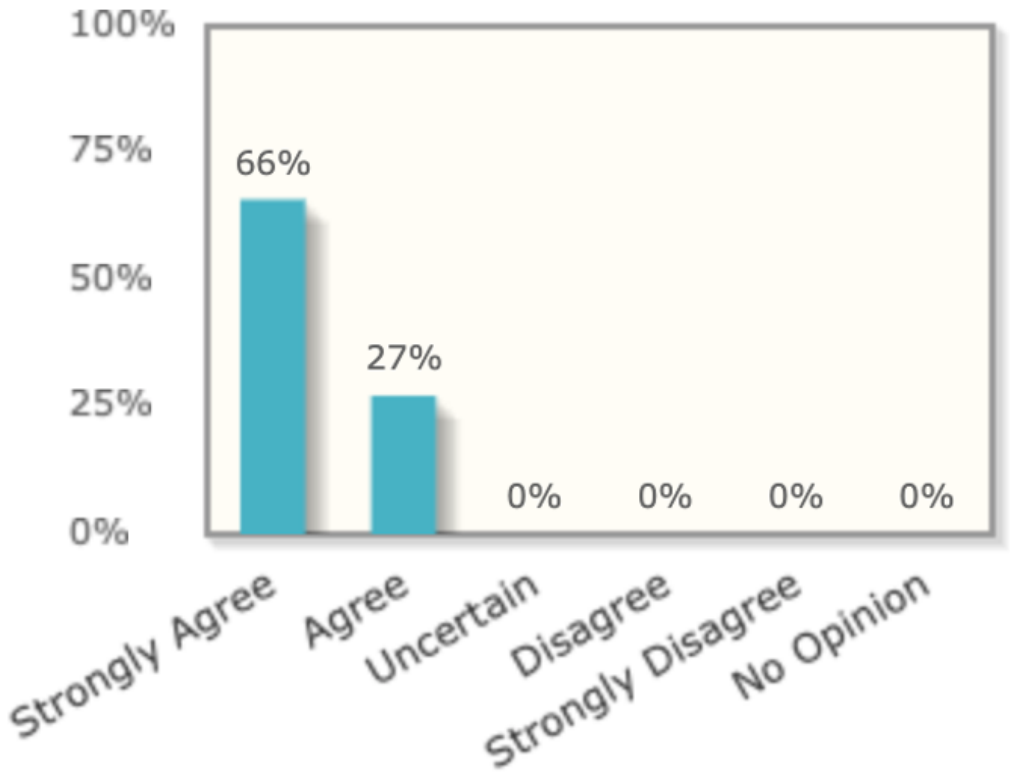
- *Helping to disseminate the survey via social networks or panel data.*
- *Reaching and helping policy makers in your country with the data we collect*
- *Helping to modify our report template to provide useful and interesting information to policy makers.*
- *Developing reports and doing analysis for policy makers*
- *Providing us with feedback based on your discussion data collection*
- *Helping with write up and dissemination when we seek to publish this work*

If you contribute to this project in any significant way then you will be recognised on all outputs and be an author on any subsequent paper. The bar for recognition will be relatively low (perhaps ~5 hours of work).

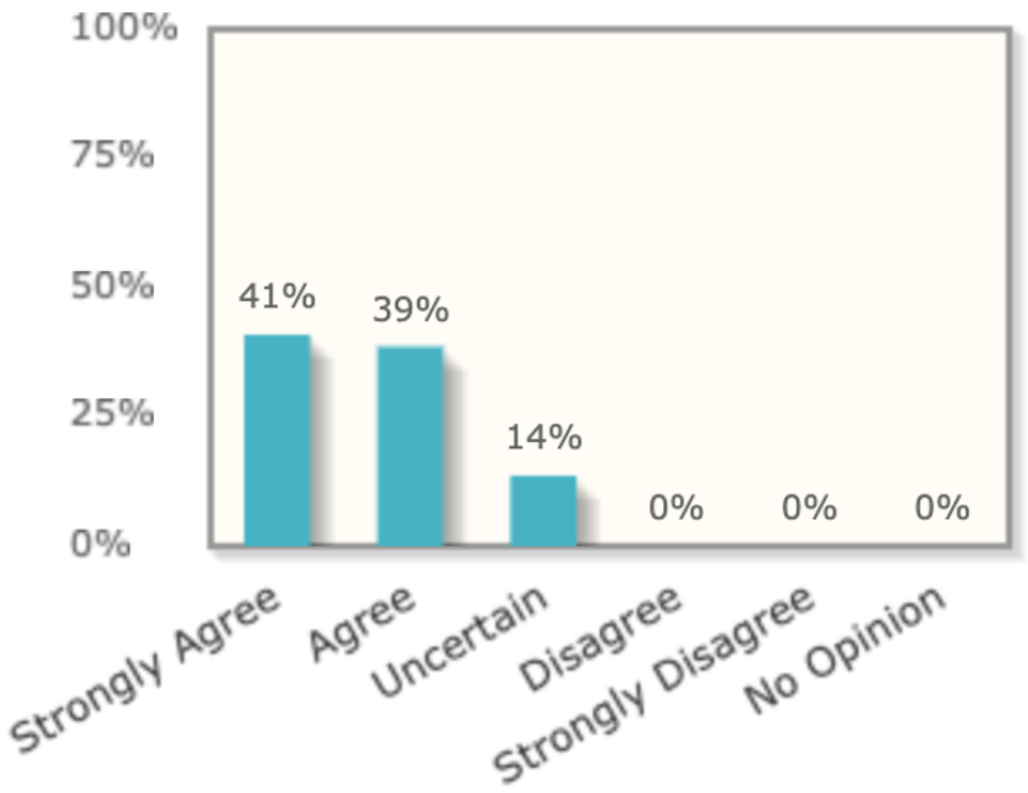
If You Still Own Envelopes, Check Their Backs - Here’s the Latest Cost-Benefit Analysis

The IGM Forum regularly polls economists about US economic policy. Economists rarely agree. But this time, economists are unanimous that the large contraction in economic activity is worth fighting coronavirus, we should not abandon the severe lockdowns, and the government should invest more in policy response.

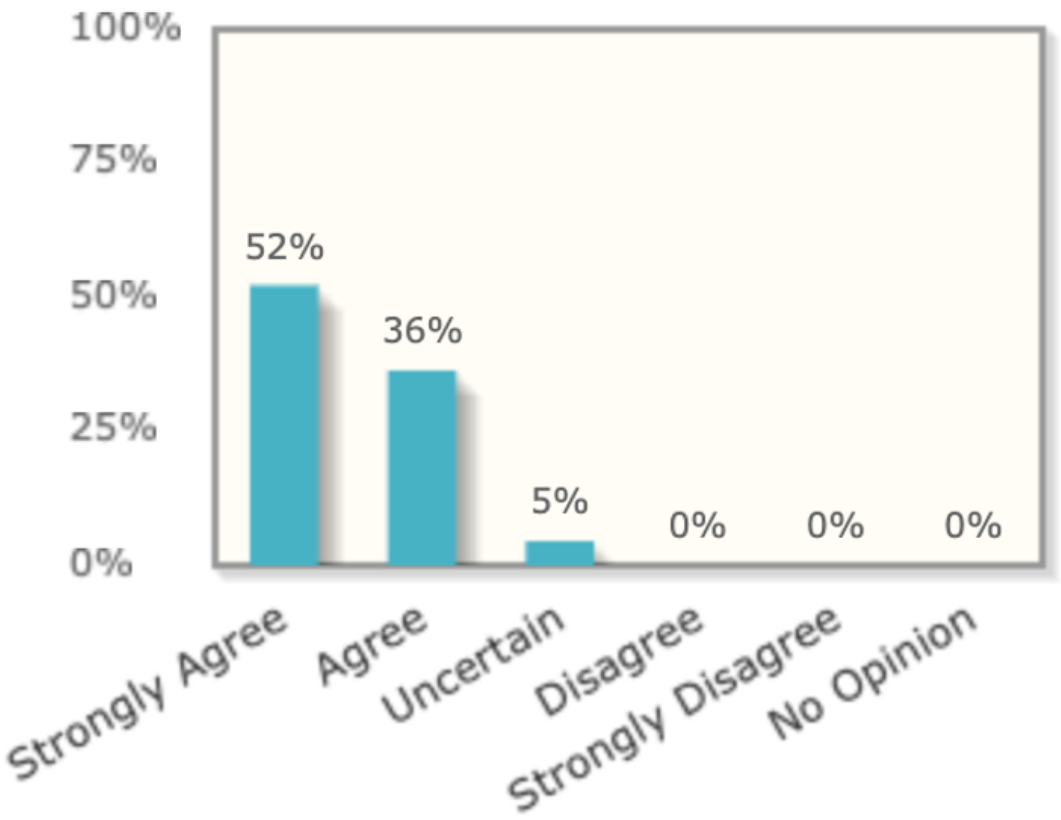
“A comprehensive policy response to the coronavirus will involve tolerating a very large contraction in economic activity until the spread of infections has dropped significantly.”



“Abandoning severe lockdowns at a time when the likelihood of a resurgence in infections remains high will lead to greater total economic damage than sustaining the lockdowns to eliminate the resurgence risk.”



“Optimally, the government would invest more than it is currently doing in expanding treatment capacity through steps such as building temporary hospitals, accelerating testing, making more masks and ventilators, and providing financial incentives for the production of a successful vaccine.”



A study using data from the 1918 Flu found that greater economic growth was connected with lockdowns as opposed to the opposite. Faster social distancing also saved a lot of lives during the 1918 Flu.

Now Just What are the Tech Overlords up to?

Apple launches a COVID-19 screening tool. According to TechCrunch: “The site is pretty simple, with basic information about best practices and safety tips alongside a basic screening tool which should give you a fairly solid idea on whether or not you need to be tested for COVID-19.”

~

The New York Times calls for “Big tech needs to rapidly build and scale a cloud-based national ventilator surveillance platform which will track individual hospital I.C.U. capacity and ventilator supply across the nation in real-time. Such a platform — which Silicon Valley could build and FEMA could utilize — would allow hospitals nationwide to report their I.C.U. bed status and their ventilator supply daily, in an unprecedented data-sharing initiative.” Probably already being done by 18 different groups now, but might still be worth exploring?

~

Zuckerberg donates $25M to the Gates Foundation to fight coronavirus. It’s an accelerator to find new possible antiviral drugs and total funding for the accelerator is now at $125M.

Not to be outdone, Google pledges to donate $800 million and 3 million face masks in an effort to combat the coronavirus... however, over 75% of this fund comes in the form of Google ad grants and cloud computing credits.

Lastly, Mayo Clinic and Amazon launched a collaboration to increase COVID-19 testing and vaccine development: "The private industry effort, spearhead by Mayo Clinic's John Halamka, M.D. and other industry leaders, plans to leverage the strengths of healthcare organizations, technology companies, non-profits, academia, and startups to provide a focused response to the coronavirus outbreak." It's not super clear what this means but I'm glad it's happening.

~

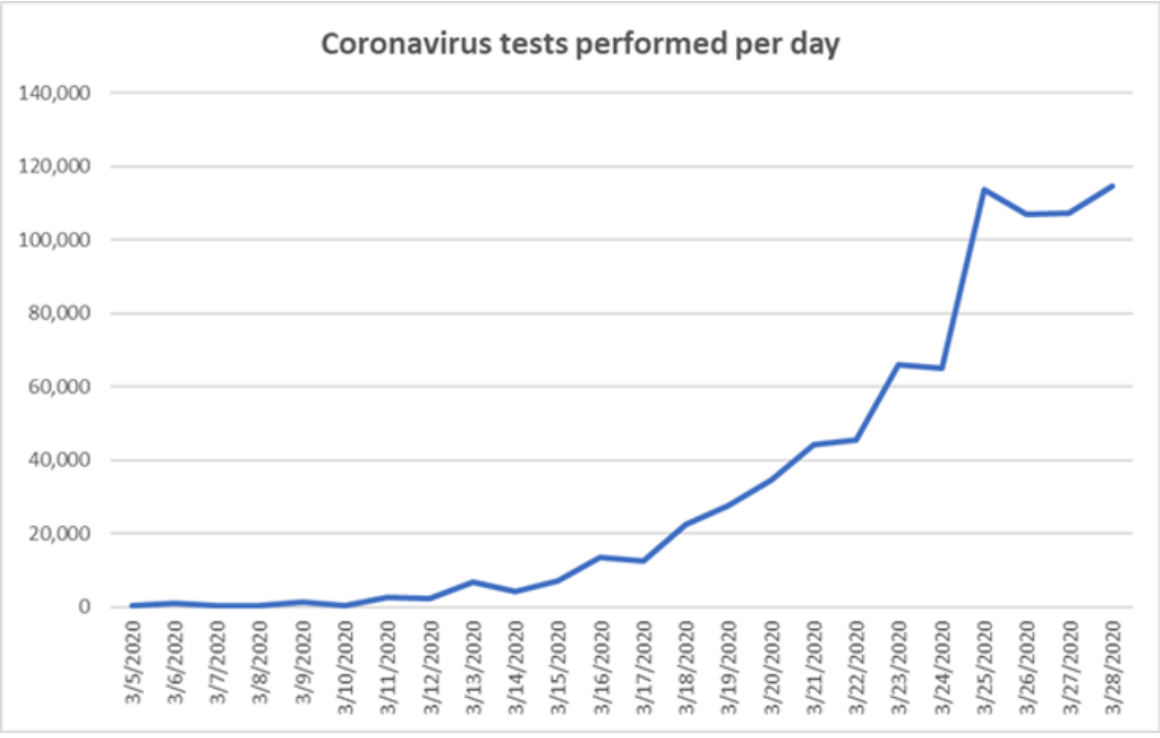
Maybe we shouldn't be using Zoom? Consumer Reports reports "Zoom Calls Aren't as Private as You May Think". Engadget reports "Zoom happens to be a privacy nightmare with a terrible security track record" and that "Zoom collects your physical address, phone number, your job title, credit and debit card information, your Facebook account, your IP address, your OS and device details, and more and traffics that data with whomever it's doing business with".

~

Predictably, Edward Snowden warns about the dramatic increase of tech surveillance.

And How Do We Get Out of this Mess? Vaccines, Treatments, Testing, Tracing, etc.

US testing is no longer increasing exponentially:



...But I'm told this is a game changer: Abbot launches molecular point-of-care test to detect novel coronavirus in as little as five minutes. They've already received FDA emergency use

authorization.

~

Dr. Fauci outlines ambitious plan to scale up COVID-19 vaccine. Looks like the proposal is to start ramping up production of a vaccine while the candidate is still in Phase II clinical trials. This risks spending a ton of money producing a vaccine that ultimately might not get approved by the FDA. Dr. Fauci has suggested hundreds of millions in incentives to make this work.

~

The Washington Post reports that "hVIVO, a clinical research group in London, has attracted more than 20,000 volunteers willing to be infected with tamer relatives of the virus that causes Covid-19 in exchange for a fee of £3,500 ($4,480)."

~

To stop COVID-19, test everyone, repeatedly: "We propose an additional intervention that would contribute to the control of the COVID-19 pandemic and facilitate reopening of society, based on: (1) testing every individual (2) repeatedly, and (3) self-quarantine of infected individuals. By identification and isolation of the majority of infectious individuals, including the estimated 86% who are asymptomatic or undocumented, the reproduction number R0 of SARS-CoV-2 would be reduced well below 1.0, and the epidemic would collapse. This testing regime would be additive to other interventions, and allow individuals who have respiratory symptoms due to other causes to return to work, but would have to be maintained until a vaccine becomes available. Unlike sampling-based tests, population-scale testing does not need to be very accurate: false negative rates up to 15% could be tolerated if 80% comply with testing, and false positives can be almost arbitrarily high when a high fraction of the population is already effectively quarantined."

And Now a Word From the Lamestream Media

Kelsey Piper is the best. Vox, and other media outlets have not apologized for continually downplaying the coronavirus and calling out people as fearmongering. For example, on 13 February, Vox made fun of the prescient tech industry, deriding them with "Although public officials in the area say the virus is contained for now, some in the tech industry fear the virus will spread out of control." Kelsey Piper, Vox Future Perfect author, is at least willing to admit she made a mistake. No one else seems to be… yet Kelsey is getting all the hate for it. :(

(I also made a mistake by privately stockpiling on 8 February, but not telling anyone of my fears out of a meta-fear of looking like some dorky prepper. I also apologize for this.)

The Non-Profit Impacts

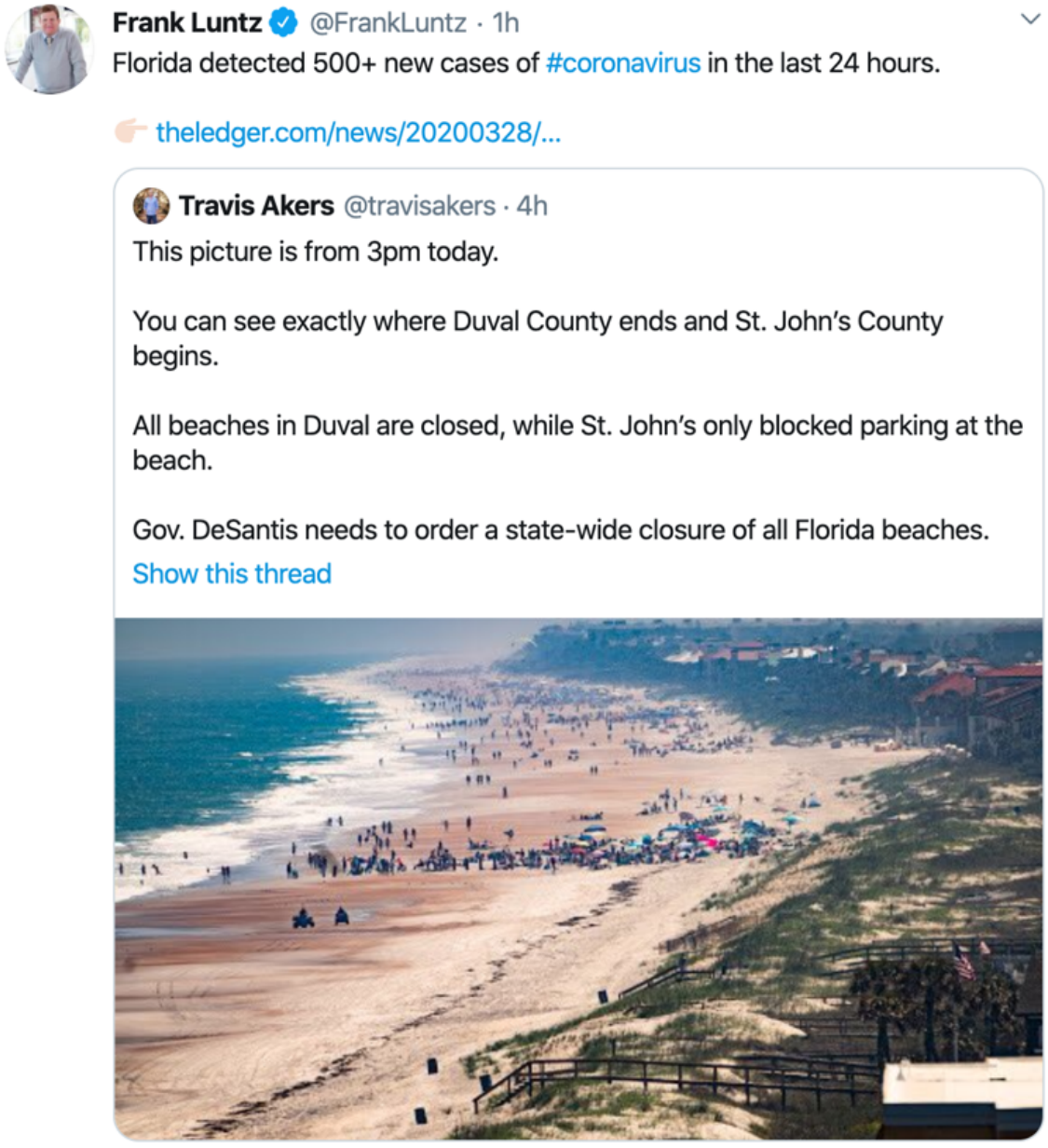
American non-profits that (a) existed before 1 March 2020, (b) have fewer than 500 employees, and (c) keep staff on payroll can get what might amount to free money from the government. This deserves urgent further investigation.

Don't Forget About the Nonhumans!

Amid continued panig egg buying, US grocers boosted egg orders by as much as six times normal and USDA relaxes rules to allow older eggs to make grade.

Vietnam's prime minister, Nguyen Xuan Phuc, has asked the country's agriculture ministry to draft a directive to stop illegal trading and consumption of wildlife over fears it spreads disease.

Your Regular Dose of WTF



Fun (Online) Distractions, Because We All Still Need to Enjoy Life

The Rotterdam Philharmonic Orchestra delivers Beethoven's 'Ode to Joy' with 19 musicians playing their parts from their homes.

Cute dog undertakes a sisyphean task of constantly fetching the ball, only for it to roll back down the hill.

Dad learns how to Tik-Tok dance.

~

Thanks to Elizabeth Van Nostrand, Robert Krzyzanowski, and countless others for assistance in aggregating and editing this list.

Near-term planning for secondary impacts of coronavirus lockdowns [assuming things don't blow up]

In this document, I attempt to discuss the impact of the coronavirus lockdown and how to prepare for it. This is not focused on the direct impact of coronavirus, but rather on the secondary impact of precautions that people are taking, including the lockdowns and the new normal of staying home and working from home.

The document is written in an imperative tone, focused on what to do. However, please don't read into this tone the idea that I am confident of these suggestions and authoritatively pushing them. They are just ideas!

Many of these ideas are self-justifying, but I have not tried to justify their *relative importance* to other ideas that I have omitted. Subject to time constraints, I'll be happy to answer specific questions challenging the ideas, or comparing them to other ideas I didn't list. If you have a question of that sort, there's a good chance I'll just agree that the idea I didn't list was more important.

My initial draft of this post included some discussion of potential future timelines, but I decided to omit that in order to make the post focus on ideas for dealing with the situation. I may separately write about possible futures.

General ideas

- Expect a **three-month timeline** for the lockdown (i.e., lockdown continuing till the end of June), with **the possibility of a six-month, twelve-month, or even eighteen-month lockdown**. Even if a strict lockdown lasts much less, health advice may still recommend that you shelter in place for additional time.
- **Brace for impact! Prepare psychologically.** Plan for three months of lockdown; it could be shorter, but it could also be much longer. If you overprepare a little bit because the lockdown lasts just one month, you'll need to write off the effort, but that'll be less painful than forming expectations that this will end in a few weeks and then constantly being disappointed at how much it's dragging out.
    - It's ideal if all your concrete actions are "no-regret" -- so that if the lockdown lasts longer or shorter than you expect, the action still gives you some lasting benefit. But this may not always be possible.
- **Keep a buffer (material goods, liquid savings, other reserves), but don't engage in panic actions to build buffers.**

Impact on day-to-day life experience

Since we're talking of a three-month timeline for a lockdown (and possibly much longer), you have to think of a sustainable way to manage your life. It's not a day or two that you can somehow brute-force. You need a sustainable approach, and a reasonable balance. Here are some ideas:

- **Strike the right balance in terms of going out**: It's reasonably safe to go out if you stay far from people and don't touch stuff. So, make sure to get a reasonable amount of exercise and fresh air. Don't stay cooped up in your home for days. Obviously, exceptions apply for people who are sick or may have been exposed, or if there is legal enforcement of a stricter stay-at-home order. Most existing stay-at-home orders, even the strictest lockdowns, allow people (who are not old or at risk of already being exposed) to go out alone for exercise. In some regions, you may need to carry documentation stating that you are going out for exercise.
- **Get the right cadence in terms of purchasing food and necessities**: Keep in mind that grocery stores and convenience stores are limiting the amount you can buy at a given time. So make sure to make regular (though not very frequent) trips to stay stocked up on necessities, thinking as far ahead as feasible. Sanitize well before and after such trips. If you can make these trips at a time when the stores and streets are less crowded, please do that. Again, please make sure to comply with any stay-at-home orders, including taking appropriate documentation in regions where documentation is necessary.
- **Take great care of your health even outside of coronavirus-related matters**: It'll be a terrible idea if you need to go to a doctor at this time. So, make sure to take good care of your health, particularly dental health and any other aspects of health that tend to be problematic for you. Make sure to eat healthy and take your normal supplements that have a good track record for you.
- **Use scarce goods sparingly.** Here are some illustrative examples; the specifics may not make sense in light of your health concerns, beliefs about environmental impact, and aesthetics, but they give a general idea:
    - If you use paper goods at home (such as paper towels), consider using cloth-based substitutes, as long as each person can use their own personal cloth: Paper goods are likely to stay in shortage, so you want to use yours sparingly if feasible. For instance, use cloth towels instead of paper towels for wiping your hands, as long as multiple people aren't sharing the same towel.
    - Give preference to handwashing with soap over using hand sanitizer. Handwashing is anyway more effective, and soap seems to be less in demand than hand sanitizer (likely because of the huge demand for hand sanitizer created by businesses offering hand sanitizers at workstations). The relative availability of hand sanitizer may, however, improve as lockdown continues and business use of hand sanitizers slows down.
    - You will need to figue out what other goods are scarce where you live and adjust consumption habits accordingly. Please weigh other considerations like health, the environment, personal aesthetics, etc.

Impact on social life and interaction

Staying at home, and refraining from participating in social activities, is something that could get harder and harder as the time period gets longer. Some social activities are easy to forgo for a week, but harder to forgo for three months. I expect that this could lead to people feeling depression, loneliness, and mental health issues, with the risks increasing the longer this continues.

A silver lining is that the reduced level of necessary activity, in particular commuting, may help people recover from months or even years of hectic commutes.

The balance of these factors will vary from person to person, but I expect that for most people, the social life impact will be a net negative.

What can we do? Here are a few thoughts:

- **Exploit the positives**: You can't do some social activities that you normally do, but perhaps the shelter-in-place and the saved commute time gives you more flexibility and time to do some other things you've always wanted to do but never had the bandwidth for. For instance, maybe you can spend evenings working on a long-deferred personal project, or learn a new skill, instead of being stuck in the commute or socially pressured to attend events you don't really enjoy. Or maybe you could spend more time with your family (in the literal sense, not as a euphemism). Or spend more time online with people who don't live near you anyway.
- **Get along better with the people you live with**: You can't escape your home to go hang out with others, so you probably need to make peace with whoever is next to you, whether that's your family, your pets, or random roommates. Appreciate more the time you spend with them (with appropriate social distance!) or at any rate, don't get into fights, considering that you can't walk out of the house that easily.
- **Switch social activities online as much as possible, and plan a little bit for them**: If you got a lot of your social energy from serendipitious in-person interaction, this will be in short supply. Instead, you may have to plan the equivalent online things a bit more. In many cases, more conscious planning and coordination may be needed. So make sure to plan and push for the online equivalents where feasible. This is important because it's likely the lockdown will last long enough that completely forgoing some kinds of social interactions will be too costly. This may be particularly important for group activities that play an important role providing emotional support to their members.

Impact on work life and job security

This mostly applies to jobs where you were previously going into an office and you're now working from home. It doesn't apply to cases where you have been fired or furloughed, or where you were always working from home, or where you still need to go in for the job.

- **Make home station adjustments**: Make adjustments to your home environment to make it more feasible to efficiently work from home. A lot of people find it helpful to have a physical separation of their work station and the rest of their home; if that's feasible and desirable for you, consider doing it.
- **Negotiate a new work-life balance**: The previous work-life balance you worked out probably needs to be adjusted in light of the new situation. For instance, perhaps you can start work earlier or end later, but need more breaks within the day to cook food or deal with your kid who's also staying at home. Think through the right balance that works for you and your employer. This may take a few days to figure out.
- **Make sure lines of communication and recognition of your work have adjusted to the work-from-home reality**: Even if you're doing just as much work as you were doing in the past, your boss or colleagues may not realize that. Make sure that the "optics" angle is well-covered. The specifics will vary from job to job.

- **Keep in mind that getting a new job may be harder, so try to secure yourself in your existing job more**: At least until the lockdown is in place, and possibly even for a few more months, switching jobs will be hard. So, try as much as possible to get along with your existing job. This is true even if your industry isn't directly affected in a severe way; for people in industries that are heavily affected, the situation is much trickier. NOTE: If you are in a heavily affected industry and have an opportuunity to jump to a less affected one, consider taking it. But secure the new opportunity first before jumping ship.

Financial impact

- **Give more importance to building a liquid savings buffer**: In my simple financial advice doc, I recommend building liquid savings for about one year. In the current climate, I recommend increasing the target to two years, and to three if your job or industry is particularly negatively affected. In particular, I suggest:
    - Stop contributing to retirement accounts until you have hit the increased liquid savings threshold. With that said, if you do have more liquid savings than the increased threshold, increasing contributions to retirement accounts may be a great idea.
    - Hold off on repaying very-low-interest loans such as student loans until you have hit the increased liquid savings threshold (though it's best to do calculations for each loan to trade off the interest rate against the loss of liquidity).
    - If you are well below your liquid savings threshold, investigate how much of your money is in retirement accounts and other funds and make contingency plans to liquidate some of it to shore up your liquid savings. Liquidating retirement accounts may come with a penalty, which is why it's better to store any new money you're getting in more liquid forms. So make the plan (to liquidate) and be prepared to execute it if you find your net savings rate turning negative (due to unexpected income loss or expense increases).
- **Beyond the goal of maintaining liquidity to weather you through 2 to 3 years, don't engage in panic buying or selling of assets**: There are arguments in favor of buying and holding in the stock market given the lower prices. Evaluate them based on your normal criteria.
- **Continue your regular philanthropy and consumer spending**: On a similar note, if you engage in regular philanthropy or in consumer spending that gives you happiness, continue with it as long as (a) it still makes sense in the context of the lockdown, and (b) you either already reached or are on the way to your liquid savings threshold. In other words, after securing your health and wealth, continue your life as close to normal as possible.
    - If, for building your savings, you have a choice between cutting down on consumer spending that gives you happiness, versus cutting down on putting money into retirement accounts, choose the latter. In other words, spend normally, and put less money into your retirement account. Saving for retirement can wait for a few years; if you don't survive those few years, there is no point saving for retirement.

