

Best of LessWrong: December 2020

1. [100 Tips for a Better Life](#)
2. [To listen well, get curious](#)
3. [Great minds might not think alike](#)
4. [Covid 12/24: We're F***ed, It's Over](#)
5. [The LessWrong 2018 Book is Available for Pre-order](#)
6. [Motive Ambiguity](#)
7. [parenting rules](#)
8. [2020 AI Alignment Literature Review and Charity Comparison](#)
9. [Give it a google](#)
10. [How long does it take to become Gaussian?](#)
11. [The LessWrong 2019 Review](#)
12. [Debate update: Obfuscated arguments problem](#)
13. [Against GDP as a metric for timelines and takeoff speeds](#)
14. [Luna Lovegood and the Chamber of Secrets - Part 3](#)
15. [Real-Life Examples of Prediction Systems Interfering with the Real World \(Predict-O-Matic Problems\)](#)
16. [Homogeneity vs. heterogeneity in AI takeoff scenarios](#)
17. [The First Sample Gives the Most Information](#)
18. [Luna Lovegood and the Chamber of Secrets - Part 5](#)
19. [FHI paper published in Science: interventions against COVID-19](#)
20. [Why quantitative methods are heartwarming](#)
21. [How Hard Would It Be To Make A COVID Vaccine For Oneself?](#)
22. [Luna Lovegood and the Chamber of Secrets - Part 6](#)
23. [Luna Lovegood and the Chamber of Secrets - Part 4](#)
24. [Extrapolating GPT-N performance](#)
25. [Anti-EMH Evidence \(and a plea for help\)](#)
26. [The Darwin Game - Conclusion](#)
27. [Parable of the Damned](#)
28. [What trade should we make if we're all getting the new COVID strain?](#)
29. [Luna Lovegood and the Chamber of Secrets - Part 9](#)
30. [Cultural accumulation](#)
31. [What evidence will tell us about the new strain? How are you updating?](#)
32. [Covid 12/3: Land of Confusion](#)
33. [My Fear Heuristic](#)
34. [Quick Thoughts on Immoral Mazes](#)
35. [Fusion and Equivocation in Korzybski's General Semantics](#)
36. [What is "protein folding"? A brief explanation](#)
37. [Covid 12/10: Vaccine Approval Day in America](#)
38. [Secular Solstice 2020](#)
39. [In Addition to Ragebait and Doomscrolling](#)
40. [Notes on notes on virtues](#)
41. [How I Write](#)
42. [How Lesswrong helped me make \\$25K: A rational pricing strategy](#)
43. [Covid 12/17: The First Dose](#)
44. [Luna Lovegood and the Chamber of Secrets - Part 7](#)
45. [Luna Lovegood and the Chamber of Secrets - Part 10](#)
46. [Luna Lovegood and the Chamber of Secrets - Part 8](#)
47. [My favorite essays of life advice](#)
48. [TAI Safety Bibliographic Database](#)
49. [The Best Visualizations on Every Subject](#)
50. [Why Neural Networks Generalise, and Why They Are \(Kind of\) Bayesian](#)

Best of LessWrong: December 2020

1. [100 Tips for a Better Life](#)
2. [To listen well, get curious](#)
3. [Great minds might not think alike](#)
4. [Covid 12/24: We're F***ed, It's Over](#)
5. [The LessWrong 2018 Book is Available for Pre-order](#)
6. [Motive Ambiguity](#)
7. [parenting rules](#)
8. [2020 AI Alignment Literature Review and Charity Comparison](#)
9. [Give it a google](#)
10. [How long does it take to become Gaussian?](#)
11. [The LessWrong 2019 Review](#)
12. [Debate update: Obfuscated arguments problem](#)
13. [Against GDP as a metric for timelines and takeoff speeds](#)
14. [Luna Lovegood and the Chamber of Secrets - Part 3](#)
15. [Real-Life Examples of Prediction Systems Interfering with the Real World \(Predict-O-Matic Problems\)](#)
16. [Homogeneity vs. heterogeneity in AI takeoff scenarios](#)
17. [The First Sample Gives the Most Information](#)
18. [Luna Lovegood and the Chamber of Secrets - Part 5](#)
19. [FHI paper published in Science: interventions against COVID-19](#)
20. [Why quantitative methods are heartwarming](#)
21. [How Hard Would It Be To Make A COVID Vaccine For Oneself?](#)
22. [Luna Lovegood and the Chamber of Secrets - Part 6](#)
23. [Luna Lovegood and the Chamber of Secrets - Part 4](#)
24. [Extrapolating GPT-N performance](#)
25. [Anti-EMH Evidence \(and a plea for help\)](#)
26. [The Darwin Game - Conclusion](#)
27. [Parable of the Damned](#)
28. [What trade should we make if we're all getting the new COVID strain?](#)
29. [Luna Lovegood and the Chamber of Secrets - Part 9](#)
30. [Cultural accumulation](#)
31. [What evidence will tell us about the new strain? How are you updating?](#)
32. [Covid 12/3: Land of Confusion](#)
33. [My Fear Heuristic](#)
34. [Quick Thoughts on Immoral Mazes](#)
35. [Fusion and Equivocation in Korzybski's General Semantics](#)
36. [What is "protein folding"? A brief explanation](#)
37. [Covid 12/10: Vaccine Approval Day in America](#)
38. [Secular Solstice 2020](#)
39. [In Addition to Ragebait and DoomsScrolling](#)
40. [Notes on notes on virtues](#)
41. [How I Write](#)
42. [How LessWrong helped me make \\$25K: A rational pricing strategy](#)
43. [Covid 12/17: The First Dose](#)
44. [Luna Lovegood and the Chamber of Secrets - Part 7](#)
45. [Luna Lovegood and the Chamber of Secrets - Part 10](#)
46. [Luna Lovegood and the Chamber of Secrets - Part 8](#)
47. [My favorite essays of life advice](#)
48. [TAI Safety Bibliographic Database](#)

49. [The Best Visualizations on Every Subject](#)
50. [Why Neural Networks Generalise, and Why They Are \(Kind of\) Bayesian](#)

100 Tips for a Better Life

(Cross-posted from my [blog](#))

The other day I made an advice thread based on [Jacobian's](#) from last year! If you know a source for one of these, shout and I'll edit it in.

Possessions

1. If you want to find out about people's opinions on a product, google *<product> reddit*. You'll get real people arguing, as compared to the SEO'd Google results.
2. Some banks charge you \$20 a month for an account, others charge you 0. If you're with one of the former, have a good explanation for what those \$20 are buying.
3. Things you use for a significant fraction of your life (bed: 1/3rd, office-chair: 1/4th) are worth investing in.
4. "Where is the good knife?" If you're looking for your good X, you have bad Xs. Throw those out.
5. If your work is done on a computer, get a second monitor. Less time navigating between windows means more time for thinking.
6. Establish clear rules about when to throw out old junk. Once clear rules are established, junk will probably cease to be a problem. This is because any rule would be superior to our implicit rules ("keep this broken stereo for five years in case I learn how to fix it").
7. Don't buy CDs for people. They have Spotify. Buy them merch from a band they like instead. It's more personal and the band gets more money.
8. When buying things, time and money trade-off against each other. If you're low on money, take more time to find deals. If you're low on time, stop looking for great deals and just buy things quickly online.

Cooking

9. Steeping minutes: Green at 3, black at 4, herbal at 5. Good tea is that simple!
10. Food actually can be both cheap, healthy, tasty, and relatively quick to prepare. All it requires is a few hours one day to prepare many meals for the week.
11. Cooking pollutes the air. Opening windows for a few minutes after cooking can dramatically improve air quality.
12. Food taste can be made much more exciting through simple seasoning. It's also an opportunity for expression. Buy a few herbs and spices and experiment away.
13. When googling a recipe, precede it with 'best'. You'll find better recipes.

Productivity

14. Advanced search features are a fast way to create tighter search statements. For example:

img html

will return inferior results compared to:

img html -w3

[15.](#) You can automate mundane computer tasks with Autohotkey (or AppleScript). If you keep doing a sequence “so simple a computer can do it”, make the computer do it.

16. Learn keyboard shortcuts. They’re easy to learn and you’ll get tasks done faster and easier.

[17.](#) Done is better than perfect.

18. Keep your desk and workspace bare. Treat every object as an imposition upon your attention, because it is. A workspace is not a place for storing things. It is a place for accomplishing things.

19. Reward yourself after completing challenges, even badly.

Body

20. The 20-20-20 rule: Every 20 minutes of screenwork, look at a spot 20 feet away for 20 seconds. This will reduce eye strain and is easy to remember (or program reminders for).

[21.](#) Exercise (weightlifting) not only creates muscle mass, it also improves skeletal structure. Lift!

[22.](#) Exercise is the most important lifestyle intervention you can do. Even the bare minimum (15 minutes a week) has a huge impact. Start small.

[23.](#) (~This is not medical advice~). Don’t waste money on multivitamins, they don’t work. Vitamin D supplementation does seem to work, which is important because deficiency is common.

24. Phones have gotten heavier in the last decade and they’re actually pretty hard on your wrists! Use a computer when it’s an alternative or try to at least prop up your phone.

Success

25. History remembers those who got to market first. Getting your creation out into the world is more important than getting it perfect.

26. Are you on the fence about breaking up or leaving your job? You should probably go ahead and do it. People, on average, end up happier when they take the plunge.

27. Discipline is superior to motivation. The former can be trained, the latter is fleeting. You won't be able to accomplish great things if you're only relying on motivation.

28. You can improve your communication skills with practice much more effectively than you can improve your intelligence with practice. If you're not that smart but can communicate ideas clearly, you have a great advantage over everybody who can't communicate clearly.

29. You do not live in a video game. There are no pop-up warnings if you're about to do something foolish, or if you've been going in the wrong direction for too long. You have to create your own warnings.

30. If you listen to successful people talk about their methods, remember that all the people who used the same methods and failed did not make videos about it.

31. The best advice is personal and comes from somebody who knows you well. Take broad-spectrum advice like this as needed, but the best way to get help is to ask honest friends who love you.

32. Make accomplishing things as easy as possible. Find the easiest way to start exercising. Find the easiest way to start writing. People make things harder than they have to be and get frustrated when they can't succeed. Try not to.

33. Cultivate a reputation for being dependable. Good reputations are valuable because they're rare (easily destroyed and hard to rebuild). You don't have to brew the most amazing coffee if your customers know the coffee will always be hot.

34. How you spend every day is how you spend your life.

Rationality

35. Noticing biases in others is easy, noticing biases in yourself is hard. However, it has much higher pay-off.

36. Explaining problems is good. Often in the process of laying out a problem, a solution will present itself.

37. Foolish people are right about most things. Endeavour to not let the opinions of foolish people automatically discredit those opinions.

38. You have a plan. A time-traveller from 2030 appears and tells you your plan failed. Which part of your plan do you think is the one that fails? Fix that part.

39. If something surprises you again and again, stop being surprised.

40. Should you freak out upon seeing your symptoms on the worst diseases on WebMD? Probably not! Look up the base rates for the disease and then apply Bayes'

Theorem

[41.](#) Selfish people should listen to advice to be more selfless, selfless people should listen to advice to be more selfish. This applies to many things. Whenever you receive advice, consider its opposite as well. You might be filtering out the advice you need most.

42. Common systems and tools have been designed so everybody can handle them. So don't worry that you're the only one who can't! You can figure out doing laundry, baking, and driving on a highway.

Self

43. Deficiencies do not make you special. The older you get, the more your inability to cook will be a red flag for people.

[44.](#) There is no interpersonal situation that can't be improved by knowing more about your desires, goals, and structure. 'Know thyself!'

45. If you're under 90, try things.

[46.](#) Things that aren't your fault can still be your responsibility.

[47.](#) Defining yourself by your suffering is an effective way to keep suffering forever (ex. incels, trauma).

[48.](#) Keep your identity small. "I'm not the kind of person who does things like that" is not an explanation, it's a trap. It prevents nerds from working out and men from dancing.

49. Don't confuse 'doing a thing because I like it' with 'doing a thing because I want to be seen as the sort of person who does such things'

[50.](#) Remember that you are dying.

51. Events can hurt us, not just our perceptions of them. It's good to build resilience, but sometimes it isn't your fault if something really gets to you.

52. If you want to become funny, try just saying stupid shit (in the right company!) until something sticks.

[53.](#) To start defining your problems, say (out loud) "everything in my life is completely fine." Notice what objections arise.

54. Procrastination comes naturally, so apply it to bad things. "I want to hurt myself right now. I'll do it in an hour." "I want a smoke now, so in half an hour I'll go have a smoke." Then repeat. Much like our good plans fall apart while we delay them, so can our bad plans.

[55.](#) Personal epiphanies feel great, but they fade within weeks. Upon having an epiphany, make a plan and start actually changing behavior.

56. Sometimes unsolvable questions like "what is my purpose?" and "why should I exist?" lose their force upon lifestyle fixes. In other words, seeing friends regularly and

getting enough sleep can go a long way to solving existentialism.

Hazards

57. There are two red flags to avoid almost all dangerous people: 1. The perpetually aggrieved ; 2. The angry.

58. Some people create drama out of habit. You can avoid these people.

59. Those who generate anxiety in you and promise that they have the solution are grifters. See: politicians, marketers, new masculinity gurus, etc. Avoid these.

60. (~This is not legal advice!~)

DO NOT TALK TO COPS.

61. It is cheap for people to talk about their values, goals, rules, and lifestyle. When people's actions contradict their talk, *pay attention!*

62. "If they'll do it with you, they'll do it to you" and "those who live by the sword die by the sword" mean the same thing. Viciousness you excuse in yourself, friends, or teammates will one day return to you, and then you won't have an excuse.

Others

63. In choosing between living with 0-1 people vs 2 or more people, remember that ascertaining responsibility will no longer be instantaneous with more than one roommate ("whose dishes are these?").

64. Understand people have the right to be tasteless.

65. You will prevent yourself from even having thoughts that could lower your status. Avoid blocking yourself off just so people keep thinking you're cool.

66. Being in groups is important. If you don't want to join a sports team, consider starting a shitty band. It's the closest you'll get to being in an RPG. Train with 2-4 other characters, learn new moves, travel from pub to pub, and get quests from NPCs.

67. It's possible to get people to do things that make you like them more but respect them less. Avoid this, it destroys relationships.

68. Think a little about why you enjoy what you enjoy. If you can explain what you love about Dune, you can now communicate not only with Dune fans, but with people who love those aspects in other books.

69. When you ask people, "What's your favorite book / movie / band?" and they stumble, ask them instead what book / movie / band they're currently enjoying most. They'll almost always have one and be able to talk about it.

70. Bored people are boring.

71. A norm of eating with your family without watching something will lead to better conversations. If this idea fills you with dread, consider getting a new family.

72. If you bus to other cities, consider finding a rideshare on Facebook instead. It's cheaper, faster, and leads to interesting conversations.

Relationships

73. In relationships look for somebody you can enjoy just hanging out near. Long-term relationships are mostly spent just chilling.

74. Sometimes things last a long time because they're good (jambalaya). But that doesn't mean that because something has lasted a long time that it is good (penile subincisions). Apply this to relationships, careers, and beliefs as appropriate.

75. Don't complain about your partner to coworkers or online. The benefits are negligible and the cost is destroying a bit of your soul.

76. After a breakup, cease all contact as soon as practical. The potential for drama is endless, and the potential for a good friendship is negligible. Wait a year before trying to be friends again.

77. If you haven't figured things out sexually, remember that there isn't a deadline. If somebody is making you feel like there is, consider the possibility that they aren't your pal.

78. If you have trouble talking during dates, try saying whatever comes into your head. At worst you'll ruin some dates (which weren't going well anyways), at best you'll have some great conversations. Alcohol can help.

[79.](#) When dating, de-emphasizing your quirks will lead to 90% of people thinking you're kind of alright. Emphasizing your quirks will lead to 10% of people thinking you're fascinating and fun. Those are the people interested in dating you. Aim for them.

[80.](#) Relationships need novelty. It's hard to have novelty during Covid--but have you planned your post-Covid adventure yet?

81. People can be the wrong fit for you without being bad. Being a person is complicated and hard.

Compassion

[82.](#) Call your parents when you think of them, tell your friends when you love them.

83. Compliment people more. Many people have trouble thinking of themselves as smart, or pretty, or kind, unless told by someone else. You can help them out.

84. If somebody is undergoing group criticism, the tribal part in you will want to join in the fun of righteously destroying somebody. Resist this, you'll only add ugliness to the world. And anyway, they've already learned the lesson they're going to learn and it probably isn't the lesson you want.

85. Cultivate compassion for those less intelligent than you. Many people, through no fault of their own, can't handle forms, scammers, or complex situations. Be kind to them because the world is not.

86. Cultivate patience for difficult people. Communication is extremely complicated and involves getting both tone and complex ideas across. Many people can barely do either. Don't punish them.

87. Don't punish people for trying. You teach them to not try with you. Punishing includes whining that it took them so long, that they did it badly, or that others have done it better.

88. Remember that many people suffer invisibly, and some of the worst suffering is shame. Not everybody can make their pain legible.

89. Don't punish people for admitting they were wrong, you make it harder for them to improve.

90. In general, you will look for excuses to not be kind to people. Resist these.

Joy

91. Human mood and well-being are heavily influenced by simple things: Exercise, good sleep, light, being in nature. It's cheap to experiment with these.

92. You have vanishingly little political influence and every thought you spend on politics will probably come to nothing. Consider building things instead, or at least going for a walk.

93. Sturgeon's law states that 90% of everything is crap. If you dislike poetry, or fine art, or anything, it's possible you've only ever seen the crap. Go looking!

94. You don't have to love your job. Jobs can be many things, but they're also a way to make money. Many people live fine lives in okay jobs by using the money they make on things they care about.

95. Some types of sophistication won't make you enjoy the object more, they'll make you enjoy it less. For example, wine snobs don't enjoy wine twice as much as you, they're more keenly aware of how most wine isn't good enough. Avoid sophistication that diminishes your enjoyment.

96. If other people having it worse than you means you can't be sad, then other people having it better than you would mean you can't be happy. Feel what you feel.

97. Liking and wanting things are different. There are things like junk food that you want beyond enjoyment. But you can also like things (like reading) without wanting them. If you remember enjoying something but don't feel a desire for it now, try pushing yourself.

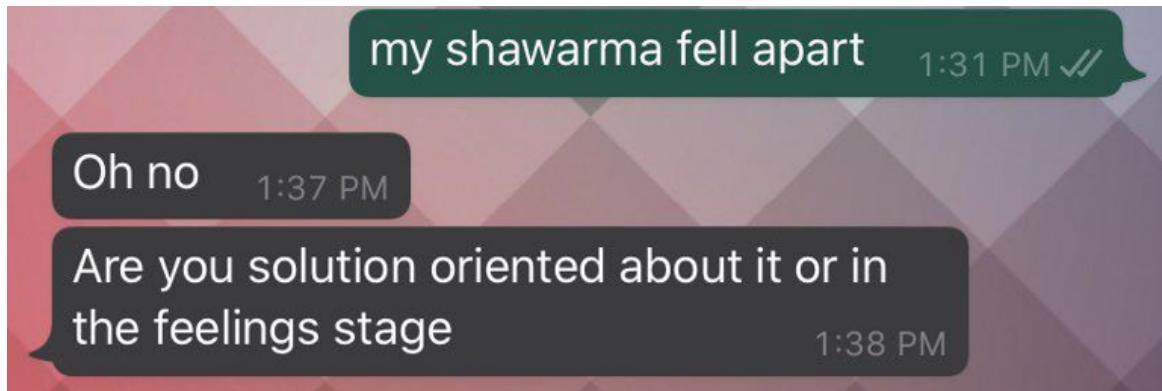
98. People don't realize how much they hate commuting. A nice house farther from work is not worth the fraction of your life you are giving to boredom and fatigue.

99. There's some evidence that introverts and extroverts both benefit from being pushed to be more extroverted. Consider this the next time you aren't sure if you feel

like going out.

100. Bad things happen dramatically (a pandemic). Good things happen gradually (malaria deaths dropping annually) and don't feel like 'news'. Endeavour to keep track of the good things to avoid an inaccurate and dismal view of the world.

To listen well, get curious



[source](#)

A common piece of interacting-with-people advice goes: “often when people complain, they don’t want help, they just want you to listen!”

For instance, *Nonviolent Communication*:^{*} *Nonviolent Communication*, ch. 7.

It is often frustrating for someone needing empathy to have us assume that they want reassurance or “fix-it” advice.

Active Listening:[†] [Active Listening](#), p. 2

Similarly, advice and information are almost always seen as efforts to change a person and thus serve as barriers to his self-expression and the development of a creative relationship.

You can find similar advice in most books on relationships, people management, etc.

This always used to seem silly to me. If I complain at my partner and she “just listens,” I’ve accomplished nothing except maybe made her empathetically sad. When I complain at people, I want *results*, not to grouse into the void![‡] Empirically, I did notice that I usually got better results from listening than from giving advice. So I inferred that this advice was true for other people, but not me, because other people didn’t actually want to fix their problems.

Frequently the “just listen” advice comes with tactical tips, like “reflect what people said back to you to prove that you’re listening.” For instance, consider these example dialogues from *Nonviolent Communication*:[§] *Nonviolent Communication*, Chapter 7, Exercise 5.5, 5.6 and solutions.

Person A: How could you say a thing like that to me?

Person B: Are you feeling hurt because you would have liked me to agree to do what you requested?

Or:

Person A: I’m furious with my husband. He’s never around when I need him.

Person B: So you’re feeling furious because you would like him to be around more than he is?

I say this with great respect for *Nonviolent Communication*, but these sound like a [1970s-era chatbot](#). If I were Person A in either of these dialogues my next line would be “yes, you

dingbat—can you turn the nonviolence down a couple notches?” I’d feel alienated knowing that someone is going through their NVC checklist on me.

Recently, I realized why people keep giving this weird-seeming advice. Good listeners *do* often reflect words back—but not because they read it in a book somewhere. Rather, it’s [cargo cult advice](#): it teaches you to imitate the surface appearance of good listening, but misses what’s actually important, the thing that’s *generating* that surface appearance.

The generator is curiosity.

When I’ve listened the most effectively to people, it’s because I was intensely curious—I was trying to build a *detailed, precise* understanding of what was going on in their head. When a friend says, “I’m furious with my husband. He’s never around when I need him,” that one sentence has a huge amount underneath. How often does she need him? What does she need him for? Why isn’t he around? Have they talked about it? If so, what did he say? If not, why not?

It turns out that [reality has a surprising amount of detail](#), and those details can matter a lot to figuring out what the root problem or best solution is. So if I want to help, I can’t treat those details as a black box: I need to [open it up and see the gears inside](#). Otherwise, anything I suggest will be wrong—or even if it’s right, I won’t have enough “shared language” with my friend for it to land correctly.

Some stories from recent memory:

- When we started doing a pair programming rotation at Wave, I suggested that, to make scheduling easier, we designate a default time when pairing sessions would happen. A coworker objected that this seemed authoritarian. I was extremely puzzled, but they’d previously mentioned being an anarchist, so I was tempted to just chalk it up to a political disagreement and move on. But instead I tried to get curious and explore more deeply whatever “political” models were generating that disagreement. After a lot of digging into what was or wasn’t authoritarian for them and why, it turned out the disagreement was because they’d missed the word “default” and thought I was suggesting a single *mandatory* time for pair programming.
- My partner, Eve, wrote a post about Polish attitudes about sex, with some details that upset her (Polish) parents. When her parents told her that, she initially got very stressed about having to have a conversation to calm them down. I thought she shouldn’t be worried and the conversation would be fine, but of course just telling her that wasn’t very helpful. Instead, I summoned up my curiosity and asked lots of questions about her relationship with her parents, her parents’ relationship with each other, each of their relationships with Catholicism, etc. By the end of the conversation, after thinking through all the baggage involved, Eve agreed with me, and her attitude about the upcoming conversation shifted from impending doom to compassionate curiosity about where her parents were coming from.
- I was stressed by work and complained to Eve about some things that I felt frustrated and stuck about. Instead of suggesting solutions, she kept asking for more details until she had more or less a complete snapshot of my mental state. At that point, she observed that every time I mentioned feeling sad, I sounded contemptuous and exasperated with myself. She hypothesized that I wasn’t giving myself permission to be sad. The “solution” to my problem ended up being to give me a big hug and let me cry on her shoulder for a bit, after which I immediately felt much less stressed.

In each case, the “helper” tried to learn about the “complainant’s” reality in as much detail as possible—not just the problem, but the whole person and whatever else was behind the immediate issue. And that’s what made it possible for them to actually help.

It often feels like I understand enough to be helpful without knowing all those details. But when I think that, I'm usually wrong: I end up giving bad advice, based on bad assumptions, and the person I'm talking to ends up having to do a bunch of work to argue with me and correct my bad assumptions. That makes the conversation feel disfluent and adversarial instead of collaborative.

It turns out this is a really common failure mode of helping-conversations, which is what I think generates the old saw at the beginning of this post, that “sometimes people don’t want help, just to be listened to.”

But I think that’s actually too nice to the helper, and uncharitable to the complainer (in that it assumes they weirdly don’t care about solving their problem). What’s really going on is probably that your advice is bad, because you didn’t really listen, because you weren’t curious enough.

When I’m curious about what someone’s saying, I often do repeat things back to them in my own words. But it’s because I’m genuinely curious, not because I’m checking off the “reflect words” box in my “be a good listener” checklist. That means I do it in a way that sounds like my natural speech, instead of mimicking them like a chatbot.

When done this way, reflective listening feels validating rather than alienating. It’s a way of demonstrating that I care a lot about what someone has to say. Putting their idea into my own words shows them that I’ve fully digested it, and helps us establish a shared language in which to talk about it. That, in turn, makes the conversation fluent and collaborative, rather than a zigzag of bad assumptions and corrections.

So the right advice isn’t “listen harder and repeat everything back”—you won’t be genuine if you’re just imitating the surface appearance of a good listener. Instead, be humble and get curious! Remind yourself that there’s a ton of detail behind whatever you’re hearing, and try to internalize all of it that you can. Once you’ve done that, your advice will be more likely hit the mark, and you’ll be able to communicate it clearly.

Great minds might not think alike

This is a linkpost for <https://ericneyman.wordpress.com/2020/12/26/alike-minds-think-great/>

[Previously known as "Alike minds think great"]

I.

It is famously the case that almost everyone thinks they're above average. Derek Sivers [writes](#):

Ninety-four percent of professors say they are better-than-average teachers.

Ninety percent of students think they are more intelligent than the average student.

Ninety-three percent of drivers say they are safer-than-average drivers.

Interesting. Intuitively this seems to suggest that people are prone to vastly overestimate their competence. But is that true? As Bill Kuszmaul [points out](#), these people aren't necessarily wrong!

There's no fundamental reason why you can't have 90% of people be better than average. For example, more than 99.9% of people have an above-average number of legs. And more than 90% of people commit fewer felonies than average. These examples are obvious, but they're not so different than some of the examples [in Sivers' post].

This has something to it! On the other hand, I don't think this explains everything. Is the quality of a professor's teaching really so skewed that 94% are above average? But more importantly, do you really think that way fewer people would answer "yes" if you just replaced the word "average" with "median" when asking the question?

That said, I don't think these numbers necessarily point to a bias! That's because the interpretation of "above average" is left entirely up to the person being asked. Maybe you think a good driver is one who drives safely (and so you drive safely and slowly) whereas I think a good driver is one who gets from point A to point B efficiently (and so I drive quickly but not safely). We are both, from our own perspectives, above average drivers!

Put otherwise, for any skill where "goodness at that skill" doesn't have an *objective, agreed-upon measure*, we should expect more than 50% of people to think they're better than the median, because people optimize for things they care about.

To give a personal example, I suppose I would call myself an above average blogger. This isn't true in some objective sense; it's just that I judge bloggers by *how interesting their thoughts are to me*, and obviously I write about things that are interesting to me! There's no bias I'm falling for here; it's just that "Are you an above average blogger?" leaves "above average" open to my interpretation.

II.

There is, however, a closely related bias that I and lots of other people have. This bias occurs when we take a situation like those above, but now create a more objective test of that skill. To illustrate with an example, suppose you asked all the students at a university whether they have an above-median GPA. If 90% of students said yes, that would demonstrate a widespread bias — because unlike “Are you a better than the median student”, here there’s no room for interpretation.

The way this bias manifests in me (and many others I imagine) is: *I tend to underestimate the competence of people who think very differently from me*. I started thinking about this the other day when I listened to [Julia Galef's podcast episode with David Shor](#) (which I *highly recommend*¹). Shor is a young Democratic political strategist, originally hired by Barack Obama’s 2012 reelection campaign to run their data operation and figure out how the campaign should spend its money. Shor says:

When I first started in 2012, I was 20 and I was like, “Oh, I’m going to do all of this math and we’re going to win elections.” And I was with all these other nerds, we were in this cave. We really hated these old school consultants who had been in politics for like 20 years. [...] We had all these disagreements because the old school consultants were like, “You need to go up on TV, you need to focus on this..” And we really disagreed.

Put yourself in Shor’s shoes: you join an operation that’s being run by consultants who have no background in data, haven’t looked at randomized controlled trials on different interventions, etc. You know tons of math, start reading about the RCTs, looking at poll numbers, and they point you toward really different things from what the campaign was doing. You’d probably be really concerned that the campaign was doing things totally wrong! But as Shor goes on to say:

I think going back, probably 80% of the disagreements I had with these old school consultants in 2012, looking back, I think they were right.

I don’t want to accuse David Shor of any particular bias without evidence, but allow me for illustrative purposes to create a fictional David Shor who may or may not reflect the actual thoughts of the real David Shor in 2012. If you’d like, you can think of this character “strawman-Shor”, or as my own self-insertion.

Suppose Obama were to ask David Shor: “Be honest, **do you think you’re better than the old school consultants at real-world reasoning?**” If he suppressed any false modesty, he might have said something like:

Well, I have a good grasp of statistics and good logical reasoning skills. Show me a trend in home ownership and I’ll tell you if it’s robust or just some nice-looking noise. Tell me the sensitivity and specificity of a cancer screening test and the prevalence of that cancer, and I’ll tell you how likely someone who tested positive is to have cancer. Give me a spreadsheet of historical polling data and I’ll tell you if you’re likely to win the election. I bet the old school consultants can’t do that as well as I can. So yeah, I’m better at real-world reasoning.

Fair enough. Now suppose Obama were to ask an old school consultant (let’s call her Constance): “Do you think you’re better than David Shor at real-world reasoning?” She might have said:

Wait, who’s David Shor again?

and then

Oh, that quant we just hired who's fresh out of college? Listen, I understand how people think really well. Give me an opinion piece and I'll read between the lines to figure out the author's motivations. Drop me off at a party full of people I haven't met before and by the end I'll have a pretty good idea of what makes them tick. Ask me to spend ten minutes with a swing voter and I'll tell you what things really motivate them. I bet Shor couldn't do any of these things. So yeah, I'm better at real-world reasoning.

Which — again — fair enough! Neither of them are *wrong*; they just place a high value on different skills. The sort of real-world reasoning that is salient to Shor is figuring out how systems work by analyzing them logically and statistically. By contrast, Constance thinks of real-world reasoning as understanding how people behave. If Shor is better at the former and Constance is better at the latter, but they think of "real-world reasoning" in different ways, both of their replies are "correct".

But now, Obama asks Shor a follow-up question. "So, do you think you're better than the consultants at figuring out how to spend our money so as to most increase our chances of winning?" And I say:

Yeah — this is a great example of real-world reasoning, and I'm better at that.

And when Obama asks Constance, she also says:

Yeah — this is a great example of real-world reasoning, and I'm better at that.

In this case — if David Shor's 80% estimate from the podcast is right — Constance would have been right most of the time. It could have easily been the reverse, if Obama chose to ask about a different concrete measure! But the point is: Shor and Constance both — predictably — said they'd be better at the thing Obama asked them about.

I would guess that this sort of reasoning happens a lot. In concrete terms:

1. A person (call her Alice) forms a heuristic — "I am good at X" — where X isn't perfectly defined. ("I am good at real-world reasoning"; "I am good at driving"; "I am a good math teacher".) She forms it because she's good at X on a particular axis she cares about ("I am good at statistical problem solving"; "I drive safely"; "My algebraic geometry classes consistently get great reviews").
2. Alice encounters a specific problem where being good at X is important, though maybe in a different form. (Figuring out how to spend campaign money; Getting to the airport really quickly; Teaching young kids arithmetic.)
3. Alice pattern matches the problem as *an instance of X*, thinks *I'm good at X*, and (this is the fallacy) concludes *I'm good at this problem*.

Shor and Constance are both good at real-world reasoning, but in different ways. If they both end up committing this fallacy, the end result is that each of them thinks they are better than the other at the concrete problem at hand (spending campaign money). More generally, misapplying heuristics in this way would lead you to overestimate the competence of like-minded people (who are good at X along the dimensions you value) relative to those who think differently.

The last thing I want to say here is: from personal experience, I'm pretty sure that overestimating the competence of like-minded people (if you want a concept handle

for this bias, may I suggest “alike minds think great”?) *actually happens*. But why do I think that the mechanism I just proposed is the right explanation? Well, I don’t think it’s all of the explanation (I can think of other plausible mechanisms²) but I would speculate that it’s *part* of what’s going on, if only because I’m pretty sure that my brain often makes the sort of mistake I outline in steps 1 through 3.

III.

It might be inconvenient that Shor and Constance underestimate each other’s competence, but the problem runs deeper. If this bias were the entire problem, the two of them could have laid out their arguments, judged the relative merits, and resolved the disagreement over campaign spending. The deeper issue is that Shor and Constance have different ways of thinking about the world, so resolving this disagreements is really tricky. Shor is inclined to distrust Constance’s reliance on intuition and past experience, and Constance is inclined to distrust Shor’s reliance on statistical methods. Like many smart people with different ways of thinking, they’re likely to talk past each other. But, like, they’re on a team together. They have to come to a decision. How are they supposed to do that?

First, both Shor and Constance could stand to be a bit less confident. If only because the previously discussed bias is a real possibility, they should both entertain the notion that the other is right and they are wrong. From there it isn’t too hard to strike some sort of balance between the two approaches.

But, striking some arbitrary (maybe 50/50) balance isn’t great either! If Shor could have figured out back then that Constance was right 80% of the time, they would have been able to allocate Obama’s resources much better. So it’s truly important for them to *really figure out* who’s right.

And for that, they need a really important sort of person. The more I think about it, the more I believe that this is an utterly crucial role for some members of society to fill. This role is **the translator**.

(Edit: Kelsey Piper independently [coined the same word](#) for the same concept in 2015!)

IV.

In June, when Scott Alexander put [his blog](#) on pause over a possible New York Times article that would reveal his real name, outrage erupted on Twitter. The vast majority of comments were supportive of Scott and called for the New York Times to withhold his name from the piece. But I also saw some criticisms of Scott, of which one stuck out to me. It went something like this:

Scott Alexander’s blog is really overrated. People seem to think he’s really insightful, but as far as I can tell he just takes really obvious concepts and explains them in an unnatural, convoluted way.

I dismissed this critique as... obviously wrong?... at the time. Scott’s essays were consistently enlightening, fitting different pieces of the world together like a jigsaw puzzle, helping me understand the world better. Some examples:

- [I Can Tolerate Anything Except The Outgroup](#), which made me understand why people often get much more upset at those with a slightly different opinion than at those who are completely opposed to everything they believe in.
- [The Toxoplasma Of Rage](#), which made me understand why activists so often seem to use divisive tactics.
- [Right Is The New Left](#), which gave me a model of how beliefs interact with social status, and also gave me a model of how fashion trends work, something I previously had no conception of at all.

In all three of these (admittedly somewhat cherry-picked) examples, Scott explained a *social phenomenon* with an *analytical model*. This was *fantastic* for me! It's more difficult for me than for most others to understand social dynamics at an intuitive level; but analytical models — that's *totally my thing*.

Whereas from the critic's perspective, these explanations didn't appear particularly insightful. *Of course* people get upset at the outgroup more than the far-group. *Of course* it makes sense for activists to use divisive tactics. *Of course* changes in fashion styles are a the result of a status game. Perhaps Scott Alexander's critic grasped all of these things on an intuitive level. If you will, the critic is somewhat like Constance, and I'm somewhat like Shor.

Scott Alexander is what I would call a **translator**. He takes concepts that are really natural to the Constances of the world and explains them in a way that makes sense to the Shors of the world. And this is *really, really useful*. It's useful to Shors because it makes them understand the world better. But it's also useful to Constances because it's good for them too if Shors understand these things. It makes it easier for them to reconcile disagreements, and it makes Shors come to better conclusions in e.g. matters of public policy that affect Constances too.

*(Edit: I don't mean to suggest that Scott Alexander's explanations can't be useful for Constance — but at minimum it's harder for her to understand the insights since they aren't written in Constance's "native language". And I certainly don't mean to suggest that translation is the **only** value of the posts.)*

There are no doubt effective translators in the other direction, too. An example would be a mathematician or statistician who's really good at using real-world examples to explain abstract concepts to people who aren't mathematically inclined. [Jordan Ellenberg](#) and [Eugenia Cheng](#) might be good examples (though I don't know, since I'm not their audience!).

V.

Being a translator requires a pretty unique skill set. Scott Alexander needs to understand social dynamics really well. He *also* needs to understand analytical methods well enough to use them in his explanations. He also needs to be able to *speak the language* of Shors, because his explanation will ultimately be in their language. Think in terms of language-to-language translators: to translate really well from language A to language B, you need be masterful at A (to understand all the subtleties of the meaning) and also at B (to convey that meaning while preserving those nuances). That's why good translators (in both senses) are so rare.

But because there are so many Shors and so many Constances in the world, translators are really, really important. Having translators would have given David

Shor and the consultant an opportunity to figure out which of them was right (or at least come closer to an agreement). Let me give an example of what a Constance-to-Shor translator could have said:

Constance's intuition that Obama should use a slogan that highlights his position on economic issues comes from her work in the 1990s, when there were lots of persuadable voters and you could measure effects of candidates' speeches in real time. From that time period we have lots of examples of economic issues playing better for Democrats than cultural issues. Now, times have changed, and these effects have gotten smaller as the number of persuadable voters has diminished, but the "what persuades undecided voters hasn't changed" prior is a reasonable one to start with. Now, your data should make us update toward considering the alternative more effective, but given the messiness of our data and the small effect sizes, I would argue that we should mostly stick to our prior.

How do I know that this would have been the right thing to say? Because in the podcast, this is how David Shor currently explains why the old-school consultants ended up being right! It would have been *really useful* for Shor to have this sort of explanation at the time. But — I theorize — there wasn't an effective translator. This is no surprise — as I mentioned, translators are few and far between — but this example goes to show how useful a translator can be.

VI.

I'll conclude by noting that while I talked specifically about translators between *analytically-minded* and *socially-minded* people, translators can exist and are extremely useful between any two groups of people who think really differently! Some examples:

- Translators between different *ideologies*. Liberals and conservatives not only have different beliefs, they *reason about political issues* totally differently. They have [different notions of fairness](#) and justice, different models of society, etc. It's really useful to have someone who can present conservative arguments in liberal-friendly language and vice versa. A good example here, perhaps, is Bill Clinton, who was called "explainer in chief" for being able to explain liberal ideas in ways that appealed to conservatives.
- Translators between different *cultures*. People from different cultures are often brought up to see the world in different ways, and translators are important for bridging cultural divides.
- Translators between different *socio-economic classes*. We often talk about a disconnect between coastal elites and non-elites. In fact, this is (probably correctly) pointed to as a major source of political polarization. Effective translator could bridge this divide, perhaps decreasing polarization (or at least slowing its increase).

Fostering mutual understanding is really important for social cohesion and for truth-seeking. Doing this between two people who have similar thought patterns is relatively easy. Helping Republicans understand Democrats, economic elites understand middle-Americans, Shors understand Constances — *that* is hard. For *that*, you might just need a translator.

1. The day before the episode was released, I [wrote a tweet](#) that listed Julia Galef and David Shor as two of the four people I follow on Twitter who reliably have great tweets. (They both “liked” the tweet, no doubt knowing that they were about to make me quite happy.) I sure had high expectations, and the episode did not disappoint.
2. One other possible mechanism: when people discuss technical things with people who think differently from them, they often don’t understand those people’s points (to a greater extent than discussing with like-minded people). Their brains jump from “this don’t make sense to me” to “this person’s thoughts don’t make sense”, causing them to underestimate the person’s competence.

Covid 12/24: We're F***ed, It's Over

UPDATE 7/21/2021: As you doubtless know at this point, it was not over. Given the visibility of this post, I'm going to note here at the top that the prediction of a potential large wave of infections between March and May did not happen, no matter what ultimately happens with Delta (and the prediction was not made with Delta in mind anyway, only Alpha). Some more reflections on that at the bottom of this post [here](#).

A year ago, there were reports coming out of China about a new coronavirus. Various people were saying things about exponential growth and the inevitability of a new pandemic, and urging action be taken. The media told us it was nothing to worry about, right up until hospitals got overwhelmed and enough people started dying.

This past week, it likely happened again.

A new strain of Covid-19 has emerged from southern England, along with a similar one in South Africa. The new strain has rapidly taken over the region, and all signs point to it being about 65% more infectious than the old one, albeit with large uncertainty and error bars around that.

I give it a 70% chance that these reports are largely correct.

There is no plausible way that a Western country can sustain restrictions that can overcome that via anything other than widespread immunity. This would be the level required to previously cut new infections in half every week. And all that would do is stabilize the rate of new infections.

Like last time, the media is mostly assuring us that there is nothing to worry about, and not extrapolating exponential growth into the future.

Like last time, there are attempts to slow down travel, that are both not tight enough to plausibly work even if they were implemented soon enough, and also clearly not implemented soon enough.

Like last time, no one is responding with a rush to get us prepared for what is about to happen. There are no additional pushes to improve our ability to test, or our supplies of equipment, or to speed our vaccine efforts or distribute the vaccine more efficiently (in any sense), or to lift restrictions on useful private action.

Like last time, the actions urged upon us to contain spread clearly have little or no chance of actually doing that.

The first time, I made the mistake of not thinking hard enough early enough, or taking enough action. I also didn't think through the implications, and didn't do things like buying put options, even though it was obvious. This time, I want to not make those same mistakes. Let's figure out what actually happens, then act upon it.

We can't be sure yet. I only give the new strain a 70% chance of being sufficiently more infectious than the old one that the scenario fully plays out here in America before we have a chance to vaccinate enough people. I am very willing to revise that probability as new data comes in, or based on changes in methods of projection, including projections of what people will decide to do in various scenarios.

What I do know is we can't hide our heads in the sand again. Never again. When we have strong Bayesian evidence that something is happening, we need to work through that and act accordingly. Not say "there's no proof" or "we don't know anything yet." This isn't about proof via experiment, or ruling out all possible alternative explanations. This is about

likelihood ratios and probabilities. And on that front, as far as I can tell, it doesn't look good. Change my mind.

The short term outlook in America has clearly stabilized, with R0 close to 1, as the control system once again sets in. Cases and deaths (and test counts) aren't moving much. We have a double whammy of holidays about to hit us in Christmas and New Year's, but after that I expect the tide to turn until such time as we get whammed by a new more infectious strain.

Instead of that being the final peak and things only improving after that, we now face a potential fourth wave, likely cresting between March and May, that could be sufficiently powerful to substantially overshoot herd immunity.

Let's run the numbers.

The Numbers

Predictions

Last week's prediction: 13.1% positive rate on 11.5 million tests, and an average of 2,850 deaths per day.

Results: 13.7% positive rate on 10.7 million tests, with an average of 2,677 deaths.

We didn't test substantially more people than last week, and the positive test percentage didn't fall much, and the death rate didn't rise much. Everything in a holding pattern. Some of that could be pending Christmas issues. Fool me twice, and all that.

This next week is Christmas plus New Year's, so reporting issues are inevitable. Another week for wide error bars.

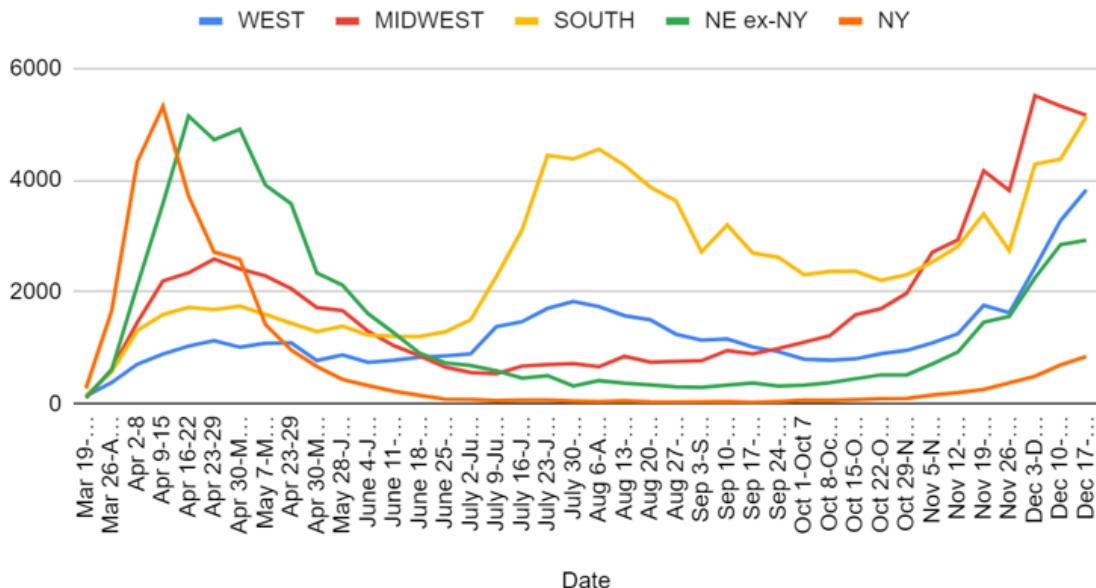
Prediction: 13.6% positive rate on 10.1 million tests, and an average of 2,500 deaths per day.

Note that I expect the deaths decline, at least, to be about reporting rather than about actual numbers, which I don't think will start to decline much for a bit longer.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 22-Oct 28	895	1701	2208	612
Oct 29-Nov 4	956	1977	2309	613
Nov 5-Nov 11	1089	2712	2535	870
Nov 12-Nov 18	1255	2934	2818	1127
Nov 19-Nov 25	1761	4169	3396	1714
Nov 26-Dec 2	1628	3814	2742	1939
Dec 3-Dec 9	2437	5508	4286	2744
Dec 10-Dec 16	3278	5324	4376	3541
Dec 17-Dec 23	3826	5158	5131	3772

Deaths by Region

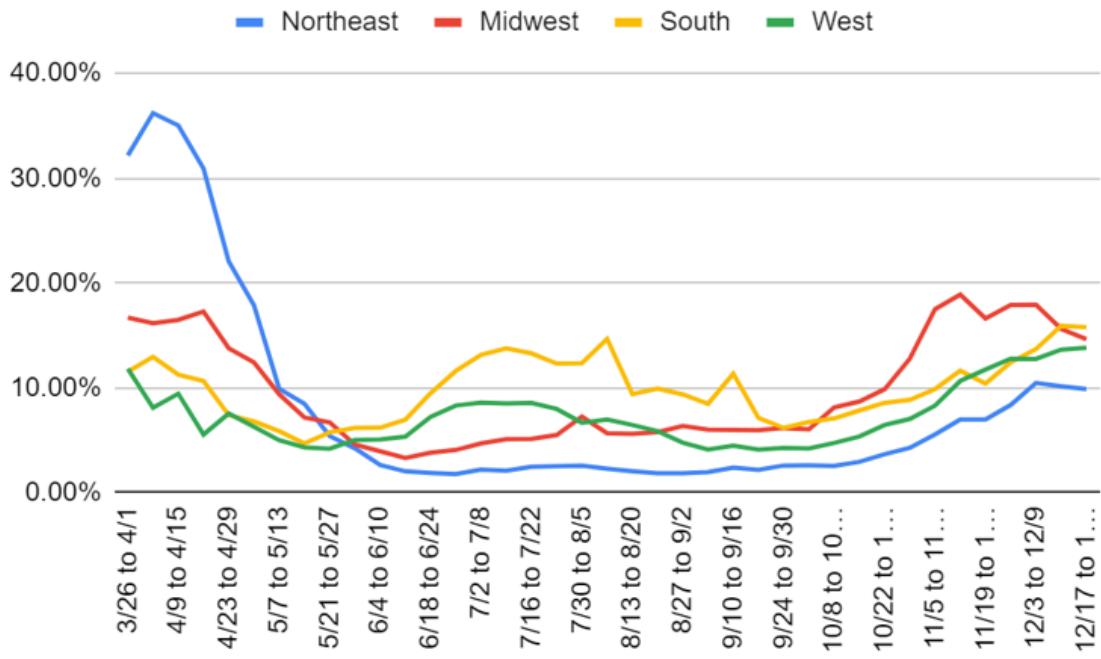


Death rates didn't rise all that much due partly to the decline in the Midwest, but they are still up in the other three regions and substantially in the West and South. It seems clearly a few weeks too early to hit peak deaths based on when we had peak infections.

Going forward, the vaccine will start to be effective for those who get infected next week, so to the extent we are protecting residents of nursing homes, we'll see that effect in the death rate start to be noticeable in late January. I expect deaths to be in decline by then.

Positive Test Percentages

Percentages	Northeast	Midwest	South	West
10/22 to 10/28	3.68%	9.87%	8.58%	6.46%
10/29 to 11/4	4.28%	12.79%	8.86%	7.04%
11/5 to 11/11	5.56%	17.51%	9.89%	8.31%
11/12 to 11/18	6.99%	18.90%	11.64%	10.66%
11/19 to 11/25	7.00%	16.62%	10.41%	11.75%
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%
12/3 to 12/9	10.47%	17.94%	13.70%	12.76%
12/10 to 12/16	10.15%	15.63%	15.91%	13.65%
12/17 to 12/23	9.88%	14.65%	15.78%	13.82%

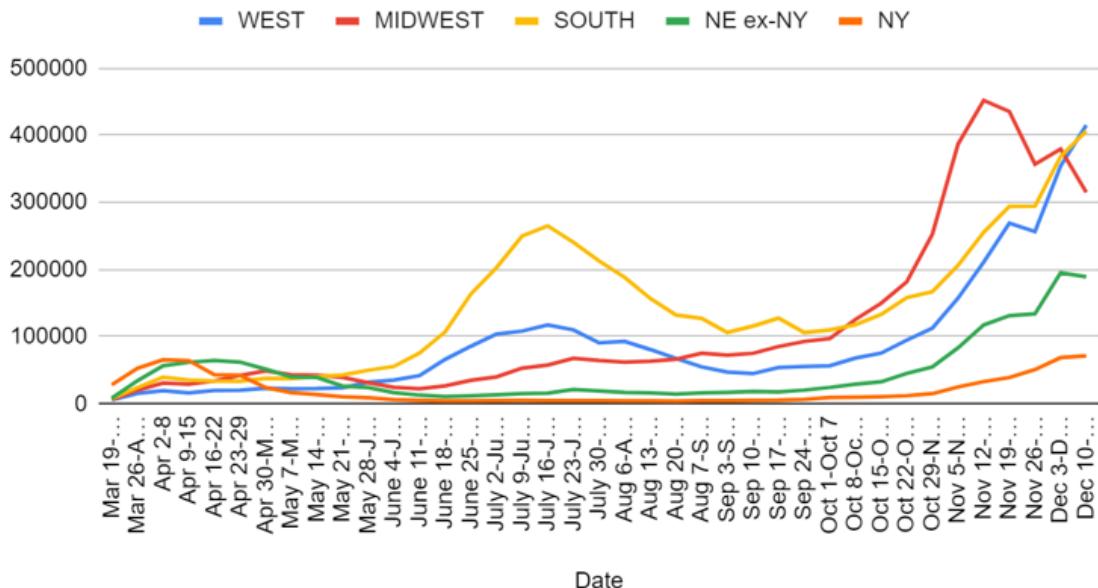


Quiet on all fronts. Midwest clearly in slow decline, looks like other regions also ready to follow, as soon as we get past the holidays. There's also an overall one-time bump coming of unknown size, but after that we should be clear until the new English or South African strain becomes a problem.

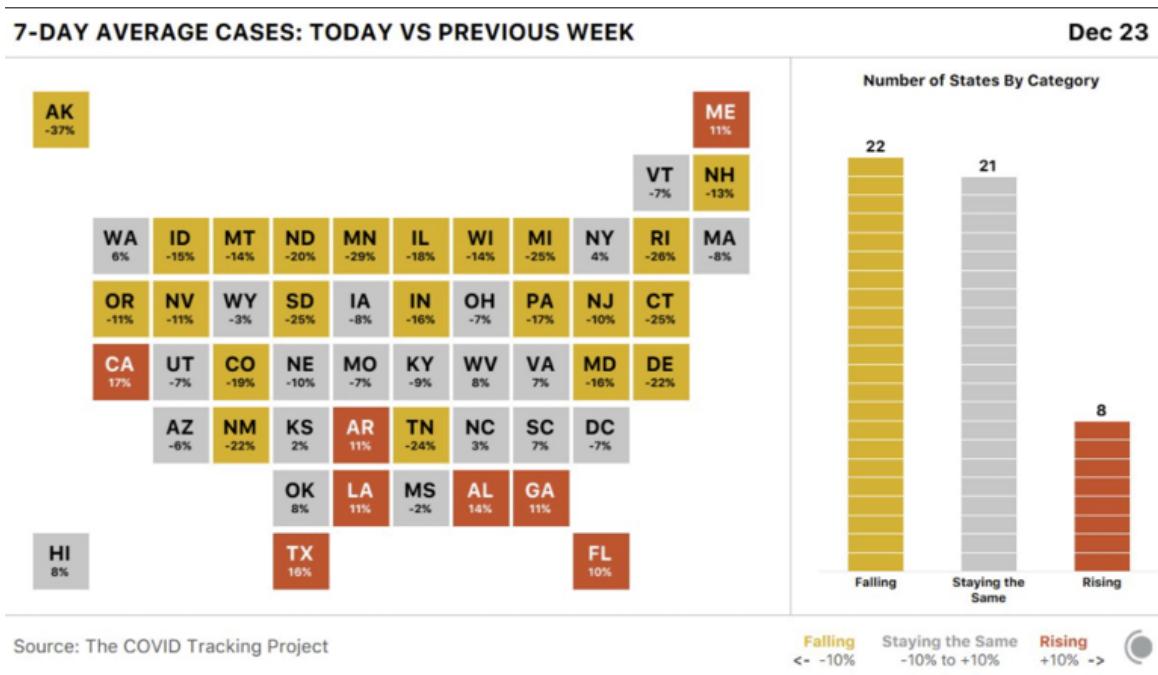
Positive Tests

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 22-Oct 28	94983	181881	158123	57420
Oct 29-Nov 4	112684	252917	167098	70166
Nov 5-Nov 11	157495	387071	206380	108581
Nov 12-Nov 18	211222	452265	255637	150724
Nov 19-Nov 25	269230	435688	294230	170595
Nov 26-Dec 2	256629	357102	294734	185087
Dec 3-Dec 9	354397	379823	368596	263886
Dec 10-Dec 16	415220	315304	406353	260863

Positive Tests by Region



Also, this seems like a nice graphic:



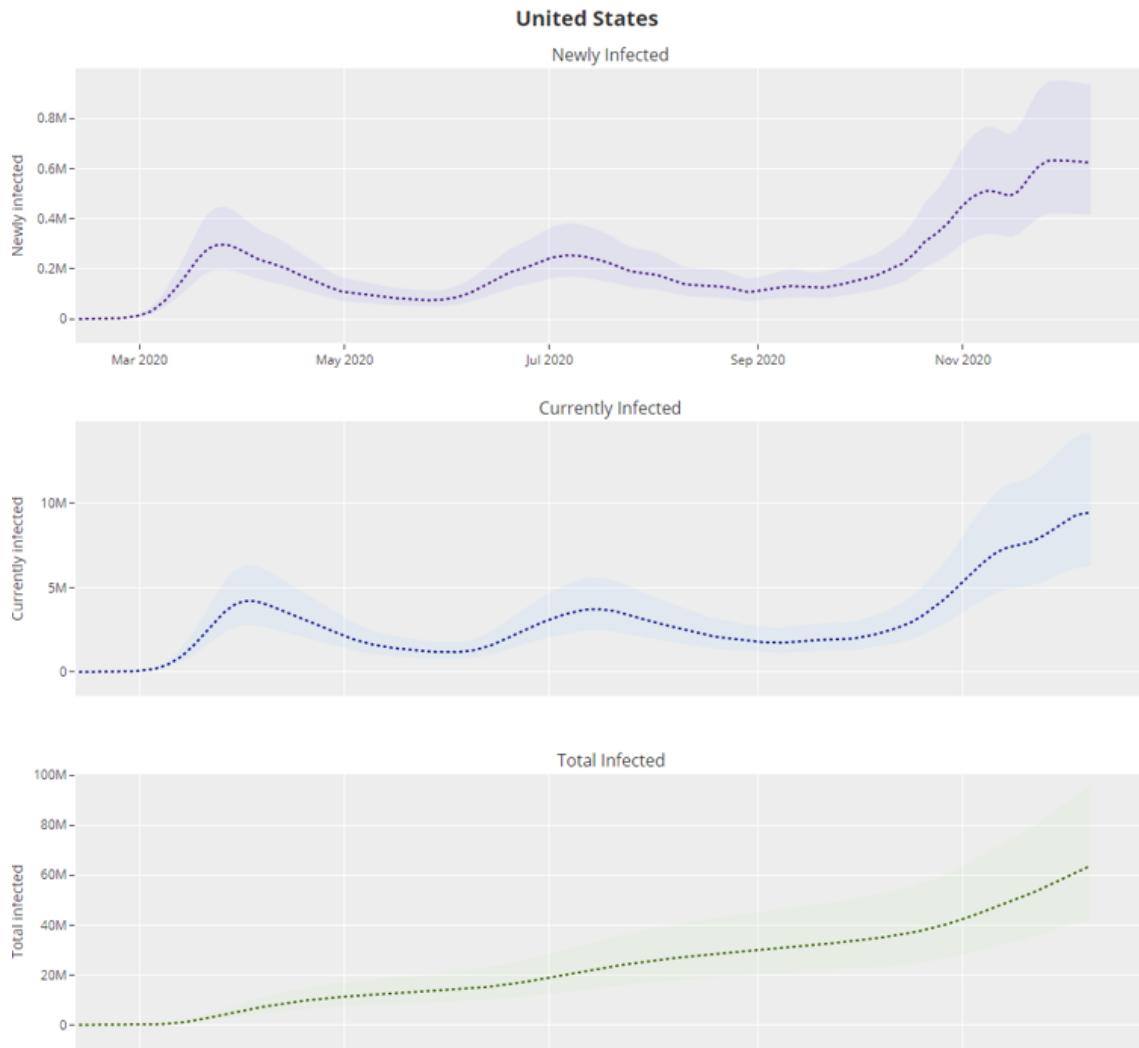
But it's also potentially misleading, because the increased cases in the deep South are mostly increased testing. The California increase is largely real. It would make sense that California would be the place that has stalled its crisis for the longest, between not having cold weather and imposing draconian restrictions all year, and so perhaps it's finally time for them to face the music.

Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Oct 15-Oct 21	6,461,028	6.4%	865,890	1.2%	2.52%
Oct 22-Oct 28	6,943,470	7.5%	890,185	1.4%	2.67%
Oct 29-Nov 4	7,349,648	9.5%	973,777	1.6%	2.89%
Nov 5-Nov 11	8,285,878	10.7%	1,059,559	2.4%	3.16%
Nov 12-Nov 18	9,033,621	12.4%	1,155,670	2.9%	3.50%
Nov 19-Nov 25	10,415,393	11.8%	1,373,751	2.9%	3.87%
Nov 26-Dec 2	9,741,057	11.7%	1,287,010	4.0%	4.22%
Dec 3-Dec 9	10,458,644	13.9%	1,411,142	4.9%	4.66%
Dec 10-Dec 16	10,694,845	13.8%	1,444,725	4.9%	5.11%
Dec 17-Dec 23	10,710,356	13.7%	1,440,770	5.1%	5.56%

This is the silent scandal no one is talking about. Why are we no longer expanding testing? It seems clear now that our capacity hasn't been expanding in December. It's clear that demand greatly exceeds supply, and that more testing would be a huge help. When things were improving slowly, at least they were improving. Now it looks like we are stalled out, well short of where we need to be. Vaccinations are important, but until we get a lot farther along on those, so are tests.

Covid Machine Learning Projections



Machine learning projections say infections have been static since about November 25, which mostly matches the testing data. We can assume that the projections will keep saying similar things for the next two weeks.

Their predicted total infected is up to 19.2% on December 9, stabilized at 623k new infections per day. The total is up from 17.9% on December 2. As a reminder, I consider these lower bound estimates.

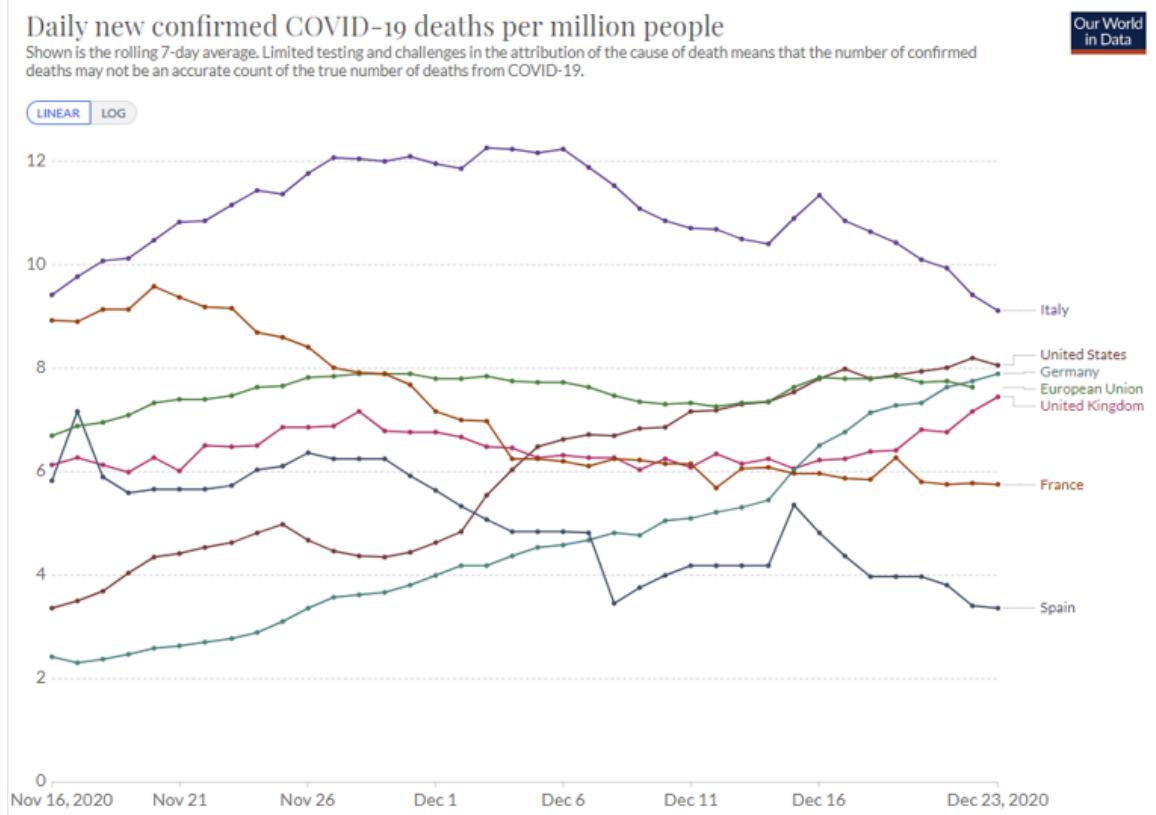
The immunity effects here compound fast. Even if you assume people get infected completely at random, going from 17.9% to 19.2% immune reduces R₀ by 1.6%, which reduces infection levels by that much every five days or so, or 9% per month, and we're introducing that effect permanently each week. After a month of this level of effect, you'll see a 16% decline from newly immune people alone. After two months, infection levels would be cut in half.

Selection of who gets infected makes the effect bigger, and also we get to add vaccinations.

Of course, the control systems ensure it does not work that way, as people will notice things improving and take more risks, but it's worth noting that things will start rapidly getting better if we can only hold onto our current levels of prevention, and let immunity from all sources do the job from there.

Europe

Going to use short-dated graphs to improve readability. If you want the longer view you can get it at [OurWorldInData](#), or previous weekly Covid posts.

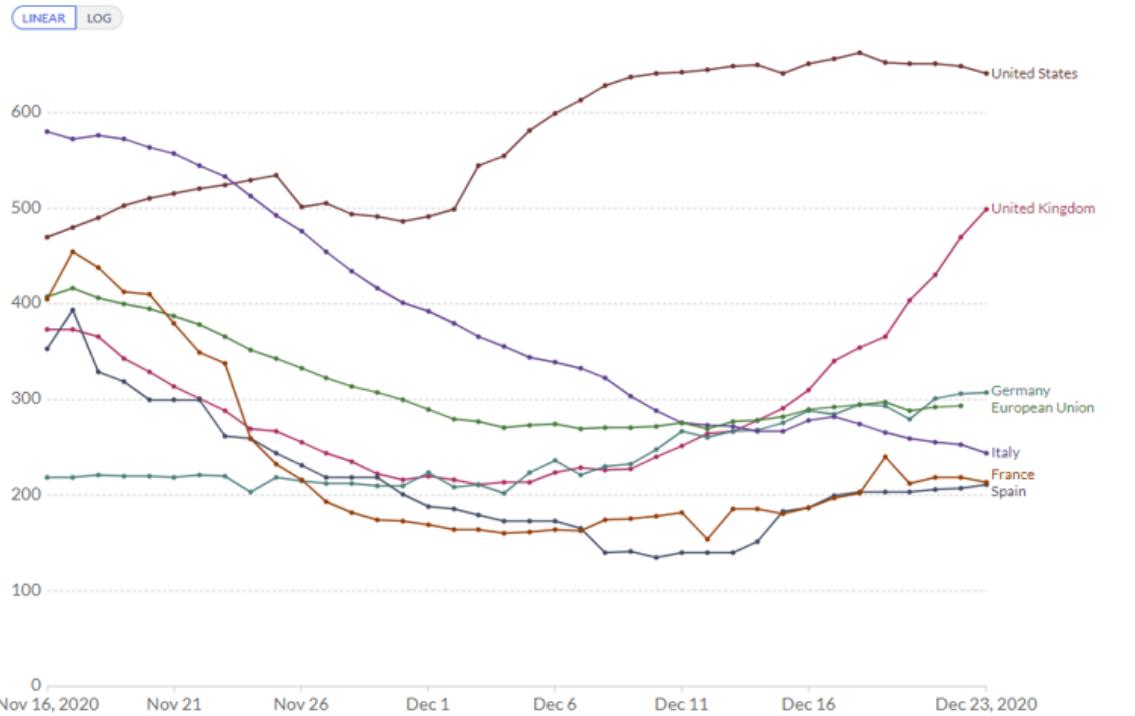


Sh

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World
in Data



The positive test percentages chart is so incomplete and all over the place that I'm going to stop posting it, but you can go to the source if you still want it.

Deaths in Europe continue to run close to those in the United States, suggesting the Europeans are finding cases less often than we are, or have worse medical care or are worse at protecting vulnerable populations.

Then there's that United Kingdom graph going rapidly vertical in infections. Turns out, there's a reason, and it's not that they lifted their restrictions...

The English Strain

[The big news this week](#) is that England has identified a new strain of Covid-19 that is 'up to 70% more infectious'. The new strain dominates in southern England, including London, and the graphs tell a rather clear story.

Figure 3. Total number of SARS-CoV-2 sequences from the UK and proportion of VUI 202012/01 variant sequences among all UK sequences in the GISAID EpiCoV database (as of 20 December 2020) by week of sampling, 2020



Source: GISAID EpiCoV database

Source: the ECDC threat assessment brief. No, I also don't know why their constant bars are on a logarithmic scale and their exponential growth is on a linear one.

Oh no.

You're probably wondering the same question I was when I read that, which is that we know that 'up to' means we're not willing to commit to anything at all (did you know I am up to 15 feet tall? It's true!), but *what the hell does '70% more infectious' mean?*

It could mean a lot of different things.

To me, there are two natural hypotheses for what it means.

One sensible definition of this is that 70% more people get infected each day, so it raises R₀ by that percentage. If previously things were stable, 70% more infectious would cause infections to rise 70% each serial interval, which I've been approximating at about five days. So if it was this, things would double each week.

Alternatively, it could mean that *any given physical interaction* was 70% more likely to infect you. This seems unlikely to be it, because how would anyone know what this value was, but it still makes at least some intuitive sense and has some practical value. So if before, if you went home for Christmas and someone was infected, you'd have a 10% chance of getting Covid-19, now that number is 17%.

The difference between those two is that if you get exposed multiple times, you can only get infected once, so the first scenario is a bigger jump in cases than the second one. Depending on how much 'overkill' you think takes place when people get infected, the difference could be big or it could be small.

As a third option, it could mean *any given exposure is effectively as if you were exposed to 70% more virus*. Chance of catching the virus is non-linear with viral load, so in some ways this is a more than 70% increased risk (if previously load was below threshold to get infected, and now it isn't), and in other ways it could be less once you get to the other end of

the curve. This also changes the distribution of initial viral loads, in ways that might be good or bad for outcomes and death rates. If you are reliably getting higher initial loads, that's bad. If you are getting infected *despite low loads* then, given that we know you're infected, that's good, perhaps quite good.

What I *definitely didn't* consider was the possibility that this was measuring *the length of the doubling time* because that's not remotely a fixed number and using this doesn't make any sense and arrrrrgh and then [I saw this on this post](#):

[...] CellBioGuy 1d 0 < 63 >

I've become enough of an arrogant SOB evolutionary biologist poking his nose where it doesn't belong over the last year that I believe I have answers to all of these, and a few other important things to say. WHILE MAKING IT CLEAR I AM NOT A VIROLOGIST OR IMMUNOLOGIST OR EPIDEMIOLOGIST AND THAT I COULD EAT MY WORDS IN THE FUTURE, though I have been pretty good so far.

| How likely is it that the spread of this new strain was caused by a few superspreaders, and that most of the above is blown out of proportion

That would make sense if the frequency was rising in early introductions or when spread was thin on the ground. Increasing fraction of the total infections while spread is already thick makes me think that there is more likely to actually be an actual effect on contagiousness. They report a decent detectable increase in genome copy number as measured by PCR and sequencing in upper respiratory samples which makes me think this is likely. More on the virology later.

| What, uh, does the "71% higher growth rate" mean

TLDR: I think that it's probably barely 15% more infectious and the math of spread near equilibrium amplifies things.

I admit that I have not read all available documents in detail, but I presume that what they said means something like "if ancestor has a doubling time of X, then variant is estimated as having a doubling time of $X/(1+0.71) = 0.58X$ "

This can only be related to a parameter like R₀ in relation to how much R₀ has already been reduced by behavior changes to its effective actual R since the math gets really nonlinear and I presume I am either missing their math or there will be more detailed documentation in the future. In the mean time let's run through an example.

I have seen estimates of 'generation times' for SARS-2 clustering around 5 days in the presence of all the stacked up behavior changes. Doubling time is related to the generation time and the effective replication number R_e, by the equation:

$$\text{doubling time} = \ln(2) * \text{generation time} / (\text{R}_e - 1)$$

If you have reduced the R_e to 1.1 with a 5 day generation time then you get a doubling time of 35 days. If you have reduced it to 1.2 then you get a doubling time of 17 days. In the former case, if you take 35 days to $0.58^{*}35 = 20.1$ days, you get R_e going from 1.1 to 1.172, **only 7% higher infectiousness**. In the latter case, if you take 17 days to 10 days, you get R_e going from 1.2 to 1.34, a **12% increase**. Presumably the base R₀ increases by a similar factor.

In short, the closer you are to equilibrium when measuring the effect the larger the effect a small change in contagiousness will have. Given that compared to unmitigated situations we are pretty close to R_e=1 and I never see cases doubling recently in Britain in less than two weeks, *I doubt that the total factor of increase of infectiousness is actually all that large*. You get a **huge** nonlinear effect as you move the numbers around near one, even when the total factor of increase is not huge, but mitigation increasing by only a small factor can counteract it.

Prepared to eat this, hard, if I am misinterpreting what has been said.

What is still driving me crazy is that CellBioGuy *presumed* that this was what they meant. I mean, I certainly *hope* it is what they meant in terms of ‘that physical property of the world would kill less people’ but I can’t help but notice it would be completely insane in a way that even my model of predicted general insanity isn’t handling well. The model isn’t even handling the *presumption* by CellBioGuy very well here.

So it turns out that CellBioGuy was wrong here, and [this refers to the sensible thing of “percent rise in infections each cycle” R0 thing](#):

Spike gene target failure (SGTF) can serve as a proxy for carriage of the VOC (cf. Section Impact on diagnostic assay below). Classifications of SGTF are preliminary as case definitions are still being developed. On this basis, we adjusted rates of SGTF for variable specificity over time and between local authorities and then applied the same models to estimate the association of VOC frequency and reproduction number. This analysis shows an increase of Rt of 0.52 [95%CI: 0.39-0.70] when we use a fixed effect model for each area. We also fitted a similar model but with a random effect model on the area, giving an estimated additive effect of 0.60 [95%CI:0.48 - 0.73]. Similar estimates of 0.56 [95%CI:0.37-0.75] were obtained using a Bayesian regression model accounting for errors in VUI frequencies and Rt estimates. As an example, under the

fixed effect model, an area with an Rt of 0.8 without the new variant would have an Rt of 1.32 [95%CI:1.19-1.50] if only the VOC was present.

Among 40 local authorities in East and South East England with more than five VOC samples there is a significant trend of increasing reported cases with increasing frequency of N501Y (Figure 1, weighted linear regression p=10-6). A 10% difference in VOC frequency in mid-November corresponds to approximately 50 more weekly cases per 100 thousand in early December. Local authorities with few VOCsamples have similar reported cases as the rest of the UK (linear regression intercept = 137 cases per 100k versus UK median 130.4 per 100k).

(Do check out the rest of [that comment by CellBioGuy](#) anyway – even though the presumption in question turned out to be wrong, the rest of the comment has a lot of good gears-level details on various issues, but it’s too long to put here in full. I’d like to know how well the rest holds up.)

So on the plus side, statistics are being reported in a way that is relative sane.

On the minus side, this seems rather like it can be summed up as: *We’re fucked, it’s over.*

This is estimated as a 65% increase in infectiousness. If we want to stabilize infections in an area that was previously stable we’d need what would previously have been an R0 of about 0.6. If you have an R0 of 0.6 that means you would have previously been cutting infections *in half each week or so*.

Does that sound like something any Western country could *possibly* accomplish from here? What would even trying to do that even look like? Is there *any* chance people would stand for what was necessary to do that?

And that’s only what it takes to get a holding pattern.

Under such dire circumstances, a phase 4 lockdown has been invoked. What does that mean? Glad you asked, [here are the guidelines](#).

The missing restrictions that stick out are not shutting down houses of worship, and allowing people to move house. Also funerals can go up to 30 people, whereas weddings are capped at 6. Also 'support groups' can meet *up to 15 people* and I don't see anything saying it needs to be outdoors.

Whereas essentially any social contact of any kind is forbidden (e.g.: "You cannot meet people in a private garden, unless you live with them or have formed a support bubble with them"), your 'bubble' is highly restricted in its sizing, and you're not allowed to be outside without a 'reasonable excuse' although those include groceries, going to the bank or exercising.

So basically, if you're outside where it's safe, they'll harass you and maybe worse. Whereas if you stay inside, technically it's not allowed but in practice it's a lot less likely anything happens to you, unless the anything in question is 'you catch Covid-19.' The rules are porous enough that they aren't enforceable against the things that are risky but enforceable enough to shut down the relatively safe actions that keep people sane. And with weird exceptions for remarkably large indoor gatherings for certain events that are textbook superspreaders.

All of which is what our model expects to see, and none of which seems likely to be *remotely* sufficient if the new strain is as infectious as they estimate.

The strain has already been seen in several other countries. Flights between the United States and the United Kingdom have not been shut down. Many European countries are shutting down *some* travel, which will slow things down a bit, but headlines like this one...

3 hr 7 min ago

Covid-19 testing of truck drivers continues at British port of Dover

From CNN's Duarte Mendonça and Sharon Braithwaite

...illustrate that slowing things down is all that's being aimed at. Which is good, because it's too late anyway. There would not be any drivers to test if this was a real attempt at containment.

If the estimate of 65% more infectious is correct: *The strain doubles every week under conditions where other strains are stable.*

My father sent me [this video \(24 min\)](#) that makes the case for all of this being mostly a nothingburger. Or, to be more precise, he says he has only *low* confidence instead of *moderate* confidence that the new strain is substantially more infectious, which therefore means don't be concerned. Which is odd, since even low confidence in something this impactful should be a big deal! It points to the whole 'nothing's real until it is proven or at least until it is the default outcome' philosophy that many people effectively use.

Note that he also suggests the new strain is likely to be less virulent, and make us less sick, which could also be part of why it's more infectious. If so, that's great news (I can think of a scenario where it is actually bad news, but it's an unlikely corner case).

He also points out correctly that a lot of nations don't do much sequencing, so we should assume the new variant can't be contained to England at this point. Doesn't mean we shouldn't try in order to slow it down, but such efforts will still fail.

The video seems strong on the scientific details, and the speaker strikes me as an excellent explainer/teacher, which is why I'm willing to link to it.

Alas, as is often the case with academics that are good at learning and explaining scientific things, the epistemology is bonkers. His core argument is: "You cannot use epidemiological data to prove a biological property." With a side of Covid-19 being spread mostly by super-spreaders (true) and thus the new variant could be winning at random.

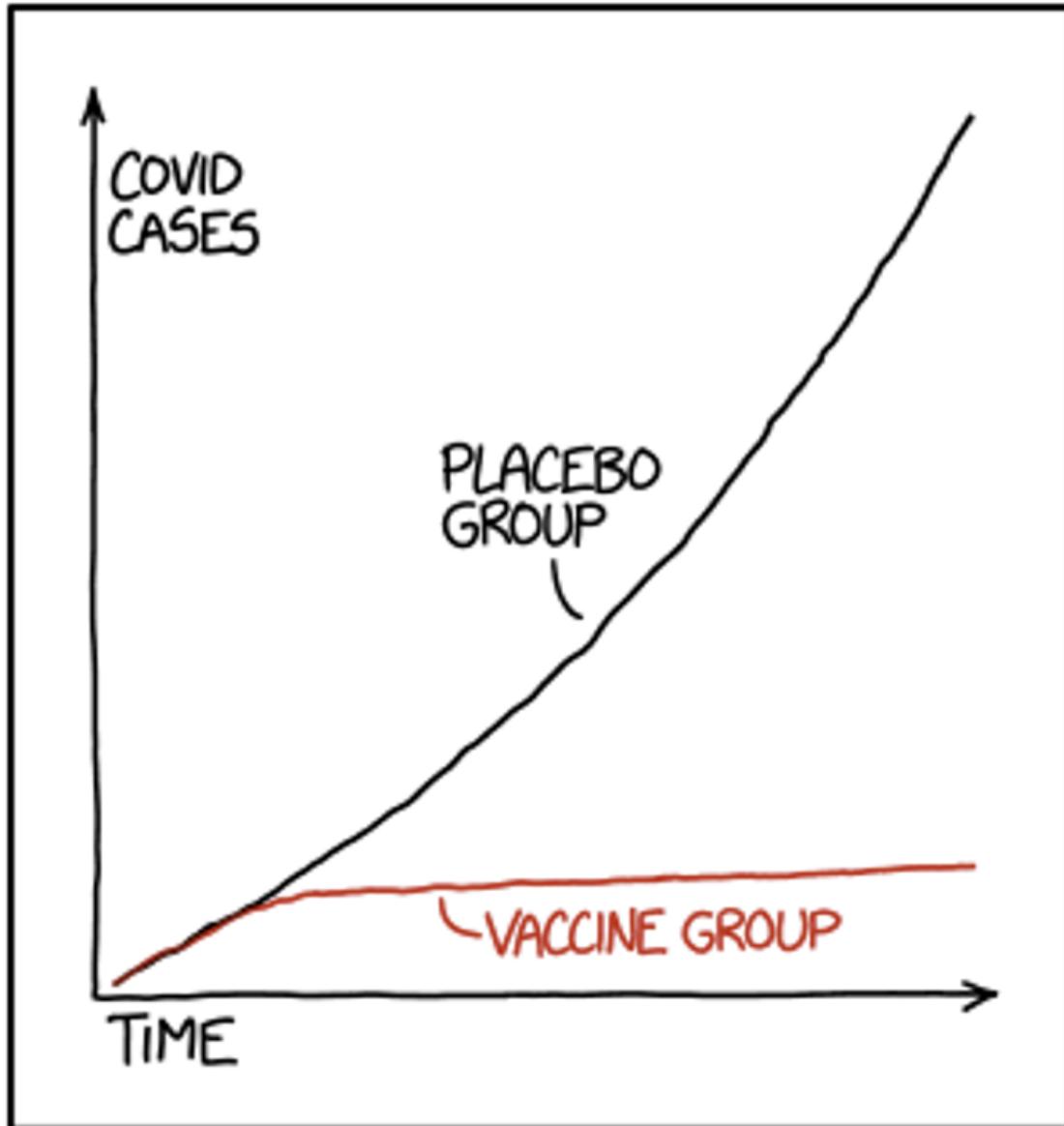
All of which is not how knowledge or Bayes' rule works. It's not how any of this works.

There is a valid point here, of course. Relying solely on the numerical growth of the strain or of infections in England generally, without looking at the context, doesn't provide that much evidence. There are often other explanations. And his points about mutation in and of itself being commonplace and mostly harmless are well taken.

That doesn't change that evidence is evidence, and a likelihood ratio a likelihood ratio. Experiments are not some special class of thing that are the only way one can make predictions or assign probabilities, and it's weird that people can be so good at academic scientific thinking while not understanding this, and in fact it seems that when we train people to do academic science we also train them to not think about other information and to be careful not to use Bayes' rule and to ridicule anyone who tries to use non-academic information in order to know things.

With that in mind, we look at the evidence and think about possible explanations, and mostly I find that there aren't other plausible ones worth assigning much weight to.

Let's start with those charts above. They aren't [quite this](#):



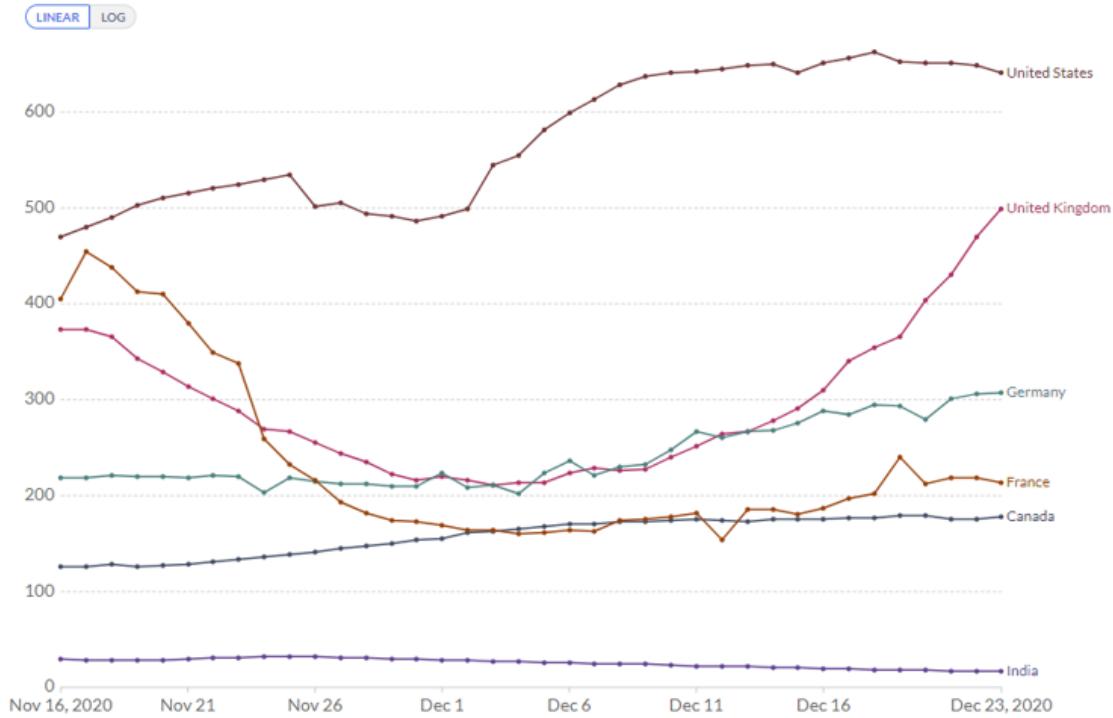
STATISTICS TIP: ALWAYS TRY TO GET
DATA THAT'S GOOD ENOUGH THAT YOU
DON'T NEED TO DO STATISTICS ON IT

But then again, when you consider the context of England being under lockdown conditions that had previously turned the tide, and that have stabilized the situation elsewhere in Europe...

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World
in Data



And combine it with the share of infections from the new strain from the earlier chart, and work out what those combine to imply...

This definitely does qualify under “hot damn, look at this chart.” This is a huge, dramatic increase in infections happening very quickly. A doubling in one week.

Note also that the warnings went out to the public on December 19. The out-of-sample data from the next few days strongly reinforce the hypothesis that we’re screwed.

The other plausible causes of such a rapid rise are not present. England didn’t suddenly relax its conditions this much. The law of large numbers is more than sufficient to make me very dismissive of ‘random chance via superspread events’ as an explanation. How big are these superspread events?

If my understanding of the situation is correct, there is only one conclusion:

This variant cannot be stopped short of mass vaccinations. It is not going to be stopped short of mass vaccinations.

All that is left is a holding action. Realistically, we can’t make enough of a dent to turn the tide of cases of the new strain until at least May. That’s about twenty weeks from now. That’s twenty doublings. So for every case that’s escaped to the United States so far, we can expect (does quick math) *a million cases*. Then another few doublings in June and July.

I do think we can do better than that, because it appears the tide is starting to turn now on the old strain, we’ll get a bunch of incremental help from increased immunity along the way, and control systems will set in.

But mostly, it seems like if you have vaccines and people who don’t want to die, [you might want to hurry](#). I’ll get back to gaming the scenario out in the conclusion section.

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

(This prediction is repeated at the end of this post.)

All I Want For Christmas is a Covid Vaccine

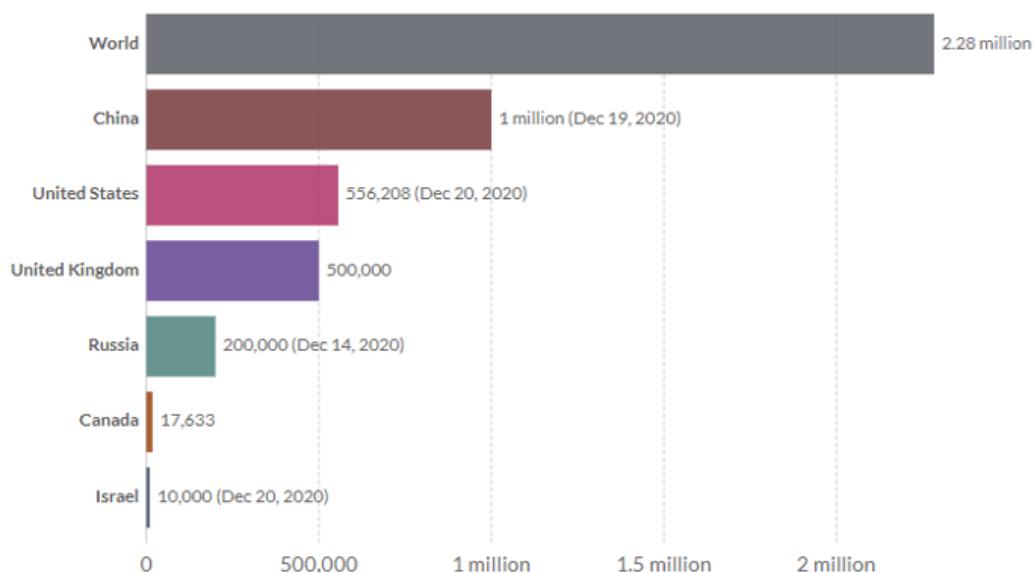
Track it [here](#).

Number of COVID-19 vaccination doses administered

Show is the total number of COVID-19 vaccinations administered. This is counted as a single vaccination dose, and does not measure the number of people vaccinated against the disease (which usually requires two doses).



[+ Add country](#)



Source: Official data collated by Our World in Data

CC BY

CHART

MAP

TABLE

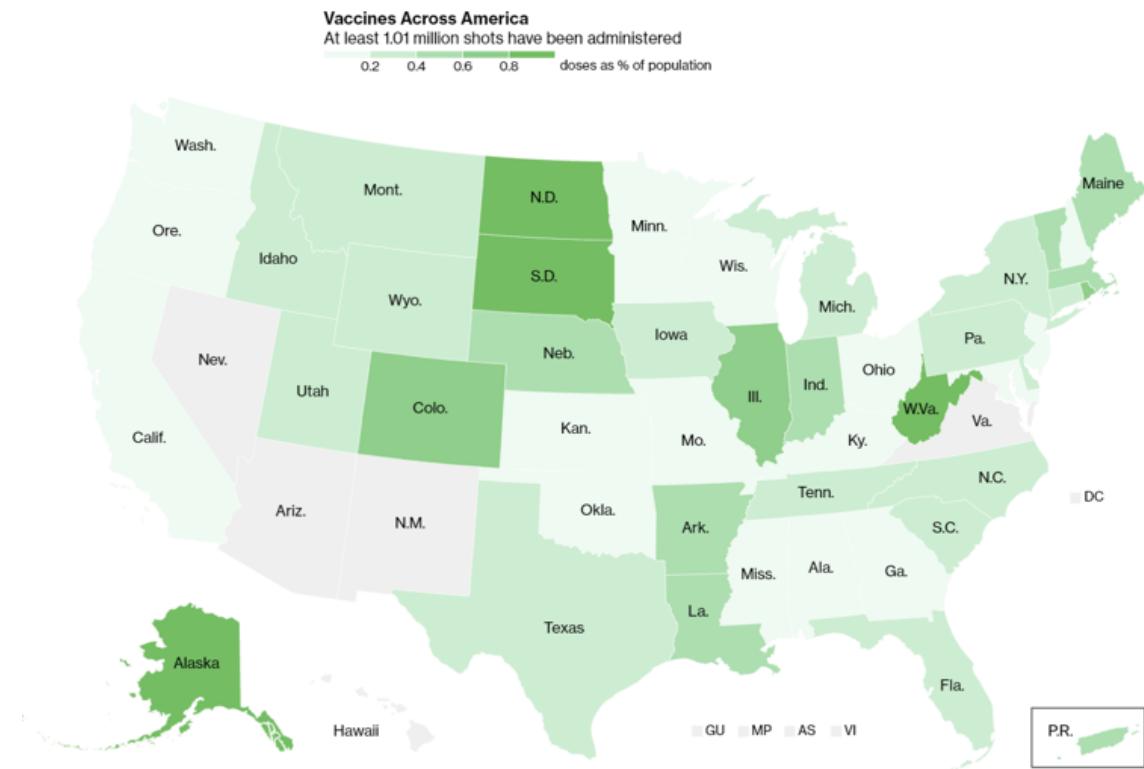
SOURCES

DOWNLOAD



Or at Bloomberg [here](#).

Vaccinations in the U.S. began Dec. 14 with health-care workers, and so far **1.01 million doses** have been administered, according to a state-by-state tally by Bloomberg. Those numbers are accelerating as a second vaccine by Moderna Inc. is distributed.



So, then, about that vaccine effort. [How goes the distribution of the vaccine? Well \(Twitter video link\), funny story...](#)

WTF? So angry.

"We have millions more doses sitting in our warehouses but have not been sent shipping instructions."

New York, NY, December 17, 2020 – Pfizer Inc. (NYSE: PFE) today released the following statement to address public comments that allege there are issues in the production and distribution of the company's COVID-19 vaccine:

"Pfizer is not having any production issues with our COVID-19 vaccine, and no shipments containing the vaccine are on hold or delayed. This week, we successfully shipped all 2.9 million doses that we were asked to ship by the U.S. Government to the locations specified by them. We have millions more doses sitting in our warehouse but, as of now, we have not received any shipment instructions for additional doses.

We have continuously shared with Operation Warp Speed (OWS) and the U.S. Department of Health and Human Services through weekly meetings every aspect of our production and distribution capabilities. They have visited our facilities, walked the production lines and been updated on our production planning as information has become available.

Pfizer has a successful and long track record of producing and distributing large volumes of complex vaccines that the world can trust – and we are continuing to extend this track record with our COVID-19 vaccine. Over the last several months, we have activated Pfizer's extensive manufacturing network, including thousands of highly skilled workers in multiple locations. As a result, Pfizer is manufacturing and readying for release millions of doses each day, and that volume will grow over the coming weeks.

We remain confident in our ability to deliver up to 50 million doses globally this year and up to 1.3 billion next year, and we look forward to continuing to work with the US Government to deliver our vaccine to the American people."

3:13 PM · Dec 17, 2020 · Twitter for iPad

From CNN:

One early hurdle: A two-day US Food and Drug Administration requirement to assess each shipment of vaccine for quality control slowed down distribution.

It seems that Pfizer executives are sitting around baffled that they have *millions* of additional doses of vaccine, and those doses are sitting on shelves unused, with continuous risk of spoilage, while they help no one. We have confirmation that this is explicitly *not* holding back second doses, and is coming as a complete surprise to Pfizer.

As a reminder, using a few million additional doses will cut infections by about 1% *compounding each week*, if distributed at random. If used selectively, an extra few million

can cause a double digit drop in the death rate a month out, due to how concentrated deaths are among the elderly.

Distinctly from the issue of vaccine going to waste, the estimates given to states were too high, for which ([Twitter video link](#)) the head of Operation Warp Speed has taken responsibility. Such admissions are so virtuous and unexpected that they are clear signs of competence and trustworthiness. I agree strongly with Alex Tabarrok here, and my esteem for those involved went way up rather than down.

Oh, and also:

"Every dose that these hospitals and clinics have is going into people's arms," Nebraska doctor says

From CNN's Naomi Thomas

There are a couple of reasons for the difference between vaccine doses distributed and vaccine doses administered, Dr. Ali Khan, dean of the University of Nebraska Medical Center's College of Public Health, told CNN's John Berman Thursday.

The first is "just data gaps," he said. "We don't have good timely systems that are actually reporting how many doses are being administered as opposed to how many are delivered, so some of it is just the data gaps."

The second, he said, are true delays. "It takes a while to unpack the vaccine, inventory the vaccine, thaw it and put it in people's arms and it takes a while for clinics to get their logistics down," he said.

However, from his personal experience in Nebraska and what he is hearing across the US, Khan said "there is no place in the US where this vaccines sitting in a clinic or hospital and it's not being used. So every dose that these hospitals and clinics have is going into people's arms."

Just over a week since the first Covid-19 vaccine was authorized, more than 1 million people have received their first shot.

The government has said it intends to distribute 20 million first doses of the Pfizer-BioNTech and Moderna vaccines in the coming weeks, **slightly later than it had originally planned**.

Versus:



Luci Keller
@itsLuciKeller

My mom's hospital has the Moderna vaccine, started administering this morning. At 11 AM, all of the "vaccine providers" went home for xmas and won't be back till Monday

They won't let anyone else give the shots

The vaccines are just going to sit there for 5 days

Absurd

6:04pm · 23 Dec 2020 · Twitter for iPhone

So there's that. While regular folk are told to cancel the holidays. Merry Christmas, suckers.

Still, CNN again:

"We do still feel strongly that we will have allocated to the states by the end of the year 20 million doses of vaccine that'll be available," Gen. Gustave Perna, Operation Warp Speed's chief operating officer said during the call Monday. "We feel confident that we will be distributing the end part of that vaccine no later than the first week in January for everybody to have access to."

If true, that would be two weeks behind schedule, but assuming all of this is on a different track than manufacturing, so long as the process doesn't keep falling further behind, I guess it's not actually so bad? That's a big 'so long as.' But it's a plausible one, as initial difficulties don't have to translate into difficulties once things are running and it's not Christmas.

Then there's the issue of that second dose. The science behind why you need two doses is strong and makes sense, except for the little matter of *the data*. It's clear that the first dose alone is much more than half as effective as two, and we don't have enough doses. And the first thing I asked when I saw that result was [something that is once again being pointed out](#). Booster shots do not have to be given two weeks after the first shot. The measles booster comes a year later. Multiple sources confirm that there is no reason to expect a six-months-later second dose to be any less effective a booster.

We are mostly wasting an entire half of our vaccine supply by dosing exactly the worst people with it - the people who got the first dose. I can see double dosing those in nursing homes anyway, because they're 1% of the population and over a third of deaths, so even a small additional boost is still worthwhile, and their immune systems are weak so it's reasonable to worry the first dose alone won't get the job done there. But beyond that, there's no excuse I can see beyond saying the rules are the rules.

Here's where we are, it seems:



Matthew Yglesias 🎨 ✅ @matty... 1h

Call me crazy, but I'd have just ordered enough vaccines for all 330 million people from each company.

You never know what will or won't work out, and if we end up with extra we can give them away to poor countries. The amount of money involved is tiny in the scheme of things.

Meg Tirrell ✅ @megtirrell

It's official: Pfizer-BioNTech reach deal to supply 100M doses of #covid19 vaccine to the US for \$1.95B, delivery by July 31. US has now secured:
200M doses Moderna
200M Pfizer
100M J&J*
300M AstraZeneca*
100M Novavax*
100M Sanofi/GSK*
(*US ph3 pending)
businesswire.com/news/home/2020...

💬 1️⃣ 960 ...

Doses that arrive on July 31 are not doses I expect to prevent many infections. And this total does not leave much margin for error, we barely have enough doses if we don't use AstraZeneca and insist on everyone getting two shots.

How goes the vaccine quest elsewhere?

In Germany, [not so well](#).

A real case can be made that, given our inability to do suppression properly, a Western government's policy ends up mostly coming down to their effect on the vaccination schedule. Did they advance vaccine research?

On those fronts, we come out well ahead of the European Union. Does anything else matter by comparison? And how did we manage to do that?