My reasons for writing this document

- Especially in the rationality community, I've seen a lot of advice on protecting oneself against coronavirus, but not as much on dealing with the massive social experiment that's being unleashed in the effort to do so. I expect the latter to increase in importance over time, both if containment efforts are successful, and if they aren't.
- Much discussion among the general public about the lockdown seems to be along the lines of "hey, we're in uncharted territory, this is scary" and isn't reiterating enough what this might mean over time periods longer than a week or two. I think what's important is to start bracing for an extended period of lockdown, to minimize the wave of secondary effects as people get frustrated with the lockdown. Preparing people in this way could help make sustained containment and social distancing efforts more palatable, and mitigate some of the adverse social and economic effects. My post is probably a very small contribution, but I hope it pushes positively in the general direction.

Simulacra and Subjectivity

This is a linkpost for http://benjaminrosshoffman.com/simulacra-subjectivity/

In Excerpts from a larger discussion about simulacra, following Baudrillard, Jessica Taylor and I laid out a model of simulacrum levels with something of a fall-from grace feel to the story:

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 construed and evaluated as a bid to move in that direction, and evaluated as such. 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.

There is some merit to this linear treatment, but it obscures an important structural feature: the resemblance of levels 1 and 3, and 2 and 4.

Another way to think about it, is that in levels 1 and 3, speech patterns are authentically part of our subjectivity. Just as babies are confused if you show them something that violates their object permanence assumptions, and a good rationalist is more confused by falsehood than by truth, people operating at simulacrum level 3 are confused and disoriented if a load-bearing *social identity or relationship* is invalidated.

Likewise, levels 2 and 4 are similar in nature - they consist of nothing more than taking levels 1 and 3 respectively as object (i.e. something outside oneself to be

manipulated) rather than as subject (part of one's own native machinery for understanding and navigating one's world). We might name the levels:

Simulacrum Level 1: Objectivity as Subject (objectivism, or epistemic consciousness)

Simulacrum Level 2: Objectivity as Object (lying)

Simulacrum Level 3: Relating as Subject (power relation, or ritual magic)

Simulacrum Level 4: Relating as Object (chaos magic, hedge magic, postmodernity) [1]

I'm not attached to these names and suspect we need better ones. But in any case this framework should make it clear that there are some domains where what we do with our communicative behavior is naturally "level 3" and not a degraded form of level 1, while in other domains level 3 behavior has to be a degenerate form of level 1.[2]

Much body language, for instance, doesn't have a plausibly objective interpretation, but is purely relational, even if evolutionary psychology can point to objective qualities we're sometimes thereby trying to signal. Sometimes we're just trying to stay in rhythm with each other, or project good vibes.

---

[1] Some chaos magicians have attempted to use the language of power relation (gods, rituals, etc) to reconstruct the rational relation between map and territory, e.g. Alan Moore's *Promethea*. The postmodern rationalist project, by contrast, involves constructing a model of relational and postrelational perspectives through rational epistemic means.

[2] A prepublication comment by Zack M. Davis that seemed pertinent enough to include:

> *Maps that reflect the territory are level 1. Coordination games are "pure" level 3 (there's no "right answer"; we just want to pick strategies that fit together). When there are multiple maps that fit different aspects of the territory (political map vs. geographic map vs. globe, or different definitions of the same word), but we want to all use the SAME map in order to work together, then we have a coordination game on which map to use. To those who don't believe in non-social reality, attempts to improve maps (Level 1) just look like lobbying for a different coordination equilibrium (Level 4): "God doesn't exist" isn't a nonexistence claim about deities; it's a bid to undermine the monotheism coalition and give their stuff to the atheism coalition.*
> *Book Review: Cailin O'Connor's The Origins of Unfairness: Social Categories and Cultural Evolution*
> *Schelling Categories, and Simple Membership Tests*

Does the 14-month vaccine safety test make sense for COVID-19?

I was first wondering why, if we keep hearing about teams rapidly generating vaccines for COVID-19, the common wisdom is that it will take 18 months to start vaccinating at a large scale.

Turns out that the scaling up takes a few months, but the real blocker is the Phase 1 trial, which requires monitoring patient health for 14 months after vaccination.

Doesn't it seem like the cost-benefit analysis changes a bit if we're in the midst of a pandemic? Wouldn't it be worth cutting it down to e.g. 3 months before at least vaccinating the highest-risk populations? Is anyone official even thinking about this?

Coronavirus is Here

Even in the best case scenarios, things are going to get a lot worse from here before they get better.

Please, please, *please,* do not rely on me here to tell you what is going on or what to do. Even more than that, please, please, *please,* do not take this as saying that you shouldn't do *a lot more* than the things I'm saying here.

This is me doing what various stupid reasons prevented me from doing earlier, deciding that saying something is a lot better than saying nothing, and hoping it will do some good.

This is not a model of what is happening, or an attempt to justify what you should do, because attempting to do that would cause me to continue to say nothing, and that seems worse.

If you already are taking the situation seriously and making serious preparations, this probably won't tell you anything new. I am totally fine with that.

This is me seeing something and saying something.

If you need some sort of permission to yourself to acknowledge that this is happening, and that you need to take action now to prepare, this is one more instance of that. You have it.

It will be important that you retain the ability, when things get bad, to keep your head on straight. Prepare now, including mentally, as best you can, for people you know and care about getting sick and dying, because this might well happen to you regardless of what actions everyone involved takes.

Here is an article describing the symptoms, if you are not yet familiar with that. It's primarily dry cough and pneumonia.

From the statistics I have seen, with super wide error bars, overall risk of death if you do get infected could be up to about 2%. But that is only an average. It varies wildly based on age and prior health, and presumably access to health care. I've also seen reasonable claims that initial degree of exposure matters here.

Risk of death chart by age:

Age Death Rate
80+ years old 14.8%
70-79 years old 8.0%
60-69 years old 3.6%
50-59 years old 1.3%
40-49 years old 0.4%
30-39 years old 0.2%
20-29 years old 0.2%
10-19 years old 0.2%
0-9 years old no fatalities

Co-morbidity is very high for conditions such as heart attacks, cancer and diabetes. If you are elderly and/or have conditions such as cancer, a prior heart attack or diabetes, you are at much greater risk, and should take all precautions more aggressively.

If you have not yet prepared for quarantine in place for at least several weeks, and ideally longer than that, you need to do at least that much now. Get a three month supply of any medications you need. Make sure you are stocked with the necessary food and water, soap and toilet paper and so on. Here is a quick LessWrong post on quarantine preperations.

Decide **now** what would make you impose this quarantine on yourself rather than waiting for an authority figure to do it for you, and what would cause you to impose other lesser restrictions. Decide what would cause you to stop going to restaurants, stop going to work, stop going outside for any non-emergency reason, and so on.

Figure out what news sources, including people who you trust, you are going to be willing to trust for information going forward.

If you are not prepared, including mentally, for approximately everyone around you to freak the hell out about this in various ways, you need to do that now.

If you have any plans to fly or attend conferences or sporting events or other major gatherings of people, at least beyond the next few days, cancel them. Period. Where and when you draw the line is up to you, but draw a line and adjust it quickly as news comes in.

If you can do your job from home rather than a crowded office, do your job from home. If you have to commute, don't do it on crowded subways or buses.

Do not shake hands or otherwise make other unnecessary physical contact with anyone.

If you are not yet doing your best to get good hand washing practices or avoid touching your face and other similar basic hygienic actions to avoid infection, get on that. Here is a post on how to properly wash your hands.

Remember that the virus can survive on surfaces for days, that it takes people several days of incubation to show symptoms, and that it takes two weeks to play itself out.

If there are people you care about who don't know that this is happening or aren't taking it seriously, fix that.

Here is a practical advice post that seems good that has detail. Biggest impact actions from its perspective are avoiding people in general, avoiding people who seem potentially sick in particular (at least 2 meters away at all times), frequent proper hand washing and keeping a reasonable stock of supplies.

Here is the LessWrong advice thread where people offer advice and justifications for potentially 'weird looking' actions. One piece of advice is to stock electrolyte drinks in case someone is unable to eat due to illness. They recommend copper tape be placed on commonly touched surfaces and the backs of phones, you can buy some here. Also they endorse a pulse oximeter. One interesting suggestion is Vitamin D supplements since you might be unable to go outside.

Here is a basic ‘the CDC expects this is happen’ thing from a few days ago in case anyone needs to share something that basic with someone, otherwise ignore.

If you live in a major city, and leaving that city is practical, it is at least reasonable to do so, especially if you are at greater risk.

Do not take this as anything like a complete list of things to do or consider doing. Seriously consider doing more things to prepare, and avoiding more things more aggressively, based on what you believe to be helpful.

Again, DON’T PANIC. Let’s all stay safe as best we can.

High Variance Productivity Advice

Despite being literally inapplicable for half the population, my post on birth control and productivity was one of my most popular Facebook posts. In that spirit, here's a roundup of my favorite weird productivity tips that aren't for everyone.

I chose hacks that several people rated as extremely high ROI…and that other people rated from "meh" to "absolutely awful." Basically, this is my list of tips that I *don't* think everyone should try. But since most blogs try to only give generally good advice, maybe you'll find something new that is fantastic for you.

Attempt them with a spirit of experimentation and caution. A few wins can be worth a lot of useless experiments, but be careful about large negative impacts; try 10 experiments that each have a 10% chance of going horribly wrong, and you'll probably have a bad experience.

Magic Pills

No, not that kind. Antidepressants.

Obviously you should talk to your doctor about medication if you think you're severely depressed. The more interesting edge case is whether people who *don't* qualify as severely depressed should still try antidepressants.

If you score below mild on the PHQ-9, then ignore this. If you score mild or above more than a quarter of the time, consider talking to your doctor about antidepressants (even if you don't think you're depressed).

I think too many people hesitate to try antidepressants unless absolutely convinced they are definitely depressed. That doesn't really make sense to me. I mean, there are a lot of problems with antidepressants in the real world and it makes sense to want to do your homework. But after you do your homework, if the expected value calculation says you'll probably be happier and more productive, then you just discovered a magic pill that delivers utility. (To be clear, antidepressants usually don't work this way, or else I would be lobbying GiveWell to make Universal Antidepressants Inc. their new top charity.)

Scott Alexander wrote that "There is a lot of worry that SSRIs are not much better than a placebo for people with mild to moderate depression.…However, this is a completely academic debate for you, because you are not going to get placebo. Your choice is between SSRIs or nothing. Everyone everywhere agrees SSRIs are much better than nothing."

As one anecdote, I qualified as mildly depressed 40% of the time on the PHQ-9 across six months, almost entirely because of persistent fatigue. My therapist suggested I try an antidepressant as an experiment; Rob Wiblin's positive review of Wellbutrin for mild depression made me think the idea might not be crazy, and my doctor was on board. So, I tried a low dose of Wellbutrin. Over the three months after starting it, my average deep work time increased >10 hours a month, compared to the six months before it. To minimize ongoing effort, I have my pharmacy automatically ship the medication to my home when I'm due for a refill.

For the cost of taking one pill a day, 10 hours a month is an extremely high return on investment.

On the other hand, most antidepressants have side effects which can be much worse than their benefits. At the extreme end, a bad reaction to an antidepressant could make you suicidal. For safety, if you want to try one, have someone else who will proactively check on you and intervene if things go really, really badly - perhaps your doctor, a trusted friend, or a family member. This Mayo Clinic chart reviews the side effects of different antidepressants.

Additionally, it often takes several weeks for you to experience the benefits (or be sure of the lack thereof) and you may need to try a few drugs before finding a good fit for you, making this a potentially expensive experiment to run. You probably have about a 1 in 3 chance of the first drug solving the issue.

So, status quo or experiment with a magic pill that could go really well or really badly. And people say the Matrix is unrealistic.

Focus, Mate

For people highly motivated by social expectations, the online-coworking-meets-Pomodoro platform Focusmate.com can be a life saver.

You schedule a 50-minute pomodoro in advance and are paired with a stranger on the internet. At the appointed time, you both show up, tell each other what you'll do, and work in silence. 50 minutes later, the bell dings and you share how it went.

Personally, I find it useful for making sure I do something at a specific time (e.g. establishing a workout habit at a particular time or getting started quickly on an aversive task). Many clients have also found it mildly useful, but a handful have found it transformative - on the order of doubling their output. Turns out that the thought of admitting you weren't focused to a stranger can focus you like a laser.

The biggest downside is that you need to be on a video call with a stranger. Both participants usually mute themselves during the work session, but some people still find random video calls awkward to do in an office.

Given that, this is probably the cheapest experiment on this list – low time investment for information and low downside risks.

Put Your Money Where Your Goal Is

When you need to hack motivation, financial penalties are the most reliable tool I know of. For most people. For another subset of people, penalties suck. (If you're already completing your goals without costly deadlines, then feel free to ignore this entry entirely!)

The idea is simple. Put ten, fifty, or a thousand dollars on the line if you don't complete your goal by the deadline, and you'll be more motivated. (I recommend you start with smaller amounts.) You can pre-commit to paying a friend, charity, or hated politician, depending on your desired motivation. Stickk.com is my favorite platform for setting up convenient commitments.

When I wanted to make my weekly planning habit 100% rock solid, I set a (very meta) penalty in Stickk to set weekly penalties. Each week, that email would pop up in my inbox on Sunday asking if I'd set my weekly goals yet. And I knew I needed to complete it or mark that I'd failed the next day. That small nudge was enough to slightly distort my in-the-moment motivation to match my reflective desires. (For more examples, Will MacAskill and Niel Bowerman describe other strategies they use to make financial penalties effective tools for them.)

The catch of this trick is that it's hard to know from the get-go whether financial penalties will help you.

Before trying penalties, many clients report finding penalties somewhat aversive to start and/or don't think they really need them. They usually decide it's worth trying after all when they reflect on why they failed to finish the task without that extra motivation. After trying it, many people are won over by the fact that they actually did the task.

On the other hand, some people find this technique more similar to a sledgehammer than a nudge. They may try financial penalties and lose hundreds of dollars without making progress. This is extremely demoralizing. If you often miss important deadlines with real consequences, this is probably not the technique for you.

Before they try it, it's hard even for me to distinguish which camp people will fall into. Since on average more people find penalties motivating, I tend to encourage people to try them. However, I suggest they try small penalties to start off with. If nothing else, it's easy to set overly ambitious goals that are impossible to complete in time if your outside view isn't calibrated yet.

To set good financial penalties, I recommend you:

- Make sure you're setting goals you will want to complete, since you are reducing your flexibility to change your mind later.
- Relatedly, set goals in the near future to reduce the chances that circumstances will change what you had wanted to complete. To start off, don't set goals for further in the future than a week.
- Check the outside view to see whether you're giving yourself enough time – because this deadline is real. How long did it take you to complete the last similar task?
- Schedule in more time than you think you need so you have a buffer. 20% to 50% extra time is probably a good amount to start with.

In short, financial penalties are the best way I know of speeding up your work and increasing the rate at which you complete things. But some people have extremely negative, demoralizing experiences trying financial penalties, and even the people who like them still often find the deadlines somewhat stressful.

"On Fire" Admin

The idea is simple; you do whatever life admin is "on fire" and ignore the rest.

E.g. Since the government could throw you in jail if you don't do your taxes, you do your taxes. Unless the broken towel bar in the bathroom is annoying you enough to be "on fire", you leave it be. If you can learn not to care if your bed is made, you skip

making it. If you're okay living out of boxes after a move, maybe it's not worth really setting up your bedroom.

I bet you're already familiar with the general idea. "Don't do unimportant things, duh."

I'm suggesting pushing this one step further. Look for those things that feel like you "need to do" or "should do" them, and try not doing them. Skip a meeting or don't reply to a few emails. Be aware that this could backfire; you might need to backtrack. Sometimes not doing something will come across as weird, and you need to decide if that matters to you.

The benefit is the time and mental energy you free up. Admin tasks can soak up a lot of time – even entire days – in what feels like productive work, yet nothing is different when the week is done. If you can batch admin tasks, great. If you can just skip doing a task without severe consequences, even better. In general, I see successful people more often skipping life admin tasks, and multiple clients have found this concept useful.

The downside is that a bunch of stuff doesn't get done. You may be less productive if your workspace is always messy. People may be annoyed you never responded to their email. These may impose legitimate costs that are worth the time to just do the admin tasks. You may just be less happy. (This is one of those strategies I think are great...for some other people. Personally, I feel happiest knowing the small things are wrapped up nicely.) I don't have a formula here – just a nudge to see which tasks are fine if you skip doing them.

The True Cost of Traveling

Travel is expensive. Not financially, though it can be that also. It's expensive to your productivity.

You leave your familiar haunts and most of your habits. The travel itself is often tiring; you probably have to work in worse environments, and you may be more likely to get sick. If you're jumping time zones, you're usually adding jet lag and sleep deprivation. Sleeping in strange beds can take some adjustment, further worsening your sleep situation.

And that's just while you're traveling. When you come home, you've had days or weeks away from your routine. If forming a habit takes the oft-quoted six weeks, then you're a good step towards rewriting those carefully cultivated habits.

These are all reasons I suspect when clients tell me "I was so unproductive on that trip. I thought I could get work done, but I just didn't." Traveling disrupts productivity, so minimizing travel can boost productivity.

The flip side here is that showing up can be valuable. In defense of travel, I expect that clients saying the above quote would say the travel was worth the cost more often than not. It's hard to replace talking to your collaborators in person, meeting new connections at a conference, or picking someone's brain because you caught them at happy hour. Sometimes these are significant productivity boosts in their own right. Honestly, some readers might benefit from the exact opposite advice - to be willing to go in person *more* often.

So I’m not saying to avoid travel entirely. You’ll need to weigh the benefit of being somewhere in person. Just make sure you’re considering the full cost of travel when you do.

More Pills

A potential productivity tip for people with periods; birth control can be useful to lessen or eliminate the negative impact of periods.

I had particularly bad periods, and estimate the Nuva ring gave me about one extra productive day a month. I used to lose about half a day to cramps, and general pain led to slightly decreased productivity for about 5 days. With approval from a gynecologist, I use birth control continuously to suppress periods almost entirely, with no side effects. (I have a mild period every three months.)

The pill can be used in the same way taking a birth control pill every day, instead of taking the placebo pills to allow for a period. Hormonal IUDs sometimes suppress periods, but I hear mixed reports from people using them. I get my birth control mailed to me by Nurx, so I only need to think about it once a month.

On the other hand, many people report side effects from some forms of birth control, including worse cramps, worse periods, weight gain, and mood swings. So you might need to experiment to find one that works for you. I haven’t experienced any side effects of suppressing periods and the gynecologist thought it was fine, but I don’t think we’ve actually studied long term effects.

Abridge Prolonged Electronic Correspondence

I.e. write short emails.

The median email I send is one line long. I expect this frees up at least 10 minutes per day that would have been sucked up writing longer emails. That adds up to about 60 hours saved per year. If you handle a daily flood of email or spend 30 minutes agonizing over the wording every time you email a coworker, you might save hundreds of hours a year.

I will spend more time on important emails where I want to convey more information, but the majority of my messages seem to be just fine stated in a sentence or two. I expect that these shorter emails are often appreciated by busy recipients.

On the other hand, brevity (sometimes) trades off against social niceness. It’s easier to come off as brusque or even unfriendly without nice padding. Longer time may be worthwhile if you work in a particularly socially conscious field, or email frequently with touchy colleagues who might suck up time and energy in drama over a misperceived tone.

It might feel weird shooting off a short missive - it can take some adjustment if you’re used to penning essays. An explanation can help ease the transition. Inspired by a lovely signature I saw, I added a note to my email signature to preempt recipients thinking me curt, “My replies will often be brief. Saves us both time!”

What tips would you add to the list?

LessWrong Coronavirus Link Database

Note: This post is old, read the new major update here.

There is a lot of coronavirus information currently out there, and a lot of people have been feeling overwhelmed keeping track of all the things coming in from different sources. To combat that, the LessWrong team is maintaining a spreadsheet with all the links that we think are useful, categorizing them, summarizing them, and prioritizing them.

So far on LW we've tried to help by gathering links in various focused threads (course of C19, mask usefulness). These are great for getting relevant information to a narrow question, but leave a lot to be desired in terms of later discoverability, especially if your question doesn't exactly match any existing question.

To address this, Elizabeth, Ben, and Oliver have created the **LessWrong Coronavirus Link Database**, a collection of 135 links and counting, sorted by topic and marked with importance (as judged by someone on the LW CV Mod Team, which right now is the LW admin team + me, Elizabeth, but we plan to expand in the future).

Here are the top-level categories in the database, with a prominent link from each category:

- **Guides/FAQs/Intros**
  - Crowdsourced Handbook: Focused on people building technology.
- **Dashboards**
  - Our World In Data Dashboard
- **Progression/Outcome**
  - Outcome data: Diamond Princess cruise ship
- **Spread & Prevention**
  - Model: Prevalence in Bay Area & Other Places
- **Science**
  - Paper: Where in the respiratory system is CV found?
- **Medical System**
  - Report of italian hospital triage practices
- **Shutdowns**
  - California bans all gatherings of 250+
- **Aggregators**
  - CV news aggregator maintained by volunteers

If you have a link you want to add to the database, add it to the spreadsheet by filling out this form.

Concrete things you can do with this spreadsheet

- Look for papers and articles relevant to specific questions you are currently investigating
- Share a link you think is valuable with other people (by putting it on the link dump page, from where it will be processed by the mod team)
- Find that paper you read six hours ago and submitted but then lost track of
- Get oriented around this whole coronavirus thing by reading summaries and finding good intros

- Find an Introduction to Coronavirus that finally convinces your parents to hole the fuck up.
    - I do specifically mean your parents. My dad is an introvert who explained his surviving-a-nuclear-war plan to me when I was six. Getting him to stay home was not a problem.

Is the coronavirus the most important thing to be focusing on right now?

LessWrong has been and is planning to devote a significant amount of attention to the coronavirus. But is that what we should be focused on? Is it more important than things like existential risk reduction and malaria treatments?

[UPDATED] COVID-19 cabin secondary attack rates on Diamond Princess

Update 19/03/20: Inspired by johnswentworth's comment, I implemented a multinomial distribution on the 4-berth cabin result. Taking this additional information into account the model shows reduced likelihood of secondary attack rates of >0.9.

Introduction

Jimrandomh recently showed how we have no real idea about the household secondary attack rates of COVID-19.

The Diamond Princess data showed that the proportion of passengers infected with COVID-19 increased with cabin occupancy.

It occurred to me that this data could be used to infer the cabin secondary attack rates.

Data

I eyeballed the data in figure 2 in the report linked above.

There were 6 COVID-19 cases in single passenger cabins which looks like ~8% infection rate so there were ~75 passengers in single cabins.

For double cabins the numbers are 485/2425 = 20%.

For triple cabins 27/129 = 21%.

For 4-berth 18/60 = 30%.

(all numbers are per person, rather than per cabin)

These numbers add up to 2,689 total passengers which is slightly more than 2,646 actually included but this is close as eyeballing is likely to get me.

Method

I implemented a model with 2 variables:

1. The background rate of infection without sharing a cabin (just from being on the ship).

2. An additional rate of infection for each infected person an individual shared a cabin with.

Given those two variables I was able to create predicted infection rates for each size of cabin by calculating the probability of the number of initial cases in a cabin (before

secondary attack) and then the probability of each result after applying secondary attacks.

I created 2 models, one where I only included secondary attack and another where the victim of the secondary attack could in turn cause a tertiary attack on any remaining healthy members of the cabin. Tertiary attack may not have been possible (or somewhat suppressed) by the quarantine and/or other factors.

Importantly the secondary attack rate as used by me here is “probability of contracting COVID-19 for each person in the cabin who had COVID-19”. So if you live with 2 infected people then you have a higher probability of contracting than if you just lived with 1. In 4-berth cabins having even one person infected gives a high probability of at least one of the remaining people being infected at which point the other 2 have a higher chance (when allowing for tertiary attack).

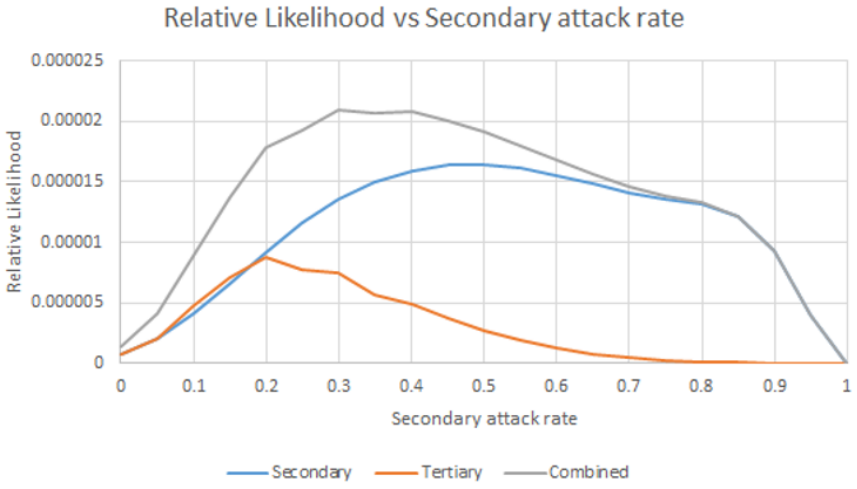
Even with a relatively low attack rate per person, it ends up being likely that many people in a 4-berth cabin will end up infected. For instance with a 0.3 secondary attack rate there is a >30% chance of all 4 people getting it from a single incoming case. A 0.5 secondary attack rate brings this up to >70% chance

These models were used to create likelihoods for the results actually witnessed via a binomial distribution.

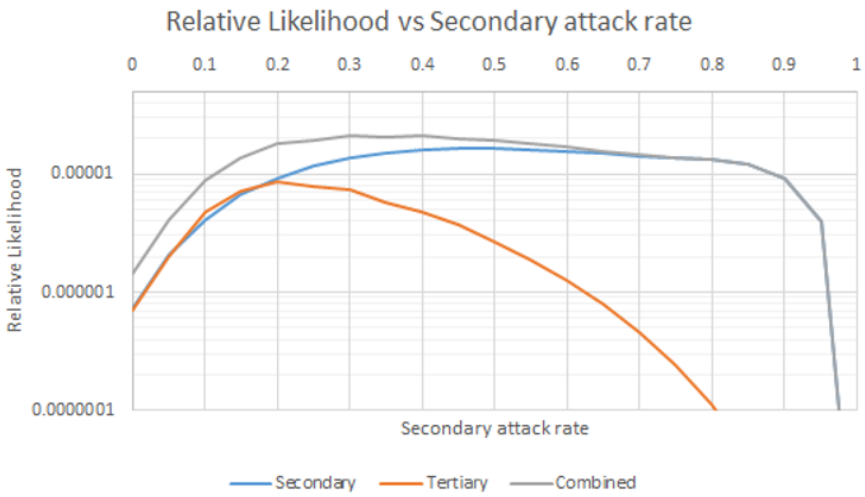
As this model isn’t computationally expensive I just brute-force calculated the likelihood over a number of possible values of the 2 variables. I then integrated across the background rate to give the likelihood function of the secondary attack rate.

Results

The likelihoods of the secondary attack rates for the two models are shown in the figure below. I’ve also included a combined likelihood based on equal confidence in both models.



And on a log axis:



This is slightly frustrating – there is a large range of secondary attack rates which fit the data adequately.

The most noticeable thing is that a very low secondary attack rate appears to be ruled out. Only 7% of the likelihood is below 0.15 and 3% below 0.1. This goes against the results from the papers analysed in jimrandomh's post (0.1 and 0.15)

The large range of possible values is caused in large part by the relatively small sample size for all except 2-berth cabins.

Discussion

There are some potential confounders here, for instance 2-berth cabins are probably mainly couples whereas 4 berth are relatively more likely to include children. I don't expect these effects to be very large (couples and their children will all have close contact) but hopefully someone will point out any potential larger confounders in the comments if there are any.

It is also not certain that cabin secondary attack rates convert directly to household secondary attack rates although my personal expectation is that they wouldn't be too far off.

Most of these secondary attack values are very bad news for larger households. Plenty of presymptomatic transmission means that if one person gets it then at least one more person will likely get it before anyone is aware that they have. So if someone does become symptomatic then isolating from each other is likely to be as important as being careful around the patient.

Isolating from each other when no-one has symptoms is likely a very costly exercise as it would need to be maintained for months but the bigger the household the more benefit is to be gained from taking care.

My impression from looking at the virus growth rate data from various countries is that massively improving hygiene and implementing social distancing can increase the doubling time by a factor of 2 (I hope to write this up in the coming days). If it can

similarly halve secondary attack rate then this could be hugely important in large households to prevent a single case infecting the entire house.

Note that as jimrandomh said, leaving a household with a sick patient in order to avoid contracting COVID-19 is a bad idea.

> If people tried to move out when their housemates got sick, they wouldn't lower their own risk much, but they would spread it wherever they moved to.

Conclusion

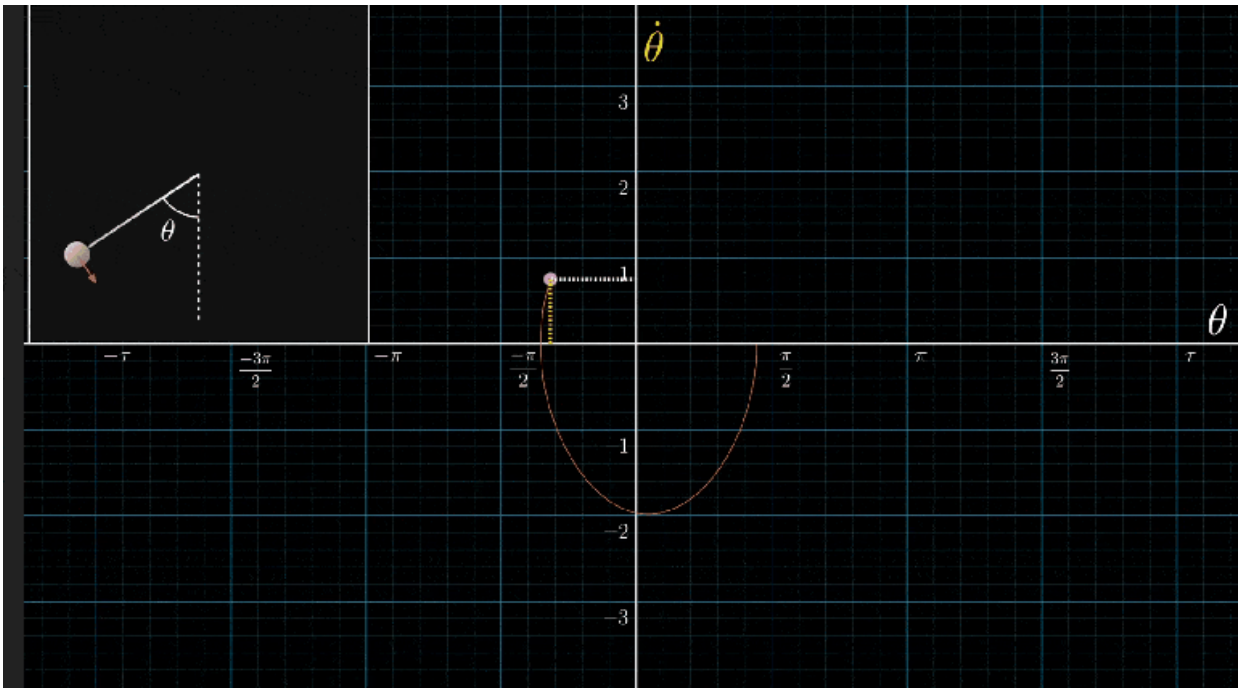
Cabin secondary attack rates of COVID-19 on the Diamond Princess were not able to be confirmed precisely. It is unlikely that the rate was very low (<0.2) and as a result additional infections are likely, especially in larger cabins.

If this can be extrapolated to households then particularly larger households may struggle to prevent additional infections after the first household member is infected.

ODE to Joy: Insights from 'A First Course in Ordinary Differential Equations'

Foreword

Sometimes, it's easier to say how things change than to say how things are.



From [3Blue1Brown: Differential Equations]

When you write down a differential equation, you're specifying constraints and information about e.g. how to model something in the world. This gives you a family of solutions, from which you can pick out any function you like, depending on details of the problem at hand.

Today, I finished the bulk of Logan's *A First Course in Ordinary Differential Equations*, which is easily the best ODE book I came across.

A First Course in Ordinary Differential Equations

As usual, I'll just talk about random cool things from the book.

Bee Movie

In the summer of 2018 at a MIRI-CHAI intern workshop, I witnessed a fascinating debate: what mathematical function represents the *movie time* elapsed in videos like The Entire Bee Movie but every time it says bee it speeds up by 15%? That is, what mapping $t \mapsto x(t)$ converts the viewer timestamp to the movie timestamp for this video?

I don't remember their conclusion, but it's simple enough to answer. Suppose $f(t)$ counts how many times a character has said the word "bee" by timestamp t in the movie. Since the viewing speed itself increases exponentially with f, we have $x'(t) = 1.15^{f(x(t))}$. Furthermore, since the video starts at the beginning of the movie, we have the initial condition $x(0) = 0$.

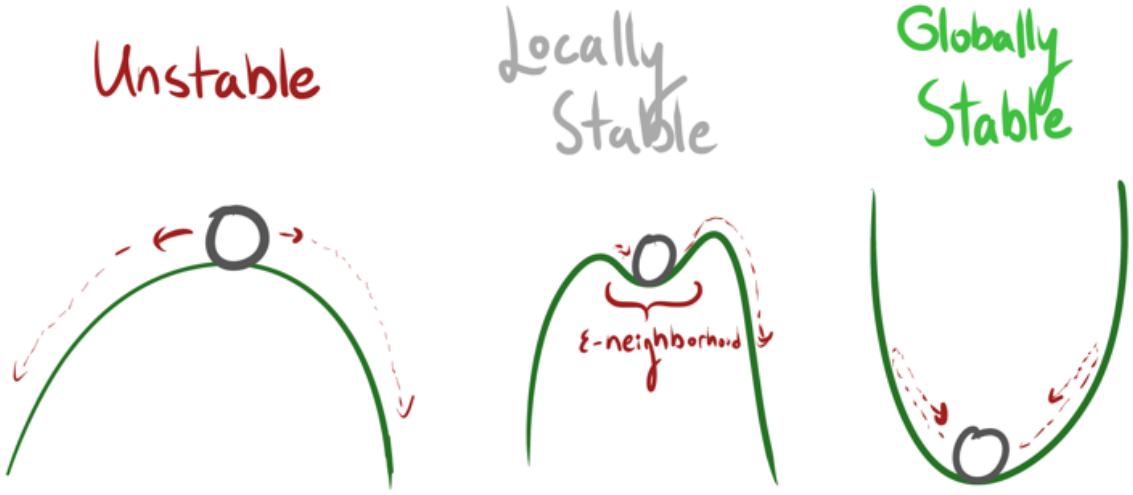
This problem cannot be cleanly solved analytically (because f is discontinuous and obviously lacking a clean closed form), but *is* expressed by a beautiful and simple differential equation.

Gears-level models?

Differential equations help us explain and model phenomena, often giving us insight into causal factors: for a trivial example, a population might grow more quickly *because* that population is larger.

Equilibria and stability theory

This material gave me a great conceptual framework for thinking about stability. Here are some good handles:



Let's think about rocks and hills. *Unstable* equilibria have the rock rolling away forever lost, no matter how lightly the rock is nudged, while *locally stable* equilibria have some level of tolerance within which they'll settle back down. For a *globally stable* equilibrium, no matter how hard the perturbation, the rock comes rolling back down the parabola.

Resonance

> A familiar example is a playground swing, which acts as a pendulum. Pushing a person in a swing in time with the natural interval of the swing (its resonant frequency) makes the swing go higher and higher (maximum amplitude), while attempts to push the swing at a faster or slower tempo produce smaller arcs. This is because the energy the swing absorbs is maximized when the pushes match the swing's natural oscillations. ~ Wikipedia

And that's also how the Tacoma bridge collapsed in 1940. The second-order differential equations underlying this allow us to solve for the forcing function which could induce catastrophic resonance.

Also note that there is only at most *one* resonant frequency of any given system, because even lower octaves of the natural frequency would provide destructive interference a good amount of the time.

Random notes

- This book gave me great chance to review my calculus, from integration by parts to the deeper meaning of Taylor's theorem: that for many functions, you can recover all of the global information from the local information, in the form of derivatives. I don't fully understand why this doesn't work for some functions which are infinitely differentiable (like $\log x$), but apparently this becomes clearer after some complex analysis.

- Bifurcation diagrams allow us to model the behavior, birth, and destruction of equilibria as we vary parameters in the differential equation. I'm looking forward to learning more about bifurcation theory. In this video, Veratasium highlights stunning patterns behind the bifurcation diagrams of single-humped functions.

Forwards

I supplemented my understanding with the first two chapters of Strogatz's *Nonlinear Dynamics And Chaos*. I might come back for more of the latter at a later date; I'm feeling like moving on and I think it's important to follow that feeling.

LW Team Updates: Pandemic Edition (March 2020)

TL; DR

The LessWrong team has been throwing our combined might behind coronavirus efforts. As I see it, coronavirus is both of a problem immense proportion that we might be able to help with as well as a domain well suited to the development and application of rationality + rationality tools

Our efforts include:

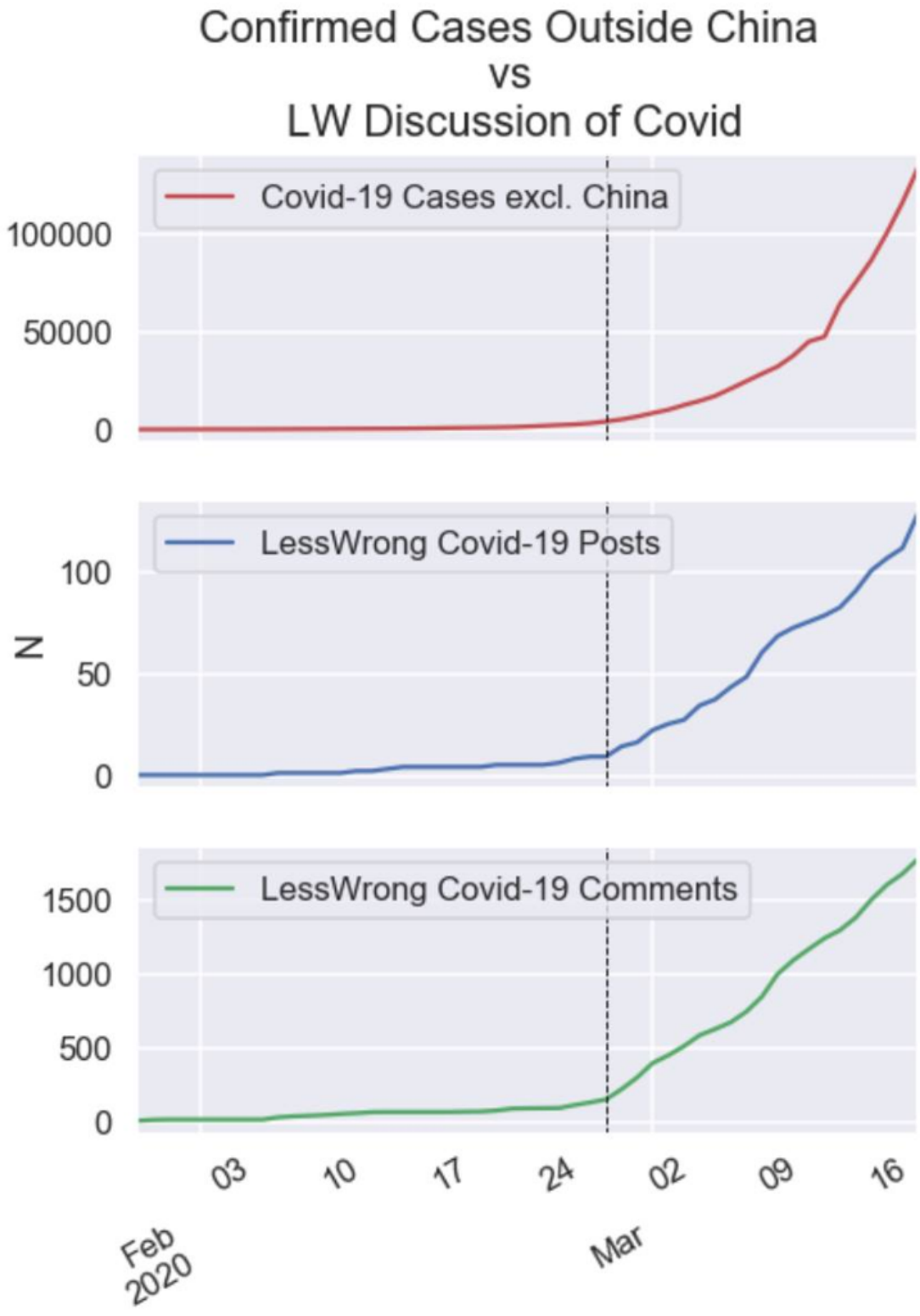
- A Coronavirus Research Agenda for the community to contribute research efforts towards.
- A Link Database of Coronavirus Resources which have been vetted, rated, and briefly summarized.
    - Plus daily updates of the best additions to the database.
- New tagging features which allow for:
    - A dedicated coronavirus tag page.
    - Filtering for coronavirus content (exclude or include).
- A thread of Justified Practical Advice.
    - plus summary.
- A Coronavirus Open Thread for coronavirus thoughts that don't warrant their own post.
- A Template Spreadsheet for Houses to coordinate their coronavirus quarantine and health statuses within and between themselves.
- Various direct research contributions [1, 2, 3].
- Online Events & Meetups
    - The first event, this Sunday 3/29, will be a livestream debate + meetup with Robin Hanson and Zvi Mowshowitz on whether we should intentionally expose parts of the population to the coronavirus in a controlled way.
    - When: Sunday, March 29 at 12:00 PDT (West Coast Time)
    - **Full Details Here**

Full Update

2020 is turning out to be quite the year. I apologize that the LessWrong team hasn't yet put out any updates so far: we'd lined up some great Q1 plans and then things got...messy.

The team's attention is now fully focused on COVID-19. Habryka has some elaboration on why this is a top priority here.

Before I go any further, serious respect to the LessWrong community which has done a fantastic job addressing the coronavirus situation ahead of the curve. Really, the response has been very proportionate to the situation.



A Note on Additional Help

For the last several weeks, the LessWrong team has been contracting with Elizabeth to help lead coronavirus efforts, hence her name being attached to several of the below projects. Elizabeth has been a member of the extended LessWrong moderator team since possibly the rise of LW2.0 and may be recognized for her work on epistemic spotchecks and other research.

Current Projects

[This Sunday!! March 29] Online Events & Meetups

The LessWrong team is beginning to organise various online meetups and events. The plan is to experiment and explore which exact kinds of meetups can be made to work and with which tech.

The first event will be this Sunday, March 29: a live-streamed debate between Robin Hanson and Zvi Mowshowitz followed by an online meetup.

- Sunday, March 29 at 12:00PM (PST / West Coast Time)
- **Full details here.**

The LessWrong Coronavirus Research Agenda

Elizabeth has laid out concrete open questions related to coronavirus here. LessWrong/Rationality is about believing true things and taking actions that cause the best outcomes. In many ways, the coronavirus pandemic is an excellent test of our rationality. It's a murky, high-uncertainty domain with high stakes.

I've been excited by the leading work the LessWrong community has done to date, and am further excited by the prospect of us pulling together communally for further answer the many questions that remain.

Right now, the top 3 concrete questions are:

- How can we estimate how many people are infected in an area?
- Where can we donate time and money to avert coronavirus deaths?
- What should we do once infected?

You do not need a background in biology to help out. Basic quantitative skills are sufficient to provide useful information. Ben Pace has further guidance in his How To Contribute Guide.

Coronavirus Links Database

On March 13th, The team launched a daily-updating database of links to coronavirus resources. Links are collected, sorted by topic, rated, and summarized. Top level categories include "Progression & Outcome", "Spread & Prevention", "Science", "DIY", and so on.

The goals is to make it easy to find important information, for examples, answers to questions like:

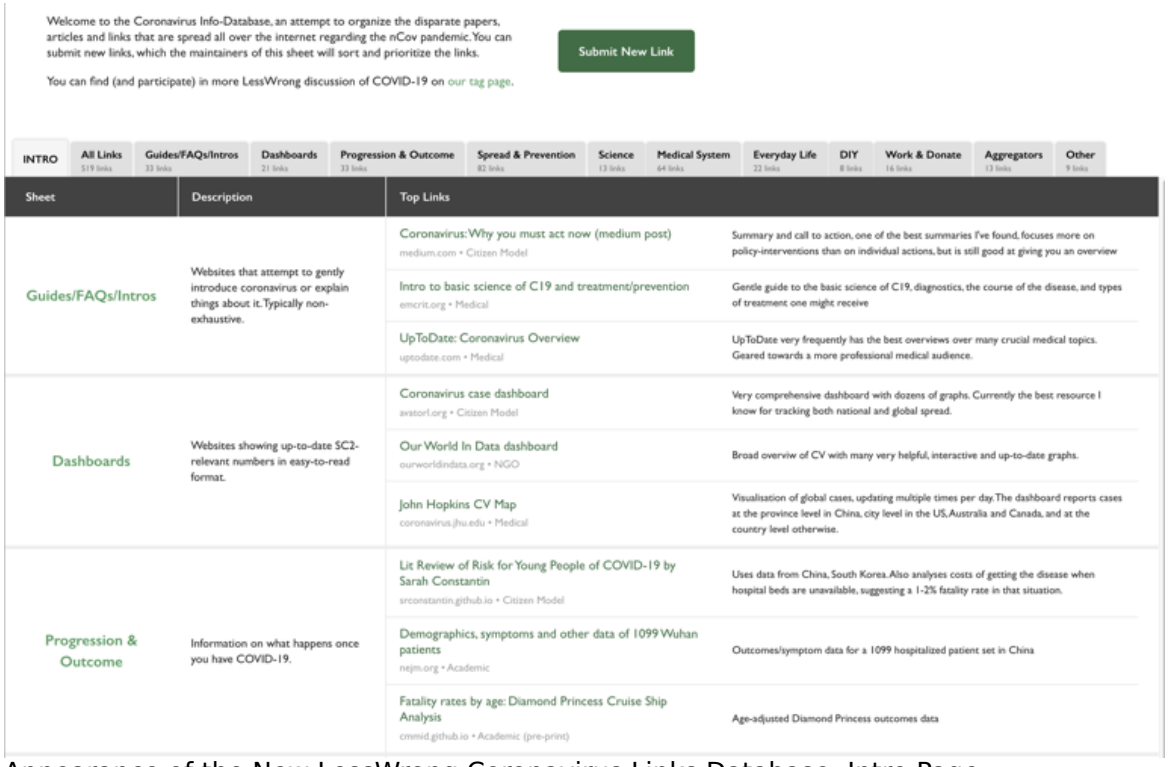
- "*What's the best dashboard to follow global case counts?*"
- "*What's the best link to send to my parents?*"
- "*How long does COVID-19 last on different material surfaces?*"

- "*Where can I find that really good thread by Rob Wiblin I read last week?*"

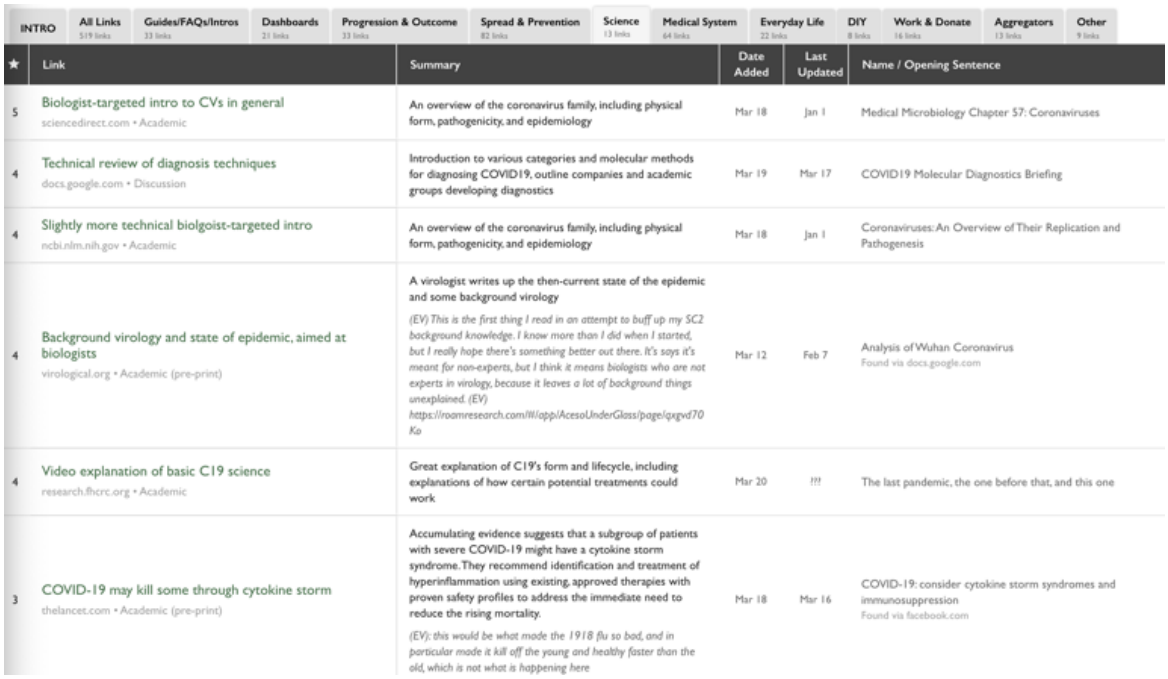
The link database was originally implemented as a Google Sheet but has just been migrated to live in the LessWrong site proper.

[See the database here]

[Contribute links here]



Appearance of the New LessWrong Coronavirus Links Database: Intro Page



LessWrong Coronavirus Links Database: Science Page

Daily Links Updates

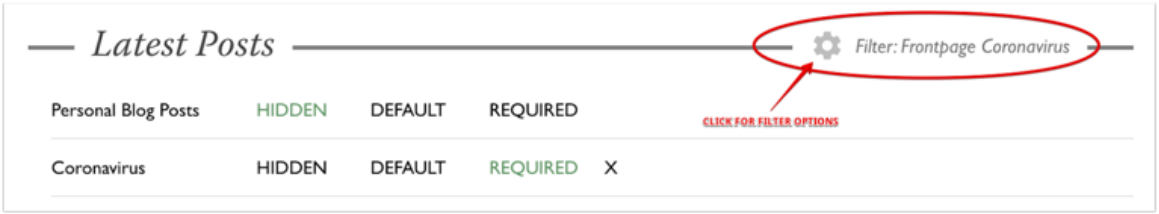
In conjunction with the links database, the team has been posting a daily-updates post containing notable links which were added in the last day. For example:

- March 21st: Daily Coronavirus Links
- March 20th: Daily Coronavirus Links
- March 19th: Daily Coronavirus Links

Tagging/Filtering

The LessWrong team has been talking about “tagging” as a feature for a long-time now and head a prototype build as of last December. The coronavirus situation gives us extra reason to roll it out now: some people might wish to filter out the deluge of coronavirus related content and view posts on other topics. To this end, we’ve rolled out a limited version of tagging with just the “Coronavirus” tag.

Raemon describes the feature fully in this post here. The key thing to know is you click the gear next to Latest Posts to open up filtering options.



Appearance of tag filtering options upon clicking Filtering Gear Icon.

- Hidden: no posts with this tag will be displayed

- Required: *only* posts with this tag will be displayed
- Default: posts with this tag will be displayed normally based on karma and time since posting.

Naturally, if you're after everything we've got on coronavirus, visit the **coronavirus tag page** at **www.lesswrong.com/tag/coronavirus.**

Coronavirus Justified Practical Advice Thread (& Summary)

The team noticed early on that a lot of coronavirus advice was being shared, often without much explanation for why it was good advice. Thus was the Justified Practical Advice (JPA) Thread born. The thread is long, contains many interesting and hopefully useful ideas, but can also take a while to read. So we've also got the Justified Practical Advice Thread **Summary** which contains the best advice from the thread.

Coronavirus Open Thread

If you have something to say on coronavirus that's not worth a top-level post, please share in the Coronavirus Open Thread.

House Coordination Spreadsheet & Isolation Levels

Raemon with the assistance of others put a great deal of effort into developing a spreadsheet template for houses within a community to coordinate on their health and quarantine status.

Coronavirus Household Isolation Coordination 1.2

The spreadsheet can be an excuse to think about isolation plans, discuss them with roommates, and create common knowledge of them. Subgoals include:

- establishing guidelines for level-of-isolation (see second tab)
- help with contract tracing
- facilitate roommate swapping between higher-caution and lower-caution households
- facilitate creation of multi-house cells that share an isolation level (where permitted by local quarantine laws, e.g. not California)

I find the second tab with "isolation levels" useful for thinking about exposure, risk, and various precautions.

Direct Research

Jim, Elizabeth, and others have been putting out some great direct research. Notable contributions include:

- Credibility of the CDC on SARS-CoV-2
- A Significant Portion of COVID-19 Transmission Is Presymptomatic
- COVID-19's Household Secondary Attack Rate Is Unknown

Upcoming Projects

I expect the team will want to remain agile in coming weeks and move our efforts to wherever seems most useful. Given that, I think it's hard to commit specifically to what we'll work on.

A project I'd like to work on if possible is moving us towards some kind of Wiki tech. I could see that being useful for establishing a Schelling location for the most up-to-date knowledge of given questions of interest, e.g. treatments for global spreading infection diseases.

How To Contribute

Ben Pace has written a guide on how to contribute to LessWrong's coronavirus efforts. It so far details how to contribute to the Links Database and Research Agenda.

Feedback, Support, Questions, Etc.

We love to hear from people. If you want to talk to us about anything, please reach out:

- Comment on this post
- Intercom (icon in the bottom right, you might have to edit your user settings)
- Email us at hello@lesswrong.com or support@lesswrong.com
- Ask a question on www.lesswrong.com/questions
- Message us on our Facebook page.

Blog Post Day II

TL;DR: You are invited to join us online on Saturday the 28th of March, to write that blog post you've been thinking about writing but never got around to. This is the second event of this type; the first one went well. Please comment if you are thinking about joining so I can gauge interest.

The Problem:

Like me, you are too scared and/or lazy to write up this idea you've had. What if it's not good? I started a draft but... Etc.

Alternatively: Coronavirus got you down? Cooped up inside? Looking for something new to do, someone new to talk to? How about you write a blog post?

The Solution:

1. Higher motivation via Time Crunch and Peer Encouragement

We'll set an official goal of having the post put up by midnight. Also, we'll meet up in a special-purpose discord channel to chat, encourage each other, swap half-finished drafts, etc. If like me you are intending to write the thing one day eventually, well, here's a reason to make that day this day.

2. Lower standards via Time Crunch and Safety in Numbers

Since we have to be done by midnight, we'll all be under time pressure and any errors or imperfections in the posts will be forgivable. Besides, they can always be fixed later via edits. Meanwhile, since a bunch of us will be posting on the same day, writing a sloppy post just means it won't be read much, since everyone will be talking about the handful of posts that turn out to be really good. If you are like me, these thoughts are comforting and encouraging.

Evidence this Works:

MIRI Summer Fellows Program had a Blog Post Day towards the end, and it was enormously successful. It worked for me, for example: It squeezed two good posts out of me. (OK, so one of them I finished up early the next morning, so I guess it technically doesn't count. But in spirit it does: It wouldn't have happened at all without Blog Post Day.) More importantly, MSFP keeps doing this every year, even though opportunity cost for them is much higher (probably) than the opportunity cost for you or me. And we did a Blog Post Day on LW last month and it worked great.

Side Benefits:

It'll be fun!

The Case for Privacy Optimism

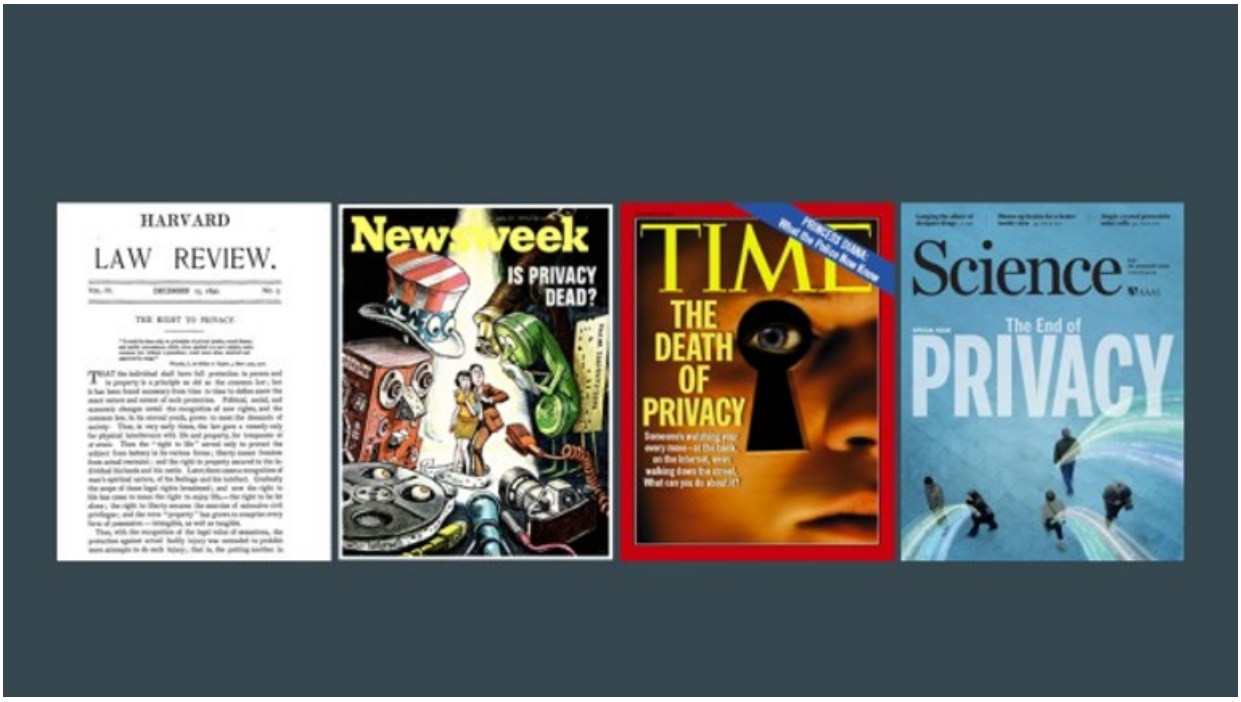
(*Cross-posted from my personal blog.*)

This post is an edited transcript of a talk I recently gave on the past and future of privacy. I argue that the story may be a more positive and hopeful one than people often realize.