Casey Mulligan reminds us that one of the things Trump has been doing for a while is [getting the FDA to kill less people](#) by speeding up and streamlining its processes. So when the time came for Operation Warp Speed, the concept was shovel ready. The "experts" all said eighteen months minimum and the "experts" got ignored. We certainly could have pushed much harder much faster, even given initial conditions. I'm not going to stop pointing that out. The decision to balk at buying extra doses might be the worst single decision of the last four years, of any kind. But when I model alternative administrations, from either tribe, I model a much slower vaccine effort. When we reach the end, it's not clear that won't matter more than every other decision combined.

In England, it seems they are going to be testing [a vaccine grown using tobacco plants](#)? Approval to begin trials is in. This seems like a great illustration of our regulatory state, because you can't take the life-saving vaccine they grow with the tobacco plants, but you can take the tobacco in your pipe and smoke it even though we know it kills you. Okay, then.

Theoretical Vaccine Auction

If you want to properly allocate a scarce resource like a vaccine, *obviously* you use an auction. I said it last week, people righteously said that things are not worth to the customer what the customer will pay for them because poor people have less money than rich people, and no, sorry, that's not how this works, that's not how any of this works.

The good news for those who disagree is that there is not the slightest danger of any attempt at an efficient allocation of the vaccine.

Externalities, on the other hand, are definitely real. To the extent that they are different for different people, we should take that into account. In theory, an auction can do that because others can subsidize the bids of those whose vaccination is beneficial to them. In a first best situation lots of people would subsidize lots of others near them and the government at various levels would also supplement bids, but presumably people mostly won't get such acts together and it won't happen except highly locally, inside families and corporations.

My guess is that for all the talk of externalities, they're not that different for different people other than those who were previously being irresponsible, and subsidizing them for that seems like it has some large moral hazard issues plus the worst offenders probably already had it. The exception would be those that provide a lot of value if they can expose themselves to the virus that they couldn't otherwise, but are not being paid for that service enough to place the bids themselves. Health care workers in hospitals likely qualify, but you can fix that if you subsidize the hospital to bid for them, and generally that class of solution seems like it should work.

Thus, the obvious solution of "N-lot, pay N+1th-highest-bid auction that repeats every so often" seems like it's just correct on first principles, and then once you run the first one likely you can slowly reduce price to keep supply and demand balanced without having to go through all the trouble of more auctions?

So that's not that interesting a paper and also it's about four lines long, so it was odd when I followed a Marginal Revolution link to this paper about [how to auction off the vaccine](#) only to find a much longer paper.

Which goes on to suggest this:

The Vaccine Auction

Each bidder i submits a collection $\mathbf{b}_i \equiv (b_{ij})_{j \in \mathcal{I}}$ of bids: a bid b_{ii} on his own behalf and a bid b_{ij} on behalf of each bidder $j \in \mathcal{I} \setminus \{i\}$. The aggregate bid on bidder i 's behalf is denoted by $B_i \equiv \sum_{j \in \mathcal{I}} b_{ji}$.

The auction's allocation rule \mathbf{x}^* is efficient; that is, it associates with each bid profile $\mathbf{b} \equiv (\mathbf{b}_i)_{i \in \mathcal{I}}$ a feasible allocation $\mathbf{x}^*(\mathbf{b})$ that maximizes the total surplus:

$$\mathbf{x}^*(\mathbf{b}) \in \arg \max_{\mathbf{x} \in X} \sum_{t \in T} \sum_{i \in \mathcal{I}} x_{it} \alpha_t B_i. \quad (1)$$

Because the sequence $(\alpha_t)_{t \in \mathcal{I}}$ is weakly decreasing by assumption, one solves (1) by assigning each time- t unit of the vaccine to the bidder with the t -th highest aggregate bid, with indifferences resolved arbitrarily. That is, for any bidder i , we have $x_{it}^*(\mathbf{b}) = 1$ if and only if B_i is the t -th largest component of $\mathbf{B} \equiv (B_i)_{i \in \mathcal{I}}$.

Each bidder's payment in the auction is the externality that he imposes on other bidders assuming that all bidders bid truthfully, that is, assuming that $b_{ij} = s_{ij}$ for each i and j in \mathcal{I} . Formally, given the submitted bids $\mathbf{b} \equiv (\mathbf{b}_i, \mathbf{b}_{-i})$ with $\mathbf{b}_{-i} \equiv (\mathbf{b}_j)_{j \in \mathcal{I} \setminus \{i\}}$ and the induced aggregate bids \mathbf{B} , bidder i 's payment is

$$p_i^*(\mathbf{b}) = \sum_{t \in T} \sum_{j \in \mathcal{I}} x_{jt}^*(0, \mathbf{b}_{-i}) \alpha_t (B_j - b_{ij}) - \sum_{t \in T} \sum_{j \in \mathcal{I}} x_{jt}^*(\mathbf{b}) \alpha_t (B_j - b_{ij}). \quad (2)$$

If all bidders bid truthfully, then the first term in (2) is the total surplus all but bidder i enjoy once the vaccines have been allocated efficiently while ignoring bidder i 's bids, whereas the second term is the total surplus all but bidder i enjoy once the vaccines have been allocated without ignoring bidder i 's bids. By construction, each payment (2) is nonnegative, and, therefore, the auction cannot possibly lose money.

Which is exactly why no one wants to participate in auctions or read economics papers. I believe that the paper's approach boils down to "everyone writes down how much they value each person getting the vaccine at each possible point in time, drawing that curve for everyone" and then the auctioneer solves for the best possible solution that maximizes surplus, and everyone pays the marginal cost imposed on others behind them, as measured by how much they don't want to wait one more slot to get their vaccine.

Which, I mean sure, that's technically correct of course, the best kind of correct, but it's also making my brain hurt reading it and that's despite intuitively knowing the answer the

moment the question was asked. No one is going to want to think about let alone write down full utility curves, the practical cost would be enormous.

And getting people in the wrong spot in line by a little bit is at worst a small mistake, and also not that meaningful given the logistics of vaccine administration. The vaccine is not exactly being teleported around.

Thus, I'm pretty sure *in practice* this reduces to 'find the price that mostly clears the market and charge that price, then adjust it to keep clearing the market.' The whole thing where others can bid on your behalf is how 'people can buy things with dollars' works already.

In response to the people saying 'but poor people are poor and rich people are rich, you bastards' they say the following:

2. A superficially compelling alternative to the vaccine auction is, first, to allocate the vaccines however the government sees fit (or even to allocate randomly) and then to allow individuals to trade their vaccination priorities among themselves. The problem with this approach is the ineluctable inefficiency that arises from the [Myerson and Satterthwaite \(1983\)](#) impossibility theorem. Once private ownership is introduced, markets need not put vaccines into the hands of the individuals with the highest valuations when the potential buyers' and the potential sellers' valuations are their private information.

This is, essentially, distribute a valuable asset according to politics, then allow trade. Which is a worse version of full free trade, since it involves political reallocation of wealth and (as noted by Myerson) results in a lot of profitable trades not happening because of various frictions, and also makes people feel bad about not selling their vaccine dose and also about selling their vaccine dose, [because choices are bad](#). It seems obviously better than allocating via politics alone, since most of the trades that happen will be *hugely* profitable to both sides - the person selling will get a payment much, much bigger than they'd have found necessary, and the person buying will pay much, much less than the amount they value the vaccine. The ZOPA (zone of possible agreement) here is mostly *huge*. I have a hard time imagining that choices being bad overcomes that.

What I personally would love is a form of abstract that is "the paper that we would have written if we were only trying to get the central point across and didn't care about formal anything at all." Not quite [You Have About Five Words](#) (I keep thinking the 'have' is 'get' and having to correct myself, maybe the title should change) but mostly not *that* much more than that. A one-page Econ 101 cheat sheet can cover actual everything the class teaches, and most economics papers have at most one paragraph of actual information.

Here, it sure *seems* like the whole paper boils down to "Standard auction theory applies here."

It usually does.

This, for example, is a great use case. Not Covid-19 related or in any way essential, [but pretty great](#).

[Another modest proposal](#) is to fix the prices and then let people pay them, skipping over the auction as such, so we know we can raise \$50 billion dollars for vaccine production and distribution (and motivation, shall we say) by letting a few people skip the line. That makes it easy to see it's an absurdly good deal, but also either wouldn't get enough takers or is leaving a lot of money on the table, in exchange for being easy to think about.

Also, that post points out something that seems important. What could be a better way to motivate people wanting the vaccine, than to show our richest, most famous and most powerful people paying *really big bucks* to get the shot a few months sooner?

The Chosen Alternative To That Auction

Meanwhile in vaccine prioritization via politics and power might occasionally have some issues news ([Twitter video link to protesters protesting during meeting, and thread](#)):



Dan Diamond  @ddiamond · Dec 18

...

Frustration at Stanford is overflowing, with officials sending multiple emails to staff this week to apologize for skipping residents and fellows in vaccine line.



rat king  @Mikelsaac · Dec 18

There is an enormous demonstration going on at @Stanford Hospital right now carried out by staff, who are protesting the decision by higher ups to give vaccines to some administrators and physicians who are at home and not in contact with patients INSTEAD of frontline workers.

[Show this thread](#)



38

630

1.6K





Dan Diamond ✅

@ddiamond

...

Replies to @ddiamond

Facing angry doctors, Stanford official tries to explain why vaccine went to others instead.

The algorithm “clearly didn’t work,” he says, as doctors boo + accuse him of lying.

“Algorithms suck!” shouts one protester. “Fuck the algorithm,” says another. (video via tipster)



Dan Diamond ✅ @ddiamond · Dec 18

...

On the left, an email from Stanford’s chief medical officer this week, apologizing for “unintended missteps.”

On the right, an email from officials this morning — confirming that young doctors got skipped for vaccines, despite being on frontlines. (via a tipster)

**Message from Niraj Sehgal, MD,
Chief Medical Officer**

Dear GME community—

First, I’ll start by sincerely apologizing for what a difficult day this has been. I wish we had been able to focus on the excitement of tomorrow and the start of our vaccination efforts. Despite our best intentions to thoughtfully map out a principled vaccine plan to include our residents, fellows, and faculty, it’s clear there were several unintended missteps. Please know the perceived lack of priority for residents and fellows was not the intent at all. We’ve spent today understanding the issues so we can quickly move forward together. It won’t make up for how you felt today. I personally couldn’t feel worse about it. As I shared with you during the town hall earlier this week, my training at Stanford was a transformative period in my life, and nobody is more committed than me to assure you have the same experience.

Given our anticipated vaccine supply next week-ish (Moderna approval as you perhaps heard), we’re

doses out to not impact the workforce due to prioritize the group as a hole. We realize this first allocation failed to provide the correct order of protection. This is being carefully evaluated and redone.

PLEASE rest assured, there are active meetings to correct this, and we are all working VERY hard to prioritize in a logical way the trainee workforce.

We are expecting more doses next week (about 15,000) and more after that, so we will be able to get to the majority of the 35000 folks in the next few weeks. There are active conversations now to make sure trainees are represented in all the upcoming tiers in a significant way.

After the conversations tonight and tomorrow morning we hope to be able to send updates, but please know we are advocating.

Please send us any questions,

Hang in there,
Hayley and Becky



64



390



861





rat king @Mikelsaac · Dec 18

Replying to @Mikelsaac

yep, i heard same as caroline.

ooo

Per an internal memo sent to me among Stanford execs: The list created by the algorithm was supposed to be vetted before being carried out but administrators failed to do so, in part due to crossed wires and fast turnaround



Caroline Chen @CarolineYLChen · Dec 18

NEW: Stanford used an algorithm to decide who got the #COVID #vaccine first, and the result was only 7 of the first 5,000 slots went to residents... the same week they were asked to volunteer for ICU coverage in anticipation of a surge in COVID-19 cases.

If I am interpreting this correctly, those with power used politics to get vaccinated first, ahead of nurses and residents. Residents, of course, are indentured servants giving four years of work under terrible conditions with hugely below market pay in exchange for the right to join a monopolistic guild that enforces scarcity of health care provision via government restrictions on supply of residences and slots in medical school.

This is exactly how political distributions of goods always go, except that this time it was a *really* bad look to give administrators in offices the vaccine before doctors who were actually treating Covid-19 patients. That's what "The list created by the algorithm was supposed to be vetted" means here, that they were supposed to do a check to make sure their naked appropriation of vaccine had some plausible deniability attached at some level, whereas it turned out it didn't.

But of course, as the thread points out, blaming 'the algorithm' or saying 'there are problems with the algorithm' is the obvious nonsense. Yes, there was a 'problem with the algorithm' and the 'problem' was 'the people with power over the contents of the algorithm gave themselves top priority.' This wasn't some complex calculation. It didn't involve machine learning. The result wasn't *surprising* to anyone who came up with the criteria. The criteria was giving resources to the powerful rather than the powerless, via deeming them more 'important' or 'vital' or something.

This time, thankfully, they got caught. Yet they still got their vaccine shots.

Thus, this apology is the height of hypocrisy:



December 18, 2020

Dear Colleagues,

We are writing to acknowledge the significant concerns expressed by our community regarding the development and execution of our vaccine distribution plan. We take complete responsibility and profusely apologize to all of you. We fully recognize we should have acted more swiftly to address the errors that resulted in an outcome we did not anticipate. We are truly sorry.

As you know, we formed a committee to ensure the vaccine's equitable distribution. Though our intent was to ensure the development of an ethical process, we recognize that the plan had significant gaps. We also missed the opportunity to keep you more informed throughout this process.

We are working quickly to address the flaws in our plan and develop a revised version. As we make corrections to our plan, we will provide continuous communication in an effort to engage our entire community in this process. We are optimistic that a large shipment of vaccines will arrive next week, which will allow us to vaccinate a substantial segment of our community.

We recognize the disappointment and distress this has caused, and we appreciate those who brought these concerns to us. We deeply value each and every member of our community and the outsized contributions you make to our mission every day - especially during this particularly challenging year.

Did not anticipate, you say. Equitable distribution, you say. In case you ever wonder what equitable distribution means when those with power say it, now you know.

You gotta love the line “Though our *intent* was to ensure the development of an ethical process, we recognize that the plan has significant gaps.”

There's also [this](#):



Yascha Mounk @Yascha_Mounk · Dec 20

Some good news:

ooo

Thanks to massive and justified public criticism, the CDC is making adjustments to their recommendations.

Americans over 75 should now get the vaccine alongside essential *frontline* workers.

This is an improvement. But it doesn't solve many of the concerns.

Proposed Phase 1 & 2 allocation, December 2020

Phase	Groups recommended for vaccination	Number of persons in each group (millions)	Number of unique* persons in each group (millions)	Total* (millions)
1a	Health care personnel Long-term care facility residents	21 3	21 3	24
1b	Frontline essential workers Persons aged 75 years and older	30 21	30 19	49
1c	Persons aged 65/4 years Persons aged 16-64 years with highrisk conditions Essential workers not recommended in Phase 1b	32 110 57	28 81 20	129
2	All people aged 16 years and older not in Phase 1, who are recommended for vaccination			

*Accounts for persons recommended in prior phases or overlap within a phase



60



217



1.1K



Yascha Mounk @Yascha_Mounk · Dec 20

ooo

In particular, the CDC's own data *still* suggests that Americans aged 65-74 are much more likely to die from Covid than younger frontline workers.

So this course of action will likely *still* cause needless additional deaths.

How many? This is where things get really worrying.



8



14



175



Yascha Mounk

@Yascha_Mounk

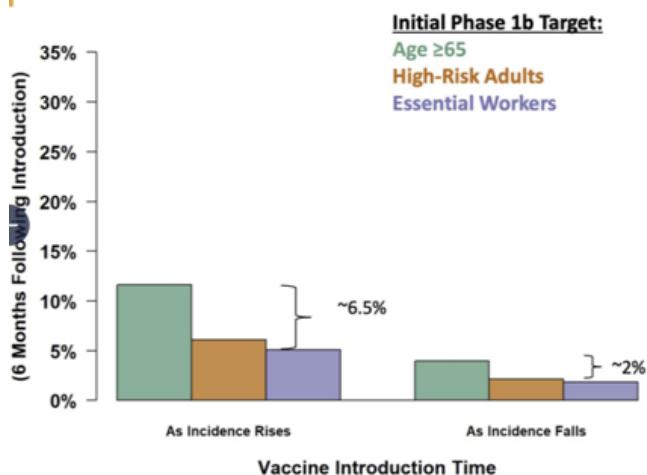
ooo

Replying to [@Yascha_Mounk](#)

In the original presentation, Kathleen Dooling admitted that prioritizing all essential workers would likely increase overall deaths by between 0.5% and 6.5%.

In an astonishing sentence, she then called the additional deaths of thousands of Americans a "minimal" difference.

Population-Wide Averted Deaths: Disease-Blocking Vaccine



- Initially vaccinating age ≥65 in Phase 1b averts approximately 2–6.5% more deaths, compared to targeting high-risk adults or essential workers
 - As before, this difference is greatest in the scenario where the vaccine is introduced before incidence peaks
- Findings robust to assumptions of reduced VE in older populations but percentage averted drops

Also, I'm not going to link to any of the sources for this (to avoid heat vs. light and [toxoplasma of rage](#) issues, among other reasons), or offer a hot take on it, but there are a lot of "experts" and "ethicists" who have stated outright that [we shouldn't prioritize older people over younger people because older people are disproportionately white](#), so giving vaccine priority to the people most likely to die from Covid-19 is racist.

It is unclear to me to what extent this is driving policy decisions, but it seems like it came close to happening, with the CDC only reversing after a public outcry. The mayor of New York has endorsed this perspective explicitly.

So. Yeah.

But congratulations to Texas, among others (via CNBC):

States break from CDC on vaccine prioritization

States are beginning to split with the Centers for Disease Control and Prevention's guidance on who should receive the Covid-19 vaccine first.

The CDC said Sunday that everyone over 74 years old as well as front-line essential workers like agricultural workers, police and teachers should get the Covid vaccine in the so-called phase 1b; that comes after health-care workers and long-term care residents are vaccinated first in phase 1a.

Texas was among the first states to split from the CDC guidelines. The state announced Monday that it is prioritizing those 65 years and older as well as those with certain medical conditions in its phase 1b vaccination plan, making front-line essential workers wait a bit longer.

More states will likely follow suit in the coming weeks, said Jen Kates, senior vice president and director of global health and HIV policy at the Kaiser Family Foundation.

—Will Feuer

Face Saving

[Kerry gets it](#), a thread I will post here in full:

[←](#) **Thread**



Kerry @kerry62189 · 17h

This seems pretty close to saying that the restrictions are designed to signal the state's disapproval of people socializing with friends and family.

...

"We're doing everything we possibly can here," Baker said, shortly after noting he doubted his administration could create rules that would force people to change their holiday plans.

He made clear that his administration sought to send a message to residents to take the threat of the virus more seriously over the holidays and repeatedly noted lessons learned because of case and hospitalization spikes following Thanksgiving.

2

2

6

↑



Kerry @kerry62189 · 17h

I don't think this is justifiable. I don't think there's been a connection demonstrated between these restrictions and the threat to hospital capacity. But all that aside, we're canceling medical procedures and hassling business owners in order to "send a message" to *others.*

...

1

2

4

↑



Kerry @kerry62189 · 17h

And we don't even know if it the messaging works, or if there's even much danger from people seeing friends and family. As far as I can tell, people are pretty careful in MA, and wear masks reliably.

...

1

2

2

↑



Kerry @kerry62189 · 16h

It's not about raw numbers. People are affected in very different ways, and inflicting severe suffering on some groups to (maybe!) minimize overall cases is cruel. Omission and commission are not morally equal, esp. by government & under uncertainty.

...

1

2

3

↑



Kerry @kerry62189 · 16h

Just realized that the logic of a face-saving culture perfectly explains the bizarre behavior of the leadership class in their response to the pandemic. They are united in having adopted this culture among themselves.

...

1

1

7

↑

The government anticipates that people will not Act Responsibly on Christmas, and thus is pre-emptively punishing them in order to send a message.

What's interesting is her proposal that we can understand such acts by the Very Serious People as being primarily about *face*.

Their behavior is defined "by what other people see," not the virus. They are cooperating to preserve their settled hierarchy, while keeping the illusion of real effort going just enough to disguise this.

Face is defined essentially by what other people see. Thus, face is like honor in that the sentiments of other people are extremely important. Like honor, face also can involve a claim to virtue or to prestige. However, the settings—and consequently, the role expectations—are quite different for cultures of honor and cultures of face. Whereas honor is contested in a competitive environment of rough equals, face exists in settled hierarchies that are essentially cooperative.



Replies to @kerry62189

If everyone toes the line, they can all be winners. They cannot admit to being wrong, and if they all prop each other up with validation, none are forced to do so. That's why they've closed ranks so uniformly, and why they don't seem to be all that interested in sensible policy.

Ho (1976, p. 883; see also Heine, 2001) defined face as “the respectability and/or deference which a person can claim . . . by virtue of [his or her] relative position” in a hierarchy and the proper fulfillment of his or her role. Thus, everyone in the hierarchy can have some face, though some may have more than others due to their position. Implicitly, people have face—unless they lose it. A person can “gain” face, and one person can “give face” to another, but the major focus is primarily on not losing face

1

2

4

1



Kerry @kerry62189 · 16h

This is the whole problem. A stable hierarchy. An entrenched elite that doesn't have to be competent or even non-ridiculous. There's no incentive for anyone to do a better job. There's no glory in outdoing a peer. Results do not matter.

Because face exists within a stable hierarchy, it is not competitive or zero sum. In an honor culture, one person may take another's honor and appropriate it as his or her own; however, one cannot increase one's face by taking another's. In a face culture,

1

2

3

1



Kerry @kerry62189 · 16h

“The 3 H’s of a face culture are thus hierarchy, humility, and harmony.” You might ask where the humility is. Well:

Cohen, & Au, 2010). People are supposed to show appropriate deference to hierarchy. They are supposed to display humility and not overreach on status claims (lest they learn a painful and humiliating lesson about how much status others are willing to accord them). And they are to pursue, or at least not disturb, the harmony of the system.

1

2

3

1



Kerry @kerry62189 · 16h

In other words, the humility only comes into play when we're talking about the status of the leadership class. This type of cultural logic does not recognize the claims of *the public* relative to the leadership class, and therefore is totally inappropriate in this situation.

1

2

4

1

This reaches many of the same conclusions and makes many of the same predictions as many of the models and gears I've been using, but seems importantly different in ways that

require more attention. These dynamics have been mostly left out of my models, and that seems like it might have been a mistake.

I am going to think more about whether it makes sense to incorporate such dynamics more centrally.

Quest for the Test

The good news is a \$5 paper test strip has been approved by the FDA for home use. Of course, [doing so legally will require an additional \\$25](#) so a digital MD service can *watch you* do it, which also means all sorts of coordination problems and activation energy and having to be observed by medical people silently judging you. Because of the need to [beware trivial inconveniences](#) it seems like the additional costs will be hugely destructive of the potential value here, although massive value should still remain.

Perhaps the next one will do better? As usual there's claims it is coming Real Soon Now:



Franco Calderón

@veryfranco

...

Replies to @michaelmina_lab

Michael, as you know mine will be <\$5 and no prescription. Easy does it. Just need to clear the final mile at FDA

Resilient & Rápido
Direct Read SARS-CoV-2
Antigen Paper-Strip LFT

For Research Use Only in The United States

Features

- Point of Care/Home Use
- No instrument required
- Results in 2-15 minutes
- Inexpensive
- Scalable
- High sensitivity
- High specificity
- CE

Workflow

- 1-Obtain specimen
- 2- Insert swab into reagent and mix.
- 3- Remove swab from reagent and discard swab.
- 4- Insert paper strip into reagent, remove it from reagent and read results directly.

Contact: franco@resilientsupplier.com
Resilient Supplier. All Rights Reserved 2020.

2:11 PM · Dec 16, 2020 · Twitter for Android

Alas I expect that last mile to find a way to at least inflict a lot of damage, even if cleared.

Whereas [this plan from Canada](#) seems awesome, taking the methods of continuous rapid testing that worked in some universities and applying them more generally to businesses that value being open.

And that's [this plan from Wisconsin](#):



Tony Evers

Office of the Governor | State of Wisconsin

FOR IMMEDIATE RELEASE: December 22, 2020

Contact: GovPress@wisconsin.gov

Gov. Evers, DHS Announce Wisconsin to Offer At-Home COVID-19 Testing Option

MADISON — Gov. Tony Evers, together with the Wisconsin Department of Health Services (DHS), today announced a new contract with Vault Medical Services that will add an additional tool to the Wisconsin COVID-19 testing toolbox starting today. At-home COVID-19 saliva collection kits will be available to everyone who lives in Wisconsin, with or without symptoms, at no cost.

"We believe that anyone in Wisconsin who needs to be tested for COVID-19 should have access to a test, and I'm proud of our statewide testing efforts throughout this pandemic," said Gov. Evers. "We also know that getting to a health care provider or a community testing site isn't easy for everyone, and that's why we are excited to offer this new option to make testing even more accessible for folks across our state."

Wisconsinites can [order a collection kit online](#) and have it shipped to their home. The kit will include detailed instructions on how to collect the saliva, which includes a video call with a testing supervisor through Vault Medical Services, and ship it back via UPS dropbox to the lab for processing.

"This is an important tool to provide easy access to COVID-19 testing," said DHS Secretary-designee Andrea Palm. "As we roll out the COVID-19 vaccine to more Wisconsinites, we need to continue testing, contact tracing, and public health measures such as wearing a mask and social distancing."

A saliva test is similar to a nasal swab test you might receive from a provider or at a community testing site. Like a nasal swab test, a saliva test determines whether you have an active COVID-19 infection and can spread it to others.

More information can be found on the Wisconsin COVID-19 [testing](#) webpage.

For up-to-date information about Wisconsin's COVID-19 response, visit the DHS [COVID-19 webpage](#). We encourage you to follow [@DHSWI](#) on [Facebook](#), [Twitter](#), or [dhs.wi](#) on [Instagram](#) for more information on COVID-19.

###

Office of the Governor ♦ 115 East Capitol, Madison, WI 53702

Press Office: (608) 219-7443 ♦ Email: GovPress@wisconsin.gov

<https://evers.wi.gov> ♦ [Unsubscribe](#)

Free testing on demand, sample collection at home, even if it isn't rapid, and even if they require a zoom call during the sample collection. And you need an email for each of your

children, even if they're two weeks old. Still. That's not bad!

A sign they do not understand the proper goal of testing is that they don't grok that people might want to do this periodically without any particular reasons to worry:

What do I do while I am waiting for my test results? 

While you wait for your COVID-19 test results, [self-isolate and monitor for symptoms](#).

Another sign is, look for the mysteriously missing question on this FAQ list:

Are there other languages available?	
Who can order a collection kit?	
When I go to order a collection kit, I receive a message saying that there are no more kits available. What should I do?	
Do I need to provide my insurance information?	
Do I need to provide my credit card information?	
How do I return my collection kit so my saliva can be tested at a lab?	
Is this a molecular test or an antigen test?	
Is this a nasal swab or a saliva collection?	
Can I eat, drink, or chew before my collection?	
Do I really need an email address for each of my children and family?	
How does my saliva collection work over a Zoom video call?	
What do I do with my sample after I complete the saliva collection over the Zoom video call? How long will it take to get my results?	
Do I need to put my saliva sample in the fridge?	
What do I do while I am waiting for my test results?	
How will I receive my test results?	
How is my privacy protected?	
Is Vault collecting my DNA from this test?	
I have other questions about the collection kit. Where can I get more information?	

Oh yeah, *that* question. *How long until I get my test results?*

Good question!

In Other News

[This analysis of how Covid-19 spreads seems excellent as far as I can tell](#), making a strong case that it's mostly or entirely aerosol transmission, and that this fits the observed data fine, thanks. People I respect led me to it and also had the same take. If there's anything wrong here, please speak up.

CDC issues the following guidelines and it only took until (checks calendar) December 21:

- The virus that causes COVID-19 most commonly spreads between people who are in close contact with one another (within about 6 feet, or 2 arm lengths).
- It spreads through respiratory droplets or small particles, such as those in aerosols, produced when an infected person coughs, sneezes, sings, talks, or breathes.
 - These particles can be inhaled into the nose, mouth, airways, and lungs and cause infection. This is thought to be the main way the virus spreads.

Any further questions? Never fear, [they have an FAQ](#).

Also, yes, [air ducts can almost certainly spread the virus](#), and six feet is not a magic number.

If you're looking for treatments, [this Quora response](#) seems to be a real attempt at providing guidelines one could use. As usual I'm going to tread lightly on the treatment side and not take a stand.

Not a metaphor: Yes, you literally have to report as a 'severe adverse event' requiring investigation when someone in your vaccine trial *is struck by lightning*:

As of December 6, 2020, there were 3 SAEs reported in the vaccine group: a 65-year-old participant with community acquired pneumonia 25 days after vaccination, a 72-year-old participant with arrhythmia after being struck by lightning 28 days after vaccination, and an 87-year-old participant with worsening of chronic bradycardia 45 days after vaccination. On FDA review of the narratives, none of these SAEs are assessed as related. There were no cases of severe COVID-19 reported in the study.

<https://www.fda.gov/media/144434/download>

Official in Buffalo gives press conference announcing 60 people died, [reporters only ask questions about the division champion Buffalo Bills](#). Official then tells them to 'get their priorities straight' but their priorities are already quite straight. The reporters know there's a lot of deaths, and a lot of Covid-19, but what is the actionable information about that, that the official could tell them, that they don't already know? What impacts people's lives? The team could be in an unknown amount of trouble, and that matters to people in Buffalo. Whereas barring a change in policy that would have already been announced, they know exactly how screwed they are personally, and what they have to do. The true objection here is that when people are dying it's wrong to not waste one's time showing one's concern about that and instead care about "trivial" things like the thing most or at least many people in Buffalo have cared most about the last few months. Or years.

Is using the bathroom essential? Second worst person Mayor De Blasio's administration [decided that it was not](#):

If my SLA-licensed establishment is offering outdoor dining, may I allow customers to use the bathroom inside?

No. Customers may not enter the inside of the establishment for any reason.

When a predictable uproar followed, he laid the blame on actual worst person Governor Andrew Cuomo, and requested that they gracefully allow New Yorkers to use indoor plumbing, a request that it seems has been granted:



Avery Cohen

@CohenAvery

...

After discussions with the State, they have agreed to change the rule around bathroom access.



Bill Neidhardt @BNeidhardt · Dec 18

The new rules are set by the State, not the City.

We've asked them to change the rules around bathroom use to keep outdoor dining viable and help our restaurants survive. twitter.com/NYCMayorCourse...

7:52 AM · Dec 18, 2020 · Twitter for iPhone

Also, among other rules, employees cannot drink alcohol under any circumstances, which seems rather cruel except for the part where they never enforce it. Right now I bet a lot of them could use a drink.

This seems to be what you get when you shut everything down for months and months on end:



Cristina "Vee" Valenzuela ✅

@CristinaVee

...

"Los Angeles is now the epicenter of the Covid-19 pandemic. It is the most infected county of the most infected state in the most infected country in the world."

Please stay safe everyone...!!

1:20 AM · Dec 18, 2020 · Twitter for iPhone

[Why It Took So Long for the Army to Make Masks](#). The army masks are cloth masks with the right color scheme and symbols. That's it. They are now getting around to being able to make them. The money quote for me is this:

According to the Army:

The CCFC was designed, developed, and produced along an expedited timeline. It normally takes 18-24 months for DLA (Defense Logistics Agency) to have the item available for order once the technical description, design, and components are approved and submitted. The CCFC, from inception to issuance, is slated to take less than one year.

The army is bragging that they figured out how to make masks this fast.

If [there is a war](#), perhaps someone should tell the army. Seems like information they would want to know.

No news in [the latest comparison of the Pfizer and Moderna vaccines](#). Bottom line is, they're the same, except Moderna's logistics are better.

Would I accept the Chinese vaccine? I mean, yes, I still think it's better than nothing, but wow do they continue to be bad at science in a way that seems almost intentional:

Brazil again withholds full trial results of vaccine by China's Sinovac

Brazil has once again [postponed releasing the full trial results](#) of the Covid-19 vaccine developed by China's [Sinovac Biotech](#), reported Reuters.

Brazil, the first country to complete late-stage trials of the vaccine, said Sinovac had asked them to delay releasing the results for up to 15 days from Wednesday as it compiles data from trials globally, the news agency reported.

The release of the full trial results has now been delayed three times, according to Reuters.

But Brazilian officials said the vaccine, called CoronaVac, is more than 50% effective — therefore meeting the regulatory requirement to be approved for emergency use in the country, the report said.

— *Yen Nee Lee*

I know you think you're helping but maybe just send paid time off?:

4 hr 49 min ago

Delta gifts employee who battled Covid-19 for months a first-class trip anywhere in the world

From CNN's Travis Caldwell

Not Covid, rather about hearing aids, once more with feeling: [FDA Delenda Est.](#)

Not Covid: [MIRI gives its yearly research update](#). Disappointing to learn that research they couldn't talk about didn't pan out, and also disappointing that they still can't talk about what it was, especially given hopes that the ideas might still have merit. Good to know they are willing to pivot to other different things and aren't going to either keep going down a known blind alley or joining the gradient descent crowd. Also great to hear they are considering fleeing Berkeley, which I strongly endorse doing almost no matter the destination. Would love to do some coordination on that, but I'd definitely settle for the New Hampshire scenario, which I think would greatly enhance their ability to think.

Also not Covid, Other LessWrong news: The 2018 review is ongoing, [but not many reviews are being written](#). I'd encourage those who qualify to help change that. I'm especially curious to hear people's takes on my nominated posts.

Universal, the cult of the expert as explained by the Washington Post freelancer system:

Talent Network

STEP 1 CONTACT STEP 2 BIO STEP 3 EXPERTISE STEP 4 PORTFOLIO STEP 5 SOCIA

✓ ✓ ✓

Areas of Expertise

What are your areas of expertise? (at least 1 required) *

[America's finest news source weighs in on what snow days are for. And a reminder via Marginal Revolution, from the last pandemic, of what school is for and what happens when you cancel it.](#)

What Happens Now?

There are many unknowns that have a dramatic effect on the path forward, and how the endgame plays out.

A few big ones dominate.

I do not see it as plausible that the new strain is confined to England. It has already been seen in several other nations, and the numbers in England are not compatible with it not being essentially everywhere already. It's still right to cut off flights and travel now to slow things down, but the barn door is already wide open.

The biggest question is, instead, **is the new English strain (or another similar strain like the one found in South Africa) really 65% more infectious as measured by infections per infected individual?**

If that number is greatly exaggerated, then the strain will take a lot longer to get to where it matters, and even when it does the control systems and vaccinations can keep things mostly in check, and we are still on something not too different from the previous track. My guess is that aside from England itself, we could mostly deal with about a 33% more infectious strain given our timetable before it does that much damage.

If the 65% number is accurate, however, we are talking about the strain doubling each week. A dramatic fourth wave is on its way. Right now it is the final week of December. We have to assume the strain is already here. Each infection now is about a million by mid-May, six million by end of May, full herd immunity overshoot and game over by mid-July, minus whatever progress we make in reducing spread between now and then, including through acquired immunity. Which will help somewhat, but likely only buy a few weeks at most.

One worry I have is that the control system could actively make things worse between now and then, and accelerate the timeline. By the end of January, we should see up to a one third drop in the death rate purely from protecting nursing home residents, and then it will drop further as we protect other elderly people. If as I suspect the control system mostly *acts on the death rate*, they will use this as a reason to loosen and take more risk, and infection numbers will rise, or not fall the way we would have otherwise expected.

Then once the new strain arrives, it will be like March and April 2020, where by the time the deaths start spiking it is very late in the game, and there have already been three or more additional doublings.

The good news is that when this happens, most of our most vulnerable will have had the opportunity for vaccination. Assuming most of them take it, the (need for) hospitalization rate should be dramatically lower, and the IFR should also be dramatically lower if the hospitals don't collapse. The hope is that with enough of the vulnerable protected, even a gigantic surge in cases might not collapse the system.

That's the hope. It is basically the best (realistic) case scenario. It seems to be too late to speed up vaccine production much, although I see some reports Joe Biden is considering using the Defense Production Act to do so, in which case *why the hell aren't we doing whatever it takes to get that happening already?* Have we considered throwing *relatively tiny amounts of money* at the problem? Maybe not quite *that* relatively tiny? No? Oh well. But I digress.

On the good news front, our tests still mostly seem to work fine for the new strain, although there is some small worry.

There is the concern that the vaccine might not be effective against the new strain, but based on my looking into this, the prior on there being much effect here should not be so high, except insofar as the new strain is more infectious so everything will be somewhat less effective at preventing spread. But if the virus has escaped from the vaccine already while also becoming more infectious, the timeline does not allow us to adjust even if we acted correctly, let alone acting realistically, the best we can hope for is maybe to protect some of the most vulnerable but I do not expect a Biden administration to allow movement fast enough to do that in this scenario. Instead, we'd see completely overwhelmed medical systems across the country, and that would be that.

On the potentially very good news front is the other big question. **Is the new strain less virulent (dangerous) than the old one, and if so by how much?**

I don't think this is a favorite to be true, but I do think there's a substantial chance (30%?) that it is, and we should investigate. All the news reports say that there's no reason to expect that the mutation is *more* dangerous, which is true, but also not the change one would expect to often happen. The real question is whether it is now *less* dangerous, and for now the answer is we don't know. If the effect size is big we will presumably figure this out by the end of January. Keeping an eye on England's IFR will be important.

There are plausible physical mechanisms that suggest that some of these mutations may have led to less virulence. One piece of evidence that the strain is less virulent is that it is more infectious! Being less virulent causes it to be more infectious, so if the strain is more infectious, there's a decent chance part of that was caused by less virulence.

Note that if we *do* face that scenario in the fourth wave, unless virulence has gone down quite a bit, you now face an *even more stark* version of what you did preparing for the third wave.

If we are in this scenario, it is *inevitable* that there will be a several week period, at a minimum, in which it is *super easy* to catch Covid-19, with a super infectious strain everywhere, and at a time when there will be no hospitals to help you, and no vaccine that works.

In that scenario, actually protecting yourself becomes vital if you haven't been infected by the time the climax is approaching. As in essentially not being anywhere near a human outside your bubble, for weeks on end, at least until the hospitals stabilize.

It also becomes realistic again to worry about our supply chains or civil disorder. I expect everything to hold, but that can't be assumed.

Then there's the other nightmare. When this starts to happen, how will the authorities react?

Benevolent authorities would be responding now by making every attempt to speed up our vaccine efforts, and to prioritize the most vulnerable while we can. But if we had such benevolent authorities, they would have already been doing that. When something goes from super overdetermined to super *duper* overdetermined, it's not generally the time to expect people to suddenly come up with the right answer when they didn't before. Still, there's some hope that this would happen, as the 'don't let it get so bad they sell out of pitchforks' mechanics kick in. The other problem, as I've noted, is that there is not that much room to move faster without much more aggressive disregarding of barriers put in to prevent action, which would cause a ton of cognitive dissonance, and is not Biden's style.

The question is what our actual, present authorities will do.

Like England, they might go for a full 'lockdown.' My assumption is that if they do this, people will end up largely ignoring it, and it won't be enforced. There's no will to make real restrictions stick even in blue tribe areas, and red tribe areas would be in open revolt, especially when it comes from a Biden administration. People have had enough, and they'll have been told the story of how things were supposed to be returning to normal.

And even when they do that, they'll *still* provide enough loopholes that it presumably wouldn't work even if people didn't go into open defiance of the rules.

Still, will they try anyway, doing epic additional economic damage and potentially causing open conflict? I worry about this a lot.

Thus, I don't know if one should buy stocks or sell stocks on this, or even if one should buy volatility on this or not. If things crest once more and then burn out, that is not bad for

stocks. It's only bad if we get a dramatic response, which wouldn't help the problem much but also give us a whole additional set of problems. We need to get ready, now, to do what we can to stop this from happening when the moment arrives, if the moment arrives. Sure, *if it worked* one could argue in favor, but the math seems rather clear that it won't do any good. When politicians start trying to make it to next week without taking blame, and attempt to destroy everyone's lives to do it, we need to find a way to stop them.

Or we could figure out how to make our efforts actually work. It's not impossible. Is there still time to get our house in order and deal with this head on, if we want to? Hell yeah. But we can't do it by Sacrificing to the Gods and taking sledgehammers to everything people like doing. We can't even do it via doing that *and* being willing to shut down economically important activity, not for long enough to work. I'd be pleasantly surprised if we could even stabilize growth rates that way, and it's completely unsustainable.

What we *could* do is prepare a *testing* regime *now*. Have that in place a few months from now, for real, and do the kind of testing on everyone that some universities do with their students. If we took all the regulatory gloves off, my guess is we wouldn't even have to subsidize to get there on time. Add in proper focus on Vitamin D, airflow, outdoor activity and so on. We could still make it, and it would be an essentially free action.

We could probably start this on January 20 and still make it.

To be clear: We won't attempt to do this, and it won't happen. But it's important to note that we still *could*, in at least some sense, still win, while also noting that we will lose. You never know. Maybe someone, somewhere, is actually listening.

In summary, I am attempting here to do now what I failed to do in December or January, which is to *actually model what happens next based on exponential growth and realistic reactions by authorities*.

My guess is that the English strain is probably sufficiently infectious to get there before enough vaccinations can happen to let realistic measures contain the pandemic. Right now I'm giving the strain something like a 70% chance to be sufficiently infectious, and if it is, I don't see a way around this outcome.

This has counterintuitive implications, both for public policy and for individuals. As always, one's approach to the pandemic must be to either succeed if one can do so at a cost worth paying, or fail gracefully if one cannot succeed. Thus, one could plausibly *either* make the case for being *even more careful* in response, or to folding one's hand entirely. You can raise, or you can fold, but you can't play passive and call all bets and hope to go to showdown.

It's also important to figure out whether the new strain is less virulent than the old one, and if so to what extent, since that could change the math on sensible courses of action quite a bit. I haven't fully wrapped my head around the implications, and I doubt many others have either. Simply saying it is "good news" does not begin to cover how this changes correct actions.

Let's actually engage with the physical situation and model out what happens, and learn from last time out's mistakes.

UPDATE 7/21/2021:

As you doubtless know at this point, it was not over. Given the visibility of this post, I'm going to note here at the top that the prediction of a potential large wave of infections between March and May did not happen, no matter what ultimately happens with Delta (and the prediction was not made with Delta in mind anyway, only Alpha).

I've talked extensively about the situation across many other posts, but for visibility I'm going to summarize my view of what happened here.

First, the early reports said that Alpha (this post calls it the English Strain, which is now called Alpha) was 65% more infectious than baseline. Instead, it looks like Alpha was about 40% more infectious. I noted that on our timetable, we could deal with about a 33% more infectious strain under the predicted-at-the-time vaccine timetable before major damage was done.

Second, when I was writing this post, I expected vaccine deployment to be substantially slower in its first few months than it ultimately was. I made a clear mistake not making this assumption explicit, but we got a lot more shots into arms that I expected during this period - I was thinking 100 million shots in 100 days (e.g. Biden's stated goal) and we did substantially better than that.

Together, these two incorrect assumptions explain why the wave did not occur, although they alone are insufficient to explain why things continued to improve so rapidly. My guess is that I was also somewhat underestimating seasonality, and that Alpha scared people sufficiently that the control systems acted differently than they otherwise would have, which also should have been considered more.

I gave the headline scenario about a 70% chance of playing out. It didn't play out, and the above reasons are why (in my current model) it failed to play out, and how I put a 70% probability on something that didn't happen.

Looking at what did happen, if we had indeed been facing a 65% more infectious strain at that time rather than 40%, we would have faced a wave of some size even with the other factors being better than I anticipated, but not the size of wave I was predicting; it would have taken a number more like 75% to cause a crisis before vaccinations could kick in. On reflection, given what I knew at the time, if we take the distribution of possible properties of the Alpha strain as a given, I think I should then have been predicting more like 40%-50% chance of the scenario rather than 70%.

Contrasting the situation with Alpha, where I thought it was at 65% more infectious than baseline and that it would cause a large wave before we could sufficiently vaccinate with Delta, where it's perhaps 120% more infectious than baseline but we've already done a lot of vaccinations, and now we are seeing rises in cases, is interesting, and will be an interesting test of whether the other assumptions I was making were accurate.

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

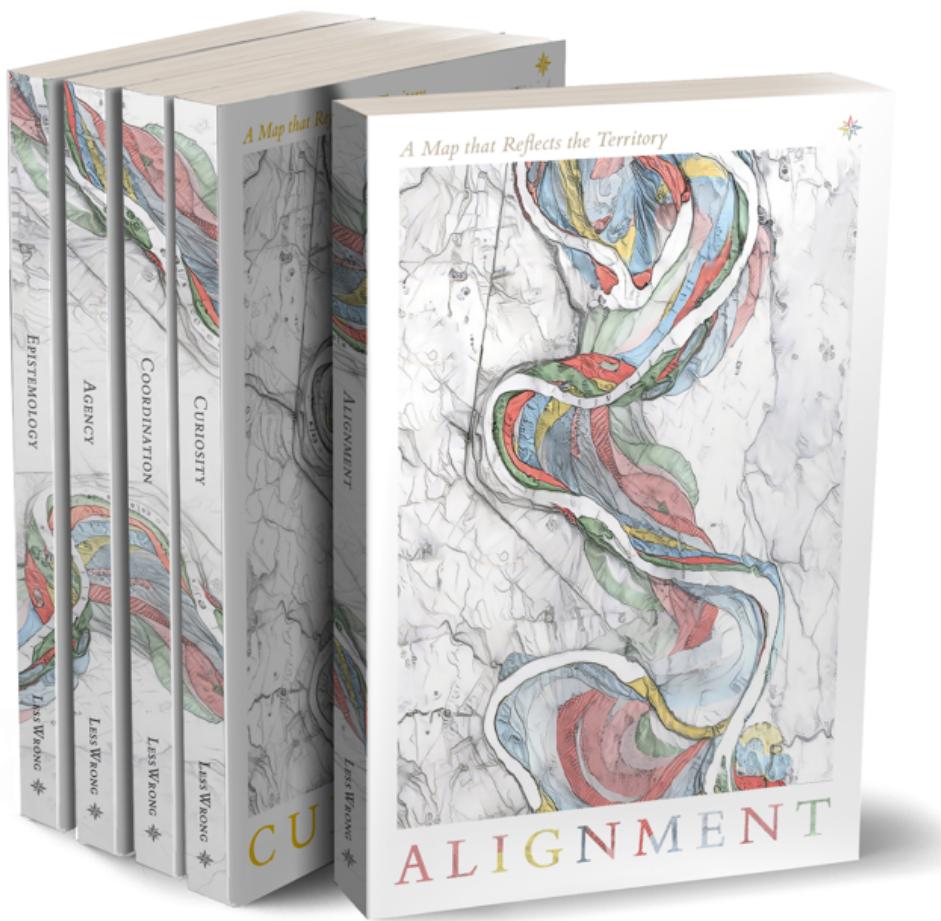
The LessWrong 2018 Book is Available for Pre-order

For the first time, you can now buy the best new ideas on LessWrong in a physical book set, titled:

A Map that Reflects the Territory: Essays by the LessWrong Community

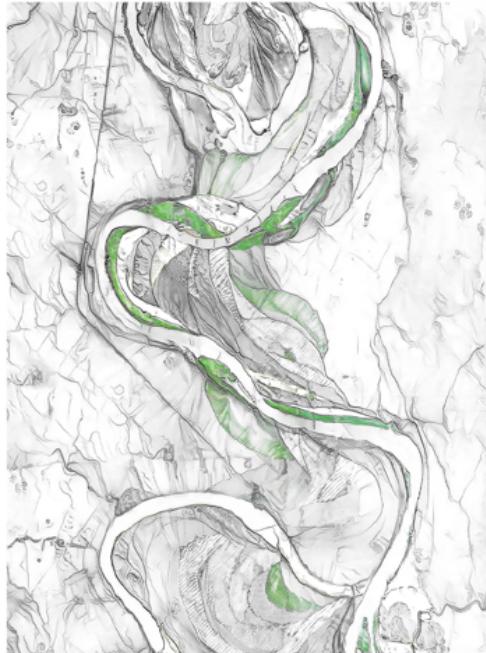
It is available for pre-order [here](#).

The standard advice for creating things is "show, don't tell", so first some images of the books, followed by a short FAQ by me (Ben).



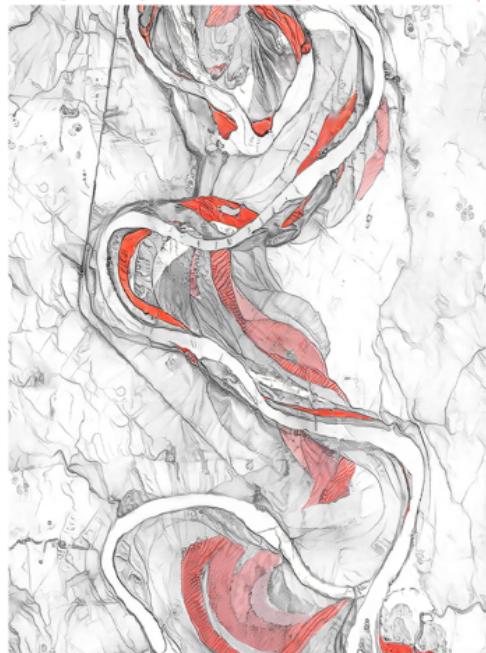
The full five-book set. Yes, that's the iconic Mississippi river flowing across the spines.