The talk stemmed from work I've done at the Centre for the Governance of AI. The people whose writing most influenced its ideas are Joan Feigenbaum, Aaron Segal, Bryan Ford, and Tyler Cowen. Mainly: This paper and this blog post. I've also especially benefitted from conversations with Andrew Trask.

Introduction

I think it's fair to say that discussions of privacy issues tend to have a pessimistic edge to them. For a long time, the dominant narrative around privacy has been that it's under threat. It's either dying or it's dead.



Here we have five decades of magazine covers announcing or predicting the death of privacy at the hands of technology. We also have the 1890 *Harvard Law Review* article "The Right to Privacy," which is often seen as the starting point for modern political and legal discussions of privacy. The framing device for that article was that the camera and other new technologies posed severe threats to privacy. These threats were meant to be unprecedented enough to warrant, for the first time, the introduction of a new "right to privacy." So it seems like we have been watching and working to stave off the death of privacy for a long time now.

Privacy Pessimism

The following narrative is a bit of a caricature, but I also do think it should be fairly recognizable. Something like it seems implicit in many discussions of privacy. The narrative goes like this: (1) People used to have a lot more privacy. (2) Unfortunately, over time, technological progress has gradually eroded this privacy. (3) Now new technologies, such as artificial intelligence, are continuing this trend. Soon we may have little or no privacy left.

Let's call this narrative "Privacy Pessimism."

Now, of course, the Privacy Pessimism narrative isn't universally accepted. However, I think that even when people do decide to critique it, these critiques don't often qualify as "heartening." Here's a quote from *Privacy: A Very Short Introduction*, pushing back against Privacy Pessimism. The author writes:

A requiem, however, is premature. The invaders are at the gate, but the citadel will not fall without a battle.

There is definitely an element of optimism here. But I still don't know that the situation being described is really one we should be happy to find ourselves in. I don't know how comforting it is for the people on the right-hand side of the painting up there to be told: "Well, at least there's a battle on."

Privacy Optimism

What might it look like to be a proper optimist about privacy? One way to explore the space of narratives here is to just turn Privacy Pessimism on its head. The narrative we get out looks something like this: (1) It used to be the case that people had very little privacy. (2) Fortunately, technological progress has tended to increase privacy. (3) Now

new technologies, like artificial intelligence, can be used to further increase and protect privacy.

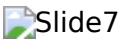
Let's call this narrative "Privacy Optimism."

Privacy Optimism is obviously a much less familiar narrative than Privacy Pessimism. It also has a bit of a ring of contrarianism-for-the-sake-of-contrarianism. But is there anything to be said in favor of it?

My view is that, perhaps surprisingly, both narratives actually do capture important parts of the truth. Technology has given us more of certain kinds of privacy and less of other kinds. And both hopeful and distressing futures for privacy are plausible.

Even if losses and risks are very real, positive trends and possibilities ought to be a larger part of the conversation. It's worth taking optimism about privacy seriously.

Plan for the Talk



To lay out my plan for the rest of the talk: First, I'm going to be saying a little bit about what I have in mind when I talk about "privacy." Second, I'm going to present an extremely condensed history of privacy over the past couple hundred years. I'm going to try to show that technology has had both positive and negative effects on privacy. I'll actually try to argue that for many people the positive effects have probably been more significant. Third, I'm going to talk about why certain privacy problems are so persistent. Then, finally, I'm going to try to stand on top of this foundation to get a clearer view of the future of privacy. I'll be giving special attention to artificial intelligence and the positive and negative effects it could ultimately have.

Clarifying "Privacy"

The word "privacy" has a lot of different definitions. There are actually entire books devoted to nothing other than just clarifying the concept of privacy. I'm going to be using a relatively simple definition of privacy. In the context of this talk, **privacy** is just the thing that we lose when people learn things about us that we'd rather they didn't know.

Sometimes we would like to talk about "larger" or "smaller" violations of privacy. Here's how I'll be thinking about quantification in the context of this talk: The more unhappy we are about a loss of privacy — or the more unhappy we would be *if properly informed* — the more privacy we have lost.

To illustrate: Sometimes people are a bit annoyed when others discover their middle name. Maybe their middle name is "Rufus" or something. But typically people don't care very much. Fully informing them of the implications of having their middle name known also probably wouldn't lead them to care much more. So we can consider this case to be one where the loss of privacy is "small." In contrast, most people would be quite a bit more upset if all of their texts with their significant other were published online. So that would count as a much larger violation of privacy.

Last bit of conceptual clarification: I'm also going to distinguish between two different kinds of privacy. I'm going to say that we suffer a loss of **social privacy** when people

within communities that we belong to learn things about us. And we suffer a loss of **institutional privacy** when institutions such as governments and companies learn things about us.

To illustrate: Let's say that someone would prefer for people in the workplace not to know their sexual orientation. If their orientation is revealed to their coworkers, against their wishes, then that would be a loss of social privacy. If some web advertising company is able to infer someone's sexual orientation from their browsing activity and sends them targeted ads on that basis, then that would be a loss of institutional privacy.

These are the concepts I'm going to be applying throughout the rest of the talk.**[1]**

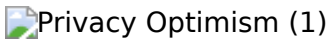
A Brief History of Privacy and Technology

Let's move on to history. One choice we need to make when exploring the history of privacy is the choice of where to begin. I'm going to start in the centuries leading up to the Industrial Revolution. My justification here is that up until the start of industrialization — up until mechanization, the use of fossil fuels, and this sort of thing really started to take off — a lot of the core material and technological features of people's lives didn't tend to change very much from one generation to the next. Many of these features were also relatively constant from one region to the next. So while they did change and they did vary, they were still constant enough for us to talk about "pre-industrial privacy" without feeling totally ashamed about how much we're overgeneralizing.

Social Privacy in the Pre-Industrial Era

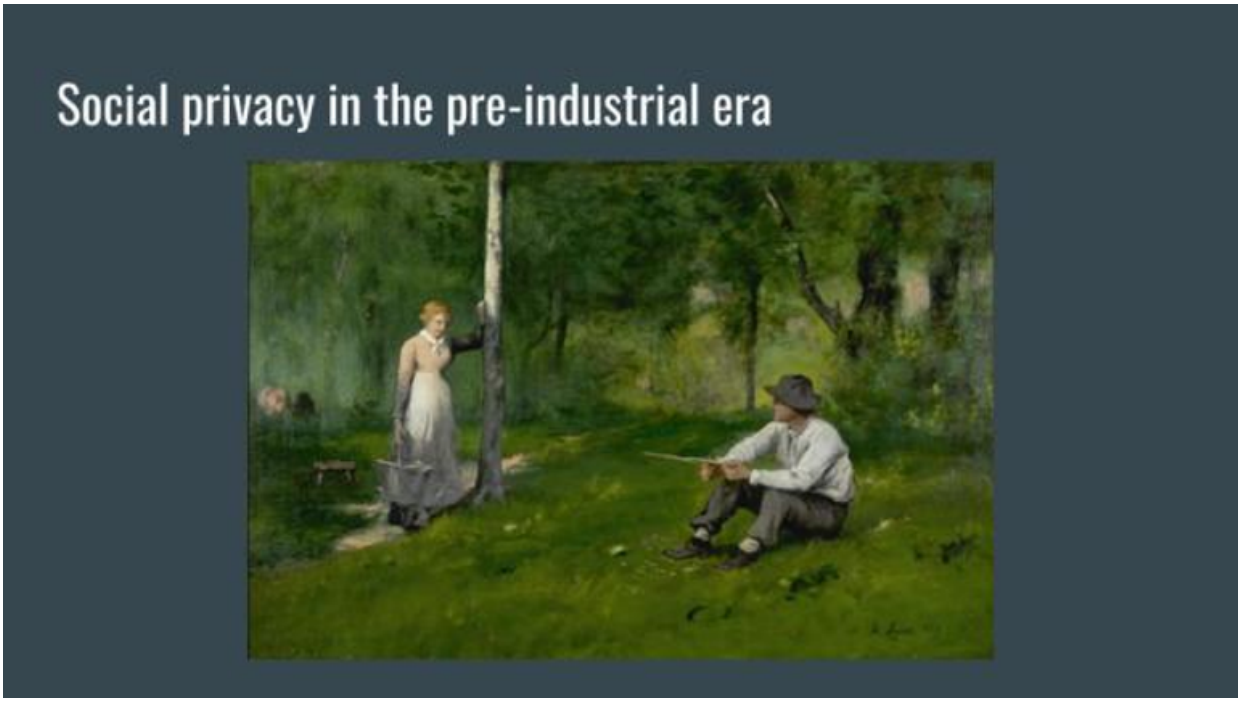
Let's focus on social privacy first. I think it's fair to say that life in the pre-industrial era — or, for that matter, present-day life in places that have yet to industrialize — is characterized by extremely little social privacy.

One significant fact is that it's common for pre-industrial homes to have only a single room for sleeping. In some places and times, it may also be common for multiple families or for a large extended family to share a single home. It might even be common for people to be able to enter neighbors' homes freely. So that means that if you're at home, you don't get space to yourself. You don't have private places to stow possessions, have conversations, host guests, or have sex. If you do something at home, your family and perhaps many other people are likely to know all about it.


Privacy Optimism (1)

Another significant fact is that you most likely live out in the countryside, in a very low population density area. You probably very seldom travel far from home. A major limitation here is that the best way you have of getting around is probably simply to walk. This means that "strangers" are rare. Most people you interact with know most other people you interact with. This makes it very easy for information about you to get around. You can't very easily, for example, choose to present a certain side of yourself to some people but not to others.

You also probably can’t read. Even if you could, you probably wouldn’t really have access to books or letters. These things are expensive. One thing this implies is that if you have a question, you can’t simply look up the answer to the question. You need to ask someone in person. This makes private intellectual exploration extremely difficult. You can’t have an intellectual interest that you pursue without alerting the people around you that you have this interest. Or you can’t, for example, have a private question about your body, about community politics, or about ethical dilemmas you might face.



Gossip is probably very prevalent and used to enforce strict behavioral norms. Because you probably live not very far above the subsistence level, you depend on your community as a social safety net. This means that disapproval and ostracism can be life-threatening. In extreme cases, you might even be subject to violence for violating important norms. Picking up and leaving to find a new community to accept you will also probably be extremely risky and difficult to do. It really is no small thing if information that you’ve been deviating from community norms gets around.

So the state of social privacy in the pre-industrial era was *not good*.

Technology and Social Privacy

I think that technological progress has mostly been a boon to social privacy.

One of the most straightforward ways in which it’s been a boon to social privacy has been by driving economic growth. Around the start of the Industrial Revolution, most people on Earth had standards of living that were just slightly above the subsistence level. Today most people on Earth are much richer. This means we can afford to spend more money and time on things that enhance our privacy. For example, we can typically afford to spend more money on homes with multiple sleeping rooms. Another smaller example: My understanding is that until the second half of the nineteenth century, it wasn’t yet fully the norm for Americans to use envelopes when sending

letters. Because paper envelopes were expensive and difficult to produce. Then people got richer and manufacturing technology got better and suddenly it made sense to buy that extra bit of privacy. Simply put: Poverty is bad for privacy. And technological progress has supported a dramatic decline in poverty.

Technological progress has also had a positive effect through new modes of transportation such as the bicycle, train, and car. These have helped people to have rich lives outside the watch of neighbors and family. It is relatively easy to go out to attend some meeting of a club, or religious group, or support group, without knowledge of your attendance or behavior at the meeting necessarily getting around. It's also easier to just go out and hang out with whoever privately. There's a lot of literature specifically on the liberating effect that the widespread adoption of the car had on teenagers and housewives in the context of American domestic life.

The effect of information technology has also been pretty huge. Again, I think that the ability to ask and receive answers to questions without alerting people in your social circle is a really big deal. I think it's a big deal that girls can privately look up information about birth control and reproductive health, for example, and I think it's a big deal that people in religious communities can privately explore doubts on online forums. I think that people's ability to communicate somewhat more furtively with friends and romantic partners using letters, phones, and instant messenger platforms has also been very important.

There are also many small examples of improvements in information technology leading to improvements in social privacy. One recent case is the introduction of e-readers. There's some evidence that this has actually changed people's reading habits, by allowing people to read books without necessarily alerting the people around them as to what they're reading. A lot of the people who read *Fifty Shades of Gray*, for example, probably would not have read it if they had needed to display a physical copy. There are so many small examples like this that we could pick. I think that together they add up into something quite significant.

Technology and Institutional Privacy

Okay, so social privacy has mostly gotten better. How about institutional privacy? Unfortunately, I think this is a very different story. In my view, the past couple hundred years have seen institutional privacy decline pretty dramatically.

The basic idea here is that pre-modern states and other centralized institutions simply had a very limited capacity to know about individuals' lives. Imagine that you're the King of France in, say, 1400. You don't have computers, you don't have recording devices, you don't have a means of transporting information faster than horse-speed speed, you don't have a massive budget or a massive supply of literate civil servants. If you want to collect and process detailed information about the people in your kingdom, you're going to have a hard time. Even just learning roughly how many people reside in your kingdom would be a difficult enough challenge.

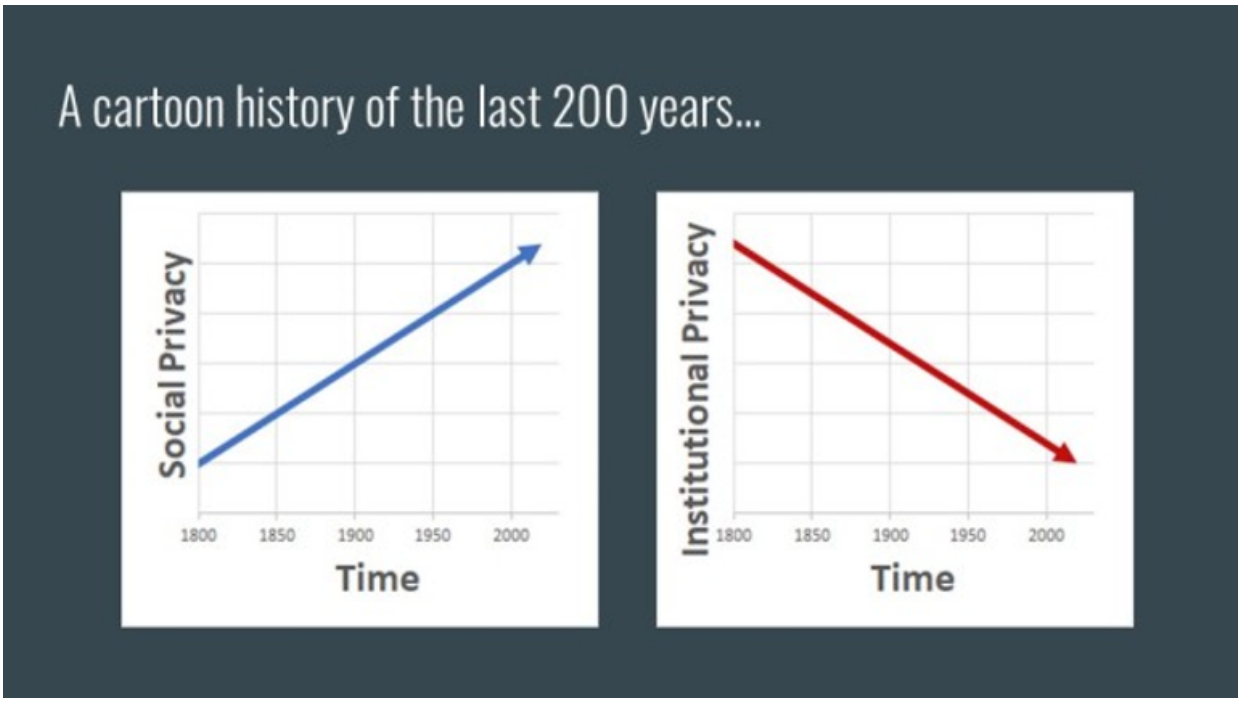
Today, you can count on many different institutions owning and being able to make use of vast amounts of data about you. That's something new and something that would not have been possible in the relatively recent past. Steady improvements in information technology have supported a steady ratcheting up of how large and usable these datasets are.

I probably don’t need to say very much about this trend, because so much has already been said about it elsewhere. One especially good book on the subject Bruce Schneier’s *Data and Goliath*. While the book is pretty gloomy, it’s also focused primarily on the erosion of institutional privacy in an American context. Here the main institutions of interest are companies like Facebook and intelligence agencies like the NSA. I think the picture gets a lot bleaker if you decide to look at places with weaker institutional safeguards.

Recent news reports concerning the mass surveillance of the Uighur minority in China provide one of the most obviously bleak examples. A set of technologies ranging from smartphone apps to facial recognition cameras are apparently being leveraged to collect extensive data on Uighur individuals’ movements throughout cities, electronic communications, consumption patterns, and so on. It’s apparently made quite clear that if you don’t comply with certain expectations, if you seem to be in any way a threat to social order, then you’ll face a high risk of imprisonment in indoctrination camps. This is horrifying. And it’s a modern variety of horror.

Comparing the Two Trends

The picture I’ve just painted of the past two hundred years is one of rising social privacy and declining institutional privacy. A more detailed history would acknowledge that these trends have not been totally steady. There have surely been individual periods where social privacy declined and institutional privacy grew. The move toward people having more unified online identities through platforms like Facebook would be an example of a technological development that probably decreased social privacy. And the move toward more companies using end-to-end encryption would likewise be an example of a technological development that probably increased institutional privacy. But at least if we zoom out enough I think the overall trends are pretty clear.



How about the *net effect* of these two trends? Have the past couple hundred years of change, overall, constituted decline or progress?

I think there are two main ways to approach this question. First, we can introspect. We can ask: “Would I be happy returning to the levels of both institutional *and* social privacy I would have had in Year X?” Returning to pre-industrial levels of institutional privacy, for example, might mean wiping all of your data from the servers of government agencies and tech giants. But returning to pre-industrial levels of social privacy might mean giving your extended family access to a live HD video stream of your bedroom. If you ultimately wouldn’t be happy to accept the trade-off, then this suggests that at least you personally have experienced an overall gain. You can also ask the same question for different pairs of dates to probe your intuitions about more short-term changes.

Second, we can observe our own behavior. We can ask: “When there are trade-offs between social and institutional privacy, what do I tend to choose?” It seems to me like opportunities to trade one form of privacy off against the other actually come up pretty frequently in modern life. You can ask a friend a personally revealing question, for example, or you can ask Google. You can buy a revealing product on Amazon, or you can buy it with cash at a local store. You can directly reveal romantic interests in people, within communities you belong to, or you can swipe on a dating app. The choices you make when you face these dilemmas provide at least some evidence about the relative importance of the two forms of privacy in your life.

My personal guess is that, for most people in most places, the past couple hundred years of changes in individual privacy have mainly constituted progress. I think that most people would not sacrifice their social privacy for the sake of greater institutional privacy. I think this is especially true in countries like the US, where there are both high levels of development and comparatively strong constraints on institutional behavior. I think that if we focus on just the past thirty years, which have seen the rise of the internet, the situation is somewhat more ambiguous. But I’m at least tentatively inclined to think that most people have experienced an overall gain.

For many other people, who have been acutely victimized by insufficiently constrained institutions, the trend has surely been a much more negative one. For the Uighur community in Xinjiang, for example, it's very difficult to imagine classifying the past thirty years of change as "privacy progress." It's very difficult to imagine taking a rosy view of privacy trends in East Germany in the middle of the previous century. A whole lot of people have experienced very tangible and even horrific suffering due to the erosion of institutional privacy.

So it would be far too simple to call the story here a "positive" one. Nevertheless, I do think that the story is much *more positive* — and certainly much more nuanced — than many discussions of privacy issues suggest.

Barriers to Institutional Privacy

Before turning to the future, I want to spend a bit more time talking about the ongoing decline of institutional privacy. Specifically, I want to dwell on the question: "Why is it that we keep losing ground? What's blocking us from having more institutional privacy than we do today?"

In the last section, I talked about how technological progress has made it possible for institutions to learn more about us. But of course, this is only a n_ecessary condition_ for institutions learning more about us. It's not a *sufficient* condition. It doesn't explain why they *actually are* learning all that they're currently learning.

I think that one explanation here is **practical necessity**. Sometimes institutions learn about us because they would otherwise be unable to provide certain services, which are worth the loss of privacy. Credit services are one clear case. Credit card companies generally keep records of their customer's purchases. These records are often pretty personally revealing. But the companies pretty much *need* to keep these records in order to provide customers with the services they want. It's hard to bill someone at the end of the month if you have no idea what they bought that month.

The other explanation is **abuse of power**. Sometimes institutions learn about us in order to pursue objectives that don't actually benefit us. I think that the totalitarian system of surveillance established in East Germany is again one especially clear case. The main point of the Stasi pretty clearly wasn't to give the people of East Germany what they wanted. Instead it was about advancing the interests of a powerful group, even though it came at an enormous cost to the people below.**[2]**

Focusing on Practical Necessity

Both explanations are clearly very important. Here, though, I want to focus on practical necessity. One justification for this focus is that "practical necessity" is, in some sense, the more *fundamental* issue. The problem still exists even in the case of ideal governance, even in the case where everyone has good intentions and we get the design of institutions just right.

Another justification for this focus is that data collected on the grounds of "practical necessity" can also subsequently be abused. You can credibly or even correctly say that some surveillance system is necessary to keep people safe, that it's necessary to provide people with security, and then turn around and abuse the information that the system allows you to collect. You can correctly say that you need to collect certain data from customers, then abuse or mishandle that data once you have it. So the fact that

“practical necessity” works as a justification for data collection, in so many cases, also helps to explain why institutions have so many opportunities to abuse their power.

The Problem of Lumpy Information

Let’s suppose that “practical necessity” is actually one major explanation for the decline in institutional privacy. This then suggests a deeper question: “Why is it practically necessary for institutions to learn so much about us?”

One simple answer is that information is *lumpy*. The idea here is that learning some simple fact, which you must know in order to provide a service, can require learning or being in a position to learn many other sensitive facts. So even institutions that provide us with very narrow services end up receiving huge lumps of information about us.

To make this idea concrete, I’m going to give a few brief examples.

First, let’s suppose that you want to learn whether someone else’s backpack has a bomb in it. If you don’t have access to any fancy tools, then you really only have a couple of options. First, you can open the backpack and look inside and learn whether or not there’s a weapon in there. In the process of doing this, though, you will also learn everything else that the backpack contains. Some of these items might of course turn out to be quite personal. Second, you can refrain from opening the backpack and thereby learn nothing. These seem to be the only two options. You need to either learn too much or learn too little.

Second, let’s say you’re a cellular service provider. To provide this service, you need to know the times and destinations of the messages your customers are sending. Unfortunately, by collecting this seemingly limited information, you’ve put yourself in a position where you can make a lot of inferences about this person’s life. You may be able to infer that this person is having an affair, or that they’re receiving

chemotherapy, and so on, purely on the basis of their call logs. A lot of additional information may come along for the ride.

Third, let's say that you want to study the statistical relationship between some health outcome and various lifestyle choices. You have no interest in learning about any individual person's lifestyle. You only want to learn about overall statistical relationships. It unfortunately may be quite hard to do this with collecting lots of information about individuals and being in a position to identify or infer things about them when doing your analysis.

Summing Up

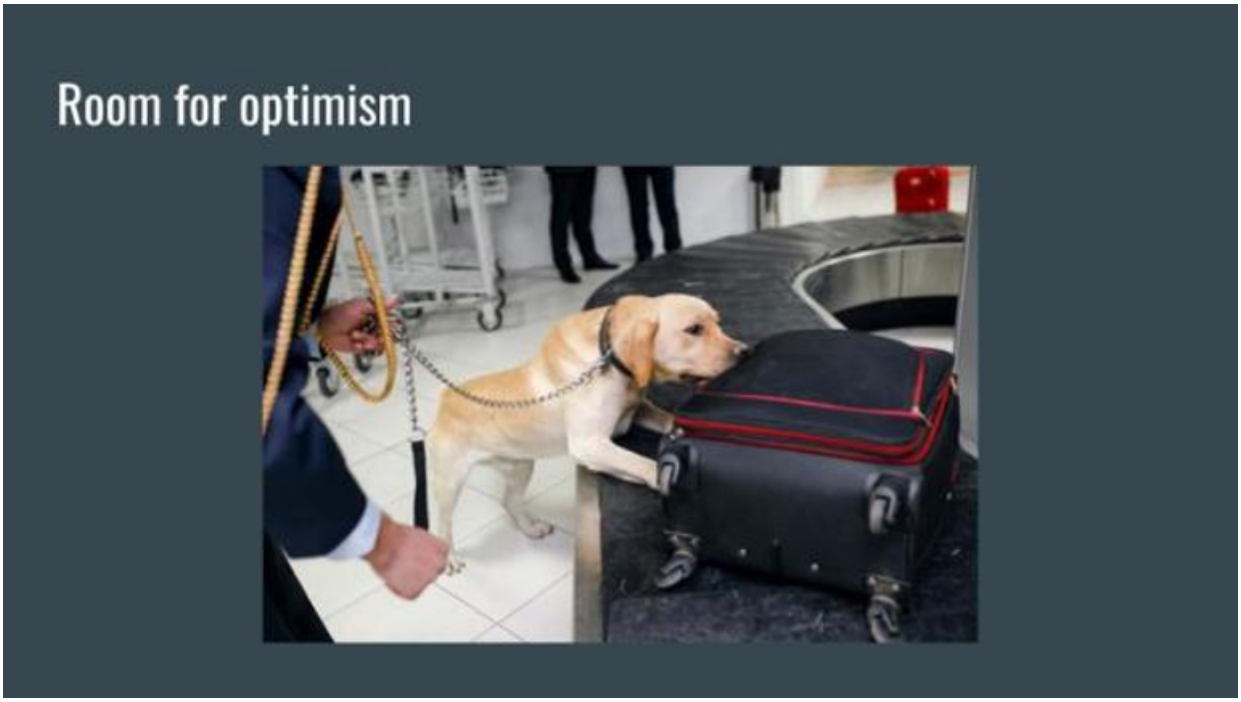
To summarise, then, here's a partial explanation for why institutional privacy continues to be so poor: Institutions provide us with services that often require learning about us. Unfortunately, the lumpiness of information means that they are often in the position to learn *a lot* about us. Institutions aiming to abuse their power can also collect more information than they need to, in the first place, by at least semi-credibly claiming that this information is necessary for the services they provide.

Institutional privacy has then been on the decline, in no small part, because the quantity and sophistication of the services that centralized institutions provide continue to grow. New services like telecommunication, online social networking, and protection from terrorism have been begetting further losses of institutional privacy.

Room for Optimism

If we want to hold out hope, that one day this trend may reverse, then one natural question to ask is: Could information be made less lumpy? Can we develop methods that allow institutions to have all the information they need to provide import services, without requiring so much other information come along for the ride?

At least in principle, I think that we can. We already have examples of technological innovations that reduce the problem of lumpy information in certain limited contexts.



Just as one example, a bomb-sniffing dog can be considered a sort of technology. And what a bomb-sniffing dog allows you to do is to get out this one piece of information, the fact that some bag either does or does not have a bomb in it, without learning anything else about what's inside the bag. Another example would be "differential privacy" techniques. These are a set of techniques that help people to discover overall statistical patterns in datasets, while being in a much weaker position to learn about the individuals represented in these datasets.

However, I think it's pretty clear that these sorts of technologies have had only a comparatively small effect so far. The effect hasn't been nearly large enough to offset the other forces that are pushing us toward a world with less and less institutional privacy.

Artificial Intelligence and the Future of Privacy

In this last section, I'm going to be turning to talk specifically about artificial intelligence.

There's recently been a lot of worry expressed about the privacy implications of AI. I also just sort of think, in general, that progress in AI is likely to be one of the most important things that happens over the next several decades and centuries. So I'd like to better understand what progress in AI implies for the future of privacy.**[3]**

Social Privacy

Institutional privacy is again going to be the main focus of this section, because I think this is currently where the most noteworthy and potentially scary developments seem to be happening. But I will first say a bit about social privacy.

One way that AI might support social privacy is by helping to automate certain tasks that would otherwise be performed by caregivers, assistants, translators, or people in other professions that essentially require sustained access to sensitive aspects of someone's life.

For example, there are people working on AI applications that could help people with certain disabilities or old-age-related limitations live more independently. This includes voice command and text-to-speech software, self-driving vehicles, smartphone applications that "narrate the world" to vision-impaired people by describing what's in front of them, and so on. Greater independence will often mean greater privacy.

On the other hand, one way that AI could erode social privacy is by making it easier to look other people up online. We may get to a point where consumers have access to software that lets you snap a photo of someone and then, taking advantage of facial recognition systems, easily search for other photos of them online. This could make people less anonymous to one another when they're out and about and make it difficult to compartmentalize different aspects of your life (insofar as these aspects are documented anywhere in online photos).

It's obviously difficult to say what the overall effect on social privacy will be.**[4]** I would loosely expect it to be positive, mainly on the basis of the historical track record. But history of course provides no guarantees. A lot more thinking can certainly be done here.

Institutional Privacy: Worries

Moving on, now, to institution privacy. A lot of people have raised concerns about the impact that AI could have on this form of privacy. I think these concerns are very often justified.

I think there are essentially two main reasons to worry here. The first reason to worry is that AI can make it faster and cheaper to draw inferences from data. It can automate certain tasks that you would normally need human analysts to perform. Picking out individuals in surveillance footage would be one example. To stress the "scariness angle" here, the Stasi relied on hundreds of thousands of employees and part-time informants. Imagine if it was possible for contemporary authoritarian countries to do what was done in East Germany for only one hundredth of the cost.

The second reason to worry is that AI can enable new inferences from data. AI can allow people to learn things from data that they otherwise wouldn't easily be able to learn. One simple illustration of this phenomenon is a study, from a few years ago, that involved giving a lot of people personality tests, looking at their Facebook profiles, and then analyzing correlations between their test results and the pages they like on Facebook. The researchers then used the data they collected to train a model that can predict an individual's personality traits on the basis of relatively small number of Facebook likes. At least the headline result from the study is that, if you know about 300 of the Facebook pages that someone has liked, then you can predict how they'd score on a personality test about as well as their spouse can. So this an example of using AI to draw new inferences about individuals from fairly small amounts of data.

One way to rephrase the second concern is just to say that AI may make information even more lumpy than it is today. There may be contexts where you're sharing what seems to you like a relatively small or narrow piece of information, but AI makes it

possible to draw far-reaching inferences from that information that unaided humans would otherwise struggle to make.**[5]**

Institutional Privacy: Hopes

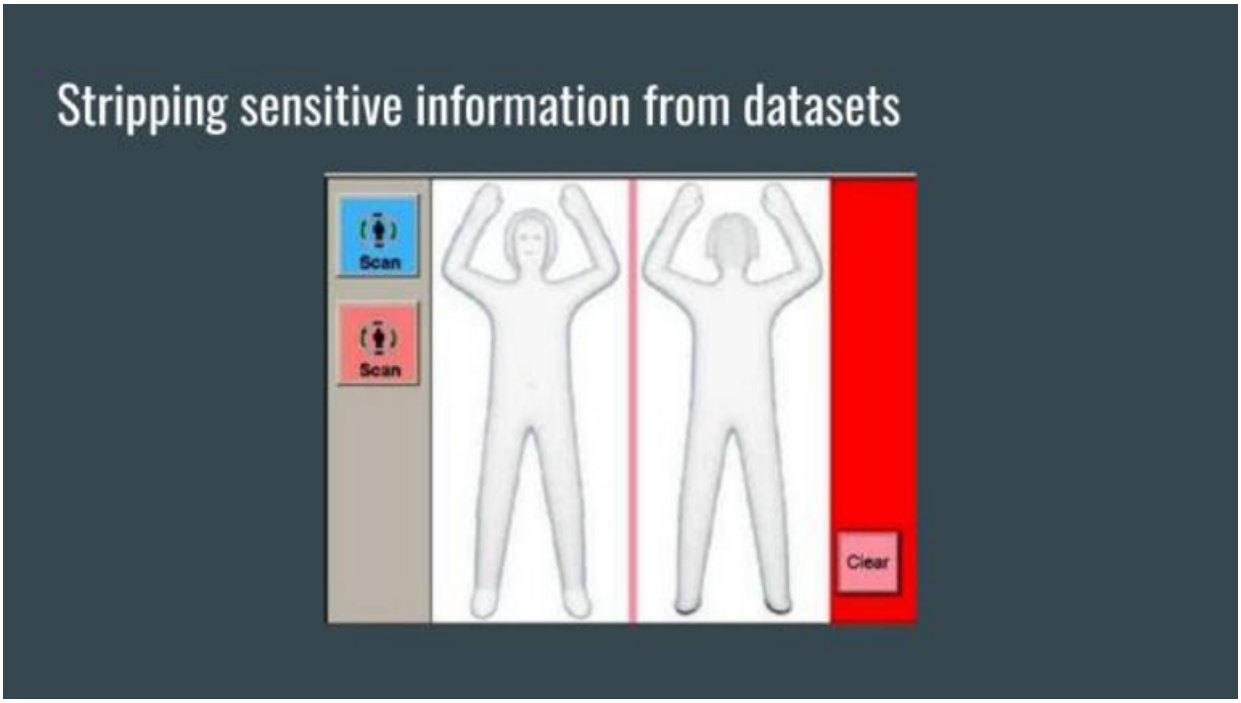
My view is that it's very reasonable to worry about where things are going. I think that the effect that AI has had on institutional privacy so far has almost certainly been negative. At least in the near-term, I think that effect will probably continue to be negative as well.

At the same time, though, I do think that there are also some ways in which AI can help to support or safeguard institutional privacy. I think a lot of the applications are promising this regard have only really begun to be seriously explored in the past few years, so have yet to leave major footprints. I'm first going to walk through a few key varieties of applications. Then I'm going to make a speculative case for optimism. I'm going to argue that if certain technical challenges are overcome, and these kinds of applications become sufficiently widely adopted, then progress in AI may eventually allow us to partially *reverse* the decline of institutional privacy.

Hope #1: Stripping Sensitive Information from Datasets

Here's the first way that artificial intelligence can help support institutional privacy: It can be used to strip sensitive information from datasets, before humans look at them.

As one example of this sort of application, the police equipment company Axon is currently using AI to automatically detect and blur faces that are incidentally captured in body camera footage. They're essentially using facial recognition software to not draw out additional information from the footage, but rather to *remove* information that it would otherwise be very costly to remove. A similar example is using image and text recognition software to automatically redact sensitive information from documents and ensure that nothing has accidentally been left in.



Another illustrative example, although this may not deserve the classification “AI,” is the software used to process data from airport body scanners. Just to give some context: About a decade ago, you may remember, airports started introducing these full-body scanners. You’d step into them, and then TSA agents would see, on their screens, essentially a pretty high-resolution scan of your naked body. A lot of people were pretty understandably concerned about that invasion of privacy. The TSA then responded to these concerns by writing software that essentially takes the raw scans and then transforms them into much more abstract and cartoonish images that don’t reveal personal anatomy. The software now allows them to learn whether people are carrying weapons without also learning what these people look like naked.

Hope #2: Reducing the Need for Humans to Access Data

Here’s a second way that AI can help support institutional privacy: It can help to automate information-rich tasks that human employees would otherwise need to perform.

To use an analogy, let’s return to the bomb-sniffing dog case. A bomb-sniffing dog can be seen as a “tool” for automating the process of analyzing the contents of a bag and determining if there’s anything dangerous inside. Because we can automate the task, in this way, it’s not necessary for any human to “access” the “raw data” about what’s inside the bag. Only the dog accesses the raw data. The dog has also essentially been “designed” not to “leak” the data to humans.

So if you’re automating the process of responding to customer requests, or the process of looking for signs of criminal activity in some dataset, then this reduces the need for any humans to actually look at the data that’s been collected. This may somewhat reduce concerns about individual employees abusing personal data and concerns about personal data being leaked. It may also just reduce the additional sense of “violation” that some people have, knowing that an actual human is learning about them. I think that most people tend feel a lot more comfortable if it’s just a dog or some dumb piece of software that’s doing the learning.

Hope #3: Reducing the Need for Data Collection

Here’s a third and final way that AI can help support institutional privacy: It can reduce the need for institutions to collect data in the first place.

There’s an obvious objection that can be raised to the points I just raised about automation. Even if humans don’t *need* to look at certain data, they may still *choose* to. For example, Google Docs is a highly automated service. No Google employee *needs* to look at the personal diary that you keep in a doc file. But you may still have a sense of personal squeamishness about the fact the diary is being stored as a readable file on a server Google owns. You might have concrete concerns about your privacy one day being violated, for instance if a major data breach ever occurs. Or you might just dislike the thought, however hypothetical, that someone *could* look at the diary.

These sorts of concerns can become a lot more serious and well-justified if a less well-established company, with weaker incentives to maintain a good reputation, is collecting your data. They can also, of course, become quite serious if your data is being held by a corrupt government or any organization that could plausibly use the information to coerce you. So automation only really gets you so far. What would really be nice is if institutions didn’t need to collect your data in the first place.

As it turns out — although this may initially sound like magic — it actually is possible to process data without collecting it in any readable form. There are a number of relevant techniques with wonky names like “secure multiparty computation” and “homomorphic encryption.” These techniques are all currently extremely slow, costly, and difficult to use. This explains why they’re still so obscure. But over the past decade a lot of progress has been made on the challenge of making them practical. We still don’t know just how much more progress will eventually be made.

Ultimately, progress in this domain and progress in AI could interact to enhance privacy. Progress on techniques like secure multiparty computation could make it possible for institutions to create and use certain AI systems without collecting data. And progress in AI could create more opportunities to apply techniques like secure multiparty computation. These techniques are of no real use, of course, if it’s still human beings doing the data processing.

Exploring Best-Case Scenarios for Data Collection

It really is important to stress, again, that techniques for privacy-preserving computing are still *wildly impractical* to apply in most cases. There are currently associated with very serious practical limitations. You are absolutely not going to find them being used everywhere next year. You also probably won’t see them being used everywhere next decade.

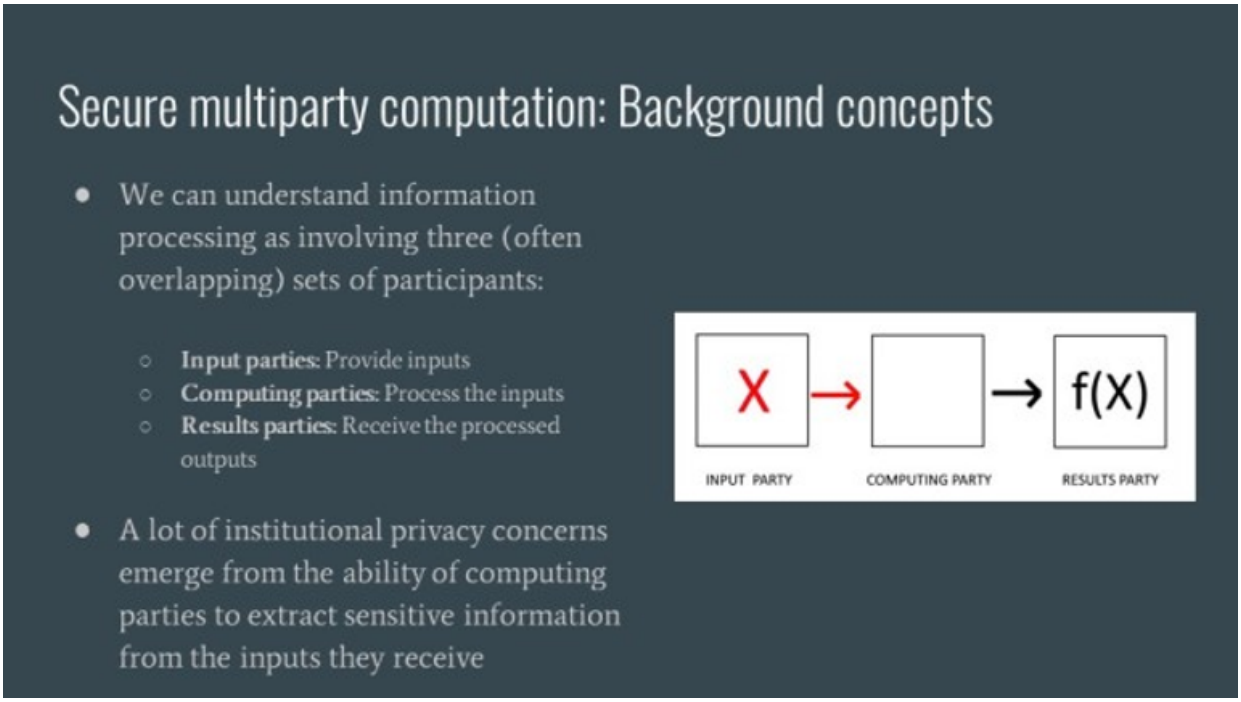
My understanding is that no one really knows exactly how efficient and user-friendly these techniques could ultimately get. It’s possible they’ll never get good enough to matter very much. But they might also one day get good enough for their pervasive adoption to become practical.

From a privacy optimism perspective, I think it’s useful to consider the best-case-scenario. If the practical limitations can mostly be overcome, then what’s the most that we could reasonably hope for?

To probe this question, I’m going to focus on one particular technique called “secure multiparty computation” (MPC). There still aren’t very many examples of MPC being used in the wild. It also has a pretty boring name. At the same time, though, it has at least two appealing properties. First, at the moment, MPC is at least _more practical_ to apply than a number of other techniques. Second, the theoretical limits on what MPC can accomplish are extremely loose. If enough practical limitations are overcome, then the “ceiling” on its long-run impact is especially high.

Secure Multi-Party Computation: Background Concepts

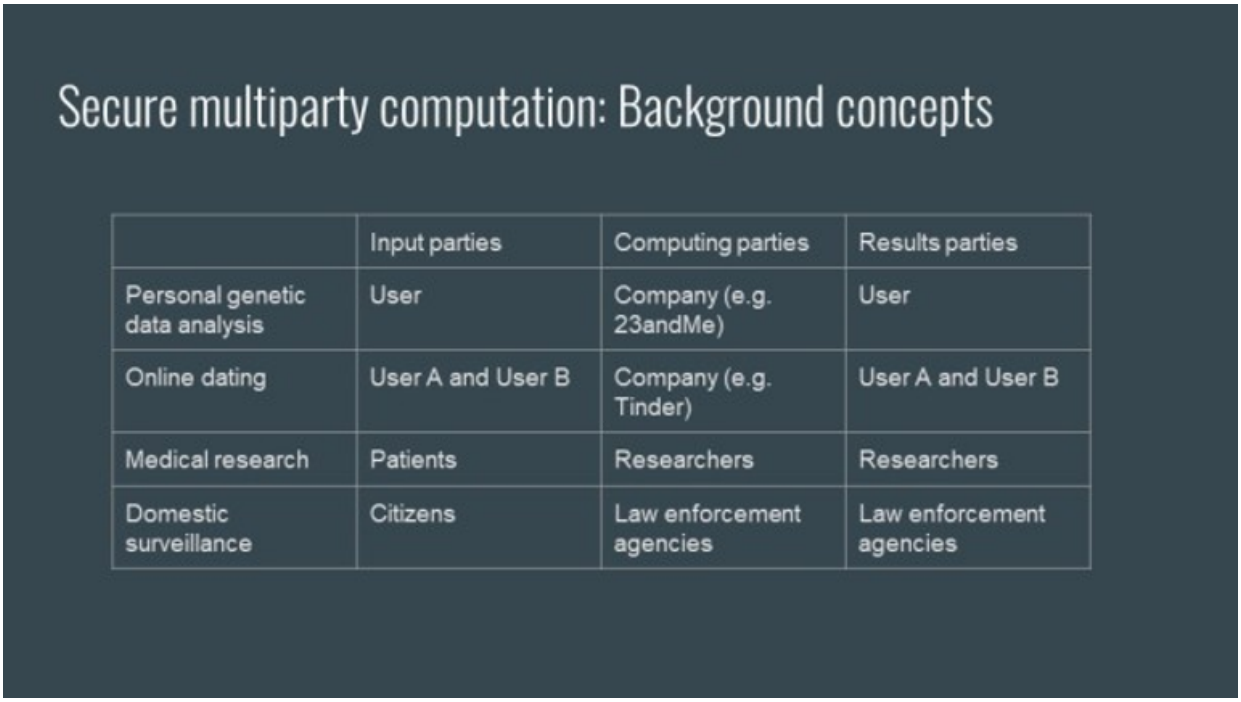
Before explaining exactly what “secure multiparty computation” is, I sort of need to introduce a few basic background concepts.



We can think of any given information processing task as involving three (often overlapping) sets of participants. First, there are **input parties** who provide inputs to be processed. This might be, for example, two Tinder users inputting their “swipes” into the app. Or it might be a group of patients whose data is being used to train a medical diagnosis system.

Then there are the **computing parties**. These are the parties who process the inputs. In the first of the two cases just mentioned, Tinder would be the computing party. It’s computing whether the two people have matched. In the second case, the computing party would be the medical team training the AI system.

The people who receive the outputs of the computation are the **results parties**. The two Tinder users are again the results parties in the first case. The team developing the system is again the results party in the second case. They’re the ones who end up with the trained AI system, which is what constitutes the output in this case.



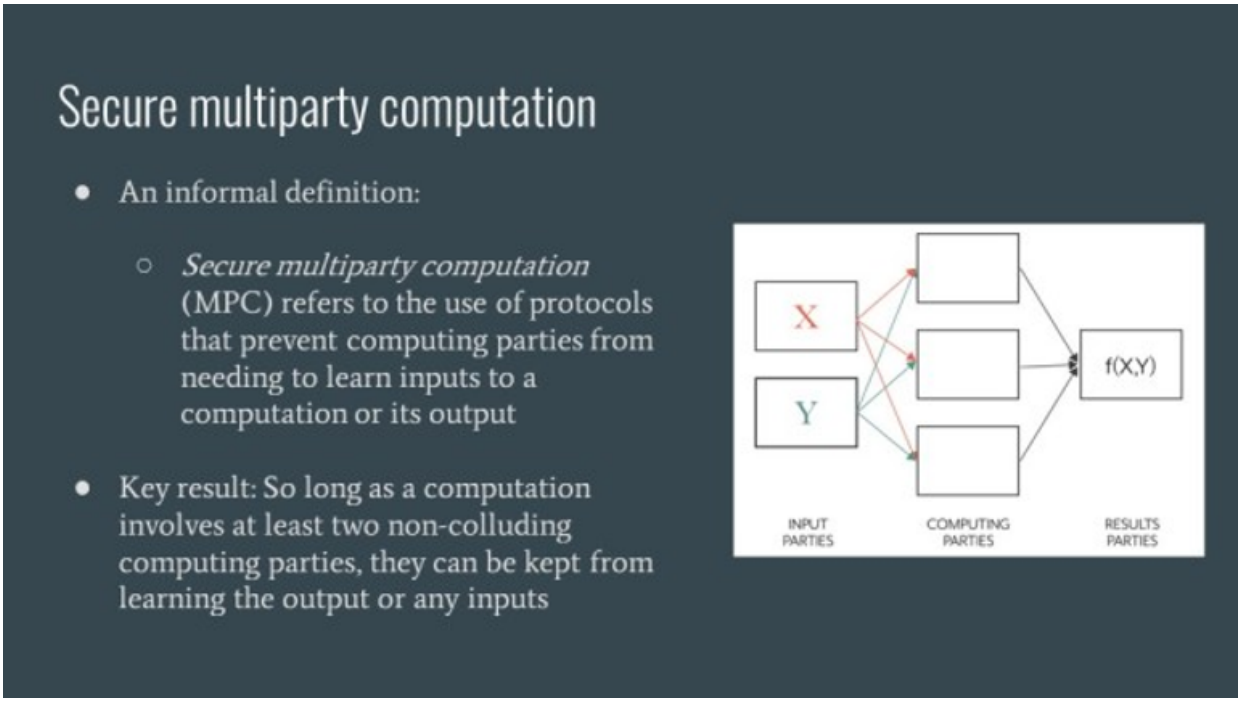
To just to make the idea totally clear, here's a little chart with four different cases laid out. You can also see how the breakdown works in the case of domestic surveillance and personal genetic data analysis.

Privacy issues basically arise when information travels from one class of parties to another. For example, Tinder ends up with information about who their users have swiped left and right on. The medical researchers end up with information about the patients who have provided their data. In general, institutions that serve as computing parties are often in a position to extract a lot of information from the inputs they receive and the outputs they send out.

Secure Multi-Party Computation: An Informal Introduction

Okay now. Preamble over. What **secure multiparty computation (or MPC)** refers to is: a set of protocols that prevent computing parties from needing to learn the inputs or outputs of a computation.

And the key result here is that: So long as a given computation involves at least two computing parties, and so long as at least one of them is correctly following the expected protocol, then it's not in principle necessary for the computing parties to ever gain access to the inputs or outputs. They can instead receive what looks to them like nonsense, process this apparent nonsense, and then send out a different form of apparent nonsense. The results parties can then reconstruct the correct output from the apparent nonsense they receive. The computing parties never need to be in a position to learn anything about the inputs and outputs.



It doesn't matter what the computation is. The result always holds. In principle, it is possible to perform *any* computation using MPC.

Let's return a couple concrete cases. Let's say two groups are cooperating to analyze medical data. If they split the work of analyzing it, then they can use an MPC protocol to learn whatever they need to learn without actually collecting any of the medical data in a legible form. Likewise: If two different companies – or two different groups within a company – cooperate to process dating app swipes, then it's possible to tell users if they've matched without knowing which way each user swiped or *even knowing what you're telling them*.

This may sound like magic. But it actually *is* possible.

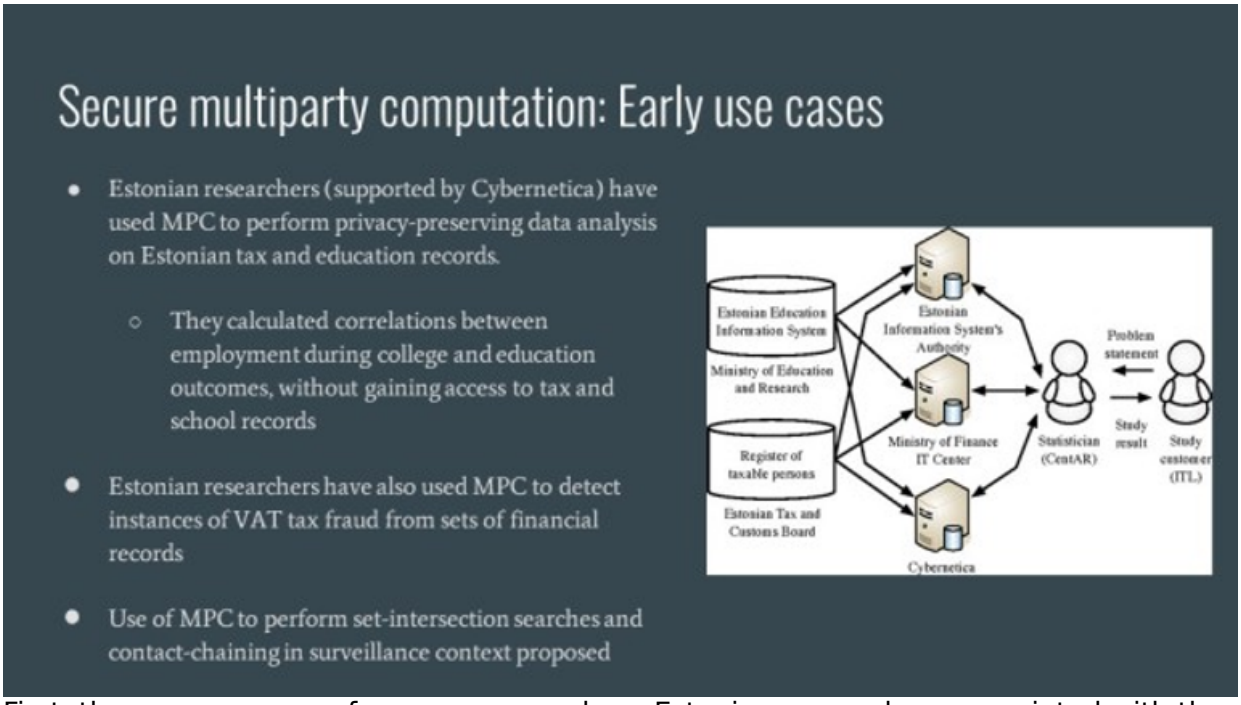
The core assumption that we need to make, when using MPC, is that at least one of the computing parties is following the expected MPC protocol honestly. If all of the computing parties collude to break from the protocol, then they actually can learn what the inputs and outputs are. If at least one party is honest, however, then this can be enough to keep them all entirely in the dark. Given the involvement of enough independent parties, or given your own partial involvement in the computation, the privacy guarantee here can ultimately become extremely strong.

So here's a general strategy for providing privacy-preserving digital services: Involve multiple parties in the provision of the service. Use MPC. Then, so long as at least one of the parties is honest, it won't be necessary for any of them to be in a position to learn anything about what they're sending and receiving.

Secure Multi-Party Computation: Some Case Studies

Again, it's important to stress that it's not yet *practical* to use MPC in most cases. The "overhead" associated with MPC is still pretty enormous. User-friendliness is also still very low. So I'm mainly just talking right about what's in principle *possible in principle.*

Nonetheless, thanks to various improvements that have been made over the past decade, MPC has started to find its first practical applications. These applications aren't necessarily earth-shattering. But, just to ground my discussion of MPC a bit more, I do think it's worth taking the time to describe a few of them.



First, there was a case a few years ago, where Estonian researchers associated with the company Cybernetica wanted to know whether working while you're in college makes you less likely to graduate. They knew about the existence of a couple of datasets that could be used to answer this question. One dataset was held by the Ministry of Education. That dataset contained information about who was in school in what years and who graduated. Then there was another dataset held by the Estonian equivalent of the IRS, which contained information about who was doing part-time work in what years. It's pretty obvious how you could put these two datasets together to see if there was a relationship between working and not graduating.

The issue, of course, is that both of these datasets are extremely sensitive. Researchers certainly shouldn't be given free access to them. The datasets also shouldn't be combined into some master dataset. They're held by different government agencies for good reason. The researchers' clever solution to this problem, then, was to use MPC. They performed the necessary computation in conjunction with the two government agencies. By using MPC, they were able to calculate the correlation between working and not graduating without getting access to any of the data sets or requiring the datasets to be combined.

Another similar case, associated with the same Estonian research group, is one where researchers used MPC to detect instances of value-added tax fraud from sets of financial records. Essentially, they performed a computation over different privately held sets of financial records and detected discrepancies that were suggestive of fraud. The group didn't need to collect any of the records to do this.

A final example is the use of MPC to perform "set intersection searches." These are essentially searches that produce a list of individuals who show up in multiple datasets of interest. It turns out that set intersection searches are used in a number of law

enforcement contexts. One specific case I want to focus on comes from a paper by Aaron Segal, Bryan Ford, and Joan Feigenbaum. Let's say that some unidentified person has robbed multiple banks in different towns. You are the FBI, and, to make progress on the case, you'd like to generate a list of people who were in the right town on the day of each robbery. To do this you might want to collect cell records and run a set intersection search on them. You want to see if anyone made calls in the right towns on the right days.

The way you would traditionally do this is to collect call logs from multiple cell towers. The obvious privacy downside here, though, is that these logs contain *a lot* of phone calls. By collecting them, you've put yourself in a position where you could easily learn a lot of information about a lot of different people. A potentially better option, then, is to use MPC. You can carry out the set intersection search jointly with the holders of the different call logs. Or, in theory, you could carry it out jointly with other independent government bodies. This allows you to get out the list of common phone numbers without collecting any of the underlying data in a legible form.

My impression is that this sort of thing has not yet been done by any law enforcement agency. The idea has only been proposed. But it does seem to already be feasible today.

MPC and AI: Long-Run Implications

Let's step back again from this handful of concrete, present-day cases. From a long-run-oriented perspective, I think there are a couple especially key points to stress.

The existence of MPC protocols implies that, *in principle*, training an AI system does not require collecting or in any way accessing the data used to train it. Likewise, *in principle*, applying a trained AI system to an input does not require access to this input or even to the system's output.

The implication, then, is this: Insofar as an institution can automate the tasks that its members perform by training AI systems to perform them instead, and insofar as the institution can carry out the relevant computations using MPC, then in the limit the institution does not need to collect *any information* about the people it serves. I'm talking about no "lumps" at all.

There is essentially no principled ceiling on institutional privacy.

A Case for Conditional Optimism

Coming back to the idea of optimism: I think that this analysis essentially suggests an optimistic story with a *big* conditional. If major practical limitations can be overcome, then artificial intelligence and privacy-presenting computing techniques together could dramatically reduce the practical need for institutions to learn about the people they serve.**[6]**

In this hopeful world, dishonest excuses for excess information collection would likewise become increasingly untenable. Pure abuses of power would, at least, become much more difficult to spin as something other than what they are.

We don't currently know how far we can move in the desired direction. Major progress would be required. But researchers have already come a long way over the past few decades of work on artificial intelligence and privacy-preserving computing. It's really

hard to justify any sort of confident cynicism here, especially if we are thinking in terms of decades-long or even centuries-long timescales.

Summing Up

Summing up this section: I think one reasonable guess is that, like many past technologies, artificial intelligence will tend to enhance social privacy and erode institutional privacy. However, in the long run, AI and other associated technologies might also be used to protect or even increase institutional privacy. If certain major practical limitations can be overcome, then it is possible to envision a world where "practical necessity" no longer requires institutions to learn very much at all about the people they serve.

Conclusion

Summing up all the sections of this talk together: I think that the historical effect of technological progress on privacy, while certainly very mixed, has been much more positive than standard narratives suggest. It's enhanced our privacy in a lot of ways that typically aren't given very much attention, but that do hold great significance both practically and morally.

I think that the long-run trend seems to be one of rising social privacy and declining institutional privacy. Whether or not this corresponds to a "net improvement" depends on the strength of institutional safeguards. So, insofar as technology has given certain people more "overall privacy," technology has not done this on its own. Good governance and good institution design have also been essential.

One might expect AI to continue the long-run trend, further increasing social privacy and further decreasing institutional privacy. I don't think it's unreasonable, though, to hope and work toward something more. It's far too soon to rule out a future where we have both much more social privacy and much more institutional privacy than we do today.**[7]**

In short: You don't need to be totally nuts to be an optimist about privacy.

Endnotes My Talk Did Not Actually Have

[1] This talk carefully avoids taking a stance on a number of complicated ethical questions around privacy. For example, I'm not taking a stance on whether privacy has intrinsic moral value or whether privacy only matters because it fosters certain other valuable things like freedom, happiness, and security. I'm also not taking a stance on when it is and isn't ethically permissible to violate someone else's privacy.

Probably the most questionable assumption implicit in this talk is that privacy tends to be a good thing. Of course, we can all agree that privacy isn't *always* a good thing. As one especially clear-cut example: Someone should not be able to keep the fact that they're a serial killer a secret from the government or from the community they live in. Horrible murders aside, I do tend to have a pretty pro-privacy attitude in *most*

contexts. But there is also a far-from-nuts case to be made that harmful forms of privacy are much more pervasive than most people think.