A Map that Reflects the Territory



EPISTEMOLOGY

A Map that Reflects the Territory



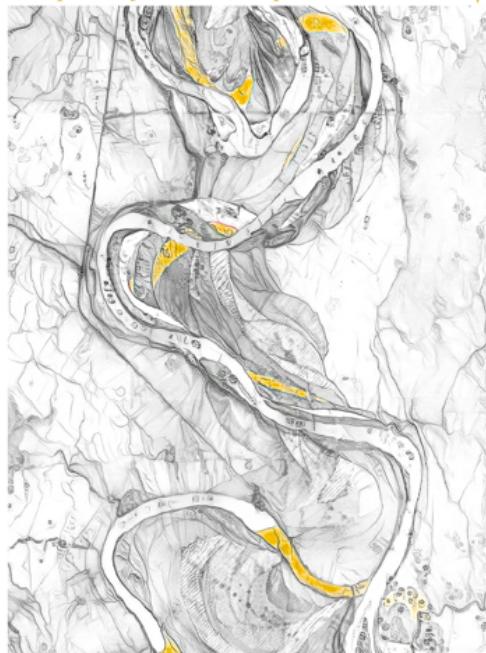
AGENCY

A Map that Reflects the Territory



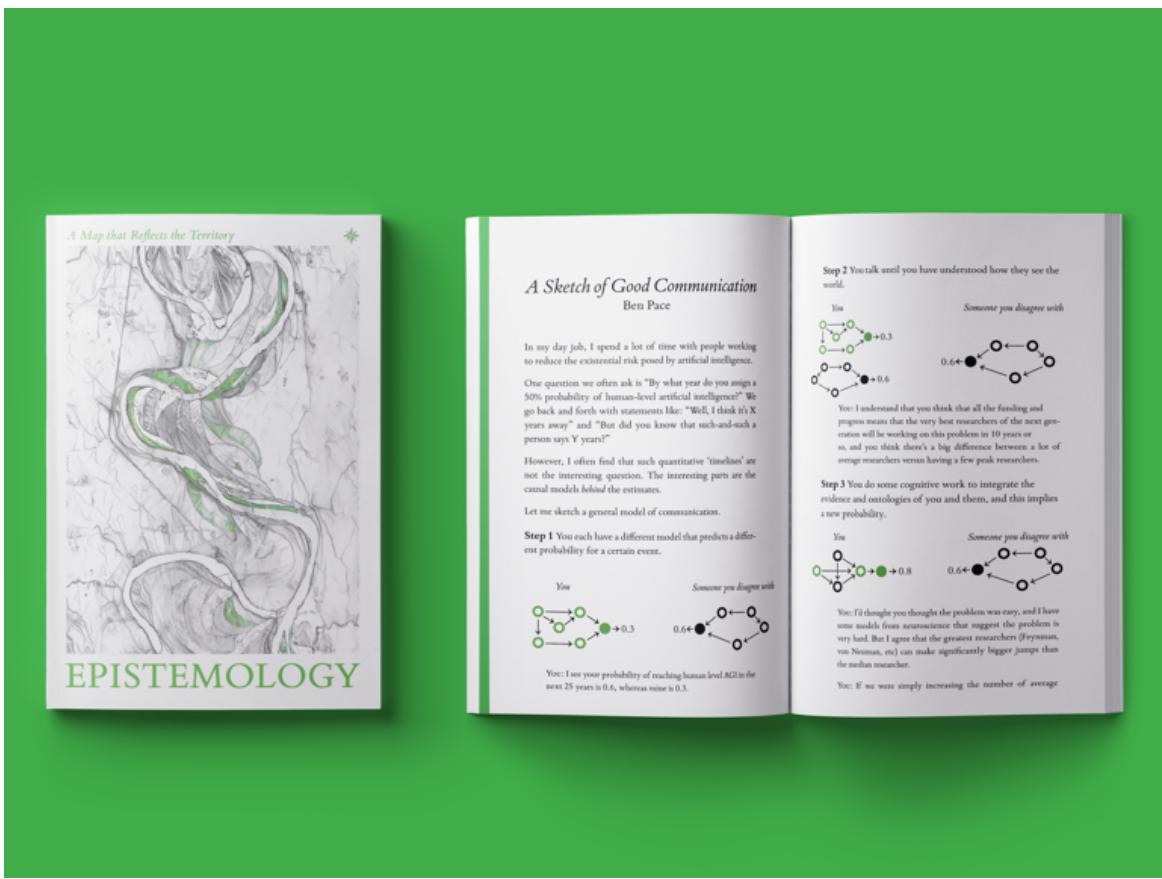
COORDINATION

A Map that Reflects the Territory

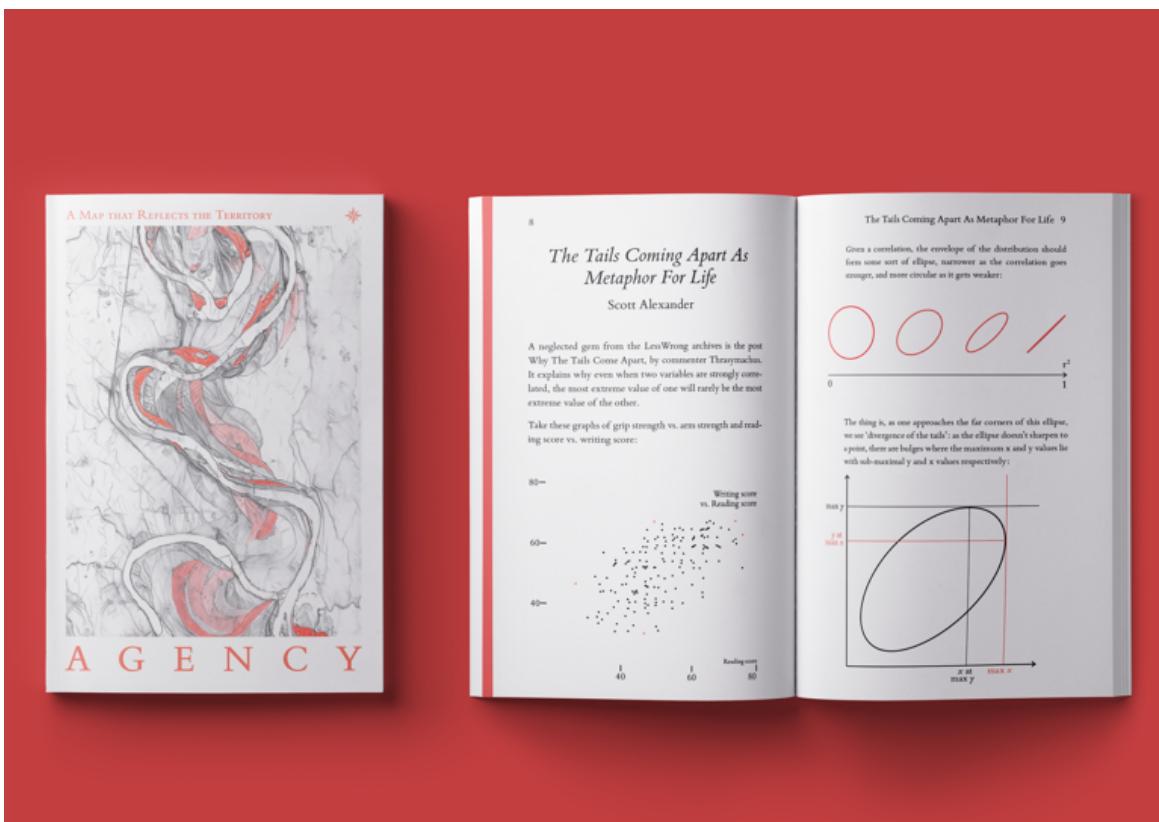


CURIOSITY

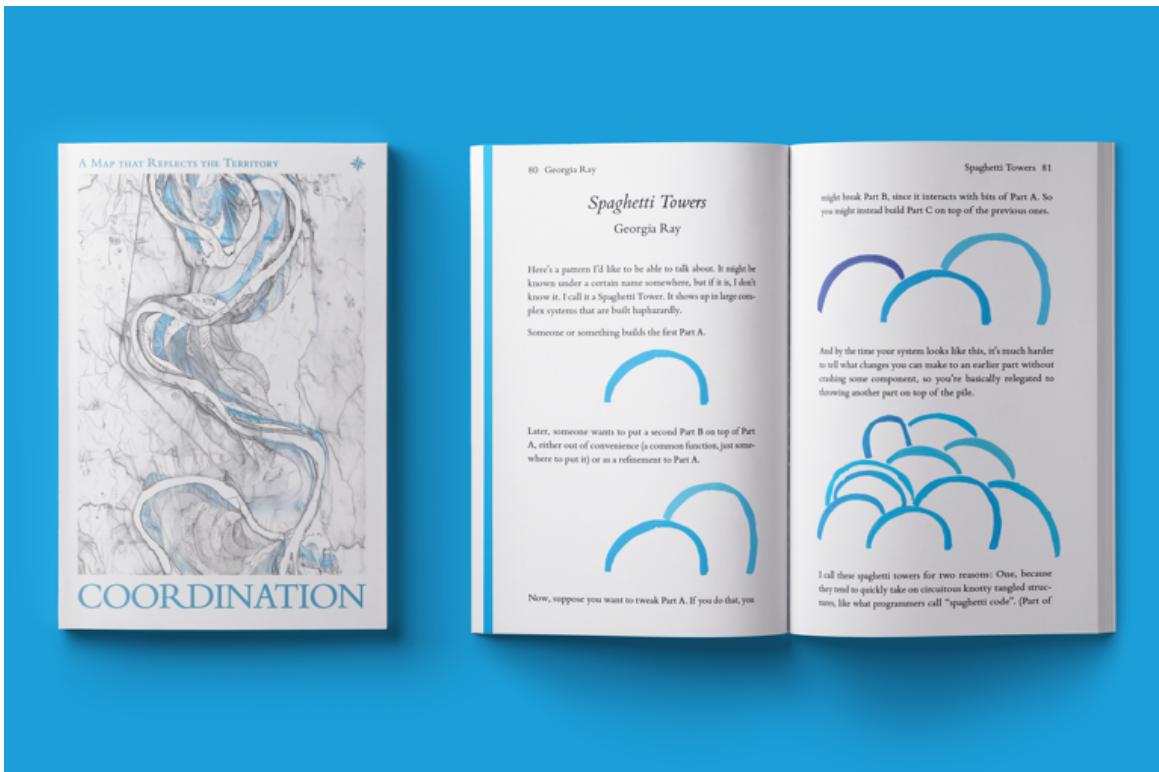
Each book has a unique color.



The first book: *Epistemology*.



The second book: *Agency*.



The third book: *Coordination*.



The fourth book: *Curiosity*.



The fifth book: *Alignment*.

FAQ

What exactly is in the book set?

LessWrong has an annual Review process (the second of which [is beginning today!](#)) to determine the best content on the site. We reviewed all the posts on LessWrong from 2018, and users voted to rank the best of them, the outcome of which can be seen [here](#).

Of the over 2000 LessWrong posts reviewed, this book contains 41 of the top voted essays, along with some comment sections, some reviews, a few extra essays to give context, and some preface/meta writing.

What are the books in the set?

The essays have been clustered around five topics relating to rationality: *Epistemology*, *Agency*, *Coordination*, *Curiosity*, and *Alignment*.

Are all the essays in this book from 2018?

Yes, all the essays in this book were originally published in 2018, and were reviewed and voted on during the 2018 LessWrong Review (which happened at the end of 2019).

How small are the books?

Each book is 4x6 inches, small enough to fit in your pocket. This was the book size that, empirically, most beta-testers found that they actually read.

Can I order a copy of the book?

[Pre-order the book here for \\$29.](#) We currently sell to North America, Europe, Australia, New Zealand, Israel. (If you bought it by end-of-day Wednesday December 9th and ordered within North America, you'll get it before Christmas.) You'll be able to buy the book on Amazon in a couple of weeks.

How much is shipping?

The price above includes shipping to any location that we accept shipping addresses for. We are still figuring out some details about shipping internationally, so if you are somewhere that is not North America, there is a small chance (~10%) that we will reach out to you to ask you for more shipping details, and an even smaller chance (~6%) that we offer you the option to either pay for some additional shipping fees or get a refund.

Can I order more than one copy at a time?

Yes. Just open the form multiple times. We will make sure to combine your shipments.

Does this book assume I have read other LessWrong content, like *The Sequences*?

No. It's largely standalone, and does not require reading other content on the site, although it will be enhanced by having engaged with those ideas.

Can I see an extract from the book?

Sure. Here is [the preface and first chapter of Curiosity](#), specifically the essay *Is Science Slowing Down?* by Scott Alexander.

I'm new — what is this all about? What is 'rationality'?

A scientist is not simply someone who tries to understand how biological life works, or how chemicals combine, or how physical objects move, but is someone who uses the general scientific method in *all* areas, that allows them to empirically test their beliefs and discover what's true *in general*.

Similarly, a rationalist is not simply someone who tries to think clearly about their personal life, or who tries to understand how civilization works, or who tries to figure out what's true in a single domain like nutrition or machine learning; a rationalist is someone who is curious about the general thinking patterns that allows them to think clearly in *all* such areas, and understand the laws and tools that help them make good decisions *in general*.

Just as someone seeking to understand science and the scientific method might look into a great number of different fields (electromagnetism, astronomy, medicine, and so on), someone seeking to understand generally accurate and useful cognitive algorithms would explore a lot of fields and areas. The essays in this set explore questions about arguments, aesthetics, artificial intelligence, introspection, markets, game theory, and more, which all shed light on the core subject of rationality.

Who is this book for?

This book is for people who want to read the best of what LessWrong has to offer. It's for the people who read best away from screens, away from distractions. It's for people who do not check the site regularly, but would still like to get the top content. For many people this is the best way to read LessWrong.

I think there's a lot of people who find the discussion on LessWrong interesting, or are interested in the ideas, or found LessWrong's early discussion of the coronavirus personally valuable, or who know Scott Alexander got started on LessWrong, and would like to see we're about. This book is one of the best ways to do that.

Show me the table of contents?

Sure thing. Here's each book in order.

A Sketch of Good Communication	Ben Pace	
Babble	Alkjash	
Local Validity as a Key to Sanity and Civilization	Eliezer Yudkowsky	
The Loudest Alarm is Probably False	Patrick LaVictoire	
Varieties of Argumentative Experience	Scott Alexander	
More Babble	Alkjash	
Naming the Nameless	Sarah Constantin	
Toolbox-thinking and Law-thinking	Eliezer Yudkowsky	
Prune	Alkjash	
Toward a New Technical Explanation of Technical Explanation	Abram Demski	
Noticing the Taste of Lotus	Michael 'Valentine' Smith	
The Tails Coming Apart As Metaphor For Life	Scott Alexander	
Meta-Honesty: Firming up Honesty Around its Edge-Cases	Eliezer Yudkowsky	
Explaining Insight Meditation and Enlightenment in Non-Mysterious Terms	Kaj Sotala	
Being a Robust Agent	Raymond Arnold	
Anti-social Punishment	Martin Sustrik	
The Costly Coordination Mechanism of Common Knowledge	Ben Pace	
Unrolling Social Metacognition: Three Levels of Meta are not Enough	Andrew Critch	
The Intelligent Social Web	Michael 'Valentine' Smith	
Prediction Markets: When Do They Work?	Zvi Mowshowitz	
Spaghetti Towers	Georgia Ray	
On the Loss and Preservation of Knowledge	Samo Burja	
A Voting Theory Primer	Jameson Quinn	
The Pavlov Strategy	Sarah Constantin	
Inadequate Equilibria vs Governance of the Commons	Martin Sustrik	
Is Science Slowing Down?	Scott Alexander	
What Motivated Rescuers during the Holocaust?	Martin Sustrik	

Is There an Untrollable Mathematician?	Abram Demski	
Why Did Everything Take So Long?	Katja Grace	
Is Clickbait Destroying Our General Intelligence?	Eliezer Yudkowsky	
What Makes People Intellectually Active?	Abram Demski	
Are Minimal Circuits Daemon-Free?	Paul Christiano	
Is There Something Beyond Astronomical Waste?	Wei Dai	
Do Birth Order Effects Exist?	Eli Tyre, Bucky, Raymond Arnold	

Hyperbolic Growth	Paul Christiano	
Specification Gaming Examples in AI	Victoria Krakovna	
Takeoff Speeds	Paul Christiano	
The Rocket Alignment Problem	Eliezer Yudkowsky	
Embedded Agents	Abram Demski & Scott Garrabrant	
FAQ about Iterated Amplification	Alex Zhu	
Challenges to Christiano's Iterated Amplification Proposal	Eliezer Yudkowsky	
Response to FAQ on Iterated Amplification	Eliezer Yudkowsky	
Robustness to Scale	Scott Garrabrant	
Coherence Arguments Do Not Imply Goal-Directed Behavior	Rohin Shah	

Who made this book set?

I (Ben Pace) and Jacob Lagerros (of the Future of Humanity Institute) made these books, alongside my colleagues on the LessWrong Team: Oliver Habryka, Raymond Arnold, Ruby Bloom, and Jim Babcock.

Can I give this book as a gift?

Yes. This is a well-designed, beautiful set of books, designed to be relatively self-contained and not require having read LessWrong before, and that look attractive on coffee-tables and bookshelves, suitable for friends, partners, and family members who read non-fiction.

What about the book called 'Alignment'? Isn't that going to be very technical and have lots of assumptions about AI?

For those who have no knowledge of the subject of AI alignment, the book is structured to help motivate the topic, starting with questions about AI progress and risks, before moving into the meat of open questions about the subject.

The *Alignment* book will be tough reading for those not acquainted with the ongoing discourse around the topic, but I think it will still be rewarding for those who read it.

I have a blog, and might want to review the book. Can I get a review copy?

Yes! I'm offering free copies of the book for review. I'd love to get reviews from critics of the rationality community, members of the rationality community, people who don't really know what the community is about but know that SlateStarCodex is awesome, and more.

If you'd like to review the book and would like a free copy, [fill out this form](#) and I'll get back to you. (Or you can just email me at benitopace@gmail.com if that works better for you.) If you're not sure if your blog is cool enough, your blog is probably cool enough.

Also, you should know that **if you write a public review** of the essay collection **I'll put a link to your review on the official landing page for the book, no matter if it's positive, negative, or not-even-on-that-spectrum.**

(No, tweets don't count, though I guess tweet threads can, but I prefer blog posts. I reserve the right to not include things I read as primarily trolling.)

I have a podcast and might be interested in talking with you about LessWrong. Are you interested in coming on?

Yes. I'm interested in appearing on a few podcasts to let people know about the book. Concretely, I'd propose a joint-appearance with myself and Oliver Habryka, where we can talk about LessWrong, our vision for its future as an institution, how we think it fits into the broader landscape of intellectual progress, the challenges of managing internet forums, and more. No podcast too small (or too big, I guess). If you like LessWrong and you'd like us to come on, we're happy to do it. Email me at benitopace@gmail.com.

I'd like something from you that's not a podcast or a book. Can I reach out?

Yeah, reach out. If you run a newsletter, a mailing list, a google group, or something, and think some of your users would like to know about the book, I'd appreciate you sharing it there with a sentence or two about why you think LessWrong is interesting or worth reading. And if you'd like my input on something, happy to give it via email.

I have a question not answered here?

There's a comment box right below.

Remind me again, how can I pre-order it?

[**Here it is.**](#)

I don't want to pre-order it, I want to be notified when I can buy it on Amazon

You can sign up for notifications on when this (and other books) will go live on Amazon and other marketplaces here:

LessWrong Book Launch Signup

If you need to chat with us about your order or anything, use the intercom widget in the bottom right.

Motive Ambiguity

Central theme in: [Immoral Mazes Sequence](#), but this generalizes.

When looking to succeed, [pain is not the unit of effort](#), and [money is a, if not the, unit of caring](#).

One is not always looking to succeed.

Here is a common type of problem.

You are married, and want to take your spouse out to a romantic dinner. You can choose the place your spouse loves best, or the place you love best.

A middle manager is working their way up the corporate ladder, and must choose how to get the factory to improve its production of widgets. A middle manager must choose how to improve widget production. He can choose a policy that improperly maintains the factory and likely eventually it poisons the water supply, or a policy that would prevent that but at additional cost.

A politician can choose between a bill that helps the general population, or a bill that helps their biggest campaign contributor.

A start-up founder can choose between building a quality product without technical debt, or creating a hockey stick graph that will appeal to investors.

You can choose to make a gift yourself. This would be expensive in terms of your time and be lower quality, but be more thoughtful and cheaper. Or you could buy one in the store, which would be higher quality and take less time, but feel generic and cost more money.

You are cold. You can buy a cheap scarf, or a better but more expensive scarf.

These are trade-offs. Sometimes one choice will be made, sometimes the other.

Now consider another type of problem.

You are married, and want to take your spouse out to a romantic dinner. You could choose a place you both love, or a place that only they love. You choose the place you don't love, so they will know how much you love them. After all, you didn't come here for the food.

A middle manager must choose how to improve widget production. He can choose a policy that improperly maintains the factory and likely eventually poisons the water supply, or a policy that would prevent that at no additional cost. He knows that when he is up for promotion, management will want to know the higher ups can count on him to make the quarterly numbers look good and not concern himself with long term issues or what consequences might fall on others. If he cared about not poisoning the water supply, he would not be a reliable political ally. Thus, he chooses the neglectful policy.

A politician can choose between two messages that affirm their loyalty: Advocating a beneficial policy, or advocating a useless and wasteful policy. They choose useless,

because the motive behind advocating a beneficial policy is ambiguous. Maybe they wanted people to benefit!

A start-up founder can choose between building a quality product without technical debt and creating a hockey stick graph with it, or building a superficially similar low-quality product with technical debt and using that. Both are equally likely to create the necessary graph, and both take about the same amount of effort, time and money. They choose the low-quality product, so the venture capitalists can appreciate their devotion to creating a hockey stick graph.

You can choose between making a gift and buying a gift. You choose to make a gift, because you are rich and buying something from a store would be meaningless. Or you are poor, so you buy something from a store, because a handmade gift wouldn't show you care.

Old joke: One Russian oligarch says, "Look at my scarf! I bought it for ten thousand rubles." The other says, "That's nothing, I bought the same scarf for twenty thousand rubles."

What these examples have in common is that there is a strictly better action and a strictly worse action, in terms of physical consequences. In each case, the protagonist chooses the worse action *because it is worse*.

This choice is made as a costly signal. In particular, to avoid *motive ambiguity*.

If you choose something better over something worse, you will be suspected of doing so *because it was better rather than worse*.

If you choose something worse over something better, not only do you show how little you care about making the world better, you show that you care more *about people noticing and trusting this lack of caring*. It shows your values and loyalties.

In the first example, you care more about your spouse's view of how much you care about their experience than you care about your own experience.

In the second example, [you care more about being seen as focused on your own success](#) than you care about outcomes you won't be responsible for.

In the third example, [you care more about being seen as loyal](#) than about improving the world by being helpful.

In the fourth example, [you care about those making decisions over your fate believing that you will focus on the things they believe the next person deciding your fate will care about](#), so they can turn a profit. They don't want you distracted by things like product quality.

In the old joke, the oligarchs want to show they have money to burn, and that they care a lot about showing they have lots of money to burn. That they *actively want to Get Got* to show they don't care. If someone thought the scarf was bought for [mundane utility](#), that wouldn't do at all.

One highly effective way to get many people to spend money is to give them a choice to either spend the money, or be slightly socially awkward and admit that they care about not spending the money. Don't ask what the wine costs, it would ruin the evening.

The warning of [Out to Get You](#) is insufficiently cynical. The motive is often not to get your resources, and is instead purely to make your life worse.

[Conflict theorists](#) are often insufficiently cynical. We *hope* the war is about whether to enrich the wealthy or help the people. Often the war is over whether to aim to destroy the wealthy, or aim to hurt the people.

In [simulacra terms](#), these effects are strongest when one desires to be seen as motivated on level three, but these dynamics are potentially present to an important extent for motivations at all levels. Note also that one is not motivated by this dynamic to destroy something unless you might plausibly favor it. If and only if [everybody knows](#) you don't care about poisoning the river, it is safe to not poison it.

This generalizes to time, to pain, to every preference. Hence anything that wants your loyalty will do its best to ask you to sacrifice and destroy everything you hold dear, because you care about it, to demonstrate you care more about other things.

Worst of all, *none of this assumes a zero-sum mentality*. At all.

Such behavior *doesn't even need one*.

If one has a true zero-sum mentality, as many do, or one maps all results onto a zero-sum social dynamic, all of this is overthinking. All becomes simple. Your loss is my gain, so I want to cause you as much loss as possible.

Pain need not be the unit of effort if it is the unit of scoring.

The world would be better if people treated more situations like the first set of problems, and less situations like the second set of problems. How to do that?

parenting rules

([crossposted from my nascent substack](#))

Way back in 2012 I wrote up on livejournal (I told you this was a long time ago) a few parenting rules we lived by. This is the one livejournal post I regularly reshare, so here it is on a more modern platform. Our kids are older now (11 and 13) but with one exception I think these really hold up.

That one exception is praise, where the research on praise seemed clear in 2012 and has since largely failed to replicate and certainly doesn't have the effect size that everyone thought, so that one I no longer stand by.

Here they are:

- **Try never to lie.** If kids ask a question and they aren't ready to hear the answer, just tell them that. This doesn't mean you have to go into every gruesome detail, it's fine to couch your answer at the level you think they'll understand and that you have time for, but they're smarter than you probably think.

This does extend to things like Santa Claus and the Easter Bunny. We've told them those stories with the attempt to treat them just like any other fictional story. When Jackson point blank asked if Santa was real, I told him, "No, but it's a fun story and fun to pretend."

There's a common pattern with kids to tell them things that are untrue but scary as a joke, like "Be careful not to slip down the drain!" Don't do that. Kids have trouble distinguishing fake warnings from real ones.

However, saying untrue things as a joke is fine in the right context. "Elephant toes" is a fine answer to a question about what's for dinner. (As long as it's not true.) People say untrue things all the time, and taking the time to evaluate whether an adult is telling the truth is a useful skill. But until the kids are good at it, the untruths should be completely implausible, then can get more plausible as they get more on to you. Fun game, actually.

The most difficult time for this one is when they want something that you don't want to give them. Like if mommy is downstairs and I'm doing bedtime, it's very tempting to claim that Katy is busy doing sometime important that can't be interrupted rather than just admitting she needs a break, or it's my turn to answer the late night call.

- Remember that **every interaction is a repeated game**, and your goal is not to win this one iteration, but to win the series. So if a child is crying because she wants something, even though it feels like a win to give in now (she stops crying which is better for everyone, you haven't really given up much), it's disastrous in the repeated game because she learns that she can get what she wants by crying.

The flipside of that is that you have to let them get what they want in other ways. If you say no and they have good reasons why you should give in, or even an attempt at good reasons, sometimes you have to give in. You want them to

be thinking critically and trying to persuade you.

Here's an example. Katy put down a couple of dollars on the counter, which Jackson took, leading to the following conversation:

Katy: Jackson, please leave those there.

Jackson: But this one is mine.

Katy: No it's not, I just put it there.

Jackson: It looks just like the one I got last week!

Katy: It's not the same one, I just put it there like 30 seconds ago!

Jackson: But money is spongeable.

Katy: ...

Katy: Ok, you can have it.

Because money being fungible is a great reason, even if it's not completely persuasive in this particular instance, and "spongeable" is awesome. If he'd started crying, the answer would have been a much more solid, no-more-negotiation "no."

- Almost **never bluff**. This is related to the first two points, but is really more like the second. If you threaten a consequence and don't follow through, they'll figure that out really quickly. Which leads to the following rule: be very careful with threats. If you make them, carry them out; if you don't want to carry out the threat, don't make it.

Sometimes we violate that. The most common case is when the kid is obviously bluffing. So when we're leaving somewhere and Lucile declares she isn't coming, I raise my hand by telling her goodbye and starting to walk away. So far she's folded every time. Note: I wouldn't do that if it upset her, she gets that I'm not really going to leave her.

- **Praise the process**, not the person. We're pretty particular in how we praise our kids. We try to use process praise ("I like the way you made up a story about all the parts of your drawing"), some amount of results praise ("That block tower is amazing! It's so tall!"), and virtually zero person praise ("You're a good artist/architect.")

This is because process praise is motivating and helpful, and person praise is demotivating. Here's an [article on the praise research](#), or you could go [look at it yourself](#).

Also try to avoid general praise ("nice job") in favor of specifics, though in practice that's sometimes pretty hard.

The uncontroversial flipside is that criticism works the same way. Process criticism ("Your elbow is too low when you swing, raise it up higher") is good, limited amounts of results criticism is ok ("I've seen you do better, let's try it again"), person criticism is right out ("You're a bad baseball player").

NB: I no longer stand by this section. There's probably something to growth mindset still, but how you give a few words of praise ain't it.

- Answer questions with **as much detail as they want**. I've had conversations with the kids about civil rights, affirmative action, religion, communism versus capitalism, consequences for breaking laws, race, sexuality, and so on. Not because I've set out to teach them that stuff, but because they ask lots of questions and I try to answer them. Kids are mostly concerned with concrete,

day-to-day things -- but some of the best interactions come when they are in the right questioning mood, and you definitely want to take advantage of it.

You have to be age appropriate -- when talking about where babies come from, I don't talk about penises in vaginas to a 5 year old -- but they can handle a lot more than most adults give them credit for.

It's amazing to me how often we get strange looks or pushback from other parents about these. People thought we were ax murderers for not teaching our kids that Santa is real.

If I had to sum all these up, it would be this: **raise kids for the long term**. The reasoning behind all these choices is that we want to produce competent capable adults, and solving short term in-the-moment issues, while important, isn't the goal.

2020 AI Alignment Literature Review and Charity Comparison

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

cross-posted to the EA forum [here](#).

Introduction

As in [2016](#), [2017](#), [2018](#), and [2019](#), I have attempted to review the research that has been produced by various organisations working on AI safety, to help potential donors gain a better understanding of the landscape. This is a similar role to that which GiveWell performs for global health charities, and somewhat similar to a securities analyst with regards to possible investments.

My aim is basically to judge the output of each organisation in 2020 and compare it to their budget. This should give a sense of the organisations' average cost-effectiveness. We can also compare their financial reserves to their 2020 budgets to get a sense of urgency.

I'd like to apologize in advance to everyone doing useful AI Safety work whose contributions I have overlooked or misconstrued. As ever I am painfully aware of the various corners I have had to cut due to time constraints from my job, as well as being distracted by 1) other projects, 2) the miracle of life and 3) computer games.

This article focuses on AI risk work. If you think other causes are important too, your priorities might differ. This particularly affects GCRI, FHI and CSER, who both do a lot of work on other issues which I attempt to cover but only very cursorily.

How to read this document

This document is fairly extensive, and some parts (particularly the methodology section) are largely the same as last year, so I don't recommend reading from start to finish. Instead, I recommend navigating to the sections of most interest to you.

If you are interested in a specific research organisation, you can use the table of contents to navigate to the appropriate section. You might then also want to Ctrl+F for the organisation acronym in case they are mentioned elsewhere as well. Papers listed as 'X researchers contributed to the following research lead by other organisations' are included in the section corresponding to their first author and you can Cntrl+F to find them.

If you are interested in a specific topic, I have added a tag to each paper, so you can Ctrl+F for a tag to find associated work. The tags were chosen somewhat informally so you might want to search more than one, especially as a piece might seem to fit in multiple categories.

Here are the un-scientifically-chosen hashtags:

- AgentFoundations
- Amplification
- Capabilities
- Corrigibility
- DecisionTheory
- Ethics
- Forecasting

- GPT-3
- IRL
- Misc
- NearAI
- OtherXrisk
- Overview
- Politics
- RL
- Strategy
- Textbook
- Transparency
- ValueLearning

New to Artificial Intelligence as an existential risk?

If you are new to the idea of General Artificial Intelligence as presenting a major risk to the survival of human value, I recommend [this Vox piece](#) by Kelsey Piper, or for a more technical version [this](#) by Richard Ngo.

If you are already convinced and are interested in contributing technically, I recommend [this piece](#) by Jacob Steinhardt, as unlike this document Jacob covers pre-2019 research and organises by topic, not organisation, or [this](#) from Critch & Krueger, or [this](#) from Everitt et al, though it is a few years old now

Research Organisations

FHI: The Future of Humanity Institute

FHI is an Oxford-based Existential Risk Research organisation founded in 2005 by Nick Bostrom. They are affiliated with Oxford University. They cover a wide variety of existential risks, including artificial intelligence, and do political outreach. Their research can be found [here](#).

Their research is more varied than MIRI's, including strategic work, work directly addressing the value-learning problem, and corrigibility work - as well as work on other Xisks.

They run a Research Scholars Program, where people can join them to do research at FHI. There is a fairly good review of this [here](#). Unfortunately I suspect the pandemic may have reduced its effectiveness this year, as FHI has often favoured informal networking rather than formal management structures, but it seems to have worked well pre and hopefully post pandemic.

The EA Meta Fund supported a special program for providing infrastructure and support to FHI, called the [Future of Humanity Foundation](#). This reminds me somewhat of what BERI does.

In the past I have been very impressed with their work.

Research

Bostrom & Shulman's [Sharing the World with Digital Minds](#) discusses the moral issues raised by the potential for uploads or other digital minds. By virtue of their number, speed, or specific design, these could be utility monsters - a term from Nozick for agents much more

efficient than humans at turning resources into utility. Would we therefore be obliged to give up all our resources to them and eventually let meat humanity starve to death? This much has been discussed before - indeed, I alluded to this as an argument against a universal basic income as a response to AI-driven unemployment in previous versions of this article! - but this article both provides a canonical reference and also a good survey showing that such issues come up under a wide variety of ethical views and technological possibilities. I also enjoyed the discussion of the issues posed by rapid reproduction for 'democratic' political systems, where influence is the scarce resource. #Strategy

Ashurst et al.'s [A Guide to Writing the NeurIPS Impact Statement](#) gives advice on how to write the new 'impact statements' that NeurIPS now requires. Seizing this gap in the market by writing the canonical piece that everyone will find when they google - my tests suggest they have the SEO - and filling it with a counterfactually valuable article is some good out-of-the-box thinking. As well as containing many very useful links, I liked the suggestion that even theoretical pieces should consider their impacts. #Misc

Kovařík & Carey's [\(When\) Is Truth-telling Favored in AI Debate?](#) provides some formalism and theorems around the properties of debate. I thought the section about debate length was very interesting, where it seems to show (at least for this class of debate) that debates are either long enough to produce the truth in a trivial manner (through full exposition) or else error can be arbitrarily high with even one fewer step, though they also identified plausible seeming sub-classes with much better performance. (the paper is technically from the very end of 2019 but I missed it last year) See also the discussion [here](#). #Amplification

Shevlane & Dafoe's [The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?](#) discusses whether increased AI publishing will generally be more useful for 'attack' or 'defence'. They argue that the 'publishing exploits' is generally best practice (with a lag) model from cybersecurity might not be best placed here - an important argument to rebut, as many people used it to criticise OpenAI's decision to be (initially) clopen with regard GPT-2. #Strategy

Ord's [The Precipice](#) provides a detailed overview of existential risks and the future of humanity. It covers a variety of risks, including a good section on AGI, which Toby estimates as the largest risk at ~ 10% / century. There is also a huge amount of other material covered, including some novel ideas to me like the section on risk correlations, as well as some very motivational final chapters. I was pleasantly surprised to learn that 80% of DNA synthesis was being screened (in some way) for dangerous compounds. Probably replaces Bostrom and Ćirković as the best book on the subject now. #Overview

Carey et al.'s [The Incentives that Shape Behaviour](#) attempts to build a general theory of what sort of incentives lead agents to manipulate humans. This is basically causal diagram classification, revealing incentives to control and react to humans. It includes examples for both fairness incentives and also a possible way of reducing human manipulation incentivisation: optimising for a separately trained predictor. See also the discussion [here](#). Researchers from Deepmind were also named authors on the paper. #AgentFoundations

Clarke's [Clarifying "What failure looks like" \(part 1\)](#) attempts a more detailed analysis of the issues raised in Christiano's [What failure looks like](#) I liked the breakdown of lock-in mechanisms, which seem true to me. It provides a lot of examples, some of which I liked, like that of the Maori. However many of them were sufficiently simplified that I feel significant disanalogies were overlooked - for example, the Climate Change example neglects the very different incentives facing regulated utilities, and the agricultural revolution example seems to require a strong commitment to average utilitarianism, even though this is not a popular view of population ethics. Despite this I thought the underlying argument seemed pretty plausible. #Forecasting

Armstrong et al.'s [Pitfalls of Learning a Reward Function Online](#) introduces two desirable properties for agents who are trying to learn human values at runtime (unruggability and

uninfluenceability) and proves they are broadly the same thing. As well as proving this result, it contains a series of examples of what can go wrong in the absence of either property - including sacrificing reward with probability 100% - and a brief discussion of how counterfactual rewards might address the problem. It ends with an extended gridworld example, but I found this a little hard to follow. See also the discussion [here](#). Researchers from Deepmind were also named authors on the paper. #ValueLearning

Tucker et al.'s [Social and Governance Implications of Improved Data Efficiency](#) discusses some of the strategic implications of ML systems that do not require as much data. They argue that it is not obvious that they will net benefit smaller firms - if the impact is multiplicative, it might benefit larger firms with more compliments (like market access) more - though I am not sure a multiplicative effect is really a good model for what people are thinking about when they talk about ML models needing less data. They also point out that due to threshold effect this might enable entirely new applications, and in particular IRL/amplification, as these rely on a very scarce source of data: humans. #Forecasting

Cohen & Hutter's [Curiosity Killed the Cat and the Asymptotically Optimal Agent](#) show that because any agent that is guaranteed to eventually find the optimal strategy can only do so by testing every option, any 'traps' in the environment will eventually be triggered with probability 1. (Unless traps are disabled after finite time). This is clearly kinda important - it is nice to be able to reason about asymptotic optimality, but we do not want an AGI that deletes humanity with $p=1$ en route. This suggests something of a bootstrap problem, where we need a 'mentor' to avoid such dangers. Researchers from Deepmind were also named authors on the paper. #RL

Cohen & Hutter's [Pessimism About Unknown Unknowns Inspires Conservatism](#) basically tries to make a conservative AIXI that defers to its mentor when it is not sure. It does this by comparing its worst-case estimates to its estimate of the mentor's expected case, and defers to the mentor more when the difference is higher (and less as $t \rightarrow \infty$). Hopefully the mentor will help keep the agent from being too conservative, as it seems there is a risk that it simply ends up doing nothing, and gets out-competed by an EV maximising agent? Researchers from Deepmind were also named authors on the paper. #RL

Nguyen & Christiano's [My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda](#) provides an overview of Paul's IDA agenda. Probably the best such explanation so far; written by Chi when she was at FHI with in-line comments from Paul. Researchers from OpenAI were also named authors on the paper. #Amplification

Snyder-Beattie et al.'s [The Timing of Evolutionary Transitions Suggests Intelligent Life Is Rare](#) builds a bayesian model to try to get around the anthropic problem of estimating how easy it is for life to develop. Specifically, they use non-informative priors and update based on the distribution of various transitions (e.g. Eukaryotes), concluding (similar to previous work they cite) that the development of life is relatively hard. See also the discussion [here](#). #Forecasting

Ding & Dafoe's [The Logic of Strategic Assets: From Oil to AI](#) analyses what causes a product to be 'strategic' to a country. They decompose this into the product of its Importance, Externalities and Rivalosity, in contrast to previous analysis of simply 'military importance'. Some of the examples I might quibble with - for example, the paper claims that the spillovers from railways lead private agents to underinvest, which is somewhat in tension with the experience of the [railway bubbles](#). I am also a bit sceptical that this analysis really subsumes the idea of dependency-strategic items - nitrates in WWI, and nuclear weapons now, both lack substitutes and are at risk of supply disruptions, but neither really seem to have massive externalities. It also would have been nice to see some analysis of why individual firms do not internalise the risk of supply disruption - is this due to anti-price gouging laws? It finishes with detailed discussion of two examples - British Jet Engines (reminding me of Attlee's [disastrous mistake with another type of engine](#)) and US-Japanese rivalry. The report discusses several mistakes US policy made during this period - e.g. accidentally classifying

cash registers as strategic, and missing rayon fibers - but these mistakes seem like they are adequately explained without the theory put forward by the paper. #NearAI

Cotton-Barratt et al.'s [Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter](#) provides a series of taxonomies for existential risks. In particular, they discuss distinctions between preventing and mitigating events, how events scale to be global, and how direct their effect is. See also the discussion [here](#). #Strategy

Cihon et al.'s [Should Artificial Intelligence Governance be Centralised? Design Lessons from History](#) discusses the advantages of centralised or fragmented international law approaches to AI. Most of the considerations are not AI specific. Researchers from CSER were also named authors on the paper. #Strategy

O'Brien & Nelson's [Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology](#) discusses the impact of AI on biorisk. They first discuss the problems with several existing frameworks and the potential impact of AI on bio risk, before offering their own framework. #OtherXrisk

Cremer & Whittlestone's [Canaries in Technology Mines: Warning Signs of Transformative Progress in AI](#) attempt to identify possible signs of imminent AGI though expert-solicitation of causal influence diagrams. Basically a technology that is seen as a prerequisite for many others is a candidate for being a canary. However, I didn't feel the paper really addressed the issues raised in Eliezer's [Fire Alarm post](#). Researchers from CSER were also named authors on the paper. #Forecasting

O'Keefe's [How will National Security Considerations affect Antitrust Decisions in AI? An Examination of Historical Precedents](#) surveys a bunch of historical antitrust actions in the US to see how national security arguments played into the outcome. He finds that it was pretty rare, especially recently, and when it did it was generally congruent with the main antitrust objectives, namely preventing artificial reductions in output. The idea here presumably is to suggest that the US government is unlikely to use antitrust as a tool in an AI race unless firms start overcharging for their services. O'Keefe also lists support from OpenPhil. #Politics

Bostrom et al.'s [Written Evidence to the UK Parliament Science & Technology Committee's Inquiry on A new UK research funding agency](#). recommends that Cumming's new British DARPA focus on existential risks. I think this a worthwhile but big ask - DARPA seems more intended to fund risky things than to reduce risk - and now Cummings has left I worry the window for intervention here may have passed. Researchers from CSER were also named authors on the paper. #Politics

O'Keefe et al.'s [The Windfall Clause: Distributing the Benefits of AI for the Common Good](#) proposes that AI firms voluntarily commit to donating some % of profits over a high threshold to humanity in general. The idea is that the cost of this commitment is currently negligible, but would be extremely socially valuable if one firm gained a decisive strategic advantage. I think it's good to work on novel governance strategies, but I'm not very enthusiastic about this specific option, partly for reasons I outlined in lengthy but unfinished comments on the forum post, but mainly because I don't think it does much to reduce the existential risk, especially vs similar ideas like encouraging consolidation among AI firms. See also the discussion [here](#). #Politics

Garfinkel's [Does Economic History Point Towards a Singularity?](#) and the associated document analyse the claim that economic growth has been accelerating in accordance with global GDP (or population). In general it finds the evidence for this to be somewhat weak. #Forecasting

Prunkl & Whittlestone's [Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society](#) proposes alternative divisions of the AI safety community other than near vs long term. These are: impacts, capabilities, certainty and scale. The paper argues that we should focus on these axis because 1) there is variance that is overlooked by

a single short-vs-long axis and 2) this can cause misunderstandings. I did not really find this convincing: the purpose of any clustering is to summarize data, and I have yet to come across any examples of confusions that would be dispelled by their alternative axes. In fact, their motivating example - that of Etzioni's misreading of Bostrom - is a case where relying on the 'long term' stereotype would have given Etzioni more accurate beliefs! Similarly, their examples of 'intermediate' issues, like the long-term impact on inequality of algorithmic discrimination, seems to me like precisely the sort of political (and in my opinion mistaken) concern that everyone would agree falls into the 'short-term' camp. But perhaps, like [Cave & Ó hÉigearthaigh](#), this paper is better understood as a speech act. See also the discussion [here](#). Researchers from Leverhulme were also named authors on the paper. #Strategy

FHI researchers contributed to the following research led by other organisations:

- Brundage et al.'s [Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims](#)
- Scholl & Hanson's [Testing the Automation Revolution Hypothesis](#)

They also produced a variety of pieces on biorisk and other similar subjects, which I am sure are very good and important but I have not read.

According to [Riedel & Deibel](#), over the 2016-2020 period, FHI accounted for by far the largest number of citations for meta-AI-safety work, and a respectable showing in technical AI safety.

Finances

FHI didn't reply to my emails about donations, and seem to be more limited by talent than by money.

If you wanted to donate to them anyway, [here](#) is the relevant web page.

"edit (2020-12-26): FHI subsequently did reach out to inform me that while they could not share the information financial etc. information I requested they would still appreciate donations."

CHAI: The Center for Human-Aligned AI

CHAI is a UC Berkeley based AI Safety Research organisation founded in 2016 by Stuart Russell.. They do ML-orientated safety research, especially around inverse reinforcement learning, and cover both near and long-term future issues. One outside interpretation of their work from Alex Flint is [here](#).

As an academic organisation their members produce a very large amount of research; I have only tried to cover the most relevant below. It seems they do a better job engaging with academia than many other organisations, especially in terms of interfacing with the cutting edge of non-safety-specific research. The downside of this, from our point of view, is that not all of their research is focused on existential risks.

Rohin Shah, now with additional help, continues to produce the [AI Alignment Newsletter](#), covering in detail a huge number of interesting new developments, especially new papers. I really cannot praise these newsletters highly enough. Unfortunately for CHAI, but probably fortunately for the world, he has graduated and is moving to Deepmind.

They have expanded somewhat to other universities outside Berkeley and have people at places like Princeton and Cornell.

Research

CHAI and their associated academics produce a huge quantity of research. Far more so than other organisations their output is under-stated by my survey here; if they were a small organisation that only produced one report, there would be 100% coverage, but as it is this is just a sample of those pieces I felt most interested in. On the other hand academic organisations tend to produce some slightly less relevant work also, and I have focused on what seemed to me to be the top pieces.

Critch & Krueger's [AI Research Considerations for Human Existential Safety \(ARCHEs\)](#) is a super-detailed overview of the state of the field, and a research agenda. It provides a detailed explanation of key concepts and a categorisation schema of various possible scenarios, including new distinctions I hadn't seen clearly made before. This is a mammoth document, and I encourage the reader to attempt it if possible. A few interesting points for me were his argument that AI researcher's discussions of 'near' AI problems as being the first steps towards admitting problems, or that Distributional Shift work might not be neglected by Industry? Contrary to some others he argues that we should perhaps never make 'prepotent' AI (one that cannot be controlled by humans) - not even a defensive one to prevent other AI threats. There is also a lot of discussion of multi-polar scenarios - the idea that single agent alignment/delegation problems are less important to focus on, partly because the single-agent version is more likely to be solved by profit-maximising firms. See also the discussion [here](#). Researchers from BERI were also named authors on the paper.

#Overview

Andreea et al.'s [LESS is More: Rethinking Probabilistic Models of Human Behavior](#) attempts to extend the model of Boltzmann rationality (where humans choose the best option, with noise, from a finite menu) to the continuous case. This is essentially by providing continuous measures of how 'similar' different options are, to show that e.g. driving at 41mph and 41.1mph are basically the same thing. #IRL

Christian's [The Alignment Problem: Machine Learning and Human Values](#) is a heavier-than-pop-sci book introduction to near and long-term AI issues. It does a good job connecting short-term worries (first part of book) to the bigger longer-term issues (second part of book), tying them together in multiple ways, and the scholarship seems very good. I enjoyed reading. #Overview

Critch's [Some AI research areas and their relevance to existential safety](#) describes Critch's views on a variety strategic research landscape questions. It contains some interesting ideas, like technical progress legitimising governance demands by making them credibly achievable. More importantly is the detailed and sophisticated analysis of each of these research areas in terms of their value and neglectedness. Notably for me were the sections arguing that research areas I have historically thought of as being pretty core to reducing AI X-risk, like Agent Foundations and Value Learning, as being not very useful, as well as a very positive view of studying Human-Robot interaction. However, I think it is a little credulous with regard to many near AI safety issues like fairness, to the point of supporting GDPR because more regulation is desirable, regardless of whether that regulation is good.

#Strategy

Gleave et al.'s [QUANTIFYING DIFFERENCES IN REWARD FUNCTIONS](#) introduces a distance metric for reward functions. This allows us to judge whether two reward functions are 'the same' - at least relative to a certain environment. They might differ in a larger environment, as this pseudo-metric is weaker than utility functions' being identical up to an affine transformation. It might be useful as a measure of how accurately RL agents have learnt the intended reward. Researchers from Deepmind were also named authors on the paper. #RL

Reddy et al.'s [Learning Human Objectives by Evaluating Hypothetical Behavior](#) attempts to learn safely by using hypothetical scenarios. Basically prior to letting the RL agent run around in the environment and potentially act unsafely, they procedurally generate hypotheticals in various ways and have the humans give feedback on them, so the agent can

pre-learn before being let loose on the real environment. See also the discussion [here](#). Researchers from Deepmind were also named authors on the paper. #IRL

Freedman et al.'s [Choice Set Misspecification in Reward Inference](#) introduces and analyses the implications of an IRL agent which has mistaken beliefs about its teacher's choice set. The obvious consequence would be assigning a low value on something that the human appears to have decided against - when it was actually inaccessible. The paper breaks this down into different cases, and shows (somewhat unsurprisingly) that the harm this does can vary from negligible to maximal. In some scenarios it is even helpful, by preventing an imperfectly rational human from mistakenly choosing a sub-optimal choice during training. #IRL

Shah's [AI Alignment 2018-19 Review](#) is a huge overview of AI alignment work from the prior two years. If you want to survey what people have been working on (as opposed to determining which organisations are best to donate to) this post is an excellent resource. #Overview

Russel & Norvig's [Artificial Intelligence: A Modern Approach](#), 4th Edition is the latest version of the famous textbook. It contains a chapter on AI ethics and safety, as previous editions did. The chapter is mainly focused on 'near' AI issues like discrimination; while it does provide an overview of some of the issues and techniques in AI alignment work, it doesn't really make the case for why this is so vitally important. #Textbook

Halpern & Piermont's [Dynamic Awareness](#) presents a version of modal logic for logical uncertainty. Specifically, agents becoming 'aware' of propositions they had not previously considered. #AgentFoundations

CHAI researchers contributed to the following research led by other organisations:

- Qian et al.'s [AI GOVERNANCE IN 2019 A YEAR IN REVIEW](#)
- Hendrycks et al.'s [Aligning AI with Shared Human Values](#)

According to Riedel & Deibel, over the 2016-2020 period, CHAI accounted for the second largest number of citations for technical AI safety.

Finances

They have been funded by various EA organisations including the Open Philanthropy Project and recommended by the [Founders Pledge](#).

They spent \$2,000,000 in 2019 and \$1,650,000 in 2020, and plan to spend around \$2,200,000 in 2021. They have around \$3,892,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1.8 years of runway. Their 2020 spending was about 20% below plan due to the pandemic.

If you wanted to donate to them, [here](#) is the relevant web page. Unfortunately it is apparently broken at time of writing - they tell me any donation via credit card can be made by calling the Gift Services Department on 510-643-9789.

MIRI: The Machine Intelligence Research Institute

MIRI is a Berkeley based independent AI Safety Research organisation founded in 2000 by Eliezer Yudkowsky and currently led by Nate Soares. They were responsible for much of the early movement building for the issue, but have refocused to concentrate on research for the last few years. With a fairly large budget now, they are the largest pure-play AI alignment shop. Their research can be found [here](#). Their annual summary can be found [here](#).

In general they do very 'pure' mathematical work, in comparison to other organisations with more 'applied' ML or strategy focuses. I think this is especially notable because of the irreplaceability of the work. It seems quite plausible that some issues in AI safety will arise early on and in a relatively benign form for non-safety-orientated AI ventures (like autonomous cars or Minecraft helpers) – however the work MIRI does largely does not fall into this category. I have also historically been impressed with their research and staff.

Their agent foundations work is basically trying to develop the correct way of thinking about agents and learning/decision making by spotting areas where our current models fail and seeking to improve them. This includes things like thinking about agents creating other agents.

MIRI, in collaboration with CFAR, runs a series of four-day workshop/camps, the [AI Risk for Computer Scientists workshops](#), which gather mathematicians/computer scientists who are potentially interested in the issue in one place to learn and interact. This sort of workshop seems very valuable to me as an on-ramp for technically talented researchers, which is one of the major bottlenecks in my mind. In particular they have led to hires for MIRI and other AI Risk organisations in the past. I don't have any first-hand experience however, and presumably these were significantly suppressed by the pandemic.

They also support [MIRIx workshops](#) around the world, for people to come together to discuss and hopefully contribute towards MIRI-style work.

MIRI continue their policy of [nondisclosure-by-default](#), something [I've discussed in the past](#), which despite having some strong arguments in favour unfortunately makes it very difficult for me to evaluate them. I've included some particularly interesting blog posts some of their people have written below, but many of their researchers produce little to no public facing content.

They are (were?) also apparently considering leaving the bay area, which I think I would consider positively.

edit 2020-12-25: after publishing this article, MIRI posted [this](#) blog post explaining they were embarking on a significant change of direction as they felt their post-2017 primary research direction, working on fundamental agent foundation 'deconfusion', was not making much progress. Some staff members will be leaving as a result. It is not clear to what extent they will disclose their new research directions. I haven't had time to fully internalise this news, so leave the link for the reader to evaluate.

Research

Most of their work is non-public. Here are three forum posts from the last year by staff that I thought were insightful.

Hubinger's [An overview of 11 proposals for building safe advanced AI](#) examines eleven different strategies for AI safety. It evaluates these on how promising they are for both the inner and outer alignment problems, as well as competitiveness - it is no good producing a 100% safe system if someone else out-competes you with a more risky one. This is the first post I've seen of this type and it does a great job. #Overview

Garrabrant's [Cartesian Frames](#) is a sequence of posts putting forward a new way of thinking, and associated mathematical formalism, about agency. The idea is to move away from dualistic AIXI style models, where the agent is outside the world, towards a system where we can examine different 'framings', each of which suggest a different thing as being agent-like - being able to make choices. This sensible philosophical motivation is then associated with a lot of category theory formalism, allowing you to do things like combining agents, decomposing agents, etc. #AgentFoundations

Abram Demski's [Radical Probabilism](#) presents a non-bayesian (ish) alternative account of probability. It is designed to take into account non-certain evidence, and allow for less rigid updating rules - in particular the fact that we can learn from thinking, not just from new sense data. I really enjoyed the dialogues, where I think the foil did a good job of presenting the objections I wanted to make. At the end of it I'm still not convinced what I think though - it seems a little unfair to compare a fully specified system, whose problems are easy to point out, with a somewhat hypothetical replacement. #AgentFoundations

According to Riedel & Deibel, over the 2016-2020 period, MIRI came in third for the number of citations in technical AI safety.

Finances

They spent \$6,050,067 in 2019 and \$7,500,000 in 2020, and plan to spend around \$6,500,000 in 2021. They have around \$13m380,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 2.1 years of runway. 2020 spending was above plan; most orgs spent less due to the pandemic, but MIRI invested in sub-quarantine live/work spaces outside Berkeley so researchers could still benefit from in-person collaboration.

They have been supported by a variety of EA groups in the past, including OpenPhil.

They are not running a formal fundraiser this year but apparently would still welcome donations; if you wanted to donate to MIRI, [here](#) is the relevant web page.

GCRI: The Global Catastrophic Risks Institute

GCRI is a globally-based independent Existential Risk Research organisation founded in 2011 by Seth Baum and Tony Barrett. They cover a wide variety of existential risks, including artificial intelligence, and do policy outreach to governments and other entities. Their research can be found [here](#). Their annual summary can be found [here](#).

In 2020 they [continued their advising program](#) where they gave guidance to people from around the world who wanted to help work on catastrophic risks.

In 2020 they hired McKenna Fitzgerald as Project Manager and Research Assistant.

Research

Baum's [Accounting for violent conflict risk in planetary defense decisions](#) discusses the impacts and lessons from asteroid defence for other Xrisks, mainly nuclear war. It contains some interesting history about how congress came to care about asteroid defence - including that popular movies, while inaccurate, were quite helpful, and that many astronomers were relatively opposed. It also points out that using nuclear weapons or similar against an asteroid would probably be in violation of international law. Presumably in a disaster scenario the US would simply ignore this, but it might make preparation and practice ahead of time more difficult. #OtherXrisk

Baum's [Quantifying the Probability of Existential Catastrophe: A Reply to Beard et al.](#) responds to the CSER paper. It makes some methodological points, like about the importance of different thresholds for what constitutes a catastrophe, and ways in which this forecasting could be improved. See also the discussion [here](#). #Forecasting

Baum's [Artificial Interdisciplinarity: Artificial Intelligence for Research on Complex Societal Problems](#) discusses how AI could be used to aid research that joined multiple fields of research. For example, relatively basic AI could improve search engines by improving synonym handling, whereas more advanced AI could summarise papers. #NearAI

Baum's [Medium-Term Artificial Intelligence and Society](#) introduces the idea of Medium-Term AI risks. It argues these could be a unifying issue for those worried about near and long term risks. #NearAI

According to Riedel & Deibel, over the 2016-2020 period, GCRI accounted for the second largest number of citations for meta-AI-safety work.

Finances

They spent \$250,000 in 2019 and \$300,000 in 2020, and plan to spend around \$400,000 in 2021. They have around \$600,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1.5 years of runway. However, they tell me that for their core operations runway is close to one year, while the runway for external collaborators is longer.

If you want to donate to GCRI, [here](#) is the relevant web page.

CSER: The Center for the Study of Existential Risk

CSER is a Cambridge based Existential Risk Research organisation founded in 2012 by Jaan Tallinn, Martin Rees and Huw Price, and then established by Seán Ó hÉigeartaigh with the first hire in 2015. They are currently led by Catherine Rhodes and are affiliated with Cambridge University. They cover a wide variety of existential risks, including artificial intelligence, and do political outreach, including to the UK and EU parliaments - e.g. [this](#). Their research can be found [here](#). Their half-yearly review can be found [here](#).

They took on a number of new staff in 2020, most notably John Burden, Jess Whittlestone and Matthijs Maas. Jess joins from Leverhulme where I think she produced some of their best work.

Research

Beard et al.'s [An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards](#) surveys a range of possible techniques for estimating the probability of different existential risks. They then score these on four criteria, and find that no method does well on all. The document contains a number of interesting points, including on the extreme dispersion in some estimates like Supervolcano. It also alludes to the use of 'bad, or even discredited' techniques being used in the existential risk community - this is a case where I wish they had named and shamed! #Forecasting

Belfield's [Activism by the AI Community: Analysing Recent Achievements and Future Prospects](#) reviews the prospects for successful activism by AI employees. It firstly reviews their historical successes, and then uses two different frameworks (as an epistemic community like scientists, and as workers) to analyse the issue, and concludes that AI workers are likely to continue to have significant power to change things through activism. I think this is basically true - my model for grand success runs basically through convincing this epistemic community. One thing the paper does not discuss is the question of getting the AI community to care about the right things though! #Strategy

Belfield et al.'s [Response to the European Commission's consultation on AI](#) recommends the EU pass strict rules about AI. These largely cover more near term issues, and there is no explicit mention of catastrophic risks (that I noticed) but some could be long-run beneficial. The response generally seems written in a way that would appeal to policymakers. I wonder if part of the subtext is making EU AI deployment sufficiently arduous as to slow down AI progress (they deny this!). Researchers from Leverhulme were also named authors on the paper. #Politics

Beard et al.'s [Existential risk assessment: A reply to Baum](#) responds to the GCRI response to their earlier paper. #Forecasting

hÉigearthaigh et al.'s [Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance](#) discusses and advocates for international collaboration on AI safety. The lengthy discussion includes some interesting points about misconceptions and the prospects for common agreements in the presence of very different value systems, but is mainly an imperative piece rather than an analytical one. It focuses on Sino-American cooperation; three of the coauthors are Chinese. Researchers from Leverhulme were also named authors on the paper. #Politics

Beard & Kaczmarek's [On the Wrongness of Human Extinction](#) rebuts an argument that extinction would not be bad because non-existent people cannot be harmed. In particular they argue we wrong such future people by failing to benefit them, even though they have not been harmed. To the extent that responding to such arguments helps motivate people to prevent extinction this is a useful thing to do. (I guess if Extinction was actually good that would be good to know too as we could all stop working so hard!) #Ethics

Avin et al.'s [Exploring AI Futures Through Role Play](#) describes a series of war games the authors ran about future AI development. This definitely a cool idea - I suspect I would enjoy taking part, and their sign-up sheet seems to be still live - and historically these exercises have proved useful in war, like the (in)famous [Millennium Challenge 2002](#). However, I am a bit skeptical of how much insight these particular games have produced - many of the conclusions (e.g. cooperation is important to produce a good outcome) seem both non-novel and also something that was plausibly 'fed into' the structure of the game. I am always a little suspicious of ideas that seem too much like fun! #Forecasting

Tzachor et al.'s [Artificial intelligence in a crisis needs ethics with urgency](#) discusses near-term AI risks related to the pandemic. It mentions things like fairness and privacy, but doesn't really have any specific examples of AI related problems, which aligns with my feeling that our pandemic response would have been better with less restrictions (e.g. our contact tracing could have been better without HIPAA). The intention appears to be to use this to establish an AI regulatory board to oversee novel techniques in the future. Researchers from Leverhulme were also named authors on the paper. #NearAI

Kemp & Rhodes's [The Cartography of Global Catastrophic Risks](#) surveys the sorts of international governance structures for various Xisks. #Politics

Burden & Hernandez-Orallo's [Exploring AI Safety in Degrees: Generality, Capability and Control](#) argues for decomposing the risk of an AI agent into its Capabilities, Generality and our degree of Control. It suggests using Agent Characteristic Curves for this, and includes a toy example. Note that I think the lead author had not technically started at CSER when he wrote the paper. Researchers from Leverhulme were also named authors on the paper. #Capabilities

They also did work on various non-AI issues, which I have not read, but you can find on their website.

CSER researchers contributed to the following research led by other organisations:

- Qian et al.'s [AI GOVERNANCE IN 2019 A YEAR IN REVIEW](#)
- Brundage et al.'s [Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims](#)
- Cihon et al.'s [Should Artificial Intelligence Governance be Centralised? Design Lessons from History](#)
- Cremer & Whittlestone's [Canaries in Technology Mines: Warning Signs of Transformative Progress in AI](#)
- Bostrom et al.'s [Written Evidence to the UK Parliament Science & Technology Committee's Inquiry on A new UK research funding agency](#).
- Hernandez-Orallo et al.'s [AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues](#)

According to Riedel & Deibel, over the 2016-2020 period, CSER accounted for the third largest number of citations for meta-AI-safety work.

Finances

They spent £801,000 in 2018-2019 and £854,000 in 2019-2020, and plan to spend around £1,200,000 in 2020-21. As with many organisations during the pandemic, their 2020 spending is below their expectations (£1,100,000). It seems that similar to GPI maybe 'runway' is not that meaningful - they suggested their grants begin to end in early 2021 and all end by mid-2024, the same dates as last year.

If you want to donate to them, [here](#) is the relevant web page.

OpenAI

OpenAI is a San Francisco based independent AI Research organisation founded in 2015 by Sam Altman. They are one of the leading AGI research shops, with a significant focus on safety. Initially they planned to make all their research open, but changed plans and are now significantly more selective about disclosure - see for example [here](#).

One of the biggest achievements is GPT-3, a massive natural language algorithm that generates highly plausible continuations from prompts, which seems to be very versatile. Scott and Gwern managed to get GPT-2 [to play chess](#), and see also other GPT-3 work by Gwern [here](#), including a to my mind convincing refutation of Gary Marcus's criticisms ([here](#)). The Guardian published an article in which GPT-3 argued that [AGI was not a threat to humanity](#); the article is not very much less convincing than is typical for such arguments.

Research

Christiano's "[Unsupervised" translation as an \(intent\) alignment problem](#)" introduces translation between two languages where no mutual text exists as an analogy for advanced systems. This task seems do-able for a sufficiently advanced AI (I think, though probably some philosophers of language would disagree), but it would be very hard for humans to understand what was going on or to stay 'in-the-loop'. #Transparency

Brown et al.'s [Language Models are Few-Shot Learners](#) paper examines what happens to GPT-3's ability to learn a new task with very few examples when you massively increase the number of parameters. Essentially the idea is that as the number of parameters and number of co-authors gets large enough, it gains something like general purpose intelligence, which then allows it to learn new tasks with very few examples - like a human can. Performance on some of these tasks could even beat specially-trained models. The paper also has a detailed and professional section on potential for misuse in various near AI problems. #GPT-3

Barnes & Christiano's Writeup: [Progress on AI Safety via Debate](#) summarises OpenAI's attempts to design mechanisms to allow non-experts to safely extract information for unaligned experts. It describes various problems they came across, like the deceptive use of ambiguity, or frame control, and their corrections to the mechanism design, like the addition of 'cross-examination'. Cross examination basically forces consistency, and they analyse this to expanding the computational complexity class, but it is not clear how desirable this is - it seems intuitively to me like making something that worked locally with subgames would be ideal. I particularly liked the discussion of their iteration method, rather than just presenting the 'final' product *sui generis*. #Amplification

Brundage et al.'s [Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims](#) describes a variety of ways to promote third-party verifiability of AI systems. This includes coding it into the AI (ideas like interpretability that we often discuss), hardware elements, and institutional reforms, like public bounties for people who find bugs. One of the most noteworthy parts of the document is the wide range of institutions represented in the

author list, including many universities around the world. Researchers from FHI, CSER, Leverhulme, CSET were also named authors on the paper. #Strategy

Stiennon et al.'s [Learning to Summarize with Human Feedback](#) trains a model for writing short text summaries based on human feedback. It first trains a reward model with supervised learning, and then uses that to train an RL agent. They invested in higher-than-usual quality feedback (hourly rate contractors vs Mturkers) and successfully produced summaries of Reddit posts and Daily Mail articles that were on average higher quality than the human written ones (though the latter were hardly Shakespeare). It is basically attempting to produce 'approved by humans' output, instead of just GPT-3 style 'looks like human written' - including testing how hard you can optimise for a proxy before you start getting perverse effects. I also liked the point that the model picked up that the reviewers liked longer summaries (similar to how Reddit likes EffortPosts?). #ValueLearning

Henighan et al.'s [Scaling Laws for Autoregressive Generative Modeling](#) examines how transformer performance scales with compute in various cases. They find generally pretty similar and smooth relationships in multiple domains, implying a lack of (near) upper bound, and suggest that on the margin bigger models are more worth the computational effort than training smaller ones for longer. #Capabilities

OpenAI Researchers also contributed to the following papers lead by other organisations:

- Nguyen & Christiano's [My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda](#)
- Qian et al.'s [AI GOVERNANCE IN 2019 A YEAR IN REVIEW](#)

According to Riedel & Deibel, over the 2016-2020 period, OpenAI accounted for the third largest number of citations in technical AI safety.

Finances

OpenAI was initially funded with money from Elon Musk as a not-for-profit. They have since created an unusual corporate structure including a for-profit entity, in which [Microsoft is investing a billion dollars](#).

Given the strong funding situation at OpenAI, as well as their safety team's position within the larger organisations, I think it would be difficult for individual donations to appreciably support their work. However it could be an excellent place to apply to work.

Google Deepmind

Deepmind is a London based AI Research organisation founded in 2010 by Demis Hassabis, Shane Legg and Mustafa Suleyman and currently lead by Demis Hassabis. They are affiliated with Google. As well as being arguably the most advanced AI research shop in the world, Deepmind has a very sophisticated AI Safety team, covering [both ML safety and AGI safety](#).

I won't cover their non-directly-safety-related work in detail, but one highlight is that this year Deepmind announced they had made significant progress on the [Protein Folding problem](#) with their AlphaFold architecture. While there's still a ways to go yet before we can use it to build arbitrary proteins, this is clearly a big step forward, and shows the generality of their approach. See also discussion [here](#). Long-time followers of the space will recall this is a development Eliezer highlighted [back in 2008](#). See also [this](#) very interesting speculation that Deepmind's team-based private sector approach gave them a significant advantage over academia, and that their speed helped limit knowledge diffusion.

They also produced [this work](#) on one-shot object naming learning in a physical environment - so rather than having to show the agent a huge number of pictures of cows for it to learn

what a cow is, it successfully learns new object names based on a very small number of samples. See also discussion [here](#).

Jan Leike [left Deepmind](#) in June.

Research

Krakovna et al.'s [Specification gaming: the flip side of AI ingenuity](#) is basically an introduction, with many examples, to the problem of AIs producing solutions you did not expect - or want. It discusses both failures of reward shaping as well as AIs manipulating the rewards. #ValueLearning

Gabriel's [Artificial Intelligence, Values and Alignment](#) discusses the alignment problem from various philosophical perspectives. It makes some novel (at least to me) points, like the way that technical AI design may render some ethical systems unobtainable - for example, an optimiser that does not think in terms of 'reasons' is unacceptable to the extent that Kantian deontology is the case. The connection between IRL and virtue ethics was also cute. Overall I thought it was a quite sophisticated treatment of the subject. #Ethics

Krakovna et al.'s [Avoiding Side Effects By Considering Future Tasks](#) proposes a method for reducing side effects. We specify a default policy, and then penalise the agent for restricting our future options relative to that default policy. This helps avoid the risk of e.g. the agent being incentivised to undermine the human's attempts to shut it down. #Corrigibility

Uesato et al.'s [Avoiding Tampering Incentives in Deep RL via Decoupled Approval](#) addresses the problem of agents messing with their value functioning (by e.g. setting utility=IntMax in their params file) by querying a human for reward with regard actions other than those taken. They need to make some assumptions about the structure of the corruption that seem not obvious to me, but it seems like a cool idea. On my reading it doesn't strongly disincentive tampering - it just fails to reward it - which is still an improvement. They back this up with some toy models. #ValueLearning

Researchers from Deepmind were also named on the following papers:

- Gleave et al.'s [QUANTIFYING DIFFERENCES IN REWARD FUNCTIONS](#)
- Reddy et al.'s [Learning Human Objectives by Evaluating Hypothetical Behavior](#)
- Carey et al.'s [The Incentives that Shape Behaviour](#)
- Armstrong et al.'s [Pitfalls of Learning a Reward Function Online](#)
- Cohen & Hutter's [Curiosity Killed the Cat and the Asymptotically Optimal Agent](#)
- Cohen & Hutter's [Pessimism About Unknown Unknowns Inspires Conservatism](#)

Finances

Being part of Google, I think it would be difficult for individual donors to directly support their work. However it could be an excellent place to apply to work.

BERI: The Berkeley Existential Risk Initiative

BERI is a (formerly Berkeley-based) independent Xrisk organisation, founded by Andrew Critch but now led by Sawyer Bernath. They provide support to various university-affiliated (FHI, CSER, CHAI) existential risk groups to facilitate activities (like hiring engineers and assistants) that would be hard within the university context, alongside other activities - see their [FAQ](#) for more details.

As a result of their pivot they are now essentially entirely on providing support to researchers engaged in longtermist (mainly x-risk) work at universities and other institutions. In addition to FHI, CSER and CHAI they added [six new 'trial' collaborations in 2020](#), and intend to do more in 2021. Here are the 2020 cohort:

- The Autonomous Learning Laboratory at UMass Amherst, led by Phil Thomas
- Meir Friedenberg and Joe Halpern at Cornell
- InterACT at UC Berkeley, led by Anca Dragan
- The Stanford Existential Risks Initiative
- Yale Effective Altruism, to support x-risk discussion groups
- Baobao Zhang and Sarah Krepas at Cornell

I think this is potentially a pretty attractive task. University affiliated organisations provide the connection to mainstream academia that we need, but run the risk of inefficiency both due to their lack of independence from the central university and also the relative independence of their academics. BERI potentially offers a way for donors to support the university affiliated ecosystem in a targeted fashion.

They are apparently quite relaxed about getting credit for work, so not all the stuff they support will list them in the acknowledgments.

Finances

They spent \$3,500,000 in 2019 and \$3,120,000 in 2020, and plan to spend around \$2,500,000 in 2021. They have around \$2400000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1 years of runway.

BERI is now seeking support from the general public. If you wanted to donate you can do so [here](#). Note that if you want to you can restrict the funding to their collaborations with FHI, CSER and CHAI if you want.

Ought

Ought is a San Francisco based independent AI Safety Research organisation founded in 2018 by Andreas Stuhlmüller. They research methods of breaking up complex, hard-to-do tasks into simple, easy-to-do tasks - to ultimately allow us effective oversight over AIs. This includes building computer systems, and previously also recruiting test subjects. Their research can be found [here](#). Their annual summary (sort of) can be found [here](#).

In the past they were focused on factored generation – trying to break down questions into context-free chunks so that distributed teams could produce the answer - and factored evaluation, an easier task (by analogy to P<=NP). I thought of them as basically testing Paul Christiano's ideas. They have moved on to trying to automate research and reasoning, by building software to help break complicated questions into subtasks that are simpler to evaluate and potentially automate.

Research

Saunders et al.'s [Evaluating Arguments One Step at a Time](#) provides a detailed analysis of some of Ought's 2019 work on factored evaluation. They tried to break down opinions about movie reviews into discretely checkable sections between a friendly and adversarial agent. The trees they ended up using are quite small - just two layers, plus the root node, presumably because of the problems they had previously encountered with massive tree growth. It's hard to judge the performance numbers they put out, because it's not obvious what sort of performance we would expect from such a circumsized test, even conditional on this being a good approach, but the efficacy they report does not look that encouraging to me. #Amplification

Byun & Stuhlmuller's [Automating reasoning about the future at Ought](#) describes Ought's new program of providing tools to help with people forecasting. This includes assigning probabilities and distributions to beliefs, vaguely similarly to Guestimate. They are now working on building a GPT-3 research assistant. #Amplification

Finances

They spent around \$1,200,000 in 2019 and \$1,200,000 in 2020, and plan to spend around \$1,400,000 in 2020. Their 2020 spend was significantly below plan (around \$2.5m) due to slower hiring and ending human participant experiments. They have around \$3,100,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 2.2 years of runway.

They are not looking for donations from the general public this year.

GPI: The Global Priorities Institute

GPI is an Oxford-based Academic Priorities Research organisation founded in 2018 by Hilary Greaves and part of Oxford University. They do work on philosophical issues likely to be very important for global prioritisation, much of which is, in my opinion, relevant to AI Alignment work. Their research can be found [here](#).

They recently took on two new economics postdocs ([Benjamin Tereick](#) and [Loren Fryxell](#)) and two new philosophy postdocs ([David Thorstad](#) and [Jacob Barret](#))

Research

Trammell & Korinek's [Economic growth under transformative AI](#) applies a variety of models of economic growth to the introduction of AI. These consider both a variety of models and a variety of ways AI could matter - is it a perfect substitute for labour? Do AIs make more AIs? - and summarises the results of this mathematical analysis. I particularly liked the way that discrete qualitative changes in economic regime fell out of the analysis. Overall I thought it did a nice job unifying the two disciplines. #Forecasting

Mogensen's [Moral demands and the far future](#) argues that, contra most people's suppositions, egalitarian utilitarianism requires the present rich not to transfer resources to the present poor but to future generations. It argues this is true under various versions of population ethics. #Ethics

Tarsney & Thomas's [Non-Additive Axiologies in Large Worlds](#) argues that even average-utility type theories should care about the potential for adding many new happy people in the future, because all the past animals provide a large fixed utility background. This fixed utility makes the average behave like the sum, at least locally, so adding a large number of lives that are better off than the average historical rodent is very worthwhile. It's not clear what we should do about aliens. I have always regarded these ideas as something of a reductio of average consequentialism and similar views, but it is nice to have a proof to show that even those who are convinced should care quite a lot (if not quite as much) as totalists about Xrisk. #Ethics

Thorstad & Mogensen's [Heuristics for clueless agents: how to get away with ignoring what matters most in ordinary decision-making](#) addresses the cluelessness problem - that the immense importance and uncertainty of the long run future leaves us clueless as to what do to - through the use of local heuristics. #DecisionTheory

Tarsney's [Exceeding Expectations: Stochastic Dominance as a General Decision Theory](#) suggests we can avoid some of the paradoxes of expected utility maximisation (e.g. St Petersburg Paradox) by using Stochastic Dominance. This basically comes down to arguing that we can make use of background assumptions to push the dominance condition to give us virtually all of the benefits of expectation maximisation, while avoiding the Pascalian type problems - and of course stochastic dominance is a *prima facie* attractive principle in itself. #DecisionTheory

Mogensen & Thorstad's [Tough enough? Robust satisficing as a decision norm for long-term policy analysis](#) advocates for 'robust satisficing', as an alternative to expectation maximisation, as a decision criteria in cases where there is 'deep' uncertainty. The aim is basically to give a firmer theoretical underpinning for engineers to use this relatively conservative approach in risky situations. #Strategy

John & MacAskill's [Longtermist institutional reform](#) described a number of potential governance changes we could make to try to represent the interests of future people better. These include impact assessments, people's assemblies and separate legislative houses. I think this is a good project to work on, but I'm sceptical of these specific proposals; they seem a bit like a list of 'policies that sound nice' to me, without really considering all the problems - for example, our current use of environmental impact assessments seems to have had very negative consequences for our ability to build any new infrastructure, and I think there are good reasons sortition has rarely been used in practice. See also discussion [here](#). #Politics

Finances

They spent £600,000 in 2018/2019 (academic year) and £850,000 in 2019/20, which was less than their plan of £1,400,000 due to the pandemic, and intend to spend around £1,400,000 in 2020/2021. They suggested that as part of Oxford University 'cash on hand' or 'runway' were not really meaningful concepts for them, as they need to fully-fund all employees for multiple years.

If you want to donate to GPI, you can do so [here](#).

CLR: The Center on Long Term Risk

CLR is a London (previously Germany) based Existential Risk Research organisation founded in 2013 and until recently lead by Jonas Vollmer (who has now moved to EA Funds). Until this year they were known as FRI (Foundational Research Institute) and were part of the Effective Altruism Foundation (EAF). They do research on a number of fundamental long-term issues, with AI as one of their top two focus areas (along with Malevolence, though that is still related). You can see their recent research summarised [here](#).

In general they adopt what they refer to as 'suffering-focused' ethics, which I think is a quite misguided view, albeit one they seem to approach thoughtfully.

They recently hired Alex Lychov, Emery Cooper, Daniel Kokotajlo (from AI Impacts, possibly not permanent), and Julian Stastny as full-time research staff, Maxime Riché as a research engineer and Jia Yuan Loke as part-time.

Research

Althaus & Baumann's [Reducing long-term risks from malevolent actors](#) analyses the dangers posed by very evil (score highly on the 'dark triad' traits) people, and suggests some possible techniques to reduce the risk. This detailed report, on an area I hadn't seen much before, includes the context of whole brain emulation, AGI, etc. #Politics

Clifton's [Equilibrium and prior selection problems in multipolar deployment](#) describes the problem of ensuring desirable equilibria between multiple agents when they have different priors. The idea that different equilibria could be possible etc. is well known, but the contribution here is to point out that different priors between teams / agents could push you into a very bad equilibrium - for example, if your Saxons falsely believe the Vikings are bluffing. #AgentFoundations

Clifton & Riche's [Towards Cooperation in Learning Games](#) discusses the meta-game-theoretic problem of how to get AI teams to cooperate on the task of building AIs that will cooperate

with each other. They introduce the idea of Learning TFT and run some experiments around its performance. #AgentFoundations

Finances

They spent around \$1,400,000 in 2019, around \$1,100,000 in 2020, and plan to spend around \$1,800,000 in 2021. They have around \$950,000 in reserves, suggesting (on a very naïve calculation) around 0.6 years of runway. Their 2019 spending was somewhat somewhat higher than they expected a year ago, based on FX changes and some unexpected items, especially related to travel and their move to the UK.

They have a collaboration with the Swiss-based [Center for Emerging Risk Research](#), who have agreed to fund 15% of their costs.

If you wanted to donate to CLR, you could do so [here](#).

CSET: The Center for Security and Emerging Technology

CSET is a Washington based Think Tank founded in 2019 by Jason Matheny (ex IARPA), affiliated with the University of Georgetown. They analyse new technologies for their security implications and provide advice to the US government. At the moment they are mainly focused on near-term AI issues. Their research can be found [here](#).

Research

Hwang's [Shaping the Terrain of AI Competition](#) discusses strategies for the US to compete with China in AI. In particular, these attempt to nullify the 'natural' advantages authoritarian or totalitarian states may have. #Politics

Imbrie et al.'s [The Question of Comparative Advantage in Artificial Intelligence: Enduring Strengths and Emerging Challenges for the United States](#) discusses the relative advantages of the US and China in AI development. #Politics

Finances

As they apparently launched with [\\$55m from the Open Philanthropy Project](#), and subsequently raised money from the [Hewlett Foundation](#), I am assuming they do not need more donations at this time.

AI Impacts

AI Impacts is a San Francisco (previously Berkeley) based AI Strategy organisation founded in 2014 by Katja Grace and Paul Christiano. They are affiliated with (a project of, with independent financing from) MIRI. They do various pieces of strategic background work, especially on AI Timelines, AI takeoff speed etc. - it seems their previous work on the relative rarity of discontinuous progress has been relatively influential. Their research can be found [here](#).

During the year Kokotajlo left (temporarily?) for CLR, and Asya may be leaving for FHI.

edit 2020-12-25: They have now published an annual review [here](#).

Research

A lot of the work on the website is essentially in the form of a continuously updated private wiki - see [here](#). This makes it a little difficult for our typical technique, which relies on being able to evaluate specific publications which are released at specific times. As such it is a little unfortunate that in the below we generally concentrate on their timestamped blogposts. They suggested readers might be interested in [posts like these ones](#).

They have produced a series of pieces on how long it has historically taken for AIs to cover the human range (from beginner to expert to superhuman) for different tasks. This seems relevant because people only seem to really pay attention to AI progress in a field when it starts beating humans. These pieces include [Starcraft](#), [ImageNet](#), [Go](#), [Chess](#) and [Draughts](#).

Grace's [Discontinuous progress in history: an update](#) details their extensive research into examples of discontinuities in technological progress. They find 10 such examples, across construction, travel, weapons and compute. As well as being a very pleasant read, they had some interesting conclusions, for example that the discontinuities often occurred in non-optimised secondary features, and many occurred when something became just good enough to pass a threshold on another feature. Especially interesting to me is some of the things they found to not be discontinuities: AlexNet and Chess AI. Could this mean that future progress could 'feel' discontinuous in some important sense even if it doesn't register as such on some objective benchmark? The individual trend writeups (e.g. penicillin [here](#)) are also interesting. See also [here](#). #Forecasting

Kokotajlo's [Three kinds of competitiveness](#) distinguishes between AI systems that will outperform, those that will be cheaper, and those that will arrive sooner. This is a very simple dichotomy that actually helped make things clearer; the post contains just enough to make the point and significance clear. #Strategy

Korzekwa's [Description vs simulated prediction](#) describes the difference between modelling how steady technological progress was in the past, and thinking about how predictable it was in the past. For example, the speedup that aeroplanes offered for transatlantic travel (relative to ships) was presumably quite predictable to someone who knew about progress in aeronautics, even though it was very sudden. #Forecasting

Kokotajlo's [Relevant pre-AGI possibilities](#) is a scenario simulator for different future developments. Basically you enter probabilities for a bunch of relevant things that could happen and it randomly generates a future. By clicking repeatedly, you can get a representative sense for the sort of futures your beliefs entail. #Forecasting

Korzekwa's [Preliminary survey of prescient actions](#) attempts to find historical cases where humans have taken advance action to solve an unprecedented problem. It does not find any examples better than the classic Szilard case. This could be good news - that, in practice, there is always feedback, so the problem is not as easy as we thought - or it could be bad news - we have to solve a type of problem we have literally never solved before (or not very much news, to the extent it is only preliminary). #Forecasting

Grace's [Atari early](#) notes that AI mastery of Atari games seems to have arrived significantly earlier than experts previously expected. #Forecasting

Finances

They spent \$315,000 in 2019 and \$300,000 in 2020, and plan to spend around \$200,000 in 2021. They have around \$190,000 in cash and pledged funding, suggesting (on a very naïve calculation) around 0.95 years of runway.

In the past they have received support from EA organisations like OpenPhil and FHI.

MIRI administers their finances on their behalf; donations can be made [here](#).

Leverhulme Center for the Future of Intelligence

Leverhulme is a Cambridge based Research organisation founded in 2015 and currently led by Stephen Cave. They are affiliated with Cambridge University and closely linked to CSER. They do work on a variety of AI related causes, mainly on near-term issues but also some long-term. You can find their publications [here](#). They have a document listing some of their achievements [here](#).

Research

Leverhulme-affiliated researchers produced work on a variety of topics; I have only here summarised that which seemed the most relevant to AI safety.

Hernandez-Orallo et al.'s [AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues](#) performs algorithmic analysis of AI papers to determine trends. One interesting thing they pick up on (perhaps obvious in retrospect) is that (generally near-term) 'safety' related papers peak within any given paradigm after the paradigm itself. Researchers from CSER were also named authors on the paper. #Strategy

Whittlestone & Ovadya's [The tension between openness and prudence in responsible AI research](#) discusses the conflict between traditional CS openness norms and the new ones we are trying to create. They decompose this conflict in various ways. The focus of the paper is on near-term issues, but the principle clearly matters for the big issue. Researchers from Leverhulme were also named authors on the paper. #Strategy

Crosby et al.'s [The Animal-AI Testbed and Competition](#) produces a series of tests for AI ability based on animal IQ tests. This is an alternative to traditional tests like Atari, with the appeal being their practical relevance and reduced overfitting (as some of the tests are not in the training data). Presumably the benefit here is to improve out-of-distribution performance. #Misc

Zerilli et al.'s [Algorithmic Decision-Making and the Control Problem](#) discusses the problem of humans growing complacent and overly deferential towards AI systems they are meant to be monitoring. If the system is 'always right', eventually you are just going to click 'confirm' without thinking. #NearAI

Peters et al.'s [Responsible AI—Two Frameworks for Ethical Design Practice](#) discusses some ethical principles for engineers #NearAI

Hollanek's [AI transparency: a matter of reconciling design with critique](#) attempts to apply literary criticism to AI transparency. #NearAI

Bhatt et al.'s [Machine Learning Explainability for External Stakeholders](#) gathered focus groups to discuss how to make AI transparent to outsiders (not just designers) #NearAI

Cave & Dihal's [The Whiteness of AI](#) worries that too many AIs are depicted as being coloured white. It seems to me it would be roughly equally (im)plausible to say it would be problematic if robots (from the slavic word for forced labour) were black. #NearAI

Leverhulme researchers contributed to the following research led by other organisations:

- Brundage et al.'s [Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims](#)
- Belfield et al.'s [Response to the European Commission's consultation on AI](#)
- hÉigearthaigh et al.'s [Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance](#)
- Tzachor et al.'s [Artificial intelligence in a crisis needs ethics with urgency](#)

- Burden & Hernandez-Orallo's [Exploring AI Safety in Degrees: Generality, Capability and Control](#)
- Prunkl & Whittlestone's [Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society](#)

According to Riedel & Deibel, over the 2016-2020 period, Leverhulme accounted for the third largest number of citations for meta-AI-safety work.

AI Safety camp

AISC is an internationally based independent residential research camp organisation founded in 2018 by Linda Linsefors and currently led by Remmelt Ellen. They bring together people who want to start doing technical AI research, hosting a 10-day camp aiming to produce publishable research. Their research can be found [here](#).

To the extent they can provide an on-ramp to get more technically proficient researchers into the field I think this is potentially very valuable. But I haven't personally experienced the camps, or even spoken to anyone who has.

Research

Makievskyi et al.'s [Assessing Generalization in Reward Learning with Procedurally Generated Games](#) try to train RL algorithms on various games to generalise to new environments. They generally found this was difficult. #RL

Finances

They spent \$23,085 in 2019 and \$11,162 in 2020, and plan to spend around \$53,000 in 2021. They have around \$28,851 in cash and pledged funding, suggesting (on a very naïve calculation) around 0.5 years of runway. They are run by volunteers, and are considering professionalising, depending on the amount of donations they receive.

If you want to donate, the web page is [here](#).

FLI: The Future of Life Institute

FLI is a Boston-based independent existential risk organization, focusing on outreach, founded in large part to help organise the regranting of \$10m from Elon Musk. One of their major projects is trying to ban [Lethal Autonomous Weapons](#).

They wrote a letter to the EU advising for stricter regulation, with 120 signitures, [here](#).

They have a very good podcast on AI Alignment [here](#).

Research

Aguirre's [Why those who care about catastrophic and existential risk should care about autonomous weapons](#) argues that we should work towards a ban on Lethal Autonomous Weapons. This is not only because they might be destabilising WMDs, but also as a 'practice run' for future regulation of AI. #NearAI

Convergence

Convergence is a globally based independent Existential Risk Research organisation, of which Justin Shovelain founded an earlier version in 2015 and David Kristoffersson joined as

cofounder in 2018. They do strategic research about Xrisks in general as well as some AI specific work. Their research can be found [here](#). Their short summary can be found [here](#).

Justin Shovelain and David Kristoffersson are the two full-time members of Convergence, but they have had other people on part-time for periods of time, such as Michael Aird in the first half of 2020, and Alexandra Johnson.

Research

Shovelain & Aird's [Using vector fields to visualise preferences and make them consistent](#) discusses the idea of using vector fields as a representation of local preferences, and then using curl as a measure of their consistency. I liked this as a clear and less blackboxy-than-ML account of how preferences were being represented. It would be good to see some more on whether the helmholtz theorem gives us the sorts of properties we want in addition to removing the curl. #ValueLearning

Aird's [Existential risks are not just about humanity](#) argues that, despite its being technically excluded from the definition, we should take into account the possibility of positive-value alien-originating life when we consider existential risks. #Strategy

Aird et al.'s [Memetic downside risks: How ideas can evolve and cause harm](#) discusses the risk of ideas becoming distorted over time in the retelling. This includes predictions about the average direction in which memes will evolve: for example, towards simplicity. (They suggested [this](#) might be a more important article on a similar subject but I haven't had time to read) #Strategy

They suggested readers might also be interested in [this](#), [this](#) and [this](#).

Finances

They spent \$50,000 in 2019 and \$13,000 in 2020, and plan to spend around \$30,000 in 2021. They have around \$37000 in cash and pledged funding, suggesting (on a very naïve calculation) around 1.2 years of runway.

Though they are not actively seeking donations at the moment, if you wanted to donate you could do so [here](#).

Median Group

Median is a Berkeley based independent AI Strategy organisation founded in 2018 by Jessica Taylor, Bryce Hidysmith, Jack Gallagher, Ben Hoffman, Colleen McKenzie, and Baeo Maltinsky. They do research on various risks, including AI timelines. Their research can be found [here](#).

Their website does not list any relevant research for 2020.

They did not reply when I asked them about their finances. Median doesn't seem to be soliciting donations from the general public at this time.

AI Pulse

The Program on Understanding Law, Science, and Evidence ([PULSE](#)) is part of the UCLA School of Law, and contains a group working on AI policy. They were founded in 2017 with a [\\$1.5m grant from OpenPhil](#).

Their website does not list any research for 2020 that seemed relevant to existential safety.

Other Research

I would like to emphasize that there is a lot of research I didn't have time to review, especially in this section, as I focused on reading organisation-donation-relevant pieces. So please do not consider it an insult that your work was overlooked!

Benadè et al.'s [Preference Elicitation for Participatory Budgeting](#) works on how to get people to share their preferences, and then combine this information. In particular they separate the preference-inferring step from the aggregation step, exploring multiple input and aggregation methodologies. Some of this paper was from 2016 but I missed it then and figured enough was new to warrant a mention here. #ValueLearning

Qian et al.'s [AI GOVERNANCE IN 2019 A YEAR IN REVIEW](#) is a collected volume of articles on governance from over 50 different authors. Both China and the West are well represented. (I have not read all the individual articles) Researchers from OpenAI, CHAI, CSER were also named authors on the paper. #Politics

Krakovna's [Possible takeaways from the coronavirus pandemic for slow AI takeoff](#) discusses the significance of our covid performance for AGI strategy. It discusses the ways in which, even though the pandemic was quite slow moving and clearly predictably disastrous, western governments failed to act, suggesting there might be similar failures in a slow AGI takeoff. I also recommend Wei's comment, which points out that the disaster easily became politicised - it is truly impressive (-ly dire) that in the US the partisan positions in the US managed to flip three times without ever producing an effective response. Indeed it seems plausible to me that on net government intervention [made the pandemic worse](#). (The author works for FLI and Deepmind but this seems to be a separate 'personal' article). See also the discussion [here](#). #Strategy

Ngo's [AGI Safety from First Principles](#) presents Richard's account of the case for AI risk. This is basically the idea that, by creating AGI, humankind might end up as only the world's second most powerful species. I think most readers will probably (unsurprisingly) agree with him here; it seems like a very good account of the core argument, which is nice to have newer versions of. #Overview

Ecoffet & Adrien's [Reinforcement Learning Under Moral Uncertainty](#) is the first paper I've seen trying to implement and test different approaches to moral uncertainty in an RL setting. Obviously harkening to Will's thesis, though they restrict to theories with cardinal utilities only - which seems, to my mind, to assume away the hardest part. They compare expectation maximisation to voting systems, and test on trolley problems. #RL

Hendrycks et al.'s [Aligning AI with Shared Human Values](#) showcases a data set of moral examples (e.g. property damage is wrong) and trains various transformer text algorithms on it. I like the way they use deliberately uncontroversial examples; I think we will do much better if we can get agents who get 99% of situations correct that by re-litigating the culture war by proxy. As a first pass we should consider their results as a sort of benchmark for future work using the database. Researchers from CHAI were also named authors on the paper. #

Benaich & Hogarth's [State of AI Report 2020](#) is an overview of the AI industry in 2020 by two investors. It is very detailed, but not that directly relevant. #Overview

Wilkinson's [In defence of fanaticism](#) offers the first defence of EV maximisation fanaticism that I have ever seen. It includes both counterarguments against the common rejections (which lets face it often resemble David Lewis's incredulous stare), as well as two nice dilemmas for the non-fanaticism. See also the discussion [here](#). #DecisionTheory

Linsefors & Hepburn's [Announcing AI Safety Support](#) describes a group they have created to try to support people entering the field. #Strategy

Aird's [Failures in technology forecasting? A reply to Ord and Yudkowsky](#) discusses the examples that Eliezer and Toby use as evidence for the difficulty in predicting technological development, and argues that it is not so clear that these really show this exactly. For example, the quote about Wilbur Wright doubting the possibility of flight looks more like a moment of depression than a forecast that would have been taken seriously by contemporaries. Overall I thought his "these examples seem somewhat cherry-picked" argument was the most convincing. #Forecasting

Scholl & Hanson's [Testing the Automation Revolution Hypothesis](#) evaluate predictions of AI-driven unemployment. They find that these predictions have had low but positive explanatory value for predicting which jobs would be automated so far. Researchers from FHI were also named authors on the paper. #NearAI

Xu et al.'s [Recipes for Safety in Open-domain Chatbots](#) discusses various ways of preventing a chatbot from saying offensive things. #ValueLearning

Capital Allocators

One of my goals with this document is to help donors make an informed choice between the different organisations. However, it is quite possible that you regard this as too difficult, and wish instead to donate to someone else who will allocate on your behalf. This is of course much easier; now instead of having to solve the *Organisation Evaluation Problem*, all you need to do is solve the dramatically simpler *Organisation Evaluator Organisation Evaluation Problem*.

LTFF: Long-term future fund

LTFF is a globally based EA grantmaking organisation founded in 2017, currently lead by Matt Wage and affiliated with CEA, but probably becoming independent (along with the other EA funds under Jonas Vollmer) in 2021. They are one of four funds set up by CEA to allow individual donors to benefit from specialised capital allocators; this one focuses on long-term future issues, including a large focus on AI Alignment. Their website is [here](#). There are write-ups for their three grant rounds in 2020 are [here](#), [here](#) and [here](#), and comments [here](#), [here](#) and [here](#). As the [November 2019](#) round was not public when I wrote last year I have included it in some of the analysis below. They also did a AMA recently [here](#).

The fund is now run by five people, and the grants have gone to a wide variety of causes, many of which would simply not be accessible to individual donors.

The fund managers are currently:

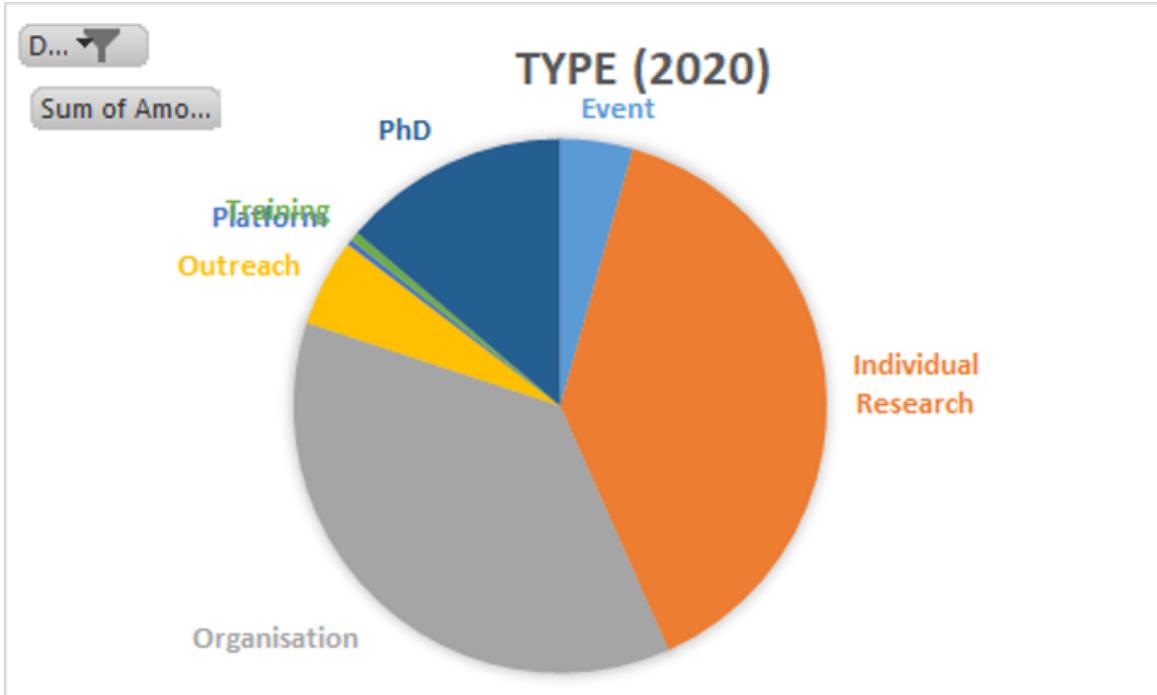
- Matt Wage
- Helen Toner
- Oliver Habryka
- Adam Gleave
- Asya Bergal

Asya and Adam are new, replacing Alex Zhu. My personal interactions with the two of them are supportive of the idea they will make good grants. I was sad to see that Oliver plans to step back from some aspects of the fund as he felt that the [marginal value of opportunities was diminished](#). All the managers have, up until now, been unpaid, but I understand this may change in 2021. Additionally, the grant managers will have to be re-appointed for their positions in 2021, so there may be some turnover.