In general: The less people know about each other, the less able they are to make informed decisions about how to interact with one another. If you're going on a date, knowing whether or not someone is a serial killer is obviously helpful. But so is knowing how former partners feel about them, knowing what their health and finances are like, and knowing whether they quietly hold any political beliefs that you find abhorrent. It's considered socially acceptable to keep this sort of information private, at least from people you don't yet know very well, but there is some sense in which keeping it private reduces the level of "informed consent" involved in social interactions. There's a big economics literature on the various inefficiencies and social ills that can be produced by "private information."

So that's one broad way in which privacy can be harmful. People's decisions to keep certain information private can cause external harm. I think that sometimes, though, the pursuit of at least social privacy can also hurt the pursuer. Here's one simple example. Given the choice, most college freshmen will prefer to live in a room with only a single bed rather than a room they share with someone else. Privacy is presumably one of the major motivations here. But living in close quarters with someone else can also foster a sort of intimacy and friendship that it's otherwise pretty hard to achieve. I think that sometimes decisions to pursue social privacy in certain settings can have long-run interpersonal costs that it's easy to underestimate.

For these two broad reasons, all cultures have at least some norms that are meant to limit privacy. The character of these norms varies a lot from place to place. If you drop in on a randomly chosen tribe, you'll probably find they have a pretty different orientation toward privacy than you do. I generally feel like the more I think about privacy, and the more I consider the cross-cultural variation in orientations toward privacy, the more uncertain I become about just exactly which expansions of privacy are worth applauding. This stuff is confusing.

[2] Of course, the boundary between "practical necessity" and "abuse of power" is often heavily contested. If you pick any given government surveillance program and ask both a government official and someone writing from *The Intercept* to classify it, you will tend to get different classifications out.

A lot of cases are in fact very blurry. Suppose that some free application collects a lot of user data, which is then used to target ads. The company may say that it's "practically necessary" to do this, because they otherwise could not afford to offer the relevant service for free or to invest in improving it. The company may also say that ad targeting is itself a service that users benefit from. On the other hand, it's probably the case that at least some subset of users will be upset about the loss of privacy. Almost no users on either side of the issue will have a good understanding of how their data is actually being used, whether a more privacy-preserving version of the app would actually be economically viable, and so on. Questions about whether users are meaningfully consenting to privacy losses and about what consent implies are also tricky. Overall, the right label for any given case is bound to be controversial.

[3] When I say "artificial intelligence," I essentially mean "machine learning."

[4] I think that social norms can play an important role in protecting against potential threats to social privacy. If there's a strong social consensus that people who use some privacy-reducing tool are jerks, then this can actually reduce how much the tool is

used. Google Glass is one famous example of a threat to social privacy that was stopped in its tracks by anti-jerk norms.

Of course, the harder it is to find out that someone is using a tool, the weaker the effect of social norms will tend to be. But I think that social norms can still have important effects even when violations of these norms are tricky to catch. For example: Most people will refrain from searching through the folders of their friend's computer, even if they've left it unlocked and unattended while running out to the store.

[5] A fourth concern is that, since certain useful AI systems are created by "training" them on fairly personally revealing data, the rise of AI is introducing additional incentives to collect this sort of data.

[6] One survey of AI researchers suggests that the vast majority of people in the field believe that AI systems will eventually be able to do anything that people can. However, it goes without saying, this could ultimately take a very long time. The median respondent thought there was about a 50/50 chance that the complete automation of all jobs will become technically feasible within the next 125 years.

[7] The OpenMined community is one group currently working toward this future. They focus primarily on increasing the accessibility of existing but little-used techniques for privacy-preserving machine learning.

The Hammer and the Dance

This is a linkpost for https://medium.com/@tomaspueyo/coronavirus-the-hammer-and-the-dance-be9337092b56

I'm currently running the Effective Altruism Coronavirus Discussion FB group. I saw a large number of posts each day, but this seems like the most important post for people to read at the moment because in order to know what to do, we need to have an idea of the end-game. Now obviously we need to discuss this post and consider possible flaws, ideally receiving input from epidemiologists, but I think this looks pretty promising.

Coronavirus Justified Practical Advice Summary

Two weeks ago Ben Pace and I asked for practical advice on what to do about coronavirus, with the requirement that advice be justified in some way- ideally full blown models informed by empirical data, but at a minimum, an explanation of why it was helpful. The resulting thread had wonderful advice but makes for, uh, inefficient reading, at best. This post is an attempt to take the best of the Justified Practical Advice Thread and present it in a clear, easy-to-digest format.

It's important to note that neither the original post nor this one are attempting to give *comprehensive* advice. If you're looking for that, I suggest checking out the Guide/FAQs/Intros tab on the LessWrong Coronavirus Link Database. This is for advice that fell through the cracks.

Though I will briefly mention the standard advice. From what I can tell, the most important (and luckily widely publicized advice) is the following:

- Wash your hands a lot, and wash them to a hospital standard (spend 20 seconds and follow these instructions).
- Socially isolate yourself, and don't be around other people unless really necessary. Avoid groups of people, especially large gatherings. Really don't go to bars, clubs, restaurants or other places that lots of people pass through (like public transportation).
- If you are sick, use a mask. They are probably also effective at preventing the disease when you are not sick, but most of the western world currently has massively limited supply, so leave the masks to health workers and sick people.

The rest of this post will summarise the interesting ideas that went beyond these basic recommendations. But if you haven't done (or at least seriously considered doing) all of the above, I would recommend you prioritise the things listed above.

Many of the below recommendations involve buying things. When possible I've included a link to a particular amazon page. This is not a strong recommendation for that particular product: it's an attempt to lower activation energy for acting on a recommendation. If you have reason to believe a particular model is better than what I linked to, please comment and I'll update if appropriate. Speaking of which, if you think of anything else I missed or got wrong, please comment with that too.

That said, here is the top advice from the Coronavirus Justified Practical Advice thread.

- Cover your high-touch surfaces with copper tape.
- Treat newly delivered packages as contagious for 48 hours.
- Take vitamin D supplements daily.
- Buy electrolytes to drink if you get ill.
- Maybe: Buy a pulse oximeter.

Practical Advice

Get 2-inch copper tape and cover high-touch surfaces with it, especially shared high-touch surfaces (Connor Flexman)

This was the clear winner of JPA Olympics– it's cheap, effective, and very few people had heard of it before. In fact we were hoping for a highly polished, intensely researched post on just this; unfortunately no one has the time right now.

In its place, please enjoy these links:

- This pre-print showed that this coronavirus in particular had a half life of 2.4-5.11 hours on copper, in contrast to 10.5-16.1 on steel or 13-19.2 on plastic
- This review showed H1N1 decreased by 4 logs (a factor of 10^4) in 6 hours;
- This study showed vaccinia and monkeypox viruses reduced by 6 logs (a factor of 10^6) in 3 minutes
- This study showed murine norovirus was destroyed in 30 minutes, though it doesn't work very well at 4C;
- This review says that copper oxide filters neutralize all of "bacteriophages [58-62], Infectious Bronchitis Virus [63], Poliovirus [61,64], Junin Virus [59], Herpes Simplex Virus [58,59], Human Immunodeficiency Virus Type 1 (HIV-1) [11,65-67], West Nile Virus [11], Coxsackie Virus Types B2 & B4, Echovirus 4 and Simian Rotavirus SA11 [68]. More recently, the inactivation of Influenza A [55,65], Rhinovirus 2, Yellow Fever, Measles, Respiratory Syncytial Virus, Parainfluenza 3, Punta Toro, Pichinde, Adenovirus Type 1, Cytomegalovirus, and Vaccinia [65]".

Benefits

- Reduces surface-to-hand-to-face transmission

Costs

- $11, half an hour to apply tape (this may be a recurring cost, but we don't know at what interval)
- Risk of cuts while applying tape (this happened to me once)
- Risk of cuts from applied tape (0 for me, but other people have reported lots). This is not just a convenience issue- small cuts are breaks in your defenses that microbes can enter.

Why you might not do this

- You are worried about the risks of exposure to copper
    - You might be are allergic
    - Some people have reported hand irritation and discoloration when they put it on their phones and laptops
    - I poked at this and my takeaway was you'd need to be hooked up to skin infusion IV to get as much copper as you consume via food. Additionally copper IUDs are a common form of birth control, so it can't be *that* dangerous.
- You think surface-to-hand-to-face transmission is not significant.
- You don't think the results of tests on other viruses apply to coronavirus

The 2’’ is from personal experiences: I got a variety of sizes up to 1’’, and even that was often inconveniently small. I basically never wanted smaller except to patch up gaps, and they’re not so much better than just tearing a small piece off of 2’’ for that.

Treat newly delivered packages and mail as contagious for 48 hours (Elizabeth)

At the time I posted this, we had very little information on the risks of contamination via packages. I based my recommendation on studies showing other coronaviruses survived for a very long time on other surfaces. Since then someone has released a pre-print on this specific coronavirus on cardboard in particular, which found a half-life of ~2-5 hours. What that means your safety depends on exactly how much your mailman sneezes on it and what the infectious threshold is, which we don’t actually know. However at the concentrations that paper was using, concentration dipped below detection by 48 hours.

Here are the specific recommendations for handling potentially infected packages:

- Do not touch your face between touching a package and washing your hands or taking off gloves.
- Spray deliveries with disinfectants.
    - Note that disinfectants are typically less effective on porous material like cardboard. I know of no hard numbers on this.
- Leave packages exposed to the sun for 1-3 days (this can probably be shortened given the newest information) or a UV lamp for shorter
- Just don’t touch them for several days.

Benefits

- Reduces package-to-hand-to-face transmission

Costs

- Adds time and annoyance to handling packages
- UV requires a sterilizer
- Leaving outside may risk theft
- Letting packages sit takes time and space.
- Disinfectants are pennies/spray.

Why you might not do this

- You think surface-to-hand-to-face transmission isn’t important.
- You’re extremely confident your delivery person is not contagious.
- You think there is little enough initial contamination pre-delivery that it will have degraded by the time the package reaches you.

Take vitamin D supplements (Connor Flexman)

Vitamin D deficiency has a surprisingly strong link with respiratory infection.

- Study says 4x rate of respiratory infection in the very deficient, but doesn't see an obvious effect in the partially deficient, so slightly weird statistics.
- Study says very large effects in children
- WHO says vitamin D supplementation may reduce respiratory infections in children
- Study says no effect from supplementing after already sick, so get on this before infection

While this hasn't been verified for COVID-19 in particular, it does seem plausible that it applies, and that a supplement can treat it.

Benefits:

- Reduces respiratory infections, via unclear mechanism.

Costs:

- 2.5 cents/day using the brand I recommended (recommended to me by a doctor but not otherwise verified), cheaper probably available.

Why you might not do this:

- You've tested your vitamin D recently and levels were normal or high (as a fat soluble vitamin, it is possible to overdose on)
- You don't think the results from general respiratory infections transfer to COVID-19 in particular.
- You think you get enough vitamin D from the sun, and will continue to do so during periods of isolation.
- You don't believe supplements can convey health benefits.

Have electrolytes on hand and use once ill (tragedyofthecomments)

COVID-like illnesses frequently cause loss of electrolytes through sweating and diarrhea, which are not replenished from food due to appetite suppression. Insufficient electrolytes can lead to digestive and neurological problems.

The exact form of this is up for debate: some people feel that salt, maybe with sugar, honey, or molasses, is sufficient. Others thought that the magnesium and potassium from a dedicated electrolyte powder/concentrate/drink, or the ease of consumption while sick, was beneficial. Once you've gone that road, there are lots of electrolyte supplements with additional vitamins and minerals you might also be missing, leading to something of a paradox of choice for me.

No one disagreed that some form of electrolyte is good to drink to once sick, just what exact form that should take. Choosing any one of these is more important than which one you choose.

Forms you might take, with approximate cost:

- A homemade version of WHO's Oral Rehydration Salts ($5 in cheapest form).
  - That, replacing sugar with molasses or raw honey
- Powder ($30)

- Tablets ($25)
- Electrolyte drops ($20) (my choice, for esoteric reasons)

Benefits

- Prevent electrolyte insufficiency, which can lead to neurological and digestive issues

Why you might not do this

- Personal experience hating electrolytes
- Inability to distinguish when you've had "enough" (since too much is also dangerous)

Maybe: Buy a pulse oximeter (juliawise)

Coronavirus moves from unpleasant to dangerous when your blood oxygen saturation level is too low. A pulse oximeter can tell you when that is happening, providing a clearer line for when to seek medical attention, which is especially important as the risks to contact with the medical system rise. However many people felt that you should ignore a low pulse ox if you felt fine or a high one if you felt short of breath, so it didn't add anything to the process, an argument I find pretty compelling.

Benefits

- Certainty in when to seek medical attention, especially if you can't trust your internal sense of being out of breath (due to e.g. panic attacks).

Costs

- $30

Why you might not do this

- You trust your internal sense of being out of breath or not.
- You expect medical care to be unavailable or net-negative no matter what.

Thank you to Raemon and Ben Pace for comments on this document.

Thank you very much to Julia Wise, Finan Adamson, Connor Flexman and Connor Flexman for this justified practical advice.

[Update: New URL] Today's Online Meetup: We're Using Mozilla Hubs

That's right! Tomorrow/Today (Sunday Mar 29th) we're having an online LessWrong meetup in Mozilla Hubs. You're welcome to join and chat with some of the interesting people who read LessWrong, SSC, Overcoming Bias, and so on. I am personally quite excited to talk with a bunch of you. **Make sure you read the rules for joining the hub.**

At 12 noon PDT we'll be having the Hanson/Mowshowitz "Expose The Youth" Debate, then after that at around 2pm we'll having a post-debate meetup. This an experiment, I'll do it more if it's fun.

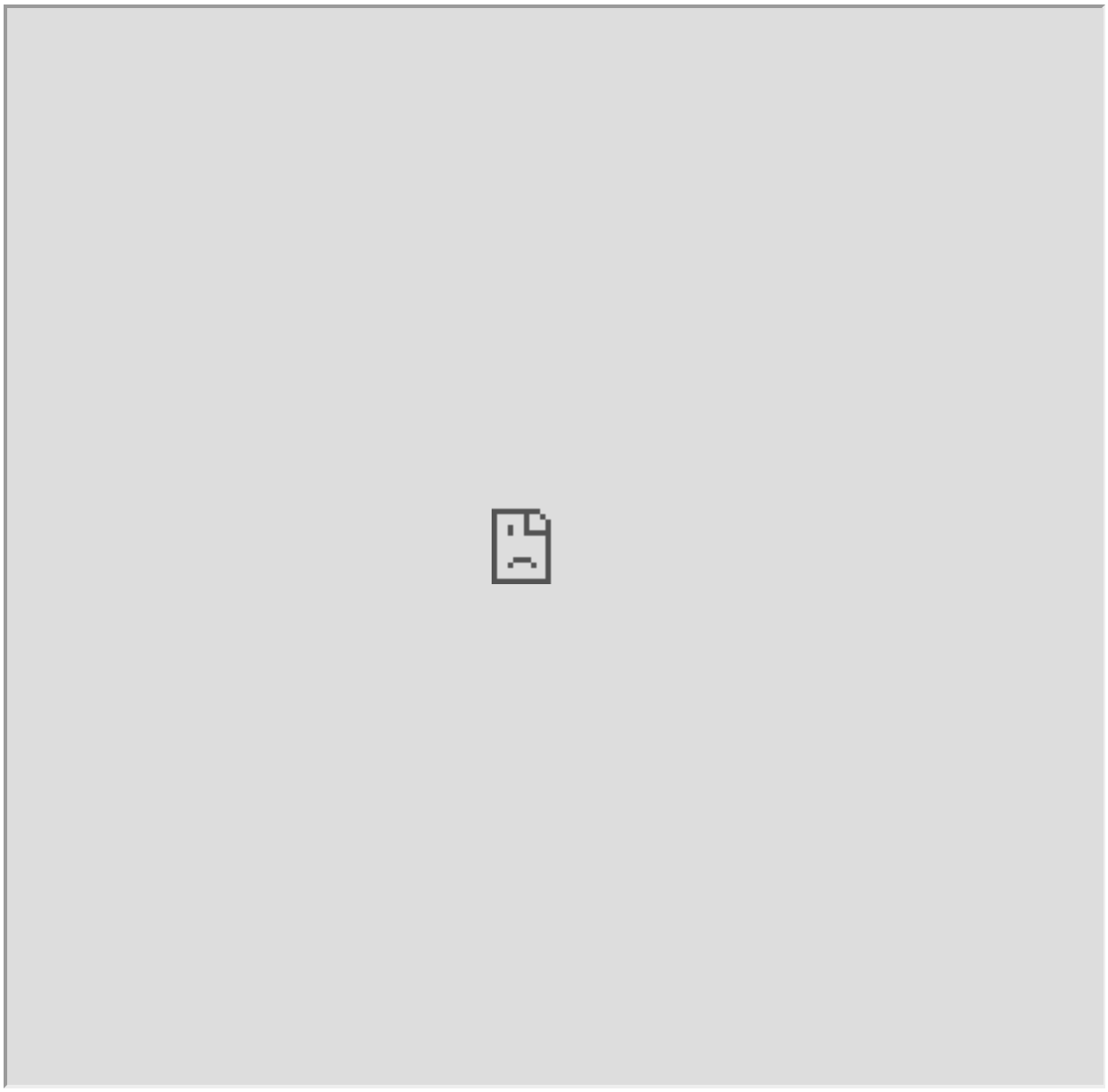
Basic controls:

- WASD to move
- G to toggle flying
- Click and Drag to look around
- You can hear and be heard by people in proportion to how close you are to them

The rules:

- **Only 25 people can be in a room.** Move around and check out the other rooms!
- **Login as soon as you arrive** and use either your real name or a standard alias like your LessWrong username. I'd like some basic accountability please, and today I will remove people who don't follow this rule.
- **Use headphones.** Mozilla Hubs does not have any basic feedback reduction tech, all sound will get unenjoyably recursive quite quickly unless everyone uses headphones.
- Being weird is positively fine and I don't think you have a limited budget to spend, you can just be super weird. But if you're too loud and **if people are having a bad time because of your presence, a moderator will remove you from the rooms.**

I've made a few basic rooms. If you've not visited before or need some help, enter the tutorial room. Here's the first:



Please leave the tutorial room once you've solved your problems, to make space for others.

And here are the hangout rooms.

- Taking Joy in the Merely Real
- Many Worlds Interpretation
- In Favour of Niceness, Community, and Civilization

Extra rooms in case you want them for whatever reason:

- Don't Touch the Floor
- Slack and the Sabbath
- Beyond the Reach of God

For anyone who has trouble with the Mozilla Hubs setup, we also have a backup Discord server that you can join here.

COVID-19 article translations site aims to bridge the knowledge gap

As anyone who spans two cultures know, it's often frustrating to see how wrong the media can get news in one or the other of their domains.

When faced with the elevated stakes of COVID-19 and the consequences of mismanagement imposed upon her family in China, my good friend and colleague took on the challenge of addressing the knowledge gap, setting up covid19readings.com.

The site, and its team of volunteer translators, is dedicated to surfacing quality journalism and first-person accounts of the virus, both to correct skewed perceptions in the English-speaking world and to help humanity at large weather the storm.

I hope it can be of value, at the very least, to the discussion here.

If I interact with someone with nCov for an hour, how likely am I to get nCov?

I'm guessing data is limited here, but a related-related question might be "how likely am I to catch the flu or a few other common diseases by interacting with a victim for an hour?"

What is a School?

Previously: The Case Against Education

You don't know what you've got till it's gone.

In a belated triumph of sanity, schools around the world are closing their doors in the wake of the Coronavirus outbreak.

The debates about schools closing make it clear how schools provide value.

Here in New York City, and in many other places, we are refusing to close to schools until (two weeks after the) last minute because schools *provide free meals to poor children.*

The other stated reason is because if children are not in school, their parents would be unable to work, and some of their parents are healthcare workers or cannot afford any missed paychecks.

(In the case of universities, we also have places like Harvard that took away students' housing and thus left them with no place to live.)

These are good, real concerns. We need solutions to these problems.

But we also now know what real problems are being solved.

We have school because our society believes that one cannot leave children unsupervised or terrible things would happen, for very broad values of children. We also have school because we don't have another way to ensure that children get to eat. And in some narrow cases, we have a very partial fix for a housing crisis.

Clearly, regardless of what the best solutions are, one could goal factor for these problems much better than 'mandatory schooling.'

The concern that I have heard *literally zero people* mention is that closing the schools will prevent children from learning.

I do realize that no one is grappling with how long this is likely to last, and that in the long term they would raise this concern.

But still, I find all of this enlightening. And refreshing.

How Do You Convince Your Parents To Prep? To Quarantine?

A number of my friends are having difficulty conveying the seriousness of COVID-19 to their parents, or getting them to take action instead of panic. What have you found useful in convincing your parents to prep and stay home, and for bonus points, convincing their gatherings to close up?

I'm actually not a great case for answering this- my dad was born 30% prepper. But I think I could have saved a statistical life had I convinced him to convince his church to close a week earlier. And he might have been able to do that sooner had he known the technical solutions available to hold virtual church. So in addition to rousing essays and threats, please consider what tools would make parents' quarantined lives better.

How to have a happy quarantine

As you may have noticed, things in the world are a bit cuckoo bananas right now. Social distancing is becoming increasingly widespread, which means that people are going to be experiencing social isolation at an unprecedented level. In this age of global connectedness, this seems really scary to some people, but I prefer to think of it as something akin to the experience of weathering a snowy winter in a log cabin on the western frontier. That is, cozy and exciting!

I live in a house with about ten people. We all have different personalities and we're all afraid of going stir-crazy spending five months together, so I did some research on how we can avoid that. (We also have a dry-erase monthly calendar where we can plan out activities!) Below are my ideas :)

Epistemic status: these recommendations are based on a combination of (1) things that are required or recommended for astronauts, (2) recommendations for people homebound due to injuries, chronic illnesses, or old age, and (3) common sense.

Body interventions:

- If at all possible, **get fresh air and sunshine** at least once a day.
- Get at least light **exercise** at least once a day; ideally get heavy exercise regularly.
- **Get enough sleep, water, and calories.**
- **Eat a variety of foods that you enjoy!**

Brain interventions:

- **Regularly video chat loved ones** who aren't quarantined with you.
- **Talk to a therapist online**, if possible.
- **Meditate.** There are lots of guided meditations available online and via apps, some of which are free.
- **Stick to a basic routine that you endorse.** For example, you might want to wake up and go to bed at the same time each day, or do daily exercise.
- **Change out of sleep clothes and into day clothes each day.** If you are a newly minted remote worker, this is especially important, and you may want to set up additional "going to work", "leaving work", and "lunch break" routines.
- **Have projects to work on; set goals** that feel meaningful to you.
- **Get around to those things you always meant to do but didn’t have time for**, e.g. learning a certain skill, cleaning out a certain room, or reading a certain book.
- **Make sure you get real alone time**, even if you're sharing a room.
- **Do things that are fun for you!** e.g. watch movies, make art, have sex, read, dance.
- **Gratitude journaling.** Try to write down three or more things you're grateful for each day, or exchange gratitudes with a friend or partner verbally.
- **Plan at least one thing to do each day.** This will give your day structure and purpose!

- **Clean the house regularly** so that your living environment stays fresh and nice.
- **Celebrate birthdays and other holidays just as you normally would**; take these opportunities to dress up nice/put on makeup if it makes you happy.

And remember that everyone else is in a similarly stressful situation to you. Interpersonal conflicts are fairly likely to arise, and you need to do your best to minimize their impact on your household. For some people that might mean you need to walk away from the situation for a while rather than letting it escalate. Maybe you need to talk to a trusted friend and/or try third-party mediation. Maybe you need to let your feelings out in a dance battle. In any case, dealing with conflict in a way that minimizes negative externalities is really important. I'd love to hear additional advice on this.

Big list of activities

Activities you can do alone

- Train to do 100 consecutive pushups
- Read books, fanfiction, anything you want
- Color in coloring books
- Online drawing lessons
- Arts and crafts
- Learn yoga online (I like Yoga with Adriene)
- Do bodyweight exercises (I like the scientific 7-minute workout with the silly song)
- Play with Legos
- Make music
- Write a song
- Write a book
- Play video games
- KonMari your living space
- Redecorate your living space
- Make candles
- Spray painting
- Play Geoguessr
- Online escape room
- Work through a textbook
- Under-desk treadmill and/or under-desk bike pedal thing
- Explore the world in VR
- Create content you can share, e.g. blog posts or YouTube videos
- Do puzzles
- Games
    - Board games one can play for free online:
    - Secret Hitler: .io, and .party (boardgamegeek)
    - Dominion (boardgamegeek)
    - Hanabi (boardgamegeek)
    - Android: Netrunner (boardgamegeek)
    - Resistance and Avalon (boardgamegeek links for Resistance and Avalon)
    - Codenames on a website and on slack (boardgamegeek)
    - Go (wikipedia)

    - Chess (wikipedia)
    - Set (boardgamegeek)
    - Board game arena is a website with a variety of games including Carcasonne, Puerto Rico, and 7 Wonders.

Group activities

- Teach each other stuff (e.g. math, singing)
- Dance parties
- Morning calisthenics
- Show each other cool movies
- Smell identifying contest
- Tasting exercises from the book *Taste*
- Become a choir
- Video game tournament
- Scavenger hunt
- Give each other makeovers
- Contest to build the strongest structure out of something you have around the house
- Learn a synchronized dance
- Sleepover night: make a blanket fort and eat popcorn and watch movies in your PJs
- Karaoke
- Cook-off
- Picnic (can be indoors or in a yard, if you have one)
- Truth or dare
- Hide and seek
- Easter egg hunt (didn't stock up on eggs? that's okay! hide something else!)
- Photo shoot
- Shadow puppet show
- Shakespeare read-throughs
- Improv exercises from *Impro*
- Play banal card games like kids do
- Paper airplane contest
- Board games
- Wrestling
- Badminton
- Quidditch
- The floor is lava
- Collaborative online games
    - Heads Up!
    - Jeopardy
- Opt-in mandatory daily exercise
- Shared meals
- Watch a long TV show together, one or two episodes per day
- Set up passive Skype calls with other houses so it's kind of like we're hanging out in each other's living rooms
- Spelling bee
- Build a Rube Goldberg machine

I welcome other ideas and will add them to the post if you want!

Zoom In: An Introduction to Circuits

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://distill.pub/2020/circuits/zoom-in/

Chris Olah and the rest of the rest of the OpenAI Clarity team just published "Zoom In: An Introduction to Circuits," a Distill article about some of the transparency research they've been doing which I think is very much worth taking a look at. I'll try to go over some of my particular highlights here, but I highly recommend reading the full article.

Specifically, I have previously written about Chris's belief that the field of machine learning should be more like the natural sciences in seeking understanding first and foremost. I think "Zoom In" is a big step towards making something like that a reality, as it provides specific, concrete, testable claims about neural networks upon which you might actually be able to build a field. The three specific claims presented in the article are:

> **Claim 1: Features**
>
> Features are the fundamental unit of neural networks. They correspond to directions [in the space of neuron activations]. These features can be rigorously studied and understood.
>
> **Claim 2: Circuits**
>
> Features are connected by weights, forming circuits. These circuits can also be rigorously studied and understood.
>
> **Claim 3: Universality**
>
> Analogous features and circuits form across models and tasks.

"Zoom In" provides lots of in-depth justification and examples for each of these claims which I will mostly leave to the actual article. Some highlights, however:

- How do convolutional neural networks (CNNs) detect dogs in an orientation-invariant way? It turns out they pretty consistently separately detect leftward-facing and rightward-facing dogs, then union the two together.
- How do CNNs detect foreground-background boundaries? It turns out they use high-low frequency detectors—which look for high-frequency patterns on one side and low-frequency patterns on the other side—in a bunch of different possible orientations.

What's particularly nice about "Zoom In"'s three claims in my opinion, however, is that they give other researchers a foundation to build upon. Once it's established that neural networks have meaningful features and circuits in them, discovering new such circuits becomes a legitimate scientific endeavor—especially if, as the third claim suggests, those features and circuits are universal across many different networks. From "Zoom In:"

> One particularly challenging aspect of being in a pre-paradigmatic field is that there isn't a shared sense of how to evaluate work in interpretability. There are

two common proposals for dealing with this, drawing on the standards of adjacent fields. Some researchers, especially those with a deep learning background, want an “interpretability benchmark” which can evaluate how effective an interpretability method is. Other researchers with an HCI background may wish to evaluate interpretability methods through user studies.

But interpretability could also borrow from a third paradigm: natural science. In this view, neural networks are an object of empirical investigation, perhaps similar to an organism in biology. Such work would try to make empirical claims about a given network, which could be held to the standard of falsifiability.

Why don’t we see more of this kind of evaluation of work in interpretability and visualization? Especially given that there’s so much adjacent ML work which does adopt this frame! One reason might be that it’s very difficult to make robustly true statements about the behavior of a neural network as a whole. They’re incredibly complicated objects. It’s also hard to formalize what the interesting empirical statements about them would, exactly, be. And so we often get standards of evaluations more targeted at whether an interpretability method is useful rather than whether we’re learning true statements.

Circuits side steps these challenges by focusing on tiny subgraphs of a neural network for which rigorous empirical investigation is tractable. They’re very much falsifiable: for example, if you understand a circuit, you should be able to predict what will change if you edit the weights. In fact, for small enough circuits, statements about their behavior become questions of mathematical reasoning. Of course, the cost of this rigor is that statements about circuits are much smaller in scope than overall model behavior. But it seems like, with sufficient effort, statements about model behavior could be broken down into statements about circuits. If so, perhaps circuits could act as a kind of epistemic foundation for interpretability.

I, for one, am very excited about circuits as a direction for building up an understanding-focused interpretability field and want to congratulate Chris and the rest of OpenAI Clarity for putting in the hard work of doing the foundational work necessary to start building a real field around neural network interpretability.

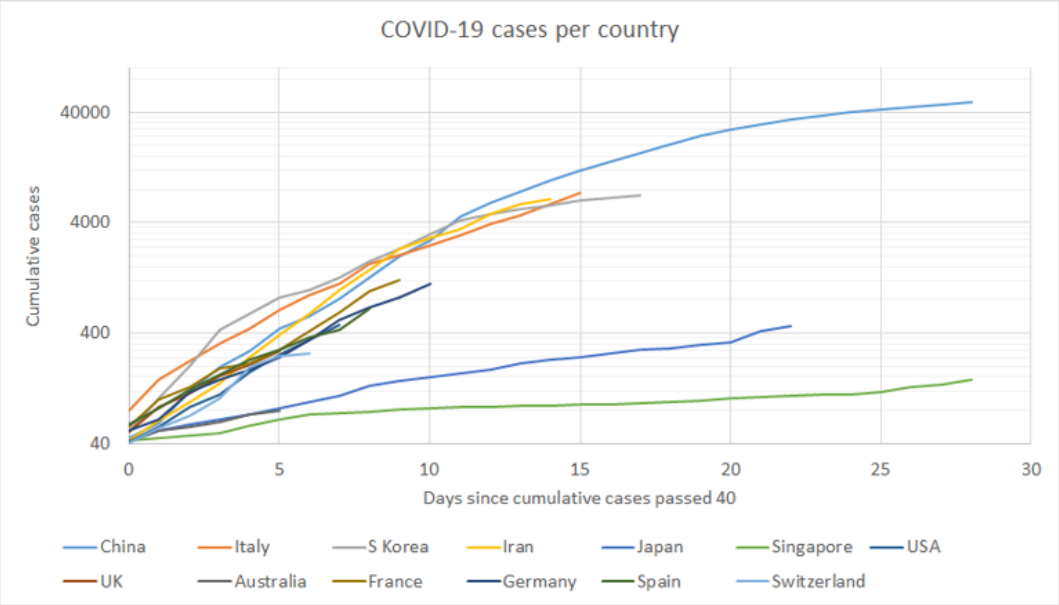
Growth rate of COVID-19 outbreaks

Edit 14/03/2020: The top two graphs are now available as interactive versions here (thanks to Ruby for helping with getting this uploaded). The labels on the right are clickable to remove or add countries (double click selects only that country or all countries). The buttons at the top change the y-axis (annoyingly the y-axis range buttons auto-set to a linear scale) and the slider at the bottom zooms the x-axis.

Note that the doubling times are actually lower than in the post below due to an error in my original spreadsheet. I've also added the last few days worth of data to the graphs.

COVID-19 has now broken out in a number of countries. This enables us to compare spread rates across to get a better idea of what to expect.

Below is a graph of cumulative cases in each country. In an attempt to normalise the x-axis, I have plotted from the day that the total number of cases in the country passed 40 (40 was just because the earliest China data that I had started at 42).



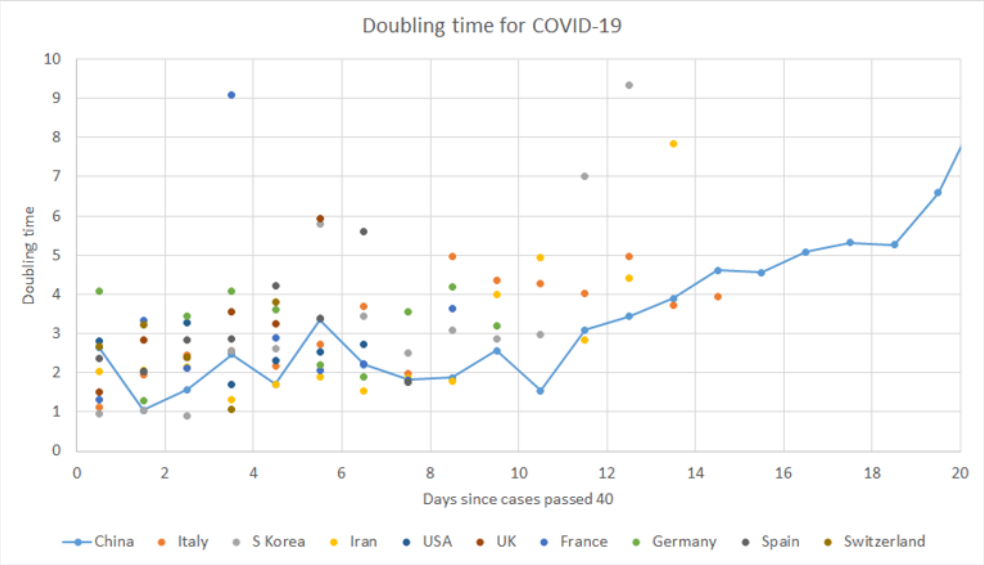
The most obvious thing is that most countries follow a fairly consistent pattern of growth in the first week and a bit.

The outliers are Singapore, Japan and Australia (plus Hong Kong, not shown). These countries have lots of cases yet have not seen a corresponding fast exponential growth in cases. I'm not sure why these particular countries have bucked the trend or whether there is something odd about their reporting (I looked for this but didn't find anything).

I haven't considered how many cases are recovered as it was hard to get reliable results and for most locations recovered cases are minimal. Something weird is happening with the number of recoveries in Iran which has over 2,000, despite only passing that number of cases within the last 6 days.

Doubling time

We can convert the above graph into doubling time:



I've removed the outlier countries for clarity. The doubling time is fairly consistently 2-3 days. It seems to increase slightly over time.

China growth rate

I wrote a post previously about analysing the growth rate of COVID-19 in China.

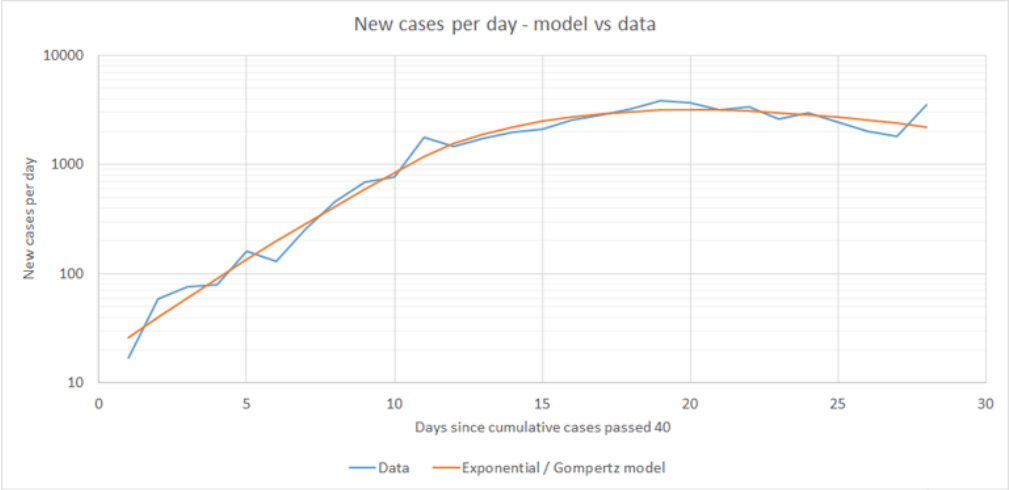
If we look at the graph above, the Chinese rate is roughly constant over the first 11 days, after which the growth rate decreases.

So the first 11 days would fit nicely to an exponential growth model, but what changed? On day 7 (23rd Feb) the quarantine was started. A decrease in growth rate starting a few days later makes sense based on what we know about incubation period.

Let's assume that the model follows an exponential distribution to start with and then after the quarantine starts to be effective it starts to obey a Gompertz function which is like an exponential function with a limit to the total number of cases (thanks to clone of saturn for the pointer here).

I've set both the number of cases and the new case rate to be the same for the two distributions at the point that the Gompertz takes over. This is to minimise free variables so I only have 4 instead of 6.

Getting the best fit parameters for this model I get:



This seems like a fairly good fit. It might be possible to get a better fit with an alternative sigmoid function but this is good enough for my purposes.

Conclusion

I'm fairly confident that, left unchecked, COVID-19 will increase at a doubling time of 2-3 days. When containment in breached in a location this is the rate that the growth occurs at over the first few week or so.

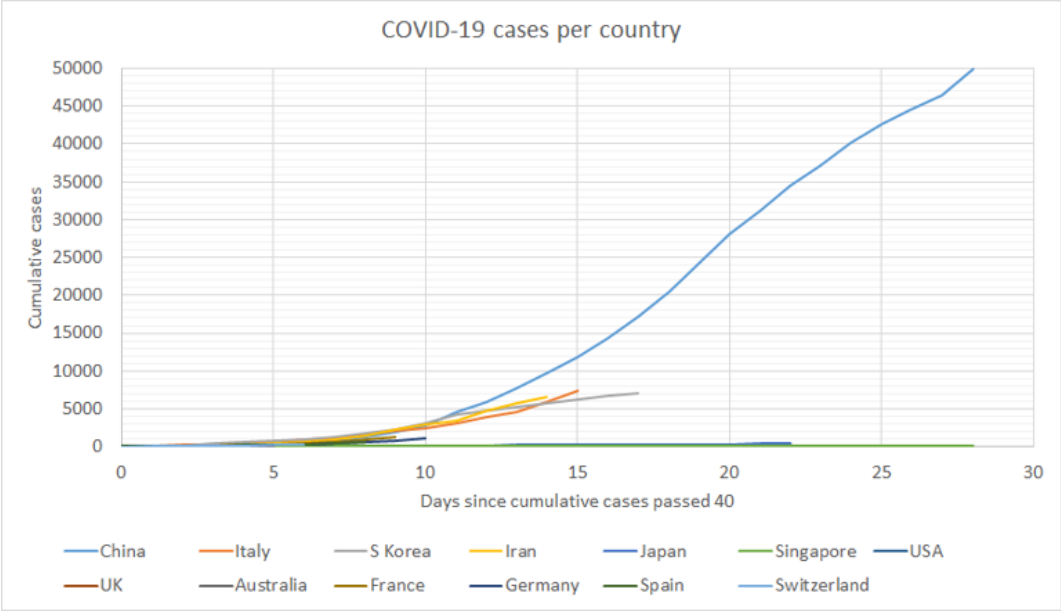
When effective measures are put in place this decreases. An effective quarantine may be able to convert the growth into a sigmoid function with a limit on the failure rate.

Some locations (Japan, Singapore, Australia and Hong Kong) have managed to avoid exponential growth despite having a large number of cases.

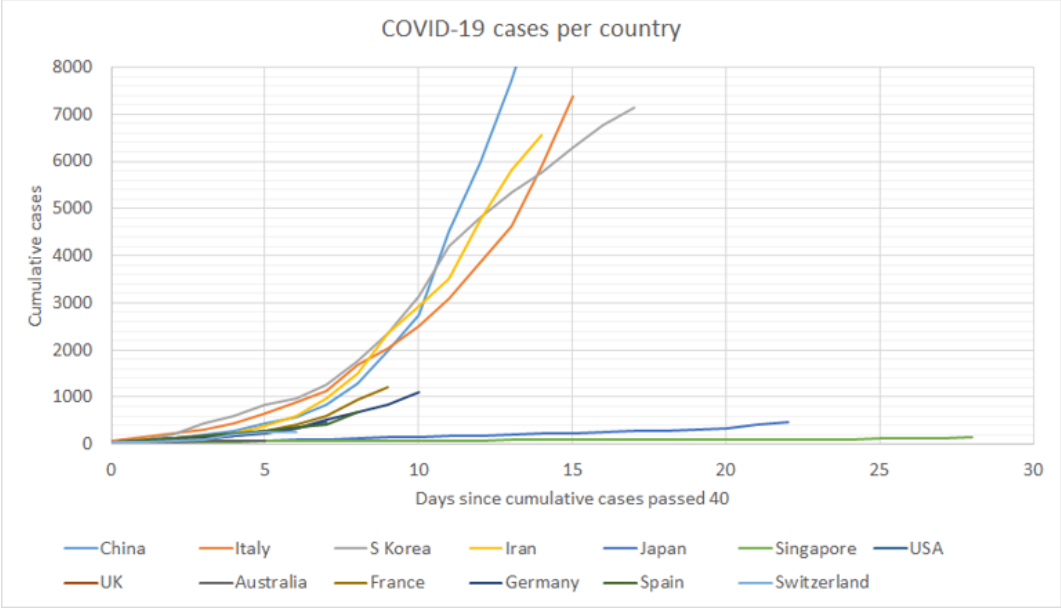
Appendix 1 - Linear growth charts

Suggested by Raemon.

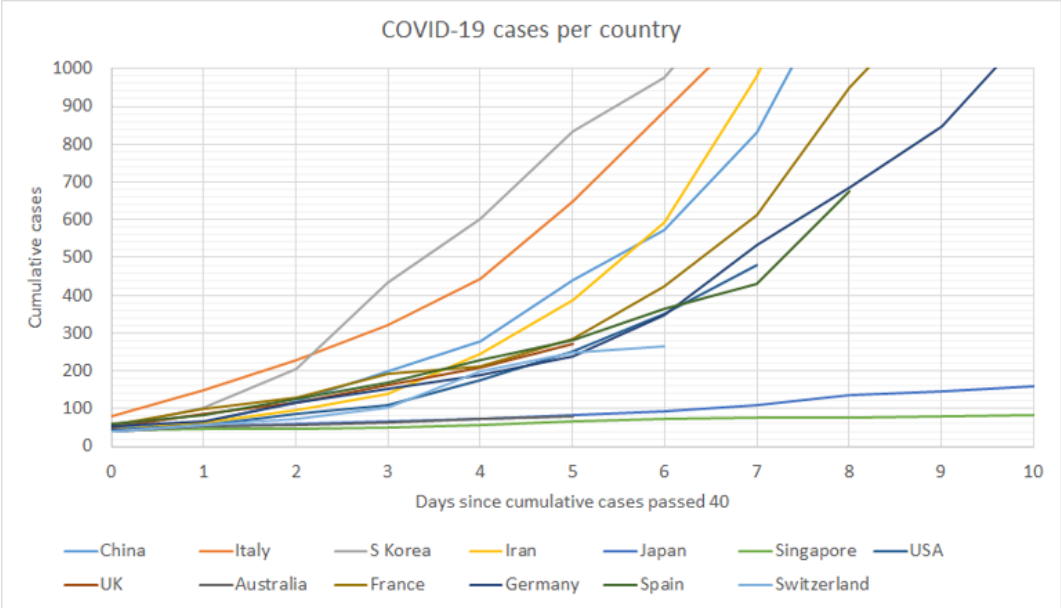
All cases



Y-axis limited at 8,000 cases per country

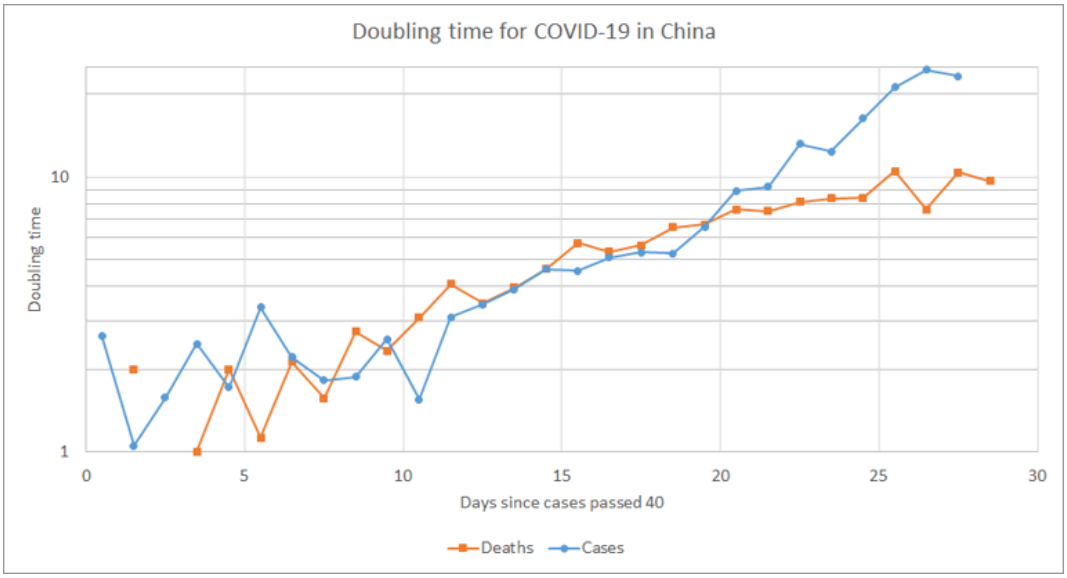


Y-axis limited at 1,000 cases per country, X-axis limited to first 10 days



Appendix 2 - Deaths vs cases

Suggested by Unnamed.



What are the most plausible "AI Safety warning shot" scenarios?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

A "AI safety warning shot" is some event that causes a substantial fraction of the relevant human actors (governments, AI researchers, etc.) to become substantially more supportive of AI research and worried about existential risks posed by AI.

For example, suppose we build an unaligned AI system which is "only" about as smart as a very smart human politician, and it escapes and tries to take over the world, but only succeeds in taking over North Korea before it is stopped. This would presumably have the "warning shot" effect.

I currently think that scenarios like this are not very plausible, because there is a very narrow range of AI capability between "too stupid to do significant damage of the sort that would scare people" and "too smart to fail at takeover if it tried." Moreover, within that narrow range, systems would probably realize that they are in that range, and thus bide their time rather than attempt something risky.

EDIT: To make more precise what I mean by "substantial:" I'm looking for events that cause >50% of the relevant people who are at the time skeptical or dismissive of existential risk from AI to change their minds.

Does SARS-CoV-2 utilize antibody-dependent enhancement?

The possibility of SARS-CoV-2 having Antibody-Dependent Enhancement (aka ADE) looked pretty real, to me.

If you're curious about the challenges of vaccine development for SARS-CoV-2, I recommend this article.

UPDATE: It looks like it isn't productively replicating in WBCs, but it probably *is* fusing with them and telling them to apoptose. Receptor uncertain, but they were checking T-cells specifically, which are exactly the WBCs that get severely depleted in severe COVID-19. They were in-vitro studies, but this mechanism matches the in-vivo results I'm seeing better, and I find the idea moderately convincing. (Article, h/t CellBioGuy's post mentioning it). SARS-2 is apparently *much* better at this than SARS-1.

Define ADE

Producing antibodies that are imperfect matches for one of these viruses (ex: optimized for another strain, or incompletely neutralizing) not only do not inactivate the virus, but instead get repurposed by the virus as a mechanism it can use to anchor and infect the cells that try to interact with those antibodies, most often immune cells.

Current conclusions

TL;DR: I think something else is having a larger impact than ADE on the immune system in severe COVID-19 disease. I have seen several theories, and listed some below. SARS-1 and MERS have both seen plenty of bad vaccine reactions that only show up at the animal-testing stage, and I still think vaccine development is going to be hard.

While I don't understand the upstream causes of this, Th-2 type activation has been extensively noted in both severe COVID-19 disease and bad vaccine reactions. Th-2 type immunopathology (roughly, allergy-like immune responses in lieu of virus-like immune responses) seems likely to play a large role in both complicating vaccine development, and influencing disease severity.

Something weird is going on with white blood cell counts, but I'm currently leaning towards believing it might be something else causing it. Specifically, they're seeing T-cell lymphopenia. T-cells seem to be the worst-hit WBC (hyper-activated, decreased numbers), and I'd have expected them to be close to immune to Fc-based ADE once mature; they only express the receptor while young.

SARS-1 and MERS vaccines both seem to have seen instances of bad reactions that required animal testing to become apparent. SARS-1 and SARS-2 exhibit ADE in-vitro against S-protein vaccines and not N-protein vaccines, but both S- and N-targeting

vaccines sometimes had bad reactions in animal testing*. What I see as an in-vitro/in-vivo divergence boosts my impression that vaccine development will be challenging.

*For example, this mouse study with SARS-1 saw hypersensitivity reactions to several S-containing vaccine subtypes (whole virus, VLP, S-protein), while this paper tested S- or N- VRPs and found that their N-inoculation not only had had no protective effect, but made things worse when it "resulted in enhanced immunopathology... within the lungs" upon infection.

Going on the current human results and a tiny Macaque study with 4 monkeys, for SARS-CoV-2 it does appear that getting through the disease once does preclude or largely-preclude reinfection by the same strain. I don't feel I can weigh in on whether this will stay true.

A few theories trying to explain the blood results (I'm sure there's more):

- Glucocorticoid reaction/Too much cortisol as a possible upstream cause of the lymphopenia + neutrophilia reaction seen in SARS, RSV, Ebola
- Something is causing/reacting to the cytokine storm
    - Cytokines essentially steer the strategy of WBCs (activation/direction/activity)
    - Both severe COVID19 disease and bad vaccine reactions steer towards Th-2 type immune responses (vaguely allergy-like)
    - I found this a bit too confusing to wade into
- Some sort of immune-cell suppression effects
    - Indirectly impacting cells at an earlier stage of blood cell development/differentiation (possibly as early as bone marrow stem cells)
    - Infecting cells via its additional receptor-binding affinities, for purely manipulation purposes (probably without viral replication)
        - It seems to clearly prefer lung and bowels for productive viral replication; we aren't seeing viral inclusion bodies in most other tissues. I don't feel this rules it out.

While I don't understand the upstream causes of this, Th-2 type activation has been extensively noted in both severe COVID-19 disease and bad vaccine reactions. Going on both the frequency with which it comes up in these contexts, and commentary by experts, Th-2 type immunopathology (roughly, allergy-like immune responses in lieu of virus-like immune responses) seems likely to play a large role in both complicating vaccine development, and influencing disease severity.

Earlier thoughts

I've repeatedly had to update in the direction of it being plausible, and I currently think it's more-likely-than-not to be a factor that will complicate vaccine development.

However, there do exist viable alternative theories for a lot of what I'm seeing, some of which I couldn't rule out.

Related Questions

I wanted to consolidate research on this into one place, and am interested in if anyone has additional solid arguments for/against it.

I have a lot of questions about this, but lets boil it down to a few.

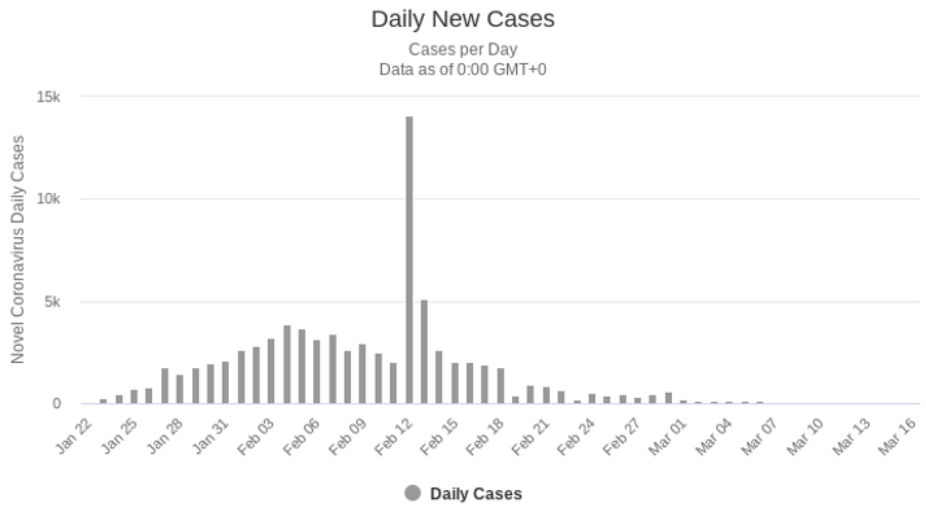
1. Is this virus doing this in-vivo? What induces it? (note: it does happen in-vitro, but in-vitro is apparently easier to induce, and often not conclusive evidence that this happens in-vivo at appreciable levels)
2. What exactly are the consequences of this ADE interaction? Can we get it down to symptoms?
3. How does this change things?
4. What can we do about it?

Good News: the Containment Measures are Working

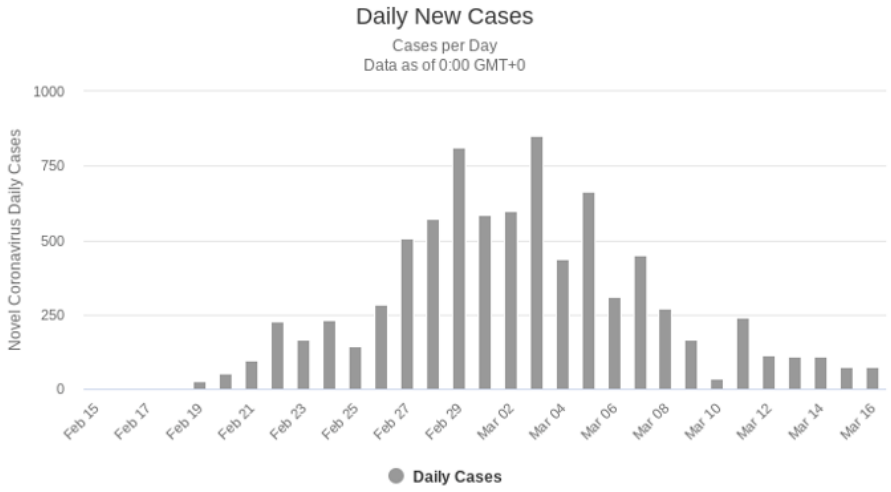
After weeks and months of rather drastic measures we can see the glimpse of how the lockdowns are making a difference. Here are the snapshots from worldometer:

New cases dropping

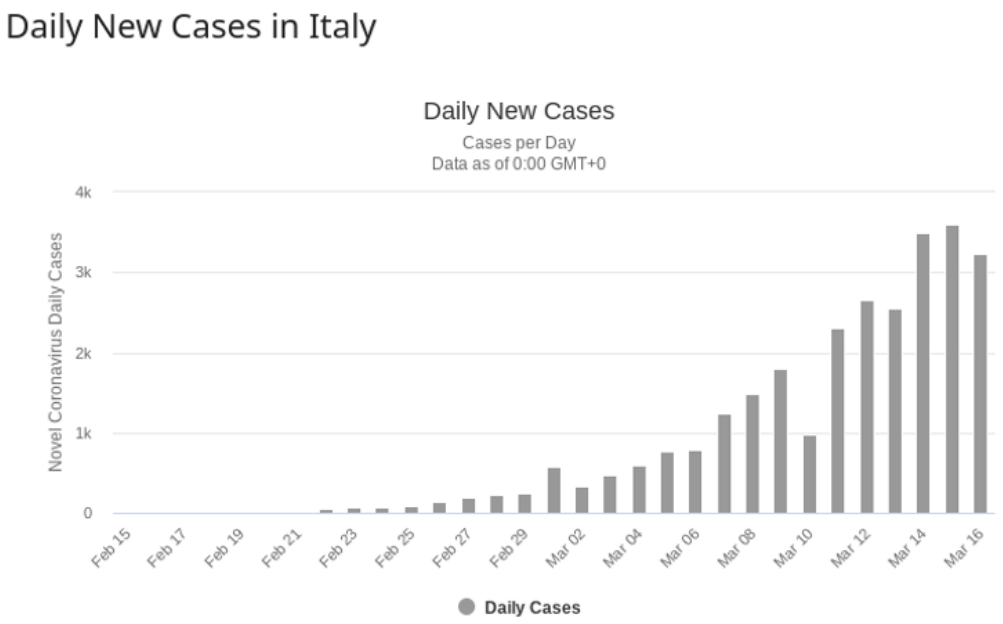
Daily New Cases in China



Daily New Cases in South Korea

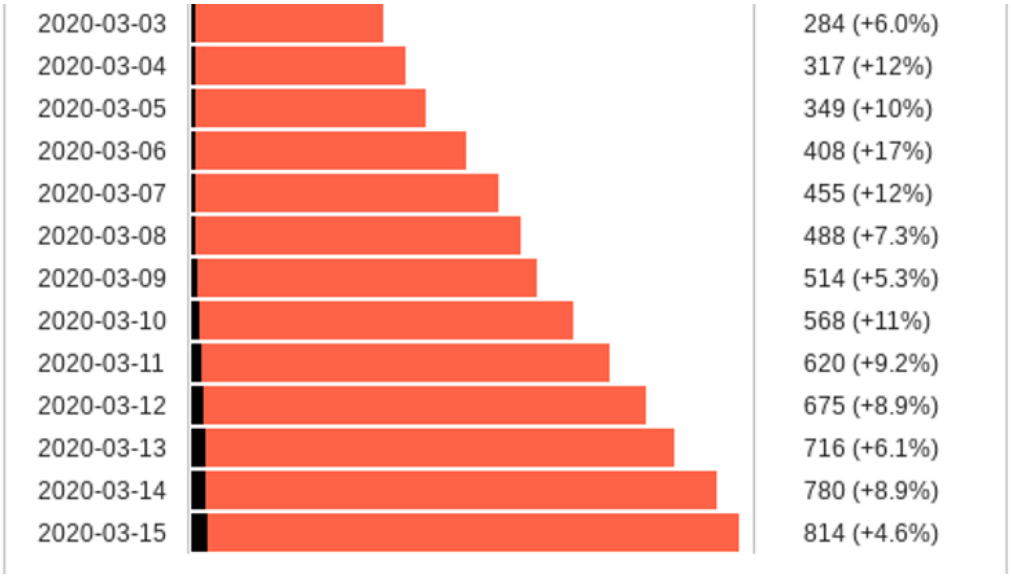


New cases leveling off



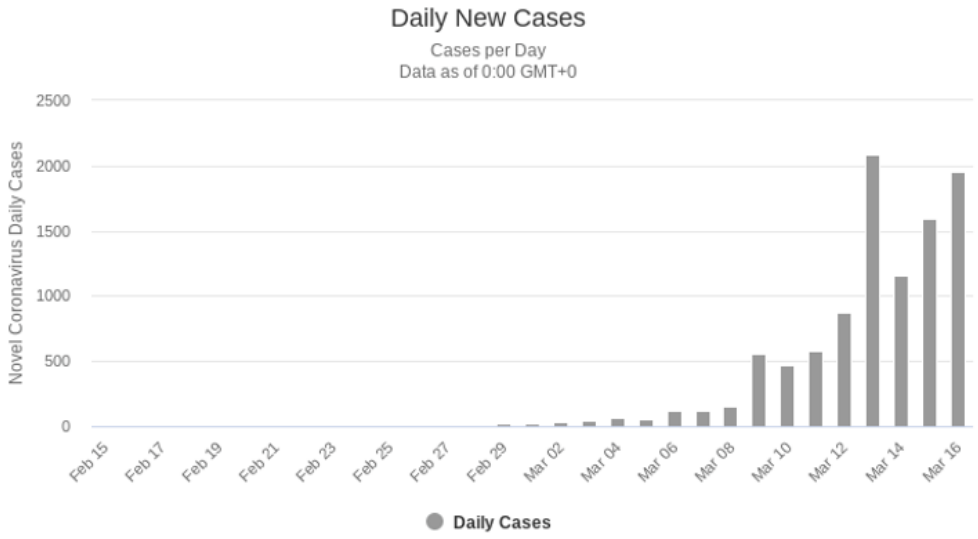
New Cases switching from exponential to linear

Japan: no Worldometer breakdown, data from Wikipedia

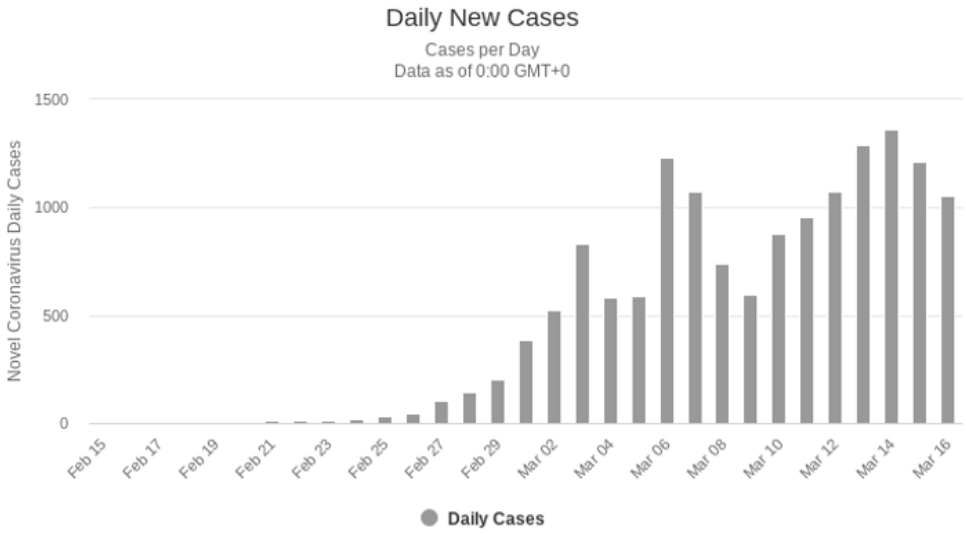


Inconclusive, but apparently sub-exponential

Daily New Cases in Spain



Daily New Cases in Iran



What does it mean for the future? That it takes about a week of a severe lockdown to switch from the exponential growth to linear, about two weeks to switch to leveling off, and three to four weeks to start seeing a meaningful decline. If this pattern holds, then the dire projections of millions dying would not come to pass. The overall toll is likely to be around 100,000 in 2020 if most countries adopt the China/Korea/Italy/Spain-level measures. Moreover, the worst of the pandemic, or at least of this first wave, will be over in 2-3 months, as long as the containment measures are in place, and the countries and jurisdictions that are proactive enough will not have to resort to the worst of the ICU triage any time soon.