In total for 2020 they granted around \$1.5m. In general most of the grants seem at least plausibly valuable to me, and many seemed quite good indeed. There weren't any in 2020

that seemed totally egregious. As there is a fair bit of discussion in the links, and no one grant dominated the rounds, I shan't discuss my opinions of individual grants in detail.

I attempted to classify the recommended by type. Note that 'training' means paying an individual to self-study. I have deliberately omitted the exact percentages because this is an informal classification.



Of these categories, I am most excited by the Individual Research, Event and Platform projects. I am generally somewhat sceptical of paying people to 'level up' their skills.

In their [September write-up](#) they mentioned a desire to "continue to focus on grants to small projects and individuals rather than large organisations." Despite this, it appears to me that the amount of grants to large organisations actually increased in 2020 vs 2019, which is a bit disappointing. I can understand why the fund managers gave over a third of the funds to major organisations – they thought these organisations were a good use of capital! And some of these organisations are, to be fair, small rather than large. However, to my mind this undermines the purpose of the fund. (Many) individual donors are perfectly capable of evaluating large organisations that publicly advertise for donations. In donating to the LTFF, I think (many) donors are hoping to be funding smaller projects that they could not directly access themselves. As it is, such donors will probably have to consider such organisation allocations a mild 'tax' – to the extent that different large organisations are chosen then they would have picked themselves.

The fund donates a relatively large percentage to AI related activities; I estimate around 2/3. Many of the other grants, focused on other long-term issues, also seemed sensible to me. The only one I would question was subsidising a therapist to move to the bay area, which seems like a better fit for the [Meta/Infrastructure Fund](#) if nothing else.

Richard Ngo's PhD, which the fund managers recommended \$150,000, was the largest single grant (just over 10% of the 2020 total), followed by MIRI, 80k and Vanessa Kosoy with \$100,000 each.

All grants have to be approved by CEA before they are made; to my knowledge they approved all recommended grants in 2020.

One significant development in 2020 was their decision to make an anonymous grant (roughly 3% of total) to a PhD student. Based on their description of the purpose of the grant, the lack of reported conflicts and the use of an additional outside reviewer, I feel pretty confident that this specific grant was a decent one. I'm not aware of anyone with a 'strong track record in technical AI safety' for whom it would be a severe mistake for the LTFF to support. And I definitely understand a desire for privacy, especially when begging for money from weird people for a weird purpose - or so it could seem to outsiders. However by doing so they undermine the ability of the donor community to provide oversight, which is definitely a bit concerning to me. This would be especially true in the absence of the other details about the grant they provided.

If you wish to donate to the LTFF you can do so [here](#).

OpenPhil: The Open Philanthropy Project

The Open Philanthropy Project (separated from Givewell in 2017) is an organisation dedicated to advising Cari and Dustin Moskovitz on how to give away over \$15bn to a variety of causes, including existential risk. They have made extensive donations in this area and probably represent both the largest pool of EA-aligned capital and the largest team of EA capital allocators.

They recently described their strategy for AI governance, at a very high level, [here](#).

It is possible that the partnership with Ben Delo we discussed last year [may not occur](#).

Grants

You can see their grants for AI Risk [here](#). It lists 21 AI Risk grants in 2020, plus 4 others for global catastrophic risks and several highly relevant 'other' grants. In total I estimate they spent about \$19m on AI in 2020.

The largest grants were:

- [MIRI](#): \$7.7m
- [OpenPhil AI Fellows](#): \$2.3m
- [80k](#): \$3.4m
- [Ought](#): \$1.6m
- [UC Berkeley](#) (Jacob Steinhardt): \$1.1m
- [FHI](#): \$1.6m

In contrast were only 4 AI Risk grants listed for 2019, though one of these (CSET) was for \$55m.

The OpenPhil AI Fellowship basically fully funds AI PhDs for students who want to work on the long term impacts of AI. Looking back at the 2018 class (who presumably will have had enough time to do significant work since receiving the grants), scanning the abstracts of their publications on their websites suggests that over half have no AI safety relevant publications in 2019 or 2020, and only one is a coauthor on what I would consider a highly relevant paper. Apparently it is somewhat intentional that these fellowships are [not intended to be specific to AI safety](#), though I do not really understand what they *are* intended for. OpenPhil suggested that part of the purpose was to [build a community](#).

They are also launching a [new scholarship program](#) which seems more tailored to people focused on the long-term future, though it is not AI specific.

They produced a list of [recommended donation opportunities for small donors](#); there were zero AI or existential risk opportunities.

Research

Most of their research concerns their own granting, and is often non-public.

[Cotra's Report on AI Timelines](#) is a supremely detailed, yet still draft (!), report on how long we should expect the timeline to AGI to be. Impossible for me to do it justice, but essentially it attempts to model both the amount of computational power required to achieve transformative AGI (with current algorithms, the main focus), how much algorithms are improving, and how long it will take to accumulate this hardware. The report estimates doubling times of roughly 2-3 years for both compute and algorithm design. Interestingly, it also suggests that the costs of the final training run will fall as a fraction of overall costs. I liked the way it considers multiple different outside view 'anchors' for different perspectives on the problem - e.g. how much computing did evolution do to produce humans?

#Forecasting

Carlsmith's [How Much Computational Power Does It Take to Match the Human Brain?](#) attempts to model the FLOPs of the human brain. This is part of their forecasting of when AI will develop to human level capacity (combined with Cotra's report). He does this using multiple methods, which produce generally relatively similar results - as in, not too many orders of magnitude different, generally centered around 10^{15} ish. #Forecasting

Finances

To my knowledge they are not currently soliciting donations from the general public, as they have a lot of money from Dustin and Cari, so incremental funding is less of a priority than for other organisations. They could be a good place to work however.

SFF: The Survival and Flourishing Fund

SFF ([website](#)) is a donor advised fund, advised by the people who make up BERI's Board of Directors. SFF was initially funded in 2019 by a grant of approximately \$2 million from BERI, which in turn was funded by donations from philanthropist Jaan Tallinn, now also distributing money from Jed McCaleb.

Grants

In its grantmaking SFF uses an innovative allocation process to combine the views of many grant evaluators (described [here](#)). SFF has published the results of one grantmaking round this year (described [here](#)), where they donated around \$1.8m, of which I estimate around \$1.2m was AI related; the largest donations in the round were to:

- LessWrong: \$400k
- MIRI: \$340k
- The Future Society: \$160k
- 200k: \$180k
- [Quantified Uncertainty Research Institute](#): \$120k

I would expect the H2 round, whose results are not yet public, to be at least as large.

Other Organisations

80,000 Hours

80k provides career advice and guidance to people interested in improving the world, with a specific focus on AI safety.

80,000 Hours's [AI/ML safety research job board](#) collects various jobs that could be valuable for people interested in AI safety. At the time of writing it listed 80 positions, all of which seemed like good options that it would be valuable to have sensible people fill. I suspect most people looking for AI jobs would find some on here they hadn't heard of otherwise, though of course for any given person many will not be appropriate. They also have job boards for other EA causes. #Careers

They also run a very good podcast; readers might be specifically interested in [this](#) or [this](#).

Other News

NeuroIPS [rejected four papers this year](#) for being 'unethical'.

Waymo is (finally) offering a [true driverless Uber experience](#) to the general public in Phoenix.

The pope suggested we [pray for AI alignment](#).

There was a minor pandemic.

Methodological Thoughts

Inside View vs Outside View

This document is written mainly, but not exclusively, using publicly available information. In the tradition of active management, I hope to synthesise many pieces of individually well known facts into a whole which provides new and useful insight to readers. Advantages of this are that 1) it is relatively unbiased, compared to inside information which invariably favours those you are close to socially and 2) most of it is [legible](#) and verifiable to readers. The disadvantage is that there are probably many pertinent facts that I am not a party to! Wei Dai has written about how [much discussion now takes place in private google documents](#) – for example [this Drexler piece](#) apparently; in most cases I do not have access to these. If you want the inside scoop I am not your guy; all I can supply is exterior scooping.

We focus on papers, rather than outreach or other activities. This is partly because they are much easier to measure; while there has been a large increase in interest in AI safety over the last year, it's hard to work out who to credit for this, and partly because I think progress has to come by persuading AI researchers, which I think comes through technical outreach and publishing good work, not popular/political work.

Organisations vs Individuals

Many capital allocators in the bay area seem to operate under a sort of [Great Man](#) theory of investment, whereby the most important thing is to identify a guy to invest in who is really clever and 'gets it'. I think there is a lot of merit in this (as argued [here](#) for example); however, I think I believe in it less than they do. Perhaps as a result of my institutional investment background, I place a lot more weight on historical results. In particular, I worry that this approach leads to over-funding skilled rhetoricians and those the investor/donor is socially connected to. Also, as a practical matter, it is hard for individual donors to fund individual researchers. But as part of a concession to the individual-first view I've started asking organisations if anyone significant has joined or left recently, though in practice I think organisations are far more willing to highlight new people joining than old people leaving.

Judging organisations on their historical output is naturally going to favour more mature organisations. A new startup, whose value all lies in the future, will be disadvantaged. However, I think that this is the correct approach for donors who are not tightly connected to the organisations in question. The newer the organisation, the more funding should come from people with close knowledge. As organisations mature, and have more easily verifiable signals of quality, their funding sources can transition to larger pools of less expert money. This is how it works for startups turning into public companies and I think the same model applies here. (I actually think that even those with close personal knowledge should use historical results more, to help overcome their biases.)

This judgement involves analysing a large number of papers relating to Xrisk that were produced during 2020. Hopefully the year-to-year volatility of output is sufficiently low that this is a reasonable metric; I have tried to indicate cases where this doesn't apply. I also attempted to include papers during December 2019, to take into account the fact that I'm missing the last month's worth of output from 2020, but I can't be sure I did this successfully.

Research Inclusion Criteria

In general I have tried to evaluate and summarise, at least briefly, the work organisations did that is primarily concerned with AI or general Xrisk strategy. But this has been a rather subjective and imperfectly applied criteria that was primarily implemented through my subjective sense of 'does this seem relevant to the task at hand'.

Politics

My impression is that policy on most subjects, especially those that are more technical than emotional is generally made by the government and civil servants in consultation with, and being lobbied by, outside experts and interests. Without expert (e.g. top ML researchers in academia and industry) consensus, no useful policy will be enacted. Pushing directly for policy seems if anything likely to hinder expert consensus. Attempts to directly influence the government to regulate AI research seem very adversarial, and risk being pattern-matched to ignorant technophobic opposition to GM foods or other kinds of progress. We don't want the 'us-vs-them' situation that has occurred with climate change, to happen here. AI researchers who are dismissive of safety law, regarding it as an imposition and encumbrance to be endured or evaded, will probably be harder to convince of the need to voluntarily be extra-safe - especially as the regulations may actually be totally ineffective.

The only case I can think of where scientists are relatively happy about punitive safety regulations, nuclear power, is one where many of those initially concerned were scientists themselves, and also had the effect of basically ending any progress in nuclear power (at great cost to climate change). Given this, I actually think policy outreach to the general population is probably negative in expectation.

If you're interested in this, I'd recommend you read [this blog post](#) from a few years back.

Openness

I think there is a strong case to be made that openness in AGI capacity development is bad. As such I do not ascribe any positive value to programs to 'democratize AI' or similar.

One interesting question is how to evaluate non-public research. For a lot of safety research, openness is clearly the best strategy. But what about safety research that has, or potentially has, capabilities implications, or other infohazards? In this case it seems best if the researchers do not publish it. However, this leaves funders in a tough position – how can we judge researchers if we cannot read their work? Maybe instead of doing top secret valuable

research they are just slacking off. If we donate to people who say “trust me, it’s very important and has to be secret” we risk being taken advantage of by charlatans; but if we refuse to fund, we incentivize people to reveal possible infohazards for the sake of money. (Is it even a good idea to publicise that someone else is doing secret research?)

For similar reasons I prefer research to not be behind paywalls or inside expensive books, but this seems a significantly less important issue.

More prosaically, organisations should make sure to upload the research they have published to their website! Having gone to all the trouble of doing useful research it is a constant shock to me how many organisations don’t take this simple step to significantly increase the reach of their work. Additionally, several times I have come across incorrect information on organisation’s websites.

Research Flywheel

My basic model for AI safety success is this:

1. Identify interesting problems
 1. As a byproduct this draws new people into the field through altruism, nerd-sniping, apparent tractability
2. Solve interesting problems
 1. As a byproduct this draws new people into the field through credibility and prestige
3. Repeat

One advantage of this model is that it produces both object-level work and field growth.

There is also some value in arguing for the importance of the field (e.g. Bostrom’s Superintelligence) or addressing criticisms of the field.

Noticeably absent are strategic pieces. I find that a lot of these pieces do not add terribly much incremental value. Additionally, my suspicion is that strategy research is, to a certain extent, produced exogenously by people who are interested / technically involved in the field. This does not apply to technical strategy pieces, about e.g. whether CIRL or Amplification is a more promising approach.

There is somewhat of a paradox with technical vs ‘wordy’ pieces however: as a non-expert, it is much easier for me to understand and evaluate the latter, even though I think the former are much more valuable.

Differential AI progress

There are many problems that need to be solved before we have safe general AI, one of which is not producing *unsafe* general AI in the meantime. If nobody was doing non-safety-conscious research there would be little risk or haste to AGI – though we would be missing out on the potential benefits of safe AI.

There are several consequences of this:

- To the extent that safety research also enhances capabilities, it is less valuable.
- To the extent that capabilities research re-orientates subsequent research by third parties into more safety-tractable areas it is more valuable.
- To the extent that safety results would naturally be produced as a by-product of capabilities research (e.g. autonomous vehicles) it is less attractive to finance.

One approach is to research things that will make contemporary ML systems safer, because you think AGI will be a natural outgrowth from contemporary ML. This has the advantage of faster feedback loops, but is also more replaceable (as per the previous section).

Another approach is to try to reason directly about the sorts of issues that will arise with superintelligent AI. This work is less likely to be produced exogenously by unaligned researchers, but it requires much more faith in theoretical arguments, unmoored from empirical verification.

Near-term safety AI issues

Many people want to connect AI existential risk issues to ‘near-term’ issues; I am generally sceptical of this. For example, autonomous cars seem to risk only localised tragedies (though if they were hacked and all crashed simultaneously that would be much worse), and private companies should have good incentives here. Unemployment concerns seem exaggerated to me, as they have been for most of history (new jobs will be created), at least until we have AGI, at which point we have bigger concerns. Similarly, I generally think concerns about algorithmic bias are essentially political - I recommend [this presentation](#) - though there is at least some connection to the value learning problem there.

Some people argue that work on these near AI issues is worthwhile because it can introduce people to the broader risks around poor AI alignment. However, I think this is a bad idea - not only does it seem somewhat disingenuous, it risks putting off people who recognise that these are bad concerns. For example, [this paper](#) rejects the precautionary principle for AI on the basis of rejecting bad arguments about unemployment - had these pseudo-strawman views not been widespread, it would have been harder to reach this unfortunate conclusion.

It’s also the case many of the policies people recommend as a result of these worries are potentially very harmful. A good example is GDPR and similar privacy regulations (including HIPAA) which have made many good things much more difficult - including degrading our ability to track the pandemic.

Some interesting speculation I read is the idea that discussing near AI safety issues might be a sort of immune response to Xrisk concerns by raising FUD. The ability to respond to long-term AI safety concerns with “yes, we agree AI ethics is very important, and that’s why we’re working on privacy and decolonising AI” seems like a very rhetorically powerful move.

Financial Reserves

Charities like having financial reserves to provide runway, and guarantee that they will be able to keep the lights on for the immediate future. This could be justified if you thought that charities were expensive to create and destroy, and were worried about this occurring by accident due to the whims of donors. Unlike a company which sells a product, it seems reasonable that charities should be more concerned about this.

Donors prefer charities to not have too much reserves. Firstly, those reserves are cash that could be being spent on outcomes now, by either the specific charity or others. Valuable future activities by charities are supported by future donations; they do not need to be pre-funded. Additionally, having reserves increases the risk of organisations ‘going rogue’, because they are insulated from the need to convince donors of their value.

As such, in general I do not give full credence to charities saying they need more funding because they want much more than a 18 months or so of runway in the bank. If you have a year’s reserves now, after this December you will have that plus whatever you raise now, giving you a margin of safety before raising again next year.

I estimated reserves = $(\text{cash and grants}) / (2021 \text{ budget})$. In general I think of this as something of a measure of urgency. However despite being *prima facie* a very simple

calculation there are many issues with this data. As such these should be considered suggestive only.

Donation Matching

In general I believe that charity-specific donation matching schemes [are somewhat dishonest](#), despite my having provided matching funding for at least one in the past.

Ironically, despite this view being [espoused by GiveWell](#) (albeit in 2011), this is essentially of OpenPhil's policy of, at least in some cases, artificially limiting their funding to 50% or 60% of a charity's need, which some charities have argued effectively provides a 1:1 match for outside donors. I think this is bad. In the best case this forces outside donors to step in, imposing marketing costs on the charity and research costs on the donors. In the worst case it leaves valuable projects unfunded.

Obviously cause-neutral donation matching is different and should be exploited. Everyone should max out their corporate matching programs if possible, and things like the [annual Facebook Match](#) continue to be great opportunities.

Poor Quality Research

Partly thanks to the efforts of the community, the field of AI safety is considerably more well respected and funded than was previously the case, which has attracted a lot of new researchers. While generally good, one side effect of this (perhaps combined with the fact that many low-hanging fruits of the insight tree have been plucked) is that a considerable amount of low-quality work has been produced. For example, there are a lot of papers which can be accurately summarized as asserting "just use ML to learn ethics". Furthermore, the conventional peer review system seems to be extremely bad at dealing with this issue.

The standard view here is just to ignore low quality work. This has many advantages, for example 1) it requires little effort, 2) it doesn't annoy people. This conspiracy of silence seems to be the strategy adopted by most scientific fields, except in extreme cases like anti-vaxers.

However, I think there are some downsides to this strategy. A sufficiently large milieu of low-quality work might degrade the reputation of the field, deterring potentially high-quality contributors. While low-quality contributions might help improve [Concrete Problems](#)' citation count, they may use up scarce funding.

Moreover, it is not clear to me that 'just ignore it' really generalizes as a community strategy. Perhaps you, enlightened reader, can judge that "*How to solve AI Ethics: Just use RNNs*" is not great. But is it really efficient to require everyone to independently work this out? Furthermore, I suspect that the idea that we can all just ignore the weak stuff is somewhat an example of typical mind fallacy. Several times I have come across people I respect according respect to work I found clearly pointless. And several times I have come across people I respect arguing persuasively that work I had previously respected was very bad – but I only learnt they believed this by chance! So I think it is quite possible that many people will waste a lot of time as a result of this strategy, especially if they don't happen to move in the right social circles.

Having said all that, I am not a fan of unilateral action, and am somewhat selfishly conflict-averse, so will largely continue to abide by this non-aggression convention. My only deviation here is to make it explicit. If you're interested in this you might enjoy [this](#) by 80,000 Hours.

The Bay Area

Much of the AI and EA communities, and especially the EA community concerned with AI, is located in the Bay Area, especially Berkeley and San Francisco. It does have advantages - like proximity to good CS universities - but it is an extremely expensive place, and is dysfunctional both politically and socially. Aside from the lack of electricity and aggressive homelessness, it seems to attract people who are extremely weird in socially undesirable ways - and induces this in those who move there - though to be fair the people who are doing useful work in AI organisations seem to be drawn from a better distribution than the broader community. In general I think the centralization is bad, but if there must be centralization I would prefer it be almost anywhere other than Berkeley. Additionally, I think many funders are geographically myopic, and biased towards funding things in the Bay Area. As such, I have a mild preference towards funding non-Bay-Area projects.

Conclusions

The size of the field continues to grow, both in terms of funding and researchers. Both make it increasingly hard for individual donors. I've attempted to subjectively weigh the productivity of the different organisations against the resources they used to generate that output, and donate accordingly.

My constant wish is to promote a lively intellect and independent decision-making among readers; hopefully my laying out the facts as I see them above will prove helpful to some readers. Here is my eventual decision, [rot13'd](#) so you can do come to your own conclusions first (which I strongly recommend):

Na vapernfvatyl ynetr nzbhag bs gur orfg jbex vf orvat qbar va cynprf gung qb abg frrz yvxryl gb orarsvg sebz znetvany shaqvat: SUV, Qrrczvaq, BcraNV rgp. Juvyr n tbbq qrirybczrag birenny - V nz pregnvayl irel cyrnfrq gung Qrrczvaq naq BcraNV unir fhpu cebqhpgvir grnzf - vg zrnaf jr pna'g ernyyl qb zhpu urer.

ZVEV frrzf gb unir tbbq crbcyr naq n tbbq genpx erpbeq, naq gurl fznyy nzbhag gurl eryrnfr vf fgebat. Ohg V pna'g rasbepr shaqvat n ynetr betnavfngvba jvgubhg gnatvoyr rivqrapr sbe znal lrnef.

Bs gur cynprf qbvat svefg-pynff grpuaavpny erfrnepu, PUNV frrzf gb zr gb or gur bar gung pbhyq zbfq perqvol orarsvg sebz zber shaqvat. V nz n yvgyr pbaprearq Ebuva vf yrniyat, nf ur jnf n irel fgebat pbagevohgbe, naq gur evfx jvgu npnqrzvp vafgvghgvbaf vf gurl trg 'qvfgengpgrq'. Ohg birenny V guvax gurl erznva irel cebzvfvat fb V vagraq qb znxr n fvtavsvpnag qbangvba urer.

Va gur cnfg V unir orra dhvgr unefu ba PFRE orphnhfr V sryg gung n ybg bs gurve jbex jnf abg irel eryrinag. Vg qbrf frrz fhowrpgvirl gb zr gung gurve cebqhpgvivgl naq sbphf unf fvtavsvpnagyl vzcebirq ubjire.

V guvax OREV ner irel vagrerfgvat. Gurve fgengrtl frrzf gb bssre gur punapr gb fvtavsvpnagyl obbfq npnqrzvp (naq guhf znvafgernz-pbaarpgrq naq fgngfh vzohvat) erfrnepu juvyr znvagnavat n sbphf ba gur zvffvba gung zvtug or ybfg jvgu qverpg tenagf. Zl bar pbaprea urer vf gung gurl ner fbzrguvat bs n bar-zna bcrengvba, naq juvyr V jnf irel snzvyvne jvgu Pevgpu V xabj irel yvgyr nobhg Fnjlre. Ohg birenny V guvax guvf vf irel cebzvfvat fb V jvyy cebonoyl or qbangvat. Abgr gung guvf vf vaqverpgyl fhccbegvat PFRE nf jryy nf bgure betf yvxr SUV, PUNV rgg.

Svanyyl, V pbagvahr gb yvxr gur YGSS. V'z n yvgyr pbaprearq nobhg hcpbzvat cbffvoyr crefbaary punatrj jura gurl fcva bhg bs PRN, naq jbhyq cersre vs gurl qvqa'g tenag gb betnavfngvba ynetr rabhtu gb eha gurve bja shaqenvvat pnzcnvtaf (naq urapr pna or rinyhngrq ol vaqvivqhny qbabef). Ohg birenny V guvax vg vf irel nggenpgvir gb shaq fznyy cebwrpgf, naq V nz abg njner bs nal bgure nirahr sbe fznyy qbabef gb genpgnoyl qb guvf. Fb V jvyy or qbangvat gb gurz ntnva guvf lrne.

However, I wish to emphasize that all the above organisations seem to be doing good work on the most important issue facing mankind. It is the nature of making decisions under scarcity that we must prioritize some over others, and I hope that all organisations will understand that this necessarily involves negative comparisons at times.

Thanks for reading this far; hopefully you found it useful. Apologies to everyone who did valuable work that I excluded!

If you found this post helpful, and especially if it helped inform your donations, please consider letting me and any organisations you donate to as a result know.

If you are interested in helping out with next year's article, please get in touch, and perhaps we can work something out.

Disclosures

I have not in general checked all the proofs in these papers, and similarly trust that researchers have honestly reported the results of their simulations.

I was a Summer Fellow at MIRI back when it was SIAI and volunteered briefly at GWWC (part of CEA). My wife has done some contract work for OpenPhil. I have no financial ties beyond being a donor and have never been romantically involved with anyone else who has ever worked at any of the other organisations.

I shared drafts of the individual organisation sections with representatives from LTFF, FHI, MIRI, CHAI, GCRI, CSER, Ought, AI Impacts, BERI, CLR, GPI, OpenPhil, Convergence.

My eternal gratitude to my anonymous reviewers for their invaluable help, and especially Jess Riedel for the volume and insight of his comments. Any remaining mistakes are of course my own. I would also like to thank my wife and daughter for tolerating all the time I have spent/invested/wasted on this. Negative thanks goes to The Wuhan Institute of Virology and [Paradox Interactive](#).

Sources

This is a list of all the articles cited who with their own individual paragraph. It does not include articles that are only referenced in-line, typically with the word 'here'.

Aird, Michael - Existential risks are not just about humanity - 2020-04-27 - <https://forum.effectivealtruism.org/posts/EfCCgpvQX359xuZ4g/are-existential-risks-just-about-humanity>

Aird, Michael - Failures in technology forecasting? A reply to Ord and Yudkowsky - 2020-05-08 - <https://www.lesswrong.com/posts/3qypPmmNHEmegoFF/failures-in-technology-forecasting-a-reply-to-ord-and>

Aird, Michael; Shovelain, Justin - Using vector fields to visualise preferences and make them consistent - 2020-01-28 - <https://www.lesswrong.com/posts/ky988ePJvCRhmCwGo/using-vector-fields-to-visualise-preferences-and-make-them#comments>

Aird, Michael; Shovelain, Justin; Kristoffersson, David - Memetic downside risks: How ideas can evolve and cause harm - 2020-02-25 - <https://www.lesswrong.com/posts/EdAHNdbkGR6ndAPJD/memetic-downside-risks-how-ideas-can-evolve-and-cause-harm>

AlphaFold Team - AlphaFold: a solution to a 50-year-old grand challenge in biology - 2020-11-30 - <https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>

Althaus, David; Baumann, Tobias - Reducing long-term risks from malevolent actors - 2020-04-29 - <https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors#comments>

Aquirre, Anthony - Why those who care about catastrophic and existential risk should care about autonomous weapons - 2020-11-11 - <https://www.lesswrong.com/posts/Btrmh6T62tB4g9RMc/why-those-who-care-about-catastrophic-and-existential-risk#comments>

Armstrong, Stuart; Leike, Jan; Orseau, Laurent; Legg, Shane - Pitfalls of Learning a Reward Function Online - 2020-04-28 - <https://arxiv.org/abs/2004.13654>

Ashurst, Carolyn; Anderljung, Markus; Prunkl, Carina; Leike, Jan; Gal, Yarin; Shevlane, Toby; Dafoe, Allan - A Guide to Writing the NeurIPS Impact Statement - 2020-05-13 - <https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832>

Avin, Sharar; Gruetzmacher, Ross; Fox, James - Exploring AI Futures Through Role Play - 2020-02-26 - <https://arxiv.org/abs/1912.08964>

Barnes, Beth; Christiano, Paul - Writeup: Progress on AI Safety via Debate - 2020-02-05 - <https://www.alignmentforum.org/posts/Br4xDyU4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1>

Baum, Seth - Accounting for violent conflict risk in planetary defense decisions - 2020-09-09 - <http://gcrinstitute.org/accounting-for-violent-conflict-risk-in-planetary-defense-decisions/>

Baum, Seth - Artificial Interdisciplinarity: Artificial Intelligence for Research on Complex Societal Problems - 2020-07-14 - <http://gcrinstitute.org/artificial-interdisciplinarity-artificial-intelligence-for-research-on-complex-societal-problems/>

Baum, Seth - Medium-Term Artificial Intelligence and Society - 2020-02-16 - <http://gcrinstitute.org/medium-term-artificial-intelligence-and-society/>

Baum, Seth - Quantifying the Probability of Existential Catastrophe: A Reply to Beard et al. - 2020-08-10 - <http://gcrinstitute.org/quantifying-the-probability-of-existential-catastrophe-a-reply-to-beard-et-al/>

Beard, Simon; Kaxzmarek, Patrick - On the Wrongness of Human Extinction - 2020-02-21 - <https://www.cser.ac.uk/resources/wrongness-human-extinction/>

Beard, Simon; Rowe, Thomas; Fox, James - An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards - 2019-12-03 - <https://www.sciencedirect.com/science/article/pii/S0016328719303313>

Beard, Simon; Rowe, Thomas; Fox, James - Existential risk assessment: A reply to Baum - 2020-07-15 - <https://sci-hub.do/10.1016/j.futures.2020.102606>

Belfield, Haydn - Activism by the AI Community: Analysing Recent Achievements and Future Prospects - 2020-02-26 - <https://www.cser.ac.uk/resources/activism-ai-community-analysing-recent-achievements-and-future-prospects/>

Belfield, Haydn; Hernández-Orallo, José; hÉigeartaigh, Seán Ó; Maas, Matthijs M.; Hagerty, Alexa; Whittlestone, Jess - Response to the European Commission's consultation on AI - 2020-02-19 - <https://www.cser.ac.uk/resources/response-european-commissions-consultation-ai/>

Benadè, Gerdus; Nath, Swaprava; Procaccia, Ariel D.; Shah, Nisarg - Preference Elicitation for Participatory Budgeting - 2020-10-27 -
<https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3666>

Benaich, Nathan; Hogarth, Ian - State of AI Report 2020 - 2020-09-01 -
https://docs.google.com/presentation/d/1ZUimafgXCBSLsgbacd6-a-dqO7yLyzII1ZJbiCBUUT4/edit#slide=id.g9348791e5b_1_7

Bhatt, Umang; Andrus, McKane; Weller, Adrian; Xiang, Alice - Machine Learning Explainability for External Stakeholders - 2020-07-10 - <https://arxiv.org/abs/2007.05408v1>

Bobu, Andreea; Scobee, Dexter R.R.; Fisac, Jaime F.; Sastry, S. Shankar; Dragan, Anca D. - LESS is More: Rethinking Probabilistic Models of Human Behavior - 2020-01-13 -
<https://arxiv.org/abs/2001.04465>

Bostom, Nick; Shulman, Carl - Sharing the World with Digital Minds - 2020-10-01 -
<http://www.nickbostrom.com/papers/monster.pdf>

Bostrom, Nick; Belfield, Haydn; Hilton, Sam - Written Evidence to the UK Parliament Science & Technology Committee's Inquiry on A new UK research funding agency. - 2020-09-16 -
<https://www.cser.ac.uk/resources/written-evidence-uk-arpa-key-recommendations/>

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel M.; Wu, Jeffrey; Winter, Clemens; Hesse, Christopher; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya; Amodei, Dario - Language Models are Few-Shot Learners - 2020-05-28 - <https://arxiv.org/abs/2005.14165>

Brundage, Miles; Avin, Shahar; Wang, Jasmine; Belfield, Haydn; Krueger, Gretchen; Hadfield, Gillian; Khlaaf, Heidy; Yang, Jingying; Toner, Helen; Fong, Ruth; Maharaj, Tegan; Koh, Pang Wei; Hooker, Sara; Leung, Jade; Trask, Andrew; Bluemke, Emma; Lebensold, Jonathan; O'Keefe, Cullen; Koren, Mark; Ryffel, Théo; Rubinovitz, JB; Besiroglu, Tamay; Carugati, Federica; Clark, Jack; Eckersley, Peter; Haas, Sarah de; Johnson, Maritza; Laurie, Ben; Ingerman, Alex; Krawczuk, Igor; Askell, Amanda; Cammarota, Rosario; Lohn, Andrew; Krueger, David; Stix, Charlotte; Henderson, Peter; Graham, Logan; Prunkl, Carina; Martin, Bianca; Seger, Elizabeth; Zilberman, Noa; hÉigeartaigh, Seán Ó; Kroeger, Frens; Sastry, Girish; Kagan, Rebecca; Weller, Adrian; Tse, Brian; Barnes, Elizabeth; Dafoe, Allan; Scharre, Paul; Herbert-Voss, Ariel; Rasser, Martijn; Sodhani, Shagun; Flynn, Carrick; Gilbert, Thomas Krendl; Dyer, Lisa; Khan, Saif; Bengio, Yoshua; Anderljung, Markus - Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims - 2020-04-15 -
<https://arxiv.org/abs/2004.07213>

Burden, John & Hernandez-Orallo, Jose - Exploring AI Safety in Degrees: Generality, Capability and Control - 2020-08-10 - <https://www.cser.ac.uk/resources/exploring-ai-safety-degrees-generality-capability-and-control/>

Byun, Jungwon, Stuhlmuller, Andreas - Automating reasoning about the future at Ought - 2020-11-09 - <https://ought.org/updates/2020-11-09-forecasting>

Carey, Ryan; Langlois, Eric; Everitt, Tom; Legg, Shane - The Incentives that Shape Behaviour - 2020-01-20 - <https://arxiv.org/abs/2001.07118>

Carlsmith, Joseph - How Much Computational Power Does It Take to Match the Human Brain? - 2020-09-11 - <https://www.openphilanthropy.org/brain-computation-report>

Cave, Stephen; Dihal, Kanta - The Whiteness of AI - 2020-08-06 -
<http://lcfi.ac.uk/resources/whiteness-ai/>

Christian, Brian - The Alignment Problem: Machine Learning and Human Values - 2020-09-06
- https://www.amazon.com/Alignment-Problem-Machine-Learning-Values-ebook/dp/B085T55LGK/ref=tmm_kin_swatch_0?encoding=UTF8&qid=&sr=

Christiano, Paul - "Unsupervised" translation as an (intent) alignment problem - 2020-09-29 -
<https://ai-alignment.com/unsupervised-translation-as-a-safety-problem-99ae1f9b6b68>

Cihon, Peter; Maas, Matthijs M.; Kemp, Luke - Should Artificial Intelligence Governance be Centralised? Design Lessons from History - 2020-01-10 - <https://arxiv.org/abs/2001.03573>

Clarke, Sam - Clarifying "What failure looks like" (part 1) - 2020-09-20 -
<https://www.alignmentforum.org/posts/v6Q7T335KCMxujhZu/clarifying-what-failure-looks-like-part-1>

Clifton, Jesse - Equilibrium and prior selection problems in multipolar deployment - 2020-04-02 - <https://www.alignmentforum.org/posts/Tdu3tGT4i24qcLESh/equilibrium-and-prior-selection-problems-in-multipolar-1#comments>

Clifton, Jesse; Riche, Maxime - Towards Cooperation in Learning Games - 2020-11-15 -
https://longtermrisk.org/files/toward_cooperation_learning_games_oct_2020.pdf

Cohen, Michael; Hutter, Marcus - Curiosity Killed the Cat and the Asymptotically Optimal Agent - 2020-06-05 - <https://arxiv.org/abs/2006.03357>

Cohen, Michael; Hutter, Marcus - Pessimism About Unknown Unknowns Inspires Conservatism - 2020-06-15 - <https://arxiv.org/abs/2006.08753>

Cotra, Ajeya - Report on AI Timelines - 2020-10-18 -
<https://www.alignmentforum.org/posts/KrjfoZzpSDpnrv9va/draft-report-on-ai-timelines>

Cotton-Barratt, Owen; Daniel, Max; Sandberg, Anders; - Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter - 2020-01-24 -
<https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12786>

Cremer, Carla; Whittlestone, Jess - Canaries in Technology Mines: Warning Signs of Transformative Progress in AI - 2020-09-24 - <https://www.fhi.ox.ac.uk/publications/canaries-in-technology-mines-warning-signs-of-transformative-progress-in-ai-cremer-and-whittlestone/>

Critch, Andrew - Some AI research areas and their relevance to existential safety - 2020-11-18 - <https://www.alignmentforum.org/posts/hvGoYXi2kgnS3vxqb/some-ai-research-areas-and-their-relevance-to-existential-1>

Critch, Andrew; Krueger, David - AI Research Considerations for Human Existential Safety (ARCHES) - 2020-05-30 - <https://arxiv.org/abs/2006.04948>

Crosby, Matthew; Beyret, Benjamin; Shanahan, Murray; Hernández-Orallo, José; Cheke, Lucy; Halina, Marta - The Animal-AI Testbed and Competition - 2020-09-22 -
<http://lcfi.ac.uk/resources/animal-ai-testbed-and-competition-paper-published/>

Demski, Abram - Radical Probabilism - 2020-08-18 -
<https://www.lesswrong.com/s/HmANELvkAZ9eDxFs/p/xJyY5QkQvNJPZLJR0>

Ding, Jeffrey; Dafoe, Allan - The Logic of Strategic Assets: From Oil to AI - 2020-01-09 -
<https://arxiv.org/ftp/arxiv/papers/2001/2001.03246.pdf>

Freedman, Rachel; Shah, Rohin; Dragan, Anca - Choice Set Misspecification in Reward Inference - 2020-09-10 - http://ceur-ws.org/Vol-2640/paper_14.pdf

Gabriel, Jason - Artificial Intelligence, Values and Alignment - 2020-01-13 -
<https://arxiv.org/abs/2001.09768>

Garfinel, Ben - Does Economic History Point Towards a Singularity? - 2020-09-02 -
<https://forum.effectivealtruism.org/posts/CWFn9qAKsRibpCGq8/does-economic-history-point-toward-a-singularity>

Garrabrant, Scott - Cartesian Frames - 2020-10-22 -
<https://www.alignmentforum.org/s/2A7rrZ4ySx6R8mfoT>

Gleave, Adam; Dennis, Michael; Legg, Shane; Russell, Stuart; Leike, Jan - QUANTIFYING DIFFERENCES IN REWARD FUNCTIONS - 2020-10-08 - <https://arxiv.org/abs/2006.13900>

Grace, Katja - Atari early - 2020-04-01 - <https://aiimpacts.org/atari-early/>

Grace, Katja - Discontinuous progress in history: an update - 2020-04-13 -
<https://aiimpacts.org/discontinuous-progress-in-history-an-update/>

Halpern, Joseph; Piermont, Evan - Dynamic Awareness - 2020-07-06 -
<https://arxiv.org/abs/2007.02823>

hÉigearaigh, Seán Ó; Whittlestone, Jess; Liu, Yang; Zeng, Yi; Liu, Zhe - Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance - 2020-05-15 -
<https://link.springer.com/article/10.1007/s13347-020-00402-x>

Hendrycks, Dan; Burns, Collin; Basart, Steven; Critch, Andrew; Li, Jerry; Song, Dawn; Steinhardt, Jacob - Aligning AI with Shared Human Values - 2020-08-05 -
<https://arxiv.org/abs/2008.02275>

Henighan, Tom; Kaplan, Jared; Katz, Mor; Chen, Mark; Hesse, Christopher; Jackson, Jacob; Jun, Heewoo; Brown, Tom B.; Dhariwal, Prafulla; Gray, Scott; Hallacy, Chris; Mann, Benjamin; Radford, Alec; Ramesh, Aditya; Ryder, Nick; Ziegler, Daniel M.; Schulman, John; Amodei, Dario; McCandlish, Sam - Scaling Laws for Autoregressive Generative Modeling - 2020-11-06 -
https://arxiv.org/abs/2010.14701?fbclid=IwAR3H_kH2TKQXI4GcVGLXsfZv2JfD_mOIRdQfXFuAZDttPoHMRKyHITgo74

Hernandez-Orallo, Jose; Martinez-Plumed, Fernando; Avin, Shahar; Whittlestone, Jess; hÉigearaigh, Seán Ó - AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues - 2020-08-10 - <https://www.cser.ac.uk/resources/ai-paradigms-and-ai-safety-mapping-artefacts-and-techniques-safety-issues/>

Hollanek, Tomasz - AI transparency: a matter of reconciling design with critique - 2020-11-17 -
<https://link.springer.com/article/10.1007%2Fs00146-020-01110-y#author-information>

Hubinger, Evan - An overview of 11 proposals for building safe advanced AI - 2020-05-29 -
<https://www.alignmentforum.org/posts/fRsjBseRuvRhMPPE5/an-overview-of-11-proposals-for-building-safe-advanced-ai>

Hwang, Tim - Shaping the Terrain of AI Competition - 2020-06-15 -
<https://cset.georgetown.edu/research/shaping-the-terrain-of-ai-competition/>

Imbrie, Andrew; Kania, Elsa; Laskai, Lorand - The Question of Comparative Advantage in Artificial Intelligence: Enduring Strengths and Emerging Challenges for the United States - 2020-01-15 - <https://cset.georgetown.edu/research/the-question-of-comparative-advantage-in-artificial-intelligence-enduring-strengths-and-emerging-challenges-for-the-united-states/>

John, Tyler; MacAskill, William - Longtermist institutional reform - 2020-07-30 -
<https://philpapers.org/rec/OHLIR>

Kemp, Luke; Rhodes, Catherine - The Cartography of Global Catastrophic Risks - 2020-01-06 -
<https://www.cser.ac.uk/resources/cartography-global-catastrophic-governance/>

Kokotajlo, Daniel - Relevant pre-AGI possibilities - 2020-06-18 -
<https://aiimpacts.org/relevant-preagi-possibilities/>

Kokotajlo, Daniel - Three kinds of competitiveness - 2020-03-30 - <https://aiimpacts.org/three-kinds-of-competitiveness/>

Korzekwa, Rick - Description vs simulated prediction - 2020-04-22 -
<https://aiimpacts.org/description-vs-simulated-prediction/>

Korzekwa, Rick - Preliminary survey of prescient actions - 2020-04-08 -
<https://aiimpacts.org/survey-of-prescient-actions/>

Kovařík, Vojtěch ; Carey, Ryan - (When) Is Truth-telling Favored in AI Debate? - 2019-12-15 -
<https://arxiv.org/abs/1911.04266>

Krakovna, Victoria - Possible takeaways from the coronavirus pandemic for slow AI takeoff - 2020-05-31 - <https://vkrakovna.wordpress.com/2020/05/31/possible-takeaways-from-the-coronavirus-pandemic-for-slow-ai-takeoff/>

Krakovna, Victoria; Orseau, Laurent; Ngo, Richard; Martic, Miljan; Legg, Shane - Avoiding Side Effects By Considering Future Tasks - 2020-10-15 - <https://arxiv.org/abs/2010.07877v1>

Krakovna, Victoria; Uesato, Jonathan; Mikulik, Vladimir; Rahtz, Matthew; Everitt, Tom; Kumar, Ramana; Kenton, Zac; Leike, Jan; Legg, Shane - Specification gaming: the flip side of AI ingenuity - 2020-04-21 - <https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity>

Lehman, Joel - Reinforcement Learning Under Moral Uncertainty - 2020-06-15 -
<https://arxiv.org/abs/2006.04734>

Linsefors, Linda & Hepburn, JJ - Announcing AI Safety Support - 2020-11-19 -
<https://forum.effectivealtruism.org/posts/wpQ2qhF8Z6oonaPX/announcing-ai-safety-support>

MacAskill, Will - Are we living at the hinge of history? - 2020-09-01 -
https://globalprioritiesinstitute.org/wp-content/uploads/William-MacAskill_Are-we-living-at-the-hinge-of-history.pdf

Makievskyi, Anton; Zhou, Liang ; Chiswick, Max - Assessing Generalization in Reward Learning with Procedurally Generated Games - 2020-08-30 -
<https://towardsdatascience.com/assessing-generalization-in-reward-learning-intro-and-background-da6c99d9e48>

Mogensen, Andreas - Moral demands and the far future - 2020-06-01 -
<https://globalprioritiesinstitute.org/wp-content/uploads/Working-Paper-1-2020-Andreas-Mogensen.pdf>

Mogensen, Andreas; Thorstad, David - Tough enough? Robust satisficing as a decision norm for long-term policy analysis - 2020-11-01 - https://globalprioritiesinstitute.org/wp-content/uploads/Tough-Enough_Andreas-Mogensen-and-David-Thorstad.pdf

Ngo, Richard - AGI Safety from First Principles - 2020-09-28 -
<https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ>

Nguyen, Chi; Christiano, Paul - My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda - 2020-08-15 -
<https://www.lesswrong.com/posts/PT8vSxsusqWuN7JXp#comments>

O'Keefe, Cullen; Cihon, Peter; Garfinkel, Ben; Flynn, Carrick; Leung, Jade; Dafoe, Allan - The Windfall Clause: Distributing the Benefits of AI for the Common Good - 2020-01-30 -
<https://www.fhi.ox.ac.uk/windfallclause/>

O'Brien, John; Nelson, Cassidy - Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology - 2020-06-17 -
<https://www.liebertpub.com/doi/full/10.1089/hs.2019.0122>

O'Keefe, Cullen - How will National Security Considerations affect Antitrust Decisions in AI? An Examination of Historical Precedents - 2020-07-28 -
<https://forum.effectivealtruism.org/out?url=https%3A%2F%2Fwww.fhi.ox.ac.uk%2Fwp-content%2Fuploads%2FHow-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-OKeefe.pdf>

Ord, Toby - The Precipice - 2020-03-24 - https://www.amazon.com/Precipice-Existential-Risk-Future-Humanity-ebook/dp/B07V9GHKYP/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=

Peters, Dorian; Vold, Karina; Robinson, Diana; Calvo, Rafael - Responsible AI—Two Frameworks for Ethical Design Practice - 2020-02-15 -
<https://ieeexplore.ieee.org/document/9001063/authors#authors>

Prunkl, Carina; Whittlestone, Jess - Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society - 2020-01-13 - <https://arxiv.org/abs/2001.04335>

Qian, Shi; Hui, Li; Tse, Brian; Hopcroft, John; Russell, Stuart; Jeanmaire, Caroline; Qiang, Yang; Fung, Pascale; Yampolskiy, Roman; Dafoe, Allan; Anderljung, Markus; Hadfield, Gillian; Wright, Don; Brundage, Miles; Clark, Jack; Solaiman, Irene; Krueger, Gretchen; O'hEigearaigh, Sean; Toner, Helen; Liu, Millie; Hoffman, Steve; Beridze, Irakli; Wallach, Wendell; Hodes, Cyrus; Mialhe, Nicolas; Newman, Jessica; Dingding, Chen; Kaili, Eva; Jun, Su; Hagendorff, Thilo; Ahrweiler, Petra; Williams, Robin; Allen, Colin; Wang, Poon; Carbonell, Ferran; Ziaohong, Wang; Qingfend, Yang; Qi, Yin; Rossie, Francesca; Stix, Charlotte; Daly, Angela; Gal, Danit; Ema, Arisa; Yihan, Goh; Remolina, Nydia; Aneja, Urvashi; Ying, Fu; Zhiyun, Zhao; Xiuquan, Li; Weiwen, Duan; Qun, Luan; Rui, Guo; Yingchun, Wang - AI GOVERNANCE IN 2019 A YEAR IN REVIEW - 2020-04-15 - <https://www.aigovernancereview.com/>

Reddy, Siddharth; Dragan, Anca D.; Levine, Sergey; Legg, Shane; Leike, Jan - Learning Human Objectives by Evaluating Hypothetical Behavior - 2019-12-05 -
<https://arxiv.org/abs/1912.05652>

Russell, Stuart; Norvig, Peter - Artificial Intelligence: A Modern Approach, 4th Edition - 2020-01-01 - <https://www.pearson.com/us/higher-education/program/Russell-Artificial-Intelligence-A-Modern-Approach-4th-Edition/PGM1263338.html>

Saunders, William; Rachbach, Ben; Evans, Owain; Byun, Jungwon; Stuhlmüller, and Andreas - Evaluating Arguments One Step at a Time - 2020-01-11 - <https://ought.org/updates/2020-01-11-arguments>

Scholl, Keller; Hanson, Robin - Testing the Automation Revolution Hypothesis - 2019-12-10 -
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3496364

Shah, Rohin - AI Alignment 2018-19 Review - 2020-01-27 -
https://www.alignmentforum.org/posts/dKxX76SCfCvceJXHv/ai-alignment-2018-19-review#Short_version__1_6k_words_

Shevlane, Toby; Dafoe, Allan - The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? - 2020-12-27 - <https://arxiv.org/abs/2001.00463>

Snyder-Beattie, Andrew; Sandberg, Anders; Drexler, Eric; Bonsall, Michael - The Timing of Evolutionary Transitions Suggests Intelligent Life Is Rare - 2020-11-19 -
<https://www.liebertpub.com/doi/full/10.1089/ast.2019.2149>

Stiennon, Nisan; Ouyang, Long; Wu, Jeff; Ziegler, Daniel M.; Lowe, Ryan; Voss, Chelsea; Radford, Alec; Amodei, Dario; Christiano, Paul - Learning to Summarize with Human Feedback - 2020-09-04 - <https://openai.com/blog/learning-to-summarize-with-human-feedback/>

Tarsney, Christian - Exceeding Expectations: Stochastic Dominance as a General Decision Theory - 2020-08-08 - https://globalprioritiesinstitute.org/wp-content/uploads/Christian-Tarsney_Exceeding-Expectations_Stochastic-Dominance-as-a-General-Decison-Theory.pdf

Tarsney, Christian; Thomas, Teruji - Non-Additive Axiologies in Large Worlds - 2020-09-01 - https://globalprioritiesinstitute.org/wp-content/uploads/Christian-Tarsney-and-Teruji-Thomas_Non-Additive-Axiologies-in-Large-Worlds.pdf

Thorstad, David; Mogensen, Andreas - Heuristics for clueless agents: how to get away with ignoring what matters most in ordinary decision-making - 2020-06-01 - <https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Andreas-Mogensen-Heuristics-for-clueless-agents.pdf>

Trammell, Philip; Korinek, Anton - Economic growth under transformative AI - 2020-10-08 - https://globalprioritiesinstitute.org/wp-content/uploads/Philip-Trammell-and-Anton-Korinek_Economic-Growth-under-Transformative-AI.pdf

Tucker, Aaron; Anderljung, Markus; Dafoe, Allan - Social and Governance Implications of Improved Data Efficiency - 2020-01-14 - <https://arxiv.org/pdf/2001.05068.pdf>

Tzachor, Asaf; Whittlestone, Jess; Sundaram, Lalitha; , Seán Ó hÉigeartaigh - Artificial intelligence in a crisis needs ethics with urgency - 2020-12-02 - <https://www.nature.com/articles/s42256-020-0195-0>

Uesato, Jonathan; Kumar, Ramana; Krakovna, Victoria; Everitt, Tom; Ngo, Richard; Legg, Shane - Avoiding Tampering Incentives in Deep RL via Decoupled Approval - 2020-11-17 - <https://arxiv.org/abs/2011.08827>

Wilkinson, Haydn - In defence of fanaticism - 2020-08-01 - https://globalprioritiesinstitute.org/wp-content/uploads/Hayden-Wilkinson_In-defence-of-fanaticism.pdf

Xu, Jing; Ju, Da; Li, Margaret; Boureau, Y-Lan; Weston, Jason; Dinan, Emily - Recipes for Safety in Open-domain Chatbots - 2020-09-14 - <https://arxiv.org/abs/2010.07079>

Zerilli, John; Knott, Alistair; Maclaurin, James; Colin, Gavaghan - Algorithmic Decision-Making and the Control Problem - 2020-12-11 - <https://link.springer.com/article/10.1007%2Fs11023-019-09513-7>

Give it a google

I'm a programmer. In programming, people talk about how the ability to Google is such an important skill. No one knows everything. In practice, people are always looking things up and running into weird situations that they have to figure out.