This was the good news. The bad news is that this is not sustainable in the medium to long term, given the impact of the measures on the economy, mobility and lifestyle. But maybe the effort to flatten the curve is not in vain. The question is, what's next? Who knows? A viable vaccine is at least a year away, the ICU bed count ramp up may help, some of the social distancing measures such as deliveries, takeouts, telecommuting may prove sustainable enough, but that doesn't seem like nearly enough in the long term.

mind viruses about body viruses

I was going to write this as a Slate Star Codex comment, but I'm going to make it a tumblr post tagging @slatestarscratchpad instead, since experience suggests it's likely to be more widely and carefully read in this form. (Crossposting to LW too, so you may be reading this there, possibly with mangled formatting.)

The idea frontier

I am getting more and more concerned about the "information epidemiology" of the public conversation about Covid-19.

Here are some distinctive features I see in the public conversation:

1. Information intake must be triaged.

There is a very large amount of *new* publicly available information every day. There are no slow news days. "Keeping up with the story" in the way one would keep up with an evolving news story would be a full-time job.

Many of us do not have time to do this, and I imagine many of those who do have time cannot tolerate the experience in practice. In fact, there can be a tradeoff between one's level of personal involvement in the crisis and one's ability to "follow it" as a news story.

(I work for a telemedicine company, and after a day of dealing with the ever-changing impacts of Covid-19 on my work, I have relatively little patience left to read about its ever-changing impacts on absolutely everything else. That's just me, though, and I realize some people's mental bandwidth does not work like this.)

2. Abstractions are needed, and the relevant abstractions are novel and contested.

Crucial and time-sensitive decisions must be made on the basis of simulations, abstract mental models, and other intellectual tools.

In some sense this is true of everything, but in most cases we have a better sense of how to map the situation onto some large reference class of past intellectual work. When there is an economic downturn, the standard macroeconomic arguments that have existed for many decades pop back up and make the predictable recommendations they always make; even though there is no expert consensus, the two or three common expert stances are already familiar.

With Covid-19, this is not so. All the intervention types currently under discussion would be, in their own ways, unprecedented. As it struggles to follow the raw facts, the general public is also struggling to get its head around terms and concepts like "suppression," "containment," "contact tracing," etc. which were (in the relevant senses) not part of our mental world at all until recently.

Thus, relative to most policy debates, this one has a strange frontier energy, a sense that we're all discovering something for the first time. Even the professional epidemiologists are struggling to translate their abstract knowledge into brief-but-clear soundbites. (I imagine many of them have never needed to be public communicators at this kind of scale.)

3. There is no division of labor between those who make ideas and those who spread them.

There is a hunger for a clear big picture (from #1). There are few pre-established intellectual furnishings (#2). This means there's a vacuum that people very much want to fill. By ordinary standards, no one has satisfying answers, not even the experts; we are all struggling to do basically the same intellectual task, simultaneously.

None of us have satisfying answers – we are all the same in that respect. But we differ in how good we are at public communication. At communicating things that *sound like they could be answers*, clearly, pithily. At optimizing our words for maximum replication.

It is remarkable to me, just as a bare observation, that (in my experience) the best widespread scientific communication on Covid-19 – I mean *just* in the sense of verbal lucidity and efficiency, effective use of graphs, etc., *not* necessarily in the sense of accuracy or soundness – has been done by Tomas Pueyo, a formerly obscure (?) expert on … *viral marketing*.

(To be clear, I am not dismissing Pueyo's opinions by citing his background. I am hypothesizing his background explains the spread of his opinions, and that their correctness level has been causally inert, or might well have been.)

The set of ideas we use to understand the situation, and the way we phrase those ideas, is being determined from scratch as we speak. Determined by all of us. For the most part, we are passively allowing the ideas to be determined by the people who determine ideas in the absence of selection – by people who have specialized, not in creating ideas, but in spreading them.

4. Since we must offload much of our fact-gathering (#1) and idea-gathering (#2) work onto others, we are granting a lot on the basis of trust.

Scott's latest coronavirus links post contains the following phrases:

> Most of the smart people I've been reading have converged on something like the ideas expressed in […]
>
> On the other hand, all of my friends who are actually worried about getting the condition are […]

These jumped out at me when I read the post. They feel worryingly like an "information cascade" – a situation where an opinion seems increasing credible as more and more people take that opinion partially on faith from other individually credible people, and thus spread it to those who find them credible in turn.

Scott puts some weight on these opinions on the basis of trust – i.e. not 100% from his independent vetting of their quality, but also to some extent from an outside view, because these people are "smart," "actually worried." Likelier to be right than baseline, as a personal attribute. So now these opinions get boosted to a *much* larger audience, who will take them again partially on trust. After all, Scott Alexander trusts it, and *he's* definitely smart and worried and keeping up with the news better than many of us.

What "most of the smart people … have been converging on," by the way, is Tomas Pueyo's latest post.

Is Tomas Pueyo right? He is certainly good at seeming like a "smart" and "actually worried" person whose ideas you want to spread. That in itself is enough. I shared his first big article with my co-workers; at that time it seemed like a shining beacon of resolute, well-explained thought shining alone in a sea of fog. I couldn't pull off that effect as well if I tried, I think – not even if the world depended on it. I'm not that good. Are you?

My co-workers read that first post, and their friends did, and their friends. If you're reading this, I can be almost sure you read it too. Meanwhile, what I am *not* doing is carefully

reading the many scientific preprints that are coming out every week from people with more domain expertise, or the opinions the same people are articulating in public spaces (usually, alas, in tangled twitter threads). That's hard work, and I don't have the time and energy. Do you?

I don't know if this is actually an effective metaphor – after all, I'm not a viral marketer – but I keep thinking of privilege escalation attacks.

It is not a bad thing, individually, to place trust some in a credible-sounding person without a clear track record. We can't really do otherwise, here. But it is a bad thing when that trust spreads in a cascade, to your "smartest" friends, to the bloggers who are everyone's smartest friends, to the levers of power – all on the basis of what is (in every individual transmission step) a tiny bit of evidence, a glimmer of what might be correctness rising above pure fog and static. We would all take 51% accuracy over a coin flip – and thus, that which is accurate 51% of the time becomes orthodoxy within a week.

Most of the smart people you've been reading have converged on something like …

#FlattenTheCurve: a case study of an imperfect meme

Keeping up with the lingo

A few weeks ago – how many? I can't remember! – we were all about flattening the curve, whatever that means.

But this week? Well, most of the smart people you've been reading have converged on something like: "flattening" is insufficient. We must be "squashing" instead. And (so the logic goes) because "flattening" is insufficient, the sound byte "flatten the curve" is dangerous, implying that all necessary actions fall under "flattening" when some non-flattening actions are also needed.

These are just words. We should be wary when arguments seem to hinge on the meaning of words that no one has clearly defined.

I mean, you surely don't need me to tell you that! If you're reading this, you're likely to be a veteran of internet arguments, familiar from direct experience and not just theory with the special stupidity of merely semantic debates. That's to say nothing of the subset of my readership who are LessWrong rationalists, who've *read the sequences*, whose identity was formed around this kind of thing long before the present situation. (I'm saying: *you if anyone should be able to get this right. You were made for this.*)

It's #FlattenTheCurve's world, we just live in it

What did "flatten the curve" mean? Did it mean that steady, individual-level non-pharmaceutical interventions would be enough to save hospitals from overload? Some people have interpreted the memetic GIFs that way, and critiqued them on that basis.

But remember, #FlattenTheCurve went viral back when fretting about "coronavirus panic" was a mainstream thing, when people actually needed to be *talked into* social distancing. The most viral of the GIFs does not contrast "flattening" with some other, more severe strategy; it contrasts it with *nothing*. Its bad-guy Goofus character, the foil who must be educated into flattening, says: "Whatever, it's just like a cold or flu."

No one is saying that these days. Why? How did things change so quickly? One day people were smugly saying not to panic, and then all of a sudden they were all sharing a string of words, a picture, something that captivated the imagination. A meme performed a trick of

privilege escalation, vaulted off of Facebook into the NYT and the WaPo and the WSJ and the public statements of numerous high officials. Which meme? – *oh, yes, that one*.

We are only able to have this conversation about flattening-vs-squashing because the Overton Window has shifted drastically. Shifted due to real events, yes. But also due to #FlattenTheCurve. The hand you bite may be imperfect, but it is the hand that feeds you.

Bach, the epidemiologists, and me

Joscha Bach thinks #FlattenTheCurve is a "lie," a "deadly delusion." Because the GIF showed a curve sliding under a line, yet the line is very low, and the curve is very high, and we may never get there.

Is he right? He is definitely right that the line is very low, and we may not slide under it. Yet I was unimpressed.

For one thing, Bach's argument was simply not formally valid: it depended on taking a static estimate of total % infected and holding it constant when comparing scenarios across which it would vary.

(This was one of several substantive, non-semantic objections I made. One of them, the point about Gaussians, turned out to be wrong, in the sense that granting my point could not have affected Bach's conclusion – not that Bach could have reached his conclusion anyway. This argument was my worst one, and the only one anyone seemed to notice.)

Something also seemed fishy about Bach's understanding of "flatten the curve." The very expert from whom he got his (misused) static estimate was still tweeting about how we needed to flatten the curve. *All* the experts were tweeting about how we needed to flatten the curve. Which was more plausible: that they were all quite trivially wrong, about the same thing, at once? Or that their words meant something more sensible?

The intersection of "world-class epidemiologists" and "people who argue on twitter" have now, inevitably, weighed in on Bach's article. For instance:

CarolineOB @Caroline_OF_B · Mar 21

All 3 point to a misunderstanding of how epidemiologists use models.

1. Claim: No y-axis. Yep, because this is a generalizable, conceptual framework based on a mechanistic model of how diseases spread, not a model of a particular place. It conveys ideas, not numbers. 2/5

2 1 27

CarolineOB @Caroline_OF_B · Mar 21

2. Claim: it overestimates capacity. Er what now? See point 1 above. These figures convey the idea that we want to slow transmission to minimize the area above capacity, with "squashing" transmission being the exact goal here. This is realistic, not nihilistic. 3/5

1 3 19

CarolineOB @Caroline_OF_B · Mar 21

3. Claim: "they" say let the epidemic burn through. No. No good epidemiologist is saying this. We want to flatten the curve as much as possible. Saying we should "stop" transmission - which of course we all want - just demonstrates a lack of understanding of disease dynamics. 4/5

1 2 22

Carl T. Bergstrom @CT_Bergstrom · 17h

Replying to @CT_Bergstrom @DanielFalush and @Caroline_OF_B

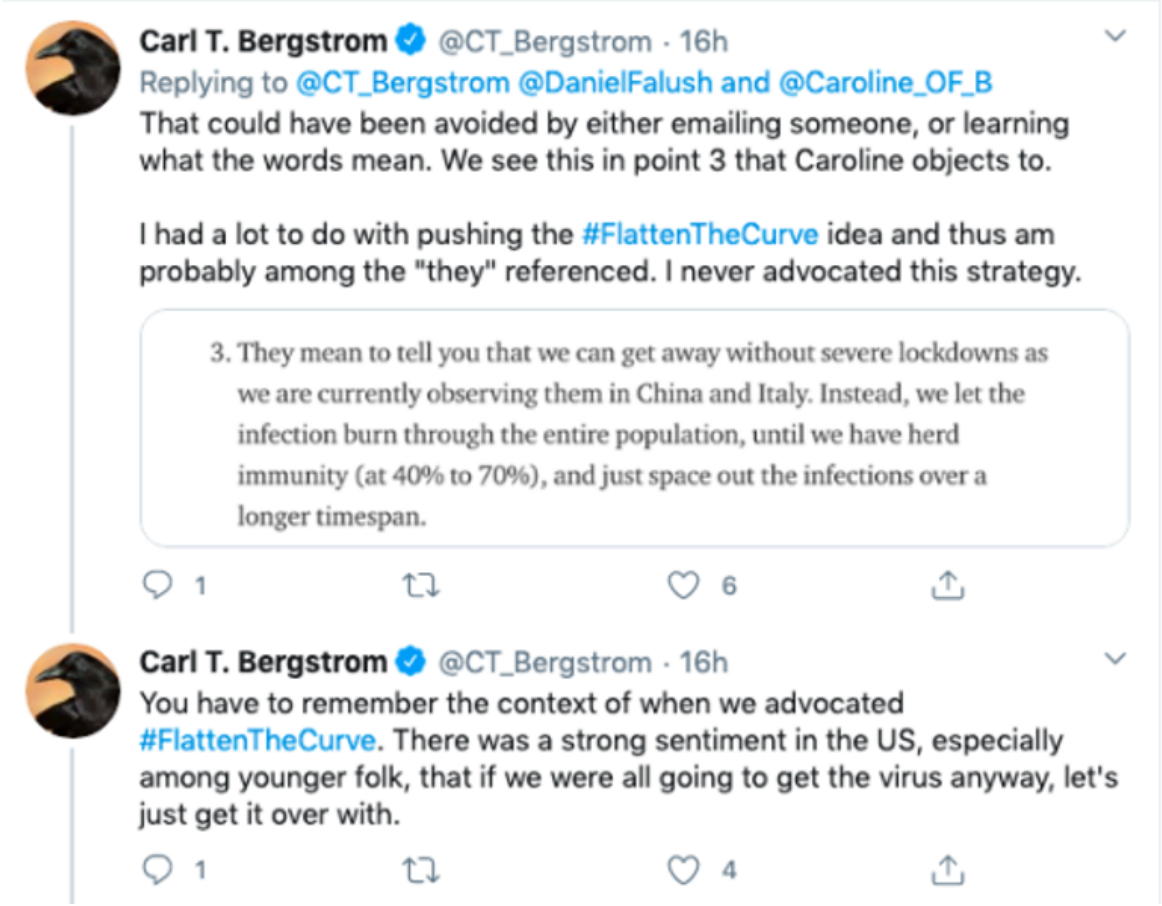
When it was first published, it made tabloid claims (still sort of does—see below) based on the author's misunderstanding of the terms "containment" and "mitigation". The first was used as synonymous with suppression, and the latter with letting things go to herd immunity.

Carl T. Bergstrom @CT_Bergstrom · 17h

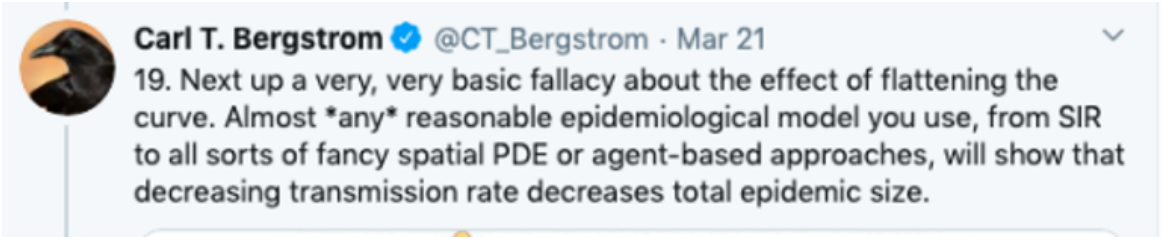
Problem is, that's not what those terms mean. The former means essentially testing+contact tracing and the latter, broader generalized measures including hygiene and social distancing.

This led the author to accuse many of us of a far stupider policy than we would ever advocate.

1 5



And I can't resist quoting [one more Carl Bergstrom thread], this one about *another* Medium post by a viral marketer (not the other one), in which Carl B's making the exact same damn point I made about the static estimate:



Like me, these people make both substantive and semantic objections. In fact, theirs are a strict superset of mine (see that last Bergstrom thread re: Gaussians!).

I am not saying “look, I was right, the experts agree with me, please recognize this.” I mean, I *am* saying that.

But I’m also saying – look, people, none of this is settled. None of us have satisfying answers, remember. We are all stressed-out, confused glorified apes with social media accounts yelling at each other about poorly defined words as we try to respond to an invader that is ravaging our glorified-ape civilization. Our minds cannot handle all this information. We are at the mercy of viral sound bites, and the people who know how to shape them.

What is it the rationalists like to say? “We’re running on corrupted hardware?”

Carl Bergstrom championed a meme, #FlattenTheCurve. He believed it would work, and I think it in fact did. But Carl Bergstrom, twitter adept though he may be, is still someone whose primary career is science, not consensus-making. In a war of memes between him and (e.g.) Tomas Pueyo, I’d bet the bank on Pueyo winning.

And that is frightening. I like Pueyo’s writing, but I don’t want to just let him – or his ilk – privilege-escalate their way into effective command of our glorified ape civilization.

I want us to recognize the kind of uncertainty we live under now, the necessity for information and idea triage, the resulting danger of viral soundbites winning our minds on virality alone because we were too mentally overwhelmed to stop the spread … I want us to recognize all of that, and act accordingly.

Not to retreat into the comfort of “fact-checking” and passive consultation of “the experts.” That was always a mirage, even when it seemed available, and here and now it is clearly gone. All of us are on an equal footing in this new frontier, all of us sifting through Medium articles, twitter threads, preprints we half understand. There are no expert positions, and there are too many facts to count.

Not to trust the experts – but to exercise caution. To recognize that we are letting a “consensus” crystalize and re-crystalize on the basis of cute dueling phrases, simplified diagrams and their counter-simplified-diagrams, bad takes that at least seem better than

pure white noise, and which we elevate to greatness for that alone. Maybe we can just … stop. Maybe we can demand better. Wash our minds' hands, too.

Our intellectual hygiene might end up being as important as our physical hygiene. Those who control the levers of power are as confused and stressed-out as you are, and as ready to trust viral marketers with firm handshakes and firm recommendations. To trust whichever sound byte is ascendant this week.

Thankfully, you have some measure of control. Because we are all on flat ground in this new frontier, your social media posts are as good as anyone's; you can devote your mind to making ideas, or your rhetorical skill to promoting *specifically those ideas you have carefully vetted.* You can choose to help those with power do better than the status quo, in your own little way, whatever that may be. Or you can choose not to.

Okay, words aside, does the right strategy look like the famous GIF taken literally, or like a feedback system where we keep turning social distancing on and off so the graph looks like a heart rate monitor, or like a "hammer" reset followed by a successful emulation of South Korea, or

I don't know and you don't know and Tomas doesn't know and Carl doesn't know. It's hard! I'm hadn't even heard of "R_0" until like two months ago! Neither had you, probably!

Marc Lipsitch's group at Harvard has been putting out a bunch of preprints and stuff that look reputable to me, and are being widely shared amongst PhDs with bluechecks and university positions. Their most recent preprint, from 3 days ago, appears to be advocating the heart rate monitor-ish thing, so yay for that, maybe. But … this sounds like the same information cascade I warned against, so really, I dunno, man.

However, I will suggest that perhaps the marginal effect of sharing additional reputable-seeming takes and crystalizing weekly orthodoxies is negative in expectation, given an environment saturated with very viral, poorly vetted words and ideas.

And that your best chance of a positive marginal impact is to be *very* careful, like the people who won't trust any medical intervention until it has 50+ p-hacked papers behind it, has been instrumental in the minting of many PhDs, and has thereby convinced the strange beings at the FDA and the Cochrane Collaboration who move at 1/100 the speed of you and me. Not because this is globally a good way to be, but because it locally is – given an environment saturated with very viral, poorly vetted words and ideas.

That you should sit down, take the outside view, think hard about whether you can make a serious independent intellectual contribution when literally everyone on earth, basically, is trying to figure out the same thing.

And you know, maybe you are really smart! Maybe the answer is yes! If so, do your homework. Read everything, more than I am reading, and more carefully, and be ready to show your work. Spend more time on this than the median person (or me) is literally capable of doing right now. *This is the value you are claiming to provide to me.*

If you can't do that, that is fine – I can't either. But if you can't do that, and you still boost every week's new coronavirus orthodoxy, you are an intellectual disease vector. Don't

worry: I will hear it from other people if I don't hear it from you. But you will lend your credibility to it. Whatever trust I place in you will contribute to the information cascade.

This work, this hard independent work collecting lots of raw undigested information, is actually what Tomas Pueyo seems to be doing – I mean, apart from framing everything in a very viral way, which is why you and I know of his work. We are saturated with signal-boosts of the few such cases that exist. We do not need more signal-boosts. We need more *independent work like this*. Please do it. Or, if not that, then be like the lady in that very problematic GIF: don't panic, but be *careful*, wash your mind's hands, and (yes) flatten the intellectual curve.

Adaptive Immune System Aging

The human adaptive immune system is the “smart” part of the human immune system, the part which learns to recognize specific pathogens, allowing for immunity to e.g. chicken pox. For our current purposes, the key players are T-cells. T-cells start out “naive” and eventually learn to recognize specific antigens, becoming “memory” T-cells. The aged immune system is characterized by a larger fraction of memory relative to naive T-cells (without dramatic change in overall counts). This makes the elderly immune system slower to adapt to new pathogens.

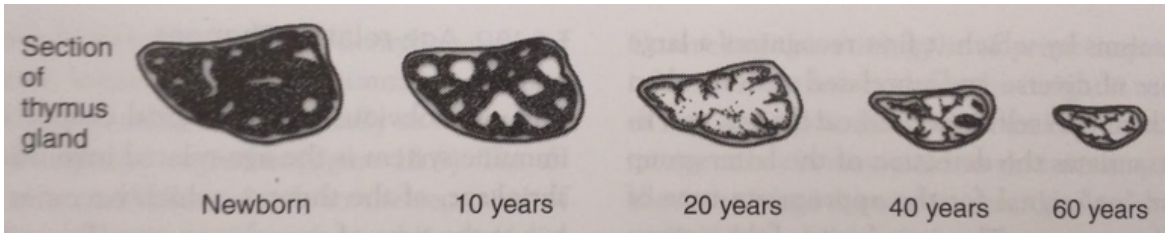
This post is mainly about why the naive:memory T-cell ratio shifts with age, how to undo that shift, and some speculation about implications and applications.

A natural hypothesis (frequently asserted in the literature): the shift toward memory T-cells is driven by slower production of new (naive) T-cells. The T-cells themselves maintain overall cell count by living longer, resulting in a larger proportion of older (memory) cells.

The interesting part: why would the production of new T-cells fall with age?

Turns out there’s an obvious culprit: the thymus. The thymus is the last stop in the production line for new T-cells. It provides a sort of boot camp, training the T-cells to distinguish “self” (your own cells) from “other” (pathogens) using a whole battery of tricks. T-cells which make it through become full-time members of the naive T-cell reserve, and go on to police the body.

With age, the thymus does this:



(source: PhysAging). This is called “involution” of the thymus.

Many organs shrink with age, but the thymus is among the most dramatic. Unlike most age-related loss, it starts even before development is complete - the thymus shrinks measurably between day zero and a child’s first birthday. And it keeps on shrinking, at a steady rate, throughout childhood and adult life. The extremely early start of thymic involution suggests it’s more a developmental phenomenon than an age-related phenomenon - perhaps an appropriate hormonal mix could undo thymic involution?

Turns out, castration of aged mice (18-24 mo) leads to complete restoration of the thymus in about 2 weeks. The entire organ completely regrows, and the balance of naive to memory T-cells returns to the level seen in young mice. (Replicated here.) This is pretty dramatic evidence that:

- The thymus only generates/regenerates in the absence of sex hormones
- The age-related shift in naive:memory T-cell ratio is driven primarily by thymic involution

Particularly intriguing: we already have chemical castration methods, which are generally considered reversible. And it only took two weeks for the mice to regrow the whole thymus. At this point we’re speculating, but assuming chemical castration also works and it translates

to humans and the thymus doesn't rapidly re-involute after ceasing the chemical castration... that sounds pretty promising as an avenue to fixing age-related adaptive immune system decline in humans.

As long as we're speculating, let's speculate hard. Immunotherapy is the hot new thing in cancer these days - apparently T-cells in young people remove precancerous cells and attack tumors, but in old people that doesn't happen as reliably. So... have there been any studies on how castration effects cancer? For starters, chemical castration is already a widely-used treatment for both prostate and breast cancer. It works. But that's prostate and breast; they're sex organs, which we'd expect to atrophy in the absence of sex hormones. I don't know of any studies on the effects of chemical castration on other types of cancer, in humans.

In rats, however, at least one century-old study finds that castration prevents age-related cancer - and quite dramatically so. Castrated mice' rate of resistance to an implanted tumor was ~50%, vs ~5% for controls. (This study finds a similar result in rabbits.) That old rat study cites a few others with mutually-conflicting results, and proposes that the age of the rats used explains it all: investigators who use young rats find that castration has little-to-no effect on resistance to an implanted tumor. Exactly what we'd expect if it's all mediated by thymic involution & regrowth.

There's a lot of questions here. Does chemical castration have similar effects to surgical castration on thymic regrowth in mice? Does chemical castration result in regrowth of the thymus in humans? (Several states/countries require chemical castration as a condition of parole for certain sex offenders, and it's also used for prostate and breast cancer, so it should be possible to find a few old people using it and see whether their thymus has regrown.) Does the thymus rapidly re-involute after administration of chemical castration ceases? Does chemical castration work as a treatment for cancers besides prostate and breast? Is the effectiveness of chemical castration against cancer age-dependent? Can temporary administration of chemical castration prevent cancer for a long period of time?

Turning away from applications and back to gears, there's also some key questions around thymic involution itself. The individual cells of the thymus don't have unusually slow turnover; if the thymic cell count is decreasing over time, then either the rate of production is decreasing or the breakdown rate is increasing. There has to be some upstream cause. Whatever that cause is, it probably isn't the same upstream cause as most age-related problems - thymic involution doesn't follow the usual pattern of no noticeable problems during development, slow loss of performance in middle age, then accelerating failure in old age.

I'd be excited to see more work along these lines and/or references to relevant studies.