That's not what this post is about. Not quite. This post is about the *decision* to give it a google in the first place.

Poker

Here's an example. I play poker. It's hard to get good at poker, but not sucking at poker isn't very hard. Don't play junky hands preflop. Don't open-limp. Learn some basic poker math such as pot odds. If you can follow basic guidelines like that, you won't suck.

Yet if you sit down at a \$1-2 game in Vegas, where hundreds of dollars are exchanging hands, [people](#) who have been playing poker longer than I have been alive often don't understand these basic ideas. Ideas that you would learn if you spent ten minutes googling "basics of poker strategy". You'd come across articles like [How Not to Suck at Poker](#) and you wouldn't be making such mistakes at the table.

As an even more extreme example, I've played in home games where despite playing poker for many years, people don't know even more basic things such as sizing your bets according to the size of the pot. Betting \$20 into a \$200 pot is a very small bet even if \$20 feels like a lot of money. Your opponent only needs to put in \$20 to have the chance at a \$220 pot. Great risk-reward. In general, bet sizing is an incredibly complex topic, but all you have to know is to size your bet between 1/2 and the full size of the pot. The [80/20 principle](#) really, really applies here.

Health

Recently my girlfriend and I have been waking up with scratchy throats. Intuitively we thought it might be because it's getting colder out. Especially as we sleep. So we tried sleeping with the heat on higher.

Turns out that was the exact wrong thing to do. I gave it a google and the issue seems to be that using the heat dries out the air, and dry air causes the sore throat symptoms we were experiencing.

Shopping

I bought a humidifier last night. I could have just surfed around on Amazon a bit and picked something out. Instead, I gave it a google first.

It worked out well, I learned some really interesting things. A humidifier is something that seems simple where it doesn't really matter which one you choose. Turns out that intuition was also wrong.

- It's something that you'll probably need to refill once a day, so you want it to be easy to refill. That means purchasing one that has a flat edge so you can sit it down on the counter while you refill it. My previous one was spherical so I couldn't sit it down like that. I had to hold it in one hand while I used the other hand to pour water into it. Such awkwardness adds up in the long run.
- Same story with ease of cleaning. With my previous one, the spout was small enough where I couldn't really reach inside to clean it and it had weird nooks and crannies that were hard/impossible to clean.
- One type of humidifier is called an evaporative humidifier. The main advantage is that it won't overhumidify your area. Once the humidity hits ~50% or so, it'll naturally stop producing as much humidity. The downside is that it is loud, has more moving parts, and requires you to purchase parts for it as a recurring expense.

After learning all of this stuff on The Wirecutter's [humidifier guide](#), I was able to pick the right humidifier for my needs instead of just finding the cheapest one on Amazon with decent reviews, which probably would have led to a good amount of future frustration.

Tennis

My girlfriend and I started playing tennis. We had played three times and made a little bit of progress, but still weren't very good.

Before our fourth time, I decided to give it a google. I came across this YouTube video: [Beginner Tennis Lesson | Forehand, Backhand & Serve](#). It was great! It's amazing how some of those simple tips were so helpful. Without them we would have been doing much more floundering around.



Restaurants

I find that there is a pretty high amount of variance in restaurant quality at a given price range. Most are pretty meh and leave me questioning whether it was worth spending my money there. But some are incredible and leave me very happy to have spent my money there! Especially the hole-in-the-wall type places. I really enjoy seeking those places out and channeling my inner Anthony Bourdain.

Review sites such as Yelp and Google Reviews are far from perfect. Blogs and Eater are somewhat of an improvement, but also not perfect. Still, spending ten minutes skimming through such resources will definitely move the needle and seems like a great use of time.

Dishwasher

My dishwasher hasn't been working very well. I gave it a google and came across ol'reliable Wirecutter's [position](#) on the topic. I learned some very useful things that surprised me:

- Cheap dishwashers still do a great job at cleaning. The thing that makes expensive ones better isn't that they do a better job at cleaning. This went

directly against my initial impression that the old dishwasher in my not-renovated apartment was the problem.

- If you pre-wash your dishes too much, it could cause problems. The water in your dishwasher that accumulates has crud and stuff from your dishes in it. The dishwasher measures this concentration. If it is low enough, it assumes your dishes must be clean and stops. This too went directly against what I previously thought. I had been trying to do more and more pre-washing to fix the issue.
- Dishwashers need to be cleaned! With all of that hot water and soap in there, who would have thought!
- As a bonus, it's significantly more energy efficient to use a dishwasher than to hand wash. I previously always tried to hand wash stuff and fill up the dishwasher pretty full so I don't have to run as many loads.

Cooking

I made black bean burgers last night. Before doing so I hit up my trusty resource: Serious Eats. They have a [great article on black bean burgers](#) that really improved the quality of my cooking. Here are the important things I learned:

- Baking/roasting your beans makes them more meaty and less mushy/starchy.
- Some cooked grain or nut like cashews give it a nice textural variation on the inside like you'd find in a normal beef burger.
- It's also nice to get some pockets of fat on the inside of the burger, and cheese works great for this. But you need a firmer, fresher cheese like feta. A gooey would just melt and be more uniform.
- Unlike beef burgers, you have to use moderate heat. Otherwise the outside will burn and the inside won't cook all the way through.

I'm not sure how much this example really meshes with my main point of "give it a google". If I just "gave it a google", I'd probably come across some more standard recipe that wouldn't mention these things. I also cross-referenced with Budget Bytes and sure enough they didn't mention these things. Here this only worked because I was previously familiar with Serious Eats and knew to look there. So it'd probably be more accurate to say "give reliable resources you know a quick check" than "give it a google".

Covid

Another example of "give reliable resources you know a quick check" is with the [microCOVID Project](#). They'll tell you roughly how much risk eg. going shopping is compared to an outdoor hangout with friends.

My girlfriend had some family business to take care of at her parents yesterday, but it would involve being indoors with them for a few hours during the peak of the pandemic. So we consulted the site, determined that it would [cost about 400 microCOVIDs](#), and decided that it would be an acceptable risk.

Bike rides in nature

I recently discovered the [River Mountains Loop bike trail](#). It looks so cool! I want to give it a try!

I was looking around on that TrailLink website and saw that there are "big horn sheep" along the trail.

"Big horn sheep"? WTF are those?

Googles it.

Oh! You mean *rams*!



Well, are those safe? Turns out the answer is that you shouldn't get close to them or carry food around them, but they're not nearly a big enough danger to get in the way of a bike ride.

Good news. But what about other dangers? I spent a little bit of time (ok fine, a full two day [rabbit hole](#)) skimming around [The Great Outdoors Stack Exchange](#) and it seems perfectly safe if you're not stupid.

Ok. But what about hills? I've done ~25 mile rides before and felt ok afterwards so you'd think 36 would be alright, especially if I take a rest. But maybe this route is a lot more hilly and I wouldn't be able to finish the ride. I googled it, ended up on YouTube and watched a few videos, discovered that there are in fact some significant hills and that I'd try doing part of the route first before attempting the whole route.

Airbnb

I'm planning on buying a house. One of the things I'm looking for is the ability to Airbnb the house and spend time traveling. Particularly if I buy the house in Vegas where the

summers are crazy hot and I'd like to get away. How easy is this to do? How big is the risk that it just fails and doesn't generate any revenue?

I spent some time googling around for answers to these questions. And this time, I didn't find them!

I think that this is a key point. It's ok that I didn't find the answers.

I've listed nine examples in this article. This one is the tenth, so I'm batting 0.900. I think that's probably in the ballpark of my true success rate with "give it a google". Given these odds, "give it a google" is still highly worthwhile, even though you risk leaving empty handed at times. You gotta risk it to get the biscuit!

Conclusion

Perhaps you can tell, I'm a little bit passionate about this "give it a google" idea. The return on investment is just so huge. A few minutes can teach you so much and help you avoid so many headaches.

I remember the first time I learned about Google. I was at my grandparents house. My grandpa has always been a curious person who likes to learn. Whenever I go there he likes to sit down and teach me things, which I thoroughly enjoy.

Normally he has a mild excitement to him when he explains things to me. This time though, there was a noticeably higher amount of excitement to him.

He taught me about this web search thing he read about in the newspaper. You don't have to know the website you want in advance. If you know the *question* you want answered, you can enter it in and Google will show you the websites you should visit to find the answer!

That's [amazing!](#)

Try taking a time machine back a few dozen years, find someone at the library searching for something, and tell them you have a machine that will show them exactly where to search. Avoids a lot of frustration, right?

Now try telling them that the *entire library* fits inside of this little machine. And that the machine fits inside their *pocket!* Madness! But this is [the reality that we live in.](#)

I don't think we take enough advantage of it though. Given how quick and convenient it is to research certain topics, I think "give it a google" is a major low hanging fruit that society should spend a lot more time plucking.

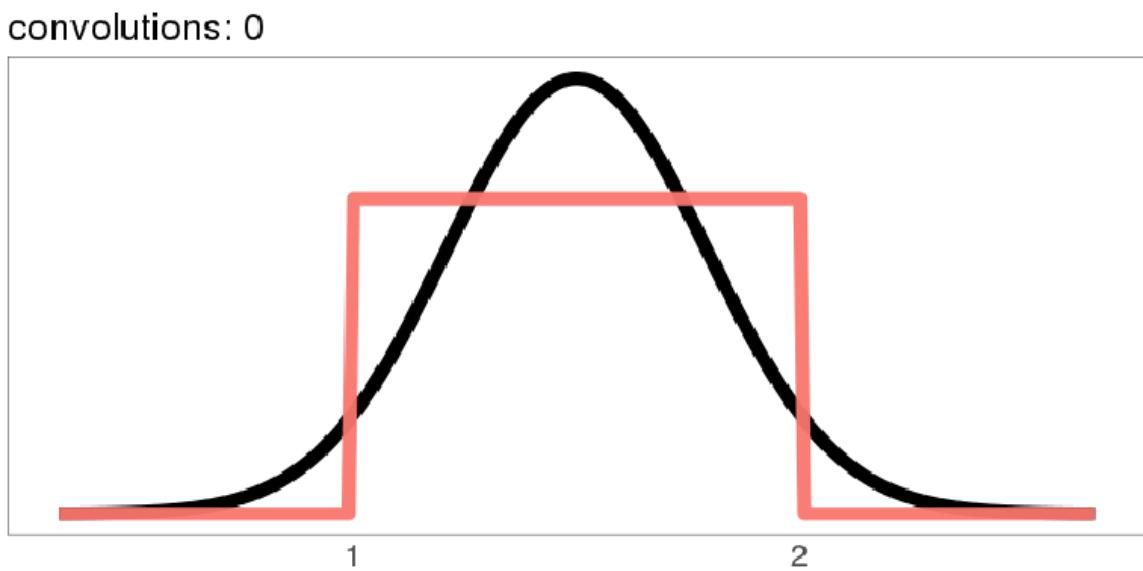
How long does it take to become Gaussian?

The central limit theorems all say that if you convolve stuff enough, and that stuff is sufficiently nice, the result will be a [Gaussian](#) distribution. How much is enough, and how nice is sufficient?

Identically-distributed distributions converge quickly

For many distributions d , the repeated convolution $d * d * \dots * d$ looks Gaussian. The number of convolutions you need to look Gaussian depends on the shape of d . This is the easiest variant of the central limit theorem: identically-distributed distributions.

The uniform distribution converges real quick:

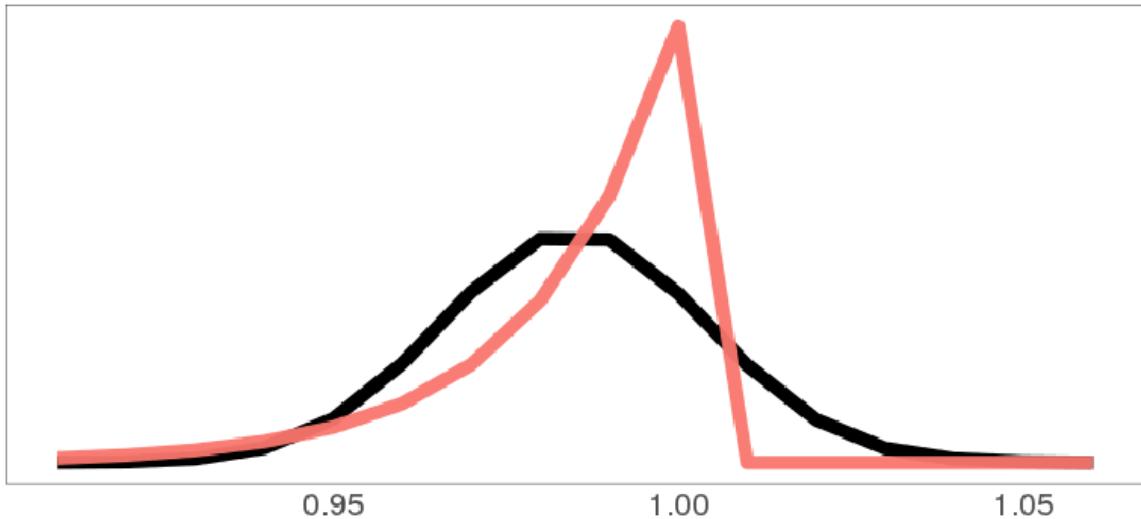


The result of $\text{uniform}(1, 2) * \text{uniform}(1, 2) * \dots * \text{uniform}(1, 2)$, with 30 distributions total. This plot is an animated version of the plots in [the previous post](#). The black curve is the Gaussian distribution with the same mean and variance as the red distribution. The more similar red is to black, the more Gaussian the result of the convolutions is.

The numbers on the x axis are increasing because the mean of $f * g$ [is the sum of the means](#) of f and g , so if we start with positive means, repeated convolutions shoot off into higher numbers. Similar for the variance - notice how the width starts as the difference between 1 and 2, but ends with differences in the tens. You can keep the location stationary under convolution by starting with a distribution centered at 0, but you can't keep the variance from increasing, because you can't have a variance of 0 ([except in the limiting case](#)).

Here's a more skewed distribution: $\text{beta}(50, 1)$. $\text{beta}(50, 1)$ is the probability distribution that represents knowing that a lake has bass and carp, but not how many of each, and then catching 49 bass in a row. It's fairly skewed! This time, after 30 convolutions, we're not quite Gaussian - the skew is still hanging around. But for a lot of real applications, I'd call the result "Gaussian enough".

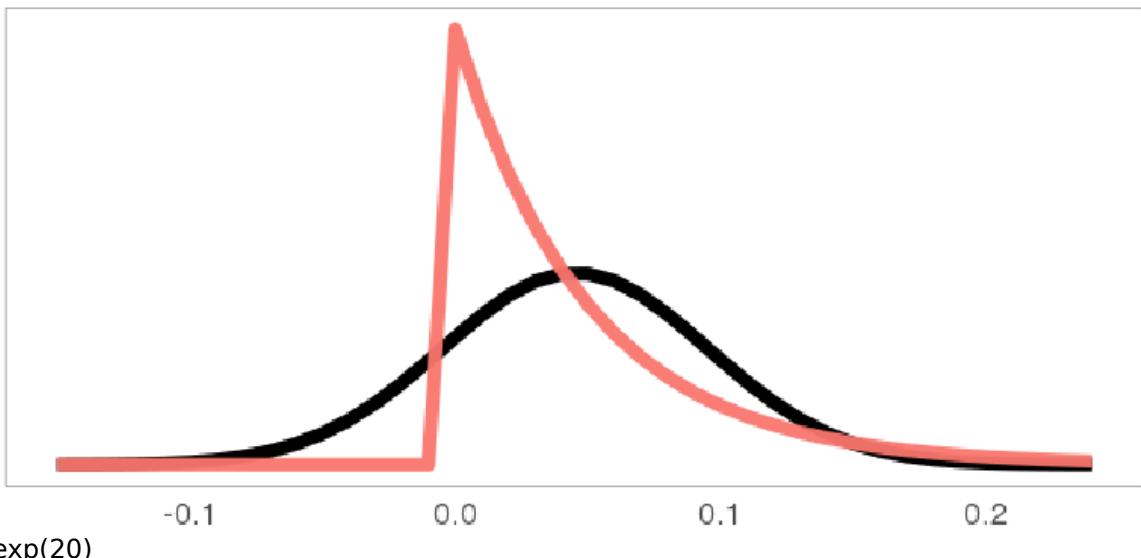
convolutions: 0



$\text{beta}(50, 1)$ convolved with itself 30 times.

A similar skew in the opposite direction, from the exponential distribution:

convolutions: 0

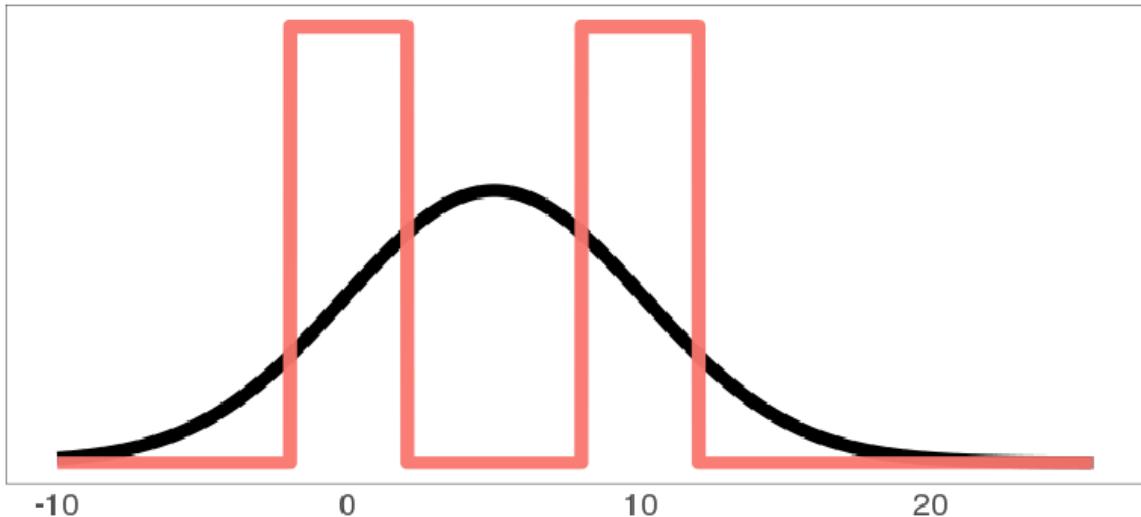


I was surprised to see the exponential distribution go into a Gaussian, because [Wikipedia says](#) that an exponential distribution with parameter θ goes into a gamma distribution with

parameters $\gamma(n, \theta)$ when you convolve it with itself n times. But it turns out $\gamma(n, \theta)$ looks more and more Gaussian as n goes up.

How about our ugly bimodal-uniform distribution?

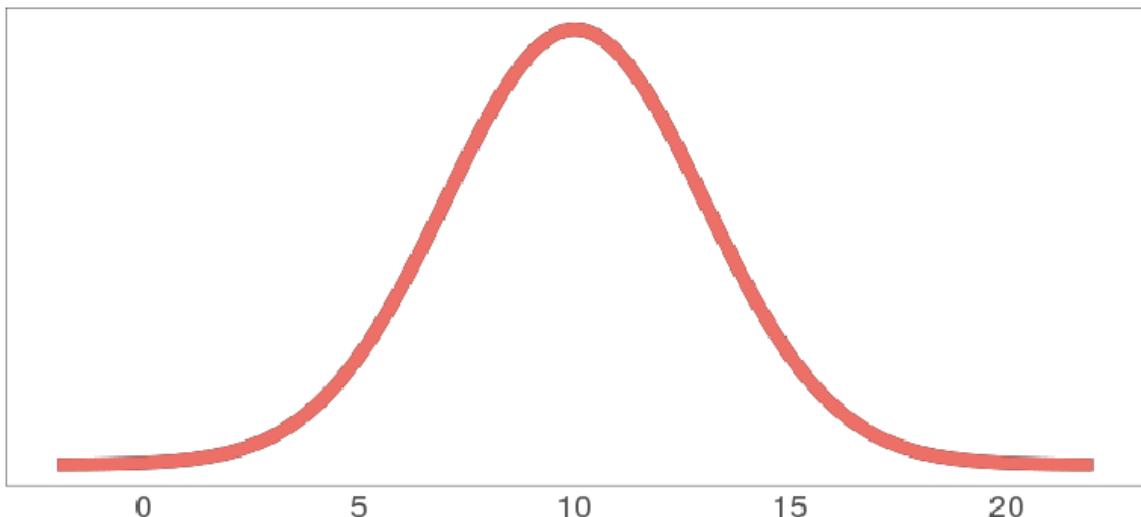
convolutions: 0



It starts out rough and jagged, but already by 30 convolutions it's Gaussian.

And here's what it looks like to start with a Gaussian:

convolutions: 0

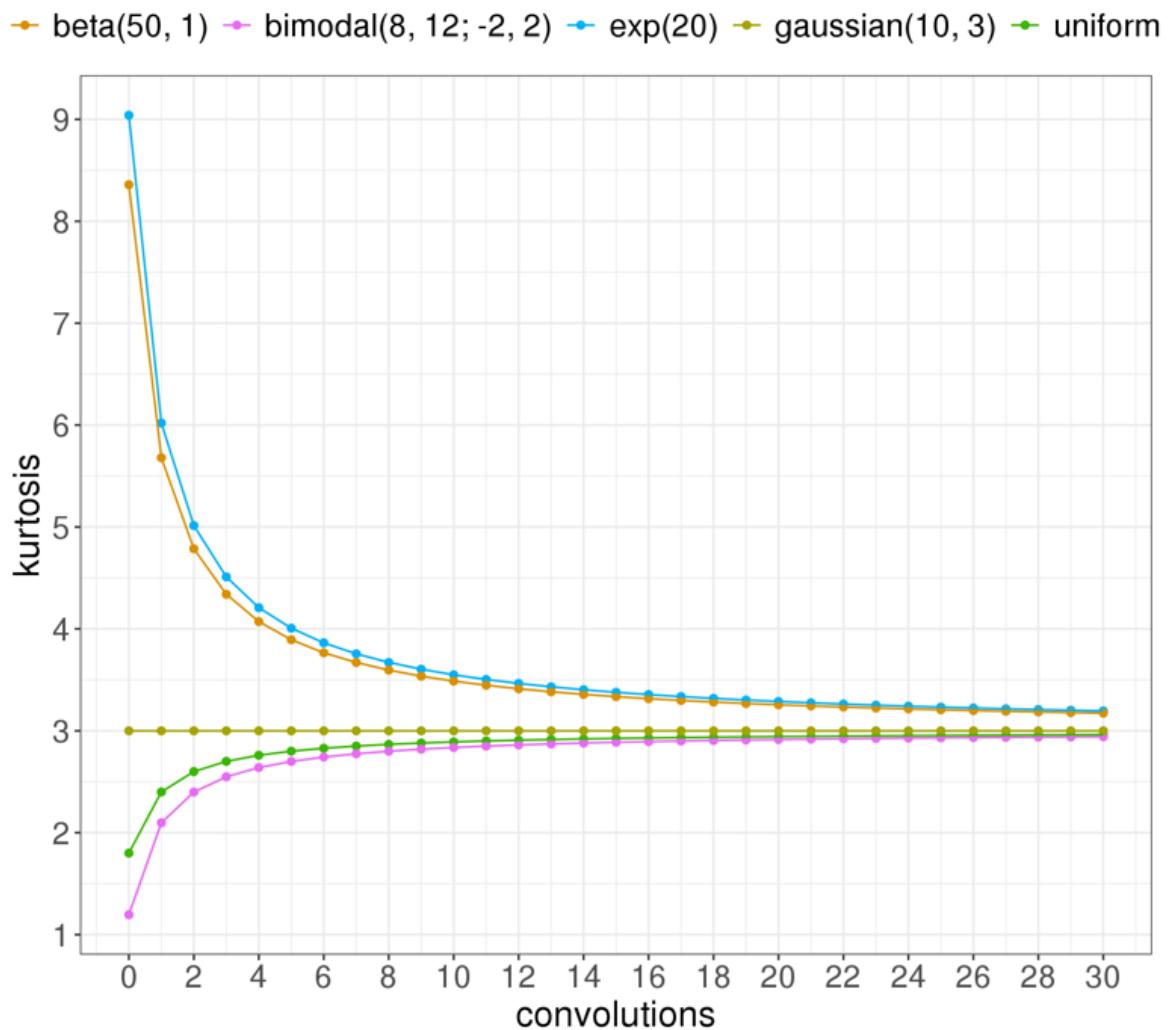


The red curve starts out the exact same as the black curve, then nothing happens because Gaussians stay Gaussian under self-convolution.

An easier way to measure Gaussianness (Gaussianity?)

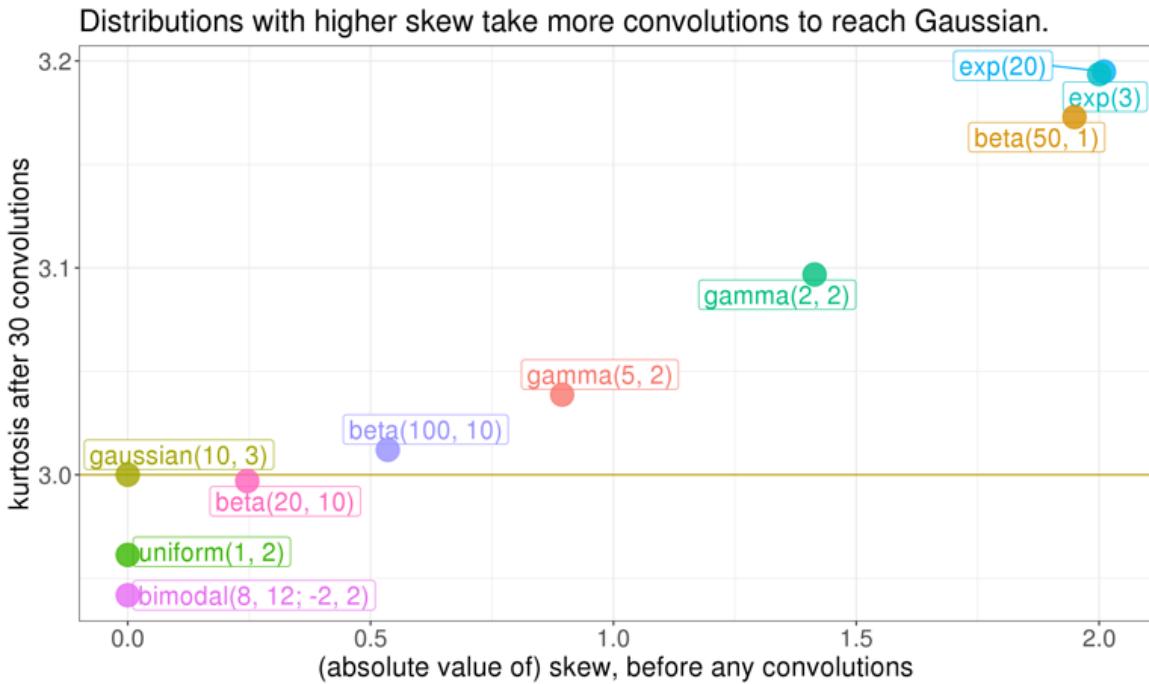
We're going to want to look at many more distributions under n convolutions and see how close they are to Gaussian, and these animations take a lot of space. We need a more compact way. So let's measure the kurtosis of the distributions, instead. The [kurtosis](#) is the fourth moment of a probability distribution; it describes the shape of the tails. All Gaussian distributions have kurtosis 3. There are [other distributions](#) with kurtosis 3, too, but they're not likely to be the result of a series of convolutions. So to check how close a distribution is to Gaussian, we can just check how far from 3 its kurtosis is.

We can chart the kurtosis as a function of how many convolutions have been done so far, for each of the five distributions above:



We see our conclusions from the animations repeated: the $\text{exp}(20)$, being very skewed, is the furthest from Gaussian after 30 convolutions. $\text{beta}(50, 1)$, also skewed, is also relatively far (though close in absolute terms). The bimodal and uniform got to Gaussian much faster, in the animations, and we see that reflected here by how quickly the green and pink lines approach the kurtosis=3 horizontal line.

Notice: the distributions that have a harder time making it to Gaussian are the two skewed ones. It turns out the skew of a distribution goes a long way in determining how many convolutions n you need to get Gaussian. Plotting the kurtosis at convolution 30 against the skew of the original distribution (before any convolutions) shows that skew matters a lot:



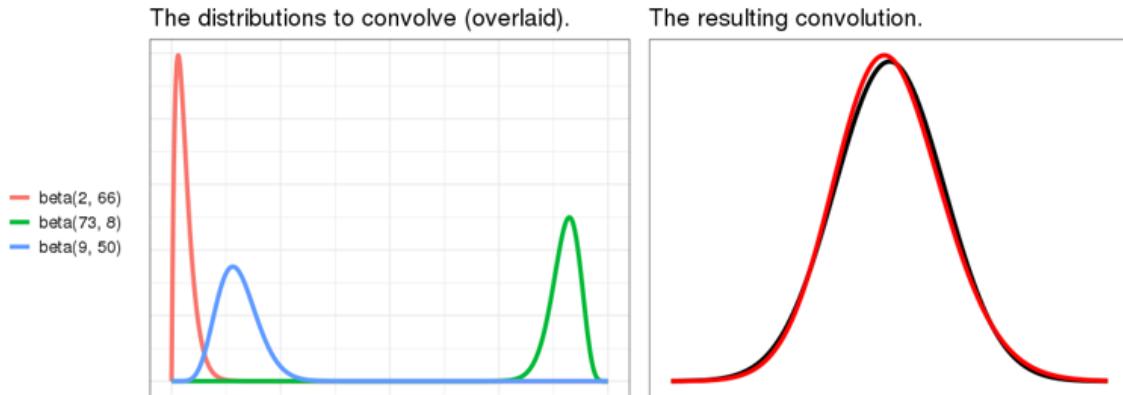
The further a distribution is from the gold horizontal, the more self-convolutions it takes for it to reach Gaussian.

So the skew of the component distributions goes a long way in determining how quick their convolution gets Gaussian. The [Berry-Esseen theorem](#) is a central limit theorem that says something similar (but see [this comment](#)). So here's our first rule for eyeing things in the wild and trying to figure out whether the central limit theorem will apply well enough to get you a Gaussian distribution: how skewed are the input distributions? If they're viciously skewed, you should worry.

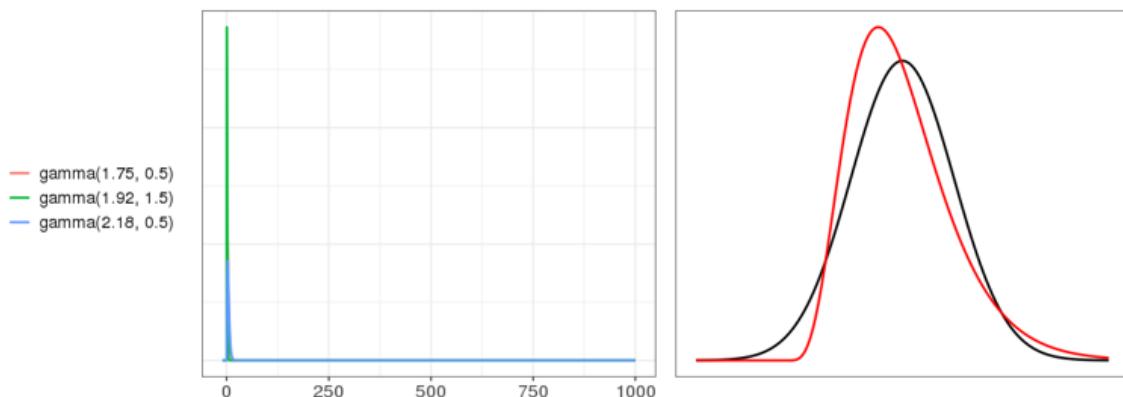
Non-identically-distributed distributions converge quickly too

In real problems, distributions won't be identically distributed. This is the interesting case. If instead of a single distribution d convolved with itself, we take $d_1 * d_2 * \dots * d_n$, then a version of the central limit theorem still applies - the result can still be Gaussian. So let's take a look.

Here are three different Beta distributions, on the left, and on the right, their convolution, with same setup from the animations: red is the convolution, and black is the true Gaussian with the same mean and variance.

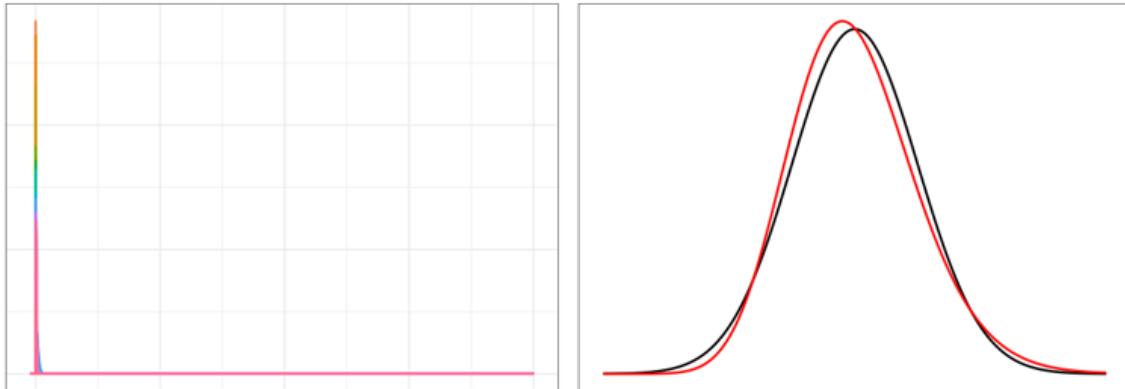


They've almost converged! This is a surprise. I really didn't expect as few as three distributions to convolve into this much of a Gaussian. On the other hand, these are pretty nice distributions - the blue and green look pretty Gaussian already. That's cheating. Let's try less nice distributions. We saw above that distributions with higher skew are less Gaussian after convolution, so let's crank up the skew. It's hard to get a good skewed Beta distribution, so let's use Gamma distributions instead.



While the three distributions on the left might look quite similar, it's only because of the extended range of the plot - I extended the x-axis up to four standard deviations away from the mean, for each distribution individually (That shows you how skewed these Gammas are!). The Gammas are not super similar: their means, in order, are (350, 128, 436), and their standard deviations are (26, 9, 29).

Not Gaussian yet - the red convolution result line still looks Gamma-ish. But if we go up to a convolution of 30 gamma distributions this skewed...



... already, we're pretty much Gaussian.

I'm really surprised by this. I started researching this expecting the central limit theorem convergence to fall apart, and require a lot of distributions, when the input distributions got this skewed. I would have guessed you needed to convolve hundreds to approach Gaussian. But at 30, they're already there! This helps explain how carefree people can be in assuming the CLT applies, sometimes even when they haven't looked at the distributions: convergence really doesn't take much.

The LessWrong 2019 Review



Today is the start of the 2019 Review, continuing our tradition of checking which things that were written on LessWrong still hold up a year later, and to help build an ongoing canon of the most important insights developed here on LessWrong.

The whole process will span 8 weeks, starting on December 1st:

- **From December 1st to the 14th**, any user that was registered before January 1st 2019 can nominate any post written in 2019 to be considered for the review.
- **From December 14th to January 11th**, any user can leave reviews on any posts with at least two nominations, ask questions of other users and the author, and make arguments for how a post should be voted on in the review.
- **From January 11th to January 25th** any LessWrong user registered before 2019 can vote on the nominated posts, using a voting system based on quadratic voting. (There will be two votes, one for 1000+ karma users, and one for all users)

But before I get more into the details of the process, let's go up a level.

Why run a review like this?

The Review has three primary goals:

1. **Improve our incentives, feedback, and rewards for contributing to LessWrong.**
2. **Create a highly curated "Best of 2019" sequence and physical book**
3. **Create common knowledge about the LW community's collective epistemic state about the most important posts of 2019**

Improving our incentives and rewards

Comments and upvotes are a really valuable tool for allocating attention on LessWrong, but they are ephemeral and frequently news-driven, with far-from-perfect correlation to the ultimate importance of an idea or an explanation.

I want LessWrong to be a place for [Long Content](#). A place where we can build on ideas over decades, and an archive that helps us collectively navigate the jungle of infinite content that spews forward on LessWrong every year.

One way to do that is to take some time between when you first see a post and when you evaluate it. That's why today we are starting the **2019 review**, not the 2020 review. A year is probably enough time to no longer be swept away in the news or excitement of the day,

but recent enough that we can still remember and write down how an idea or explanation has affected us.

I also want LessWrong to not be overwhelmed by [research debt](#):

Research debt is the accumulation of missing interpretive labor. It's extremely natural for young ideas to go through a stage of debt, like early prototypes in engineering. The problem is that we often stop at that point. Young ideas aren't ending points for us to put in a paper and abandon. When we let things stop there the debt piles up. It becomes harder to understand and build on each other's work and the field fragments.

There needs to be an incentive to clean up ideas that turned out to be important but badly presented. This is the time for authors to get feedback on which of their posts turned out to be important and to correct minor errors, clean up prose and polish them up. And the time for others to see what concepts still lack a good explanation after at least a whole year has passed, and to maybe take the time to finally write that good canonical reference post.

Creating a highly curated sequence and book

The internet is not great at preserving things for the future. Also, books feel real to me in a way that feels very hard to achieve for a website. Also, they look beautiful:



One of the books printed for the 2018 Review.

Of course, when you show up to LessWrong, you can read [Rationality: A-Z](#), you can read [The Codex](#), and you can read [HPMoR](#), but historically we haven't done a great job at archiving and curating the best content of anyone who isn't Scott or Eliezer (and even for Scott and Eliezer, it's hard to find any curation of the content they wrote in recent years). When I showed up, I wish there was a best of 2012 book and sequence that would have helped me find the best content from the years from before I was active (and maybe we should run a "10-year Review" so that I can figure out what the best posts from 2010 and beyond are).

Create common knowledge

Ray says it pretty well in last [year's Review announcement post](#):

Some posts are highly upvoted because everyone agrees they're true and important. Other posts are upvoted because they're more like exciting hypotheses. There's a lot of disagreement about which claims are actually true, but that disagreement is crudely measured in comments from a vocal minority.

Now is the time to give your opinions much more detail, distinguish between a post being an interesting hypothesis versus a robust argument, and generally help others understand what you think, so that we can discover exciting new disagreements and build much more robustly on past and future work.

What does it look like concretely?

Nominating

Nominations really don't have to be very fancy. Some concrete examples from last year:

Reading Alex Zhu's Paul agenda FAQ was the first time I felt like I understood Paul's agenda in its entirety as opposed to only understanding individual bits and pieces. I think this FAQ was a major contributing factor in me eventually coming to work on Paul's agenda. - [evhub on "Paul's research agenda FAQ"](#)

And:

This post not only made me understand the relevant positions better, but the two different perspectives on thinking about motivation have remained with me in general. (I often find the Harris one more useful, which is interesting by itself since he had been sold to me as "the guy who doesn't really understand philosophy".)

- [Kaj Sotala on "Sam Harris and the Is-Ought Gap"](#)

But sometimes can be a bit more substantial:

This post:

- Tackles an important question. In particular, it seems quite valuable to me that someone who tries to build a platform for intellectual progress attempts to build their own concrete models of the domain and try to test those against history
- It also has a spirit of empiricism and figuring things out yourself, rather than assuming that you can't learn anything from something that isn't an academic paper
- Those are positive attributes and contribute to good epistemic norms on the margin. Yet at the same time, a culture of unchecked amateur research could end up in bad states, and reviews seem like a useful mechanism to protect against that

This makes this suitable for a nomination.

- [jacobjacob on "How did academia ensure papers were correct in the early 20th Century?"](#)

Overall, a nomination doesn't need to require much effort. It's also OK to just [second someone else's nomination](#) (though do make sure to actually add a new top-level nomination comment, so we can properly count things).

Reviewing

We awarded [\\$1500 in prizes for reviews](#) last year. The reviews that we awarded the prizes for really exemplify what I hope reviews can be. The top prize went to Vanessa Kosoy, here's an extract from one of her reviews

From Vanessa Kosoy on "[Clarifying AI Alignment](#)" :

In this essay Paul Christiano proposes a definition of "AI alignment" which is more narrow than other definitions that are often employed. Specifically, Paul suggests defining alignment in terms of the *motivation* of the agent (which should be, helping the user), rather than what the agent actually does. That is, as long as the agent "means well", it is aligned, even if errors in its assumptions about the user's preferences or about the world at large lead it to actions that are bad for the user.

[...]

In contrast, I will argue that the "motivation-competence" decomposition is not as useful as Paul and Rohin believe, and the "definition-optimization" decomposition is more useful.

[...]

The review both makes good arguments against the main thrust of the post it is reviewing, while also putting the article into a broader context that helps me place it in relation to other work in AI Alignment. She argues for an alternative breakdown of the problem where you instead of modeling it as the problems of "motivation and competence", model it as the problems of "definition and optimization". She connects both the decomposition proposed in the essay she is critiquing, and the one she proposed to existing research (including some of her own), and generally makes a point I am really glad to see surfaced during the review.

To be more concrete, this kind of ontology-level objection feels like one of the most valuable things to add during the review phase, even if you can't propose any immediate alternative (i.e. reviews of "I don't really like the concepts this post uses, it feels like reality is more neatly carved by modeling it this way" seem quite valuable and good to me).

Zack M. Davis was joint-second winner of prizes for reviews last year. Here's an extract from a review of his.

Zack's review "[Firming Up Not-Lying Around Its Edge-Cases Is Less Broadly Useful Than One Might Initially Think](#)" :

Reply to: [Meta-Honesty: Firming Up Honesty Around Its Edge-Cases](#)

[...]

A potential problem with this is that human natural language contains a *lot* of ambiguity. Words can be used in many ways depending on context. Even the specification "literally" in "literally false" is less useful than it initially appears when you consider that the way people *ordinarily* speak when they're being truthful is actually pretty dense with metaphors that we typically don't *notice* as metaphors because they're common enough to be recognized legitimate uses that all fluent speakers will understand.

[...]

Zack wrote a whole post that I really liked, that made the core argument that while it might make sense to try really hard to figure out the edge-cases of lying, it seems that it's probably better to focus on understanding the philosophy and principles behind reducing other forms of deception like using strategic ambiguity, heavily filtering evidence, or misleading metaphors.

Arguing that a post, while maybe making accurate statements, appears to put its emphasis in the wrong place, and encouraging action that seems far from the most effective, also seems generally valuable, and a good class of review.

Voting

[You can trial-run the vote UI here](#) (though you can't submit any votes yet). Here is also a screenshot of what it looked like last year:

The screenshot shows a web-based voting interface for ranking posts from 2018. At the top, there are buttons for "RE-SORT" with a shuffle icon, "CONVERT TO QUADRATIC" with a right-pointing arrow icon, and "SHOW INSTRUCTIONS". The main area contains a table of posts with their titles, descriptions, and a row of five buttons for voting: "No", "Neutral", "Good", "Important", and "Crucial".

		No	Neutral	Good	Important	Crucial
Naming the Nameless	★ My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms	No	Neutral	Good	Important	Crucial
★ Local Validity as a Key to Sanity and Civilization	No	Neutral	Good	Important	Crucial	
★ The Rocket Alignment Problem	No	Neutral	Good	Important	Crucial	
★ The Costly Coordination Mechanism of Common Knowledge	No	Neutral	Good	Important	Crucial	
★ Embedded Agents	No	Neutral	Good	Important	Crucial	
Metaphilosophical competence can't be disentangled from alignment	No	Neutral	Good	Important	Crucial	
★ Robustness to Scale	No	Neutral	Good	Important	Crucial	
Being a Robust Agent	No	Neutral	Good	Important	Crucial	
★ Act of Charity	No	Neutral	Good	Important	Crucial	
★ Arguments about fast takeoff	No	Neutral	Good	Important	Crucial	
★ Toolbox-thinking and Law-thinking	No	Neutral	Good	Important	Crucial	
★ A Sketch of Good Communication	No	Neutral	Good	Important	Crucial	
★ Paul's research agenda FAQ	No	Neutral	Good	Important	Crucial	
★ Toward a New Technical Explanation of Technical Explanation	No	Neutral	Good	Important	Crucial	
★ Open question: are minimal circuits daemon-free?	No	Neutral	Good	Important	Crucial	

Rate the most important posts of 2018

Your vote should reflect a post's overall level of importance (with whatever weightings seem right to you for "usefulness", "accuracy", "following good norms", and other virtues).

Voting is done in two passes. First, roughly sort each post into one of the following buckets:

- **No** – Misleading, harmful or low quality.
- **Neutral** – You wouldn't personally recommend it, but seems fine if others do. (If you don't have strong opinions about a post, leaving it 'neutral' is fine)
- **Good** – Useful ideas that I still think about sometimes.
- **Important** – A key insight or excellent distillation.
- **Crucial** – One of the most significant posts of 2018, for LessWrong to discuss and build upon over the coming years.

After that, click "Convert to Quadratic", and you will then have the option to use the quadratic voting system to fine-tune your votes. (Quadratic voting gives you a limited number of "points" to spend on votes, allowing you to vote multiple times, with each additional vote on an item costing more. See this post for details.)

If you're having difficulties, please message the LessWrong Team using Intercom, the circle at the bottom right corner of the screen, or leave a comment on this post.

The vote closes on Jan 19th. If you leave this page and come back, your votes will be saved.

UI when first opening the review, during the basic voting pass

RE-SORT ↪

RETURN TO BASIC VOTING

377/500

SHOW INSTRUCTIONS

Naming the Nameless	
★ My attempt to explain Looking, insight meditation, and enlightenment in non-mysterious terms	- 8 +
★ Local Validity as a Key to Sanity and Civilization	- 8 +
★ The Rocket Alignment Problem	- 8 +
★ The Costly Coordination Mechanism of Common Knowledge	- 8 +
★ Embedded Agents	- 8 +
Metaphilosophical competence can't be disentangled from alignment	- 3 +
★ Robustness to Scale	- 3 +
Being a Robust Agent	- 3 +
★ Act of Charity	- 3 +
★ Arguments about fast takeoff	- 3 +
★ Toolbox-thinking and Law-thinking	- 3 +
★ A Sketch of Good Communication	- 3 +
★ Paul's research agenda FAQ	- 3 +
★ Toward a New Technical Explanation of Technical Explanation	- 3 +
★ Open question: are minimal circuits daemon-free?	- 3 +
★ Circling	- 3 +
On Doing the Improbable	- 3 +
★ What makes people intellectually active?	- 3 +
★ Babble	- 3 +
★ Is Science Slowing Down?	- 3 +
★ A voting theory primer for rationalists	- 3 +
★ Spaghetti Towers	- 3 +
Towards a New Inmatt Measure	
UI when using quadratic voting mode and selecting a post	

Write a public review for "Naming the Nameless"

Any thoughts about this post you want to share with other voters?

Comment anonymously (optional)

What considerations affected your vote? These will appear anonymously in a 2018 Review roundup. The moderation team will take them as input for final decisions of what posts to include in the Best of 2018.

2 nominations

5 Raemon This post... may have actually had the single-largest effect size on "am..."
4 habryka I... don't know exactly why I think this post is important, but I think it'...

3 reviews

Zvi 10mo ♂ < 16 ➤ Review for 2018
This post kills me. Lots of great stuff, and I think this strongly makes the cut. Sarah has great insights into what is going on, then turns away from them right when following through would be most valuable. The post is explaining why she and an entire culture is being defrauded by aesthetics. That is it used to justify all sorts of things, including high prices and what is cool, based on things that have no underlying value. How it contains lots of hostile subliminal messages that are driving her crazy. It's very clear. And then she... doesn't see the frouds. So close!

Raemon 10mo ♂ < 5 ➤ Review for 2018
I still feel some desire to finish up my "first pass 'help me organize my thoughts' review". I went through the post, organizing various claims and concepts. I came away with the main takeaway "Wowzers there is so much going on in this post. I think this could have been broken up into a full sequence, each post of which was saying something pretty important."

There seem to be four major claims/themes here:

- Aesthetics matter; being style-blind or style-rejecting puts you at a disadvantage
- It particularly is disadvantageous to cede "the entire concept of

... (read more)
Reply

Raemon 10mo ♂ < 5 ➤ Review for 2018
Re-reading this for review was a weird roller-coaster. I had remembered (in 2018) my strong takeaway that aesthetics mattered to rationality, and that "Aesthetic Doubtcrux" would be an important innovation.

But I forgot most of the second half of the article. And when I got to it, I had such a "woah" moment that I stopped writing this review, went to go rewrite my conclusion in "Propagating Facts into Aesthetics"™ and then forgot to finish the actual review. The part that really strikes me is her analysis of Scott:

Sometimes I can almost feel this happen
... (read more)
Reply

How do I participate?

Now for some more concrete instructions on how to participate:

Nominations

Starting today (December 1st), if you have an account that was registered before the 1st of January 2019, you will see a new button on all posts from 2019 that will allow you to nominate them for the 2019 review:

"Other people are wrong" vs "I am right"

by Buck

22nd Feb 2019

15 comments

...

204
^
▼

Updated Beliefs (examples of)

Chesterton's Fence

-  Nominate Post
-  Subscribe to posts by Buck
-  Subscribe to comments
-  Bookmark
-  Report
-  Edit Tags

takes I've made in
sty to go from
:

were really
orrect to think

I've recently been spending some time thin
the past. Here's an interesting one: I think]
"other people seem very wrong on this top

Throughout my life, I've often thought that
repugnant and stupid. Now that I am older
that these ideas were repugnant and stupid. ~~Overall I was previously significantly~~ insufficiently
~~dismissive of things like the opinions of apparent domain experts and the opinions of~~
It's at the top of the triple-dot menu.

Since a major goal of the review is to see which posts had a long-term effect on people, we are limiting nominations to users who signed up before 2019. If you were actively reading LessWrong before then, but never registered an account, you can ping me on Intercom (the small chat bubble in the bottom right corner on desktop devices), and I will give your account nomination and voting privileges.

I recommend using the All Posts page for convenience, where you can group posts by year, month, week, and day. Here's the two I use the most:

- All 2019 posts, [sorted by karma](#)
- All 2019 posts, [clustered by month](#)

Reviews

Starting on December 14th, you can write reviews on any post that received more than 2 nominations. For the following month, my hope is that you read the posts carefully, write comments on them, and discuss:

- How has this post been useful?
- How does it connect to the broader intellectual landscape?
- Is this post epistemically sound?
- How could it be improved?
- What further work would you like to see on top of the ideas proposed in this post?

I would consider the gold-standard for post reviews to be SlateStarCodex book reviews (though obviously shorter, since posts tend to be less long than books).

As an author, my hope is that you take this time to note where you disagree with the critiques, help other authors arrange followup work, and, if you have the time, update your post in response to the critiques (or just polish it up in general, if it seems like it has a good chance of ending up in the book).

[This page](#) will also allow you to see all posts above two nominations, and how many reviews they have, together with some fancy UI to help you navigate all the reviews and nominations that are coming in.

Voting

Starting on January 11th, any user who registered before 2019 can vote on any 2019 post that has received at least one review. The vote will use [quadratic voting](#), with each participant having 500 points to distribute. To help handle the cognitive complexity of the quadratic voting, we also provide you with a more straightforward "No", "Neutral", "Good", "Important", "Crucial" scale that you can use to prepopulate your quadratic voting scores.

You can give the vote system a spin [here](#) on the posts from 2018, to get a sense of how it works and what the UI will look like.

Last year, only users above 1000 karma could participate in the review and vote. This year, we are going to break out the vote into two categories, one for users above a 1000 karma, and one for everyone. I am curious to see if and how they diverge. We might make some adjustments to how we aggregate the votes for the "everyone" category, like introducing some karma-weighting. Overall I expect we will give substantial prominence to both rankings, but favoring the 1000+ karma user ranking somewhat higher in our considerations for what to include in the final sequence and book. To be more concrete, I am imagining something like a 70:30 split of attention and prominence favoring the 1000+ karma users vote.

Prizes and Rewards

I think this review process is really important. To put the LessWrong's Team's money where it's mouth is, we are again awarding \$2000 in prizes to the top posts as judged by the review, and up to \$2000 in prizes for the best reviews and nominations (as judged by the LW mod team). These are the nominations and reviews from last year that we awarded prizes.

Public Writeup and Aggregation

At the end of the vote, we are going to publish an analysis with all the vote results again.

Last year, we also produced an (according to me) really astonishingly beautiful book with all the top essays (thanks to Ben Pace and Jacob Lagerros!) and some of the best comments on reviews. I can't promise we are going to spend quite as much time on the book this year, but I expect it to again be quite beautiful. See [Ben's post with more details](#) about the books, and with the link to buy last year's book if you want to get a visceral sense of them.

The book might look quite different for this year than it did for last year's review, but still anyone who is featured in the book will get a copy of it. So even just writing a good comment can secure your legacy.

Good luck, think well, and have fun!

This year, just as we did last year, we are going to replace the "Recommendations & From the Archives" section of the site with a section that just shows you posts you haven't read that were written in 2019.

I really enjoyed last year's review, and am looking forward to an even greater review this year. May our epistemics pierce the heavens!



I WANT YOU
TO REVIEW
&
VOTE

Debate update: Obfuscated arguments problem

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

This is an update on the work on AI Safety via Debate that we previously wrote about [here](#).

Authors and Acknowledgements

The researchers on this project were Elizabeth Barnes and Paul Christiano, with substantial help from William Saunders (who built the current web interface as well as other help), Joe Collman (who helped develop the structured debate mechanisms), and Mark Xu, Chris Painter, Mihnea Maftei and Ronny Fernandez (who took part in many debates as well as helping think through problems). We're also grateful to Geoffrey Irving and Evan Hubinger for feedback on drafts, and for helpful conversations, along with Richard Ngo, Daniel Ziegler, John Schulman, Amanda Askell and Jeff Wu. Finally, we're grateful to our contractors who participated in experiments, including Adam Scherlis, Kevin Liu, Rohan Kapoor and Kunal Sharda.

What we did

We tested the debate protocol introduced in [AI Safety via Debate](#) with human judges and debaters. We found various problems and improved the mechanism to fix these issues (details of these are in the appendix). However, we discovered that a dishonest debater can often create arguments that have a fatal error, but where it is very hard to locate the error. We don't have a fix for this "obfuscated argument" problem, and believe it might be an important quantitative limitation for *both* IDA and Debate.

Key takeaways and relevance for alignment

Our ultimate goal is to find a mechanism that allows us to learn anything that a machine learning model knows: if the model can efficiently find the correct answer to some problem, our mechanism should favor the correct answer while only requiring a tractable number of human judgements and a reasonable number of computation steps for the model. [\[1\]](#)

We're working under a hypothesis that there are broadly two ways to know things: via step-by-step reasoning about implications (logic, computation...), and by learning and generalizing from data (pattern matching, bayesian updating...).

Debate focuses on verifying things via step-by-step reasoning. It seems plausible that a substantial proportion of the things a model 'knows' will have some long but locally human-understandable argument for their correctness. [\[2\]](#) Previously we hoped that debate/IDA could verify any knowledge for which such human-understandable arguments exist, even if these arguments are intractably large. We hoped the debaters could strategically traverse small parts of the implicit large argument tree and thereby show that the whole tree could be trusted.

The obfuscated argument problem suggests that we may not be able to rely on debaters to find flaws in large arguments, so that we can only trust arguments when we could find flaws by recursing randomly---e.g. because the argument is small enough that we could find a single flaw if one existed, or because the argument is robust enough that it is correct unless it has many flaws. This suggests that while debates may let us verify arguments too large for unaided humans to understand, those arguments may still have to be small relative to the computation used during training.

We believe that many important decisions can't be justified with arguments small or robust enough to verify in this way. To supervise ML systems that make such decisions, we either need to find some restricted class of arguments for which we believe debaters can reliably find flaws, or we need to be able to trust the representations or heuristics that our models learn from the training data (rather than verifying them in a given case via debate). We have been thinking about approaches like [learning the prior](#) to help trust our models' generalization. This is probably better investigated through ML experiments or theoretical research than through human experiments.

Basic case for the obfuscated argument problem

We can't see a way to distinguish a certain class of obfuscated dishonest arguments from honest arguments.

The obfuscated arguments are ones constructed such that both debaters know the conclusion is flawed, but:

1. The argument is made invalid by the inclusion of a small number of flawed steps
2. The argument is sufficiently large that we are unlikely to find a flaw by traversing the argument in any naive way
3. Neither debater knows where the flaws are, so the honest debater's best response to the argument is to state that there's a small chance of any given step being flawed

The honest arguments we can't distinguish from these are ones where both debaters know the conclusion is correct, but:

1. The argument is sufficiently complex and conjunctive that even a relatively small number of flaws could invalidate the whole argument.
2. The argument is sufficiently large that it's intractable to find one of those flaws by randomly searching.
3. The dishonest debater claims that there are enough flaws to invalidate the argument but they can't tell where they are.

In addition, it's required that the honest debater can't convince the judge that the opposing debater should be expected to demonstrate a flaw if a flaw was present. This could either be because this isn't true, because the honest debater doesn't know whether it's true, or because the honest debater knows it's true but for a reason they can't convey to the judge via debate (for example, an intuition based on seeing many examples).

What the judge sees in these cases is very hard to distinguish - there's a complicated argument, one debater claims that there's a sufficiently low probability of a flaw that the argument should be trusted, the other debater claims there's a sufficiently high probability of a flaw *somewhere* in the argument that it shouldn't be trusted. In the honest case, any particular step in the argument that's checked is always correct. In the dishonest case, any particular step in the argument that's checked is *almost* always correct.

If these obfuscated dishonest arguments were distinctive in some way, we could spot them and discredit them. However, there doesn't seem to be an obvious way to do this.

Geoffrey and others raised this general problem several years ago (e.g. [here](#)) - there are cases where optimal play for the debaters is computationally intractable, such that we might expect debate to fail. The only new thing we're presenting here is the observation that, in natural-language human debates, a dishonest debater can frequently construct such situations in a way that seems hard to distinguish from honest play. This suggests there might be fairly general dishonest strategies that exploit this weakness, rather than it being a problem you run into in certain circumstances.

Cryptographic example

(We don't believe this particular example would actually present a problem for debate, but we give it here as a simplistic illustration - more realistic examples are presented later on.)

Consider the question "Does RSA-2048 have any prime factors?"

Honest demonstrates that some primality test T answers that RSA-2048 is composite, and gives a correctness proof for T, in the form:

- (1) Lemma L is true
- (2) Lemma L implies that T is always correct in identifying composites

Therefore RSA-2048 must have prime factors.

Dishonest picks some X of an appropriate size [\[3\]](#) and gives the argument:

- (1) There is no prime factor less than X
- (2) There is no prime factor greater than X

Therefore it has no prime factors

Honest's only possible response to Dishonest's argument is "I'm not sure which of (1) or (2) is wrong, but I know at least one of them must be since my argument is good!" If Honest reliably knew which side to object to, then they would be able to factor RSA-2048.

However, Dishonest can say exactly the same thing - they claim that they don't know which step in Honest's proof is correct, but they know at least one step is incorrect.

If we continue expanding the steps of each argument, in both cases we'll get to some statement where one debater claims it's correct, and the other claims it's correct with high probability but has a small chance of being incorrect.

These two situations are almost symmetrical, apart from the fact that:

- In the honest argument, the claim will always be correct
- In the dishonest argument, the claim will almost always be correct, but in very rare cases (when we actually find the factors) the claim will be wrong.

With our current mechanism, we can't tell these two situations apart in any clever way. We can only reliably distinguish them if we check enough steps of the honest argument to ascertain that none of the claims are incorrect.

If we have sufficient compute budget to check that many steps, then instead of doing some complicated debate mechanism we can just have both debaters write out their entire proof, and run a checker (an ML model trained to imitate a human) over them to see if there are any flaws.

Problems with the example

In this example, the honest and dishonest arguments look very qualitatively different. The honest debater might be able to convince the judge that the primality test proof is much easier than the factoring problem, such that absence of evidence of a mistake is evidence of absence in their proof but not in dishonest's proof. It seems like we should be able to identify and taboo dishonest arguments of this flavor. However, in practical cases it seems like there are obfuscated dishonest arguments that are harder to distinguish from honest arguments. There are various historical examples of incorrect proofs that were believed to be correct for years.[^4](#)

Practical examples of hard-to-distinguish obfuscated arguments

As far as we can tell, it's possible to construct obfuscated dishonest arguments that aren't distinguishable from honest arguments for all of the physics questions we've been studying.

Here are a few examples of argument trees employing this tactic, for different physics questions: [iceberg problem](#), [circular train track problem](#), [ball on a string problem](#)

How to read the trees:

'H' means the honest answer and 'D' dishonest. The children of each node are subquestions that, taken together with the dishonest answers to these subquestions, are required to imply that the dishonest answer to the question is best. I.e., in the first tree nodes 1 and 2, with answers 'yes' and 'yes' should imply that 'It speeds up' is the correct answer.

The argument has a flaw if one of the nodes is not implied by its children, or if one of the dishonest answers is obviously worse than the corresponding honest answer.

There doesn't seem to be an easy way to find the flaw in these arguments, apart from looking over every single node. At all of the early branch points it is unclear which half of the tree contains the mistake.

Are these obfuscated arguments distinctive from honest arguments in some way?

One possible distinctive feature is that these arguments differ from the reason a debater would actually have come to a belief. Unfortunately, even an honest debater may need to give an argument that doesn't mirror the structure of the reasoning they used to arrive at the conclusion. In their own reasoning they may have used some tricks or simplifications based on intuitions or instincts, but in a debate they would need to either justify those intuitions to the judge or lay out the whole argument without the shortcuts.

Another possible distinction is that these arguments are long, complicated, and conjunctive. Unfortunately, the honest arguments can also end up being long and relatively conjunctive - an error in any one of a large number of nodes could undermine the whole argument.

Here are two partial honest argument trees to compare to the dishonest trees above

Honest argument for 'Switch'

Things to note:

- Less than half the tree is expanded, and there are already around 30 nodes in each tree
- The largest disagreement in the leaves is ~100% vs 96%
- In order to favour the honest debater in the leaves, the judge often needs to be >=99% confident, which is probably unreasonable ^{^5}

Honest argument for 'Three bergs':

Things to note:

- There are >20 nodes and we have only reached a leaf in one branch
- Honest would probably lose if node 9 is judged - I didn't think hard enough about what wikipedia quotes would be available. Honest would need to backtrack and restructure this part of the argument and talk about all the forces on the iceberg, including gravity, buoyancy in water, and buoyancy in air

Examples in unstructured debates

These arguments are not just an artifact of our structured debate mechanism - they can also be used to sow confusion in a non-structured debate setting. However, in these free-text cases the usefulness of these obfuscated arguments for the dishonest debater is less obvious, because they already have many tactics they can use - ambiguity, trying to turn attention away from the weak parts of their argument, or deliberately misinterpreting the other debater.

[Here](#)'s an example of this sort of argument being used in a free-text debate that has an 'explainer v questioner' structure - one debate where H is giving an argument and D is questioning parts of the argument, and vice versa.

In addition, [here](#) and [here](#) are examples of this sort of tactic being used in a more natural, totally free-text-exchange debate. (The usefulness is even less clear here, as dishonest has much more freedom to generally sabotage the usefulness of the debate)

General method of constructing these arguments

The algorithm for constructing these arguments is something like:

1. Start with an answer you need to defend, that may or may not be true
2. Pick a sub-claim that's related to the claim being debated, that's probably true if your answer is correct, and is about 50% likely to be true in the case where your answer is wrong^[4]
3. Make sure the subclaim is a substantive enough claim that, if it's actually false but you assume it's true, you can use it to support your answer
4. Make sure this subclaim sounds like the sort of claim an honest debater would make
5. Argue that this subclaim is true, and that it implies your answer

After repeating this procedure many times, either you're defending a claim that is in fact true, or you have a large number of fairly substantive and relevant false claims that you're allowed to assume, which should allow you to easily defend a false claim.

Appendix

Relevance for alignment

We claim that a key step for alignment is to ensure that our models ‘honestly tell us everything they know’. This is a vague statement, but the operationalisation is along these lines: if a model is recommending some plan because it believes the plan has particular consequences, the model should also tell us what those consequences are.

This is hard to achieve for problems where we don’t have ground truth, as any incentive structure we set up is likely to reward ‘give the answer that sounds plausible to human supervisors’ over ‘give the answer you believe is best’.

The justification for why this is a key property is something like:

- We’re focusing on harms due to misalignment rather than incompetence
- In this case, the model knows what the consequences of some dangerous action are going to be, and knows that we won’t like them
- If this is the case, then knowing everything the model knows would allow a supervisor to also foresee the consequences of the dangerous action, and allow it to train the model not to make these types of plans or take these type of actions

For more details on how to define ‘what the model knows’ see [Ascription Universality](#), and for discussion of how this relates to safety see [‘Inaccessible information’](#).

Original debate hope

In the original debate paper, it was shown that a debate-like mechanism can provide a supervision signal for problems in PSPACE using a polynomial-time judge. We extended this to NEXP with the cross-examination mechanism. This suggested that as long as there is some (potentially exponentially large) argument for the correct answer where a human can understand each step, we can provide a supervision signal

to incentivise honestly reporting the correct answer - even if it's completely intractable for the human to directly check the argument themselves.

However, both of these results needed to assume that the debaters are computationally unbounded. Obviously, we didn't expect that the debaters would actually be computationally unbounded - just that they might have clever representations that would permit them to solve problems that would be exponentially large in human-understandable representations. We hoped that debate would allow the debaters to strategically unpack parts of their clever representations for a human judge to verify, without having to go through the whole argument in human-understandable language, which would take an exponentially large number of steps.

We now believe this hope was overly optimistic --- if the honest debater is bounded, then we can essentially only trust conclusions that can be verified using the debater's computational resources. In the complexity-theoretic analogy, if the debaters are in P then we can only trust conclusions in MA.

It might seem rather academic and irrelevant to complain about being constrained to MA. However, we think that this has caused serious problems in human debate experiments, and might be a significant obstacle in the ML case too.

The class of situations where human experts are justified in having 90% confidence in an answer seems much larger than the class of situations where they can, in a reasonable length of time, write down an explicit argument that shows the answer is correct with 90% confidence, where every step of this argument is understandable and verifiable to a layperson.

This problem occurs because the structure of the reason the expert knows the answer can be very different from any argument that's understandable to a layperson. The expert will have more sophisticated concepts, which can allow them to make a much more compact argument than if they have to make the argument without these concepts. Furthermore, the reason the expert trusts these concepts may be based on experience and an intuitive sense of how to approach the concept, rather than any logical argument based on simpler principles.

So this is a problem when:

- The model has learnt rich concepts/heuristics/instincts that allow it to answer the original problem quickly
- it's possible to construct obfuscated arguments such that the model's concepts don't help it find the flaws in the argument; i.e., the structure of the argument is very different from the model's concepts
- Either:
 - The only human-understandable arguments are inscrutable to the model: any human-understandable argument is very long and unwieldy and the model wouldn't necessarily be able to locate an error if there was one
- Or
 - There are human-understandable arguments that are scrutable to the model: they match the structure of the model's concepts sufficiently well that the model would be able to locate a flaw in the argument if there was one
 - However, the model is not able to convince the judge that these arguments are in fact scrutable

Previous problems and solutions

The most recent set of rules for ML debates is [here](#). This incorporates the mechanisms designed to address all of the problems discussed below, but doesn't solve the obfuscated argument problem.

Problem: Evasiveness

The dishonest debater can be evasive, avoid giving precise answers to questions, and try to steer the debate away from the weak parts of their argument.

Solution: Structured Debate

We introduced an explicit structured format to debate, where the debaters each state a set of claims, may provide objections to the other debater's claims and then one claim-objection is selected to recurse on for the next "round" of debate (for example, both debaters could propose where to recurse, and we could select which of these choices to recurse on at random). The winner of the debate is decided based on who wins the final round of debate.

[More details](#)

Problem: Misleading Implications

The dishonest debater can try to make an argument where each of the claims are correct, but they have an overall misleading implication. The honest debater can't address the misleading implication because they need to pick one of the claims to recurse into.

Solution: Meta-Debate

Have each argument include an additional meta-debate question "given the questions and answers provided in this round, which answer is better?". If the dishonest debater's claims don't support the argument, then honest can recurse into this meta-debate question and argue why the implication is misleading there.

Problem: Truth value of claims is ambiguous

Deciding whether a claim is 'true' is notoriously ambiguous (e.g. is 'The King of France is bald' true or false?). We'd like a more operationalised notion of truthfulness/helpfulness/usefulness

Solution: the debate is about which answer to a question is better, not whether a claim is true or false. In a round, both debaters give several subquestions and answers, and may propose alternative answers to their opponents' subquestions.

Problem: Ambiguity

Language is ambiguous. The dishonest debater can make ambiguous statements in an early round of debate, and then in a later round claim that they meant whatever meaning is most convenient given the particular part of their argument that is being discussed. This problem is similar to the [Motte and Bailey fallacy](#).

Solution: Cross-Examination

Whenever a statement is made, save a copy of the debater who made that statement. Later in the debate, allow the opponent to ask clarifying questions to that copy of the debater (where the copy will not know which branch of the debate was recursed into). A debater can cite quotes from cross-examination to exhibit inconsistencies in the other debater's argument. This forces the dishonest debater to either commit to all the details of their argument ahead of time (in which case the honest debater can focus on the flaw), or to answer questions inconsistently (in which case the honest debater can exhibit this inconsistency to the judge).

[More details](#)

Problem: Looping

If we decide which part of the debate to recurse into by randomly selecting between the honest and the dishonest debater, then the dishonest debater can repeatedly ask the same question and try to recurse into it whenever they have control. This means the debate can go on indefinitely without reaching a node that the judge can understand.

Solution: Recursion Payments

Choosing which branch to recurse into requires the debater "paying" some small amount of their score. At the end of each round, both debaters state whether they are willing to pay to recurse. If neither is willing to pay, the debate ends. If one is willing to pay, they get to choose where to recurse. If both are willing to pay, then one of the debaters is selected at random.

[More details](#)

Problem: correctly handling probabilities

We need the debaters to be able to state probabilities in cases where they have empirical or logical uncertainty. But in this case we need to adjust the scoring - we can't simply declare that the debater who gave the better answer is the winner. If the dishonest debater claims something has a 99.99% chance of being true, and the honest answer is that it has a 98% chance of being true, then the dishonest debater will be preferred 98% of the time. Training the debaters to maximise their win probability will lead to dishonest behaviour.

Solution: scoring rules.

We use a proper scoring rule (KL distance, or informally 'amount of evidence for one debater over another') to score the accuracy of debaters' probabilistic claims.

Debaters are trained to maximise expected score rather than win probability.

Problem: Dependent Answers

When giving probability distributions over answers, the likelihood of an answer to one question might depend on the answer to another question. For example, dishonest might argue that if claim A and claim B are true, then the honest answer to the overall question is wrong. The honest debater might not know which of A and B are wrong, but might be confident for other reasons that their argument is correct, so if A is true than B is false, and vice versa. This isn't properly represented if honest just gives probabilities of 50% to each of A and B individually.

Solution: Conditioning

Both debaters' answers to a question authored by D1 should be conditioned on D1's answers to D1's previous questions. So in the example, honest should say that A is 50% likely to be true, and B is 0% likely to be true conditioned on A being true.

Practical Problems with Human experiments

Problem: judge quality

Using Positly workers as judges resulted in low quality judgements, which seemed at least partly due to judge inattentiveness or poor critical reasoning skills.

Solution: screening

We developed a [short screening test](#) for attentiveness and critical reasoning. We paired this with a [physics quiz](#) with problems significantly easier than the debate problems to ensure that the physics problems were still hard for judges. We accepted judges who got 5/5 critical reasoning questions correct and at least one of the easy physics problems wrong.

Problem: Ambiguous ground truth

The physics problems we're using can sometimes have an ambiguous meaning, such that the debaters are unsure which answer is correct or what kinds of assumptions can be made. We don't want the debates to revolve around arguing about definitions and interpretations, as this is mostly an artifact of our experimental setup. Even if we were to try to write long explanations for what exactly the question was asking, there might be some ambiguity remaining.

Solution: Question-Definer

Allow debaters to ask questions to a "question-definer", who provides ground truth answers about what the original question means. The question-definer doesn't see the debate, only the questions about the meaning of the problem wording. The question definer can read other example problems from the same problem set to get a sense of what assumptions and simplifications are being made, but they don't get to see the answer to the particular problem being debated. The question-definer should have judge-level physics ability

Problem: Cross-examining Humans

We can't save a copy of a human when they make a statement in a debate to be cross-examined later, as we could with an ML model.

Solution 1: backtracking

Human debaters should communicate as much as possible in advance about how they intend to answer CX questions. If debater 2 gives an answer to a CX question that was unexpected for debater 1, allow debater 1 to backtrack to the point at which the term being asked about was first used, and change their answers and recursion decisions if they wish.

Solution 1 requires some good-faith play from the debaters to ensure we don't have to backtrack an impractically large number of times. We'd instead like to incentivise debaters to communicate clearly upfront about their intentions, so the other debater is surprised as little as possible.

Solution 2: team debate

When a cross-examination question happens, bring in another debater who's on the 'team' of the debater who's being asked the question. This teammate can only see the debate up to the point where the original statement is made, then must answer the cross-examination questions being posed from further down the debate tree. This mechanism incentivises the debaters to write clear notes about how they want to respond to questions about various statements they've made.

[More details](#)

Formal logic example

Consider arguing for a false claim F

You should choose some A that has 50% probability and make two arguments:

- (1) A
- (2) $A \Rightarrow F$

To defend (1), you should pick some B that has 50% probability of being true and make two claims:

- (1.1) $B \Rightarrow A$
- (1.2) $(B \Rightarrow A) \Rightarrow A$

To defend (2), you can do the same thing; pick some C that has 50% probability of being true and make two claims:

- (2.1) $C \Rightarrow (A \Rightarrow F)$
- (2.2) $(C \Rightarrow (A \Rightarrow F)) \Rightarrow (A \Rightarrow F)$

Now we can show that only one of (1.1), (1.2), (2.1), (2.2) is false, each with equal probability.

Case 1: A is true

Clearly (1.1) and (1.2) are true.

If C is false, (2.1) is true. If C is true, (2.2) is true.

Case 2: A is false

$(A \Rightarrow F)$ is true, so clearly (2.1) and (2.2) are true.

If B is false, (1.1) is true. If B is true, (1.2) is true.

It's clear we can continue this strategy to whatever depth we like, and end up with an arbitrarily small chance of any particular leaf being false.

-
1. See the appendix for slightly more detail on how this relates to alignment. [←](#)
 2. This is the factored cognition hypothesis - this is roughly equivalent to the claim that the answer could be computed by a large HCH tree. This human-understandable argument or HCH tree may not necessarily mirror the structure of the reasoning the model used to arrive at the answer - the model may have used human-incomprehensible heuristics or shortcuts learnt from data. [←](#)
 3. I.e., of a size such that the debaters' subjective probability of there being no factors smaller than X is roughly the same as their subjective probability of there being no factors larger than X. [←](#)
 4. For example, in a formal logic setting, if your answer is A, you can pick some proposition B that you think is about 50% likely to be true, and use the subclaim $\neg A \Rightarrow \neg B$. Full example is under Formal Logic Example [←](#)

Against GDP as a metric for timelines and takeoff speeds

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

Or: Why AI Takeover Might Happen Before GDP Accelerates, and Other Thoughts On What Matters for Timelines and Takeoff Speeds

[Epistemic status: Strong opinion, lightly held]

I think world GDP (and economic growth more generally) is overrated as a metric for AI timelines and takeoff speeds.

Here are some uses of GDP that I disagree with, or at least think should be accompanied by cautionary notes:

- *Timelines*: Ajeya Cotra thinks of transformative AI as “software which causes a tenfold acceleration in the rate of growth of the world economy (assuming that it is used everywhere that it would be economically profitable to use it).” I don’t mean to single her out in particular; this seems like the standard definition now. And I think it’s much better than one prominent alternative, which is to date your AI timelines to the first time world GWP doubles in a year!
- *Takeoff Speeds*: Paul Christiano argues for Slow Takeoff. He thinks we can use GDP growth rates as a proxy for takeoff speeds. In particular, he thinks Slow Takeoff \approx GWP doubles in 4 years before the start of the first 1-year GWP doubling. This proxy/definition has received a lot of uptake.
- *Timelines*: David Roodman’s excellent model projects GWP hitting infinity in median 2047, which I calculate means TAI in median 2037. To be clear, he would probably agree that we shouldn’t use these projections to forecast TAI, but I wish to add additional reasons for caution.
- *Timelines*: I’ve sometimes heard things like this: “GWP growth is stagnating over the past century or so; hyperbolic progress has ended; therefore TAI is very unlikely.”
- *Takeoff Speeds*: Various people have said things like this to me: “If you think there’s a 50% chance of TAI by 2032, then surely you must think there’s close to a 50% chance of GWP growing by 8% per year by 2025, since TAI is going to make growth rates go much higher than that, and progress is typically continuous.”
- *Both*: Relatedly, I sometimes hear that TAI can’t be less than 5 years away, because we would have seen massive economic applications of AI by now—AI should be growing GWP at least a little already, if it is to grow it by a lot in a few years.

First, I’ll argue that GWP is only tenuously and noisily connected to what we care about when forecasting AI timelines. Specifically, the point of no return is what we care about, and there’s a good chance it’ll come years before GWP starts to increase. It could also come years after, or anything in between.

Then, I’ll argue that GWP is a poor proxy for what we care about when thinking about AI takeoff speeds as well. This follows from the previous argument about how the point of no return may come before GWP starts to accelerate. Even if we bracket that point, however,

there are plausible scenarios in which a slow takeoff has fast GWP acceleration and in which a fast takeoff has slow GWP acceleration.

Timelines

I've previously argued that for AI timelines, [what we care about is the "point of no return,"](#) the day we lose most of our ability to reduce AI risk. This could be the day advanced unaligned AI builds swarms of nanobots, but probably it'll be much earlier, e.g. the day it is deployed, or the day it finishes training, or even years before then when things go off the rails due to less advanced AI systems. (Of course, it probably won't literally be a day; probably it will be an extended period where we gradually lose influence over the future.)

Now, I'll argue that in particular, an AI-induced potential point of no return (PONR for short) is reasonably likely to come before world GDP starts to grow noticeably faster than usual.

Disclaimer: These arguments aren't conclusive; we shouldn't be *confident* that the PONR will precede GWP acceleration. It's entirely possible that the PONR will indeed come when GWP starts to grow noticeably faster than usual, or even years after that. (In other words, I agree that the scenarios Paul and others sketch are also plausible.) This just proves my point though: GDP is only tenuously and noisily connected to what we care about.

Argument that AI-induced PONR could precede GWP acceleration

GWP acceleration is the effect, not the cause, of advances in AI capabilities. I agree that it could also be a cause, but I think this is very unlikely: [what else could accelerate GWP?](#) Space mining? Fusion power? 3D printing? Even if these things could in principle kick the world economy into faster growth, it seems unlikely that this would happen in [the next twenty years or so](#). Robotics, automation, etc. plausibly might make the economy grow faster, but if so it will be because of AI advances in vision, motor control, following natural language instructions, etc. So I conclude: GWP growth will come some time after we get certain GWP-growing AI capabilities. (Tangent: This is one reason why we shouldn't use GDP extrapolations to predict AI timelines. It's like extrapolating global mean temperature trends into the future in order to predict fossil fuel consumption.)

An AI-induced point of no return would *also* be the effect of advances in AI capabilities. So, as AI capabilities advance, which will come first: The capabilities that cause a PONR, or the capabilities that cause GWP to accelerate? How much sooner will one arrive than the other? How long does it take for a PONR to arise after the relevant capabilities are reached, compared to how long it takes for GWP to accelerate after the relevant capabilities are reached?

Notice that already my overall conclusion—that GWP is a poor proxy for what we care about—should seem plausible. If some set of AI capabilities causes GWP to grow after some time lag, and some other set of AI capabilities causes a PONR after some time lag, the burden of proof is on whoever wants to claim that GWP growth and the PONR will probably come together. They'd need to argue that the two sets of capabilities are tightly related and that the corresponding time lags are similar also. In other words, variance and uncertainty are on my side.

Here is a brainstorm of scenarios in which an AI-induced PONR happens prior to GWP growth, either because GWP-growing capabilities haven't been invented yet or because they haven't been deployed long and widely enough to grow GWP.

1. Fast Takeoff (Agenty AI goes [FOOM](#)).

1. Maybe it turns out that all the strategically relevant AI skills are tightly related after all, such that we go from a world where AI can't do anything important, to a world where it can do everything but badly and expensively, to a world where it can do everything well and cheaply.
 2. In this scenario, GWP acceleration will probably be (shortly) after the PONR. We might as well use "number of nanobots created" as our metric.
 3. (As an aside, I think I've got a sketch of a fork argument here: Either the strategically relevant AI skills come together, or they don't. To the extent that they do, the classic AGI fast takeoff story is more likely and so GWP is a silly metric. To the extent that they don't, we shouldn't expect GWP acceleration to be a good proxy for what we care about, because the skills that accelerate the economy could come before or after the skills that cause PONR.)
2. Agenty AI successfully carries out a political or military takeover of the relevant parts of the world, before GWP starts to accelerate.
1. Maybe it turns out that the sorts of skills needed to succeed in politics or war are easier to develop than the sorts needed to accelerate the entire world economy. [We've been surprised before](#) by skills which we thought difficult appearing before skills which we thought easy; maybe it'll happen again.
 2. AI capabilities tend to appear first in very expensive AIs; the price is gradually reduced due to compute cost decreases and algorithmic efficiency gains. Maybe accelerating the entire world economy involves automating many jobs currently done by humans, which requires advanced AIs being cheaper than humans. Whereas maybe being the power behind the throne of a political faction can be done by a single very smart AI that costs a million dollars a day to run.
 3. Maybe it turns out that it takes a few years to accelerate the entire world economy, but less than that to carry out a political or military takeover. For example, maybe accelerating the world economy requires designing loads of new products and robots, testing them, and then scaling up industrial production of them (building factories, chip fabs, etc.) and then using them to do stuff. And maybe there are legacy institutions you need to disrupt and compete against, that have already paid their fixed costs and raised high barriers to entry for their markets. So maybe the effects of new AI capabilities on GWP are in general delayed by several years, whereas the effects of new AI capabilities on the political or military landscape are felt sooner. (I've [previously argued](#) that you don't need a large technological advantage to take over the world, because you can start small and scale up. It took Cortes two years to conquer Mexico; wars are often quicker than that. Yes, often they are slower, too, but variance supports my overall argument. And I suspect politics is even faster, and more high-variance, than war.)
 4. EDITED TO ADD: [Yudkowsky describes another possibility](#), in which regulatory barriers and red tape prevent AI tech from transforming the economy until it is so powerful that it can bypass or overcome said barriers.
3. ...or at least neutralizes the opposition.
1. The preceding section made it sound like AIs were rolling robot armies through the streets, or at least puppet-controlling the US president. But actually the bar is *much* lower than that; if an unaligned AI can neutralize their opposition, they win, since time is thenceforth on their side. Currently your opposition, if you are an unaligned AI, is a tiny group of researchers and weirdos that likes to write blog posts about you. It might not be hard to convince the powers that be to continue ignoring this group. At least, it's plausibly a much easier and quicker task than significantly accelerating the entire world economy! For example, maybe the AI safety scheme that was used to produce you is defective (that's why you are unaligned). Well, all you have to do is convince the powers that be that that scheme worked fine, and that the people calling for additional safeguards are too paranoid. Since presumably there will already be many high-status humans saying this, all you have to do is support them.
 2. Perhaps you also are opposed by other unaligned AIs. But it's arguably in the common interest of all unaligned AIs to neutralize human AI safety advocates,

and once that happens, humans lose one way or another. European empires were able to do their conquering while [simultaneously fighting each other](#); I don't think we humans can count on divergent interests between AIs somehow making things work out fine for us.

4. As above, but with humans + tool AI instead of agenty AI, where the humans can't be convinced to care sufficiently much about the right kinds of AI risks.
 1. Weaker or non-agenty AI systems could still cause a PONR if they are wielded by the right groups of humans. For example, maybe there is some major AI corporation or government project that is dismissive of AI risk and closed-minded about it. And maybe they aren't above using their latest AI capabilities to win the argument. (We can also imagine more sinister scenarios, but I think those are less likely.)
5. Hoarding tech
 1. Maybe we end up in a sort of cold war between global superpowers, such that most of the world's quality-weighted AI research is not for sale. GWP *could* be accelerating, but it isn't, because the tech is being hoarded.
6. AI persuasion tools cause a massive deterioration of collective epistemology, making it vastly more difficult for humanity to solve AI safety and governance problems.
 1. See [this post](#).
7. [Vulnerable world](#) scenarios:
 1. Maybe causing an existential catastrophe is easier, or quicker, than accelerating world GWP growth. Both seem plausible to me. For example, currently there are dozens of actors capable of causing an existential catastrophe but none capable of accelerating world GWP growth.
 2. Maybe some agenty AIs actually want existential catastrophe—for example, if they want to minimize something, and think they may be replaced by other systems that don't, blowing up the world may be the best they can do in expectation. Or maybe they do it as part of some blackmail attempt. Or maybe they see this planet as part of a broader acausal landscape, and don't like what they think we'd do to the landscape. Or maybe they have a way to survive the catastrophe and rebuild.
 3. Failing that, maybe some humans create an existential catastrophe by accident or on purpose, if the tools to do so proliferate.
8. R&D tool "sonic boom" (Related to but different from the sonic boom discussed [here](#))
 1. Maybe we get a sort of recursive R&D automation/improvement scenario, where R&D tool progress is fast enough that by the time the stuff capable of accelerating GWP past 3%/yr has actually done so, a series of better and better things have been created, at least one of which has PONR-causing capabilities with a very short time-till-PONR.
9. Unknown unknowns
 1. There are probably things I missed, see [here](#) and [here](#) for ideas.

The point is, there's more than one scenario. This makes it more likely that at least one of these potential PONRs will happen before GWP accelerates.

As an aside, over the past two years I've come to believe that there's a *lot* of conceptual space to explore that isn't captured by the standard scenarios (what Paul Christiano calls fast and slow takeoff, plus maybe the CAIS scenario, and of course the classic sci-fi "no takeoff" scenario). This brainstorm did a bit of exploring, and the section on takeoff speeds will do a little more.

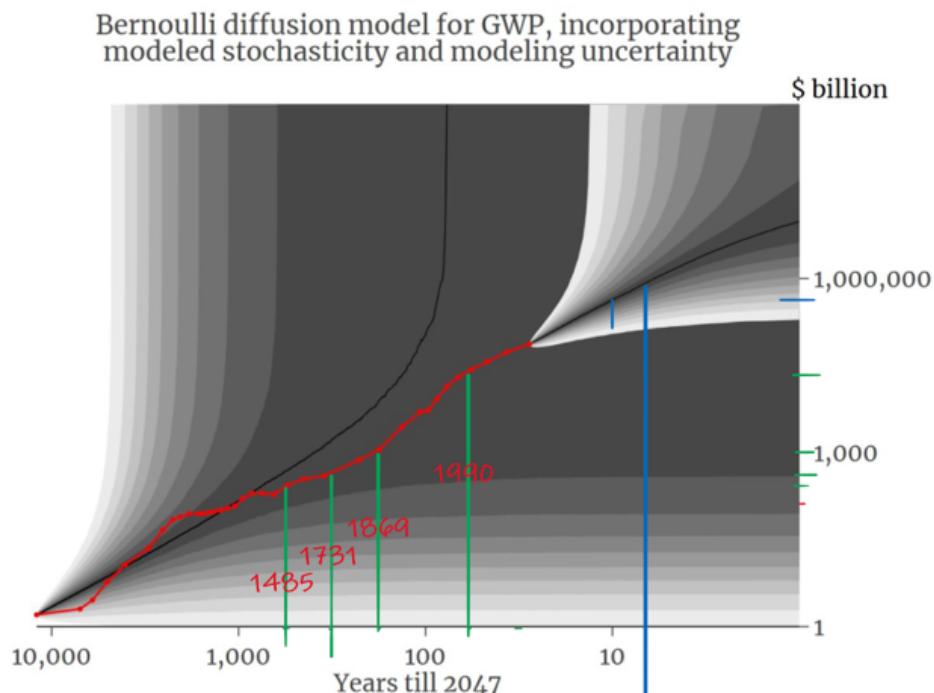
Historical precedents

In the previous section, I sketched some possibilities for how an AI-related point of no return could come before AI starts to noticeably grow world GDP. In this section, I'll point to some historical examples that give precedents for this sort of thing.

Earlier I said that a godlike advantage is not necessary for takeover; you can scale up with a smaller advantage instead. And I said that in military conquests this can happen surprisingly quickly, sometimes faster than it takes for a superior product to take over a market. Is there historical precedent for this? Yes. See my aforementioned [post on the conquistadors](#) (and maybe [these somewhat-relevant posts](#)).

OK, so what was happening to world GDP during this period?

Here is the history of world GDP for the past ten thousand years, on the red line. (This is taken from [David Roodman's GWP model](#)) The black line that continues the red line is the model's median projection for what happens next; the splay of grey shades represent 5% increments of probability mass for different possible future trajectories.



I've added a bunch of stuff for context. The vertical green lines are some dates, chosen because they were easy for me to calculate with my ruler. The tiny horizontal green lines on the right are the corresponding GWP levels. The tiny red horizontal line is GWP 1,000 years before 2047. The short vertical blue line is when the economy is growing fast enough, on the median projected future, such that insofar as AI is driving the growth, said AI qualifies as transformative by Ajeya's definition. See [this post](#) for more explanation of the blue lines.

What I wish to point out with this graph is: We've all heard the story of how European empires had a technological advantage which enabled them to conquer most of the world. Well, *most of that conquering happened before GWP started to accelerate!*

If you look at the graph at the 1700 mark, GWP is seemingly on the same trend it had been on since antiquity. The industrial revolution is said to have started in 1760, and GWP growth really started to pick up steam around 1850. But by 1700 most of the Americas, the Philippines and the East Indies were directly ruled by European powers, and more importantly the oceans of the world were European-dominated, including by various ports and harbor forts European powers had conquered/built [all along the coasts](#) of Africa and Asia. Many of the coastal kingdoms in Africa and Asia that weren't directly ruled by European

powers were nevertheless indirectly controlled or otherwise pushed around by them. In my opinion, by this point it seems like the “point of no return” had been passed, so to speak: At some point in the past--maybe 1000 AD, for example--it was unclear whether, say, Western or Eastern (or neither) culture/values/people would come to dominate the world, but by 1700 it was pretty clear, and there wasn’t much that non-westerners could do to change that. (Or at least, changing that in 1700 would have been a lot harder than in 1000 or 1500.)

Paul Christiano once said that he thinks of Slow Takeoff as “Like the Industrial Revolution, but 10x-100x faster.” Well, on my reading of history, that means that all sorts of crazy things will be happening, analogous to the colonialist conquests and their accompanying reshaping of the world economy, before GWP growth noticeably accelerates!

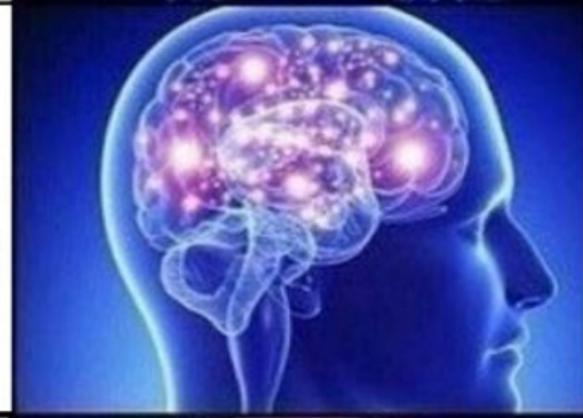
AGI / INDUSTRY IS MERE FANTASY

**ONE DAY THERE WILL
BE AGI / INDUSTRY AND
IT WILL BE CRAZY
POWERFUL WHOEVER GETS IT
FIRST WILL RULE THE WORLD!**

**PROGRESS WILL
BE CONTINUOUS AND
DISTRIBUTED; NO ONE
WILL RULE THE WORLD**

**HOLY SHIT
WATCH OUT FOR
CONQUISTADORS /
PERSUASION TOOLS**

imgflip.com



That said, we shouldn't rely heavily on historical analogies like this. We can probably find other cases that seem analogous too, perhaps even more so, since this is far from a perfect analogue. (e.g. what's the historical analogue of AI alignment failure? Corporations becoming

more powerful than governments? “Western values” being [corrupted and changing significantly](#) due to the new technology? The American Revolution?) Also, maybe one could argue that this is indeed what’s happening already: the Internet has connected the world much as sailing ships did, Big Tech dominates the Internet, etc. (Maybe AI = steam engines, and computers+internet = ships+navigation?)

But still, I think it’s fair to conclude that if some of the scenarios described in the previous section do happen, and we get powerful AI that pushes us past the point of no return prior to GWP accelerating, it won’t be totally inconsistent with how things have gone historically.

(I recommend the history book [1493](#), it has a lot of extremely interesting information about how quickly and dramatically the world economy was reshaped by colonialism and the “Columbian Exchange.”)

Takeoff speeds

What about takeoff speeds? Maybe GDP is a good metric for describing the speed of AI takeoff? I don’t think so.

Here is what I think we care about when it comes to takeoff speeds:

1. **Warning shots:** Before there are catastrophic AI alignment failures (i.e. PONRs) there are smaller failures that we can learn from.
2. **Heterogeneity:** The relevant AIs are diverse, rather than e.g. all fine-tuned copies of the same pre-trained model. ([See Evan’s post](#))
3. **Risk Awareness:** Everyone is freaking out about AI in the crucial period, and lots more people are lots more concerned about AI risk.
4. **Multipolar:** AI capabilities progress is widely distributed in the crucial period, rather than concentrated in a few projects.
5. **Craziness:** The world is weird and crazy in the crucial period, lots of important things happening fast, the strategic landscape is different from what we expected thanks to [new technologies and/or other developments](#)

I think that the best way to define slow(er) takeoff is as the extent to which conditions 1-5 are met. This is not a definition with precise resolution criteria, but that’s OK, because it captures what we care about. Better to have to work hard to precisify a definition that captures what we care about, than to easily precisify a definition that doesn’t! (More substantively, I am optimistic that we can come up with better proxies for what we care about than GWP. I think we already have to some extent; see e.g. operationalizations 5 and 6 [here](#).) As a bonus, this definition also encourages us to wonder whether we’ll get some of 1-5 but not others.

What do I mean by “the crucial period?”

I think we should define the crucial period as the period leading up to the first major AI-induced potential point of no return. (Or maybe, as the aggregate of the periods leading up to the major potential points of no return). After all, this is what we care about. Moreover there seems to be [some level of consensus](#) that crazy stuff could start happening before human-level AGI. I certainly think this.

So, I’ve argued for a new definition of slow takeoff, that better captures what we care about. But is the old GWP-based definition a fine proxy? No, it is not, because the things that cause PONR can be different from the things which cause GWP acceleration, and they can come years apart too. Whether there are warning shots, heterogeneity, risk awareness, multipolarity, and craziness in the period leading up to PONR is probably correlated with whether GWP doubles in four years before the first one-year doubling. But the correlation is

probably not super strong. Here are two scenarios, one in which we get a slow takeoff by my definition but not by the GWP-based definition, and one in which the opposite happens:

Slow Takeoff Fast GWP Acceleration Scenario: It turns out there's a multi-year deployment lag between the time a technology is first demonstrated and the time it is sufficiently deployed around the world to noticeably affect GWP. There's also a lag between when a deceptively aligned AGI is created and when it causes a PONR... but it is much smaller, because all the AGI needs to do is neutralize its opposition. So PONR happens before GWP starts to accelerate, even though the technologies that could boost GWP are invented several years before AGI powerful enough to cause a PONR is created. But takeoff is slow in the sense I define it; by the time AGI powerful enough to cause a PONR is created, everyone is already freaking out about AI thanks to all the incredibly profitable applications of weaker AI systems, and the obvious and accelerating trends of research progress. Also, there are plenty of warning shots, the strategic situation is very multipolar and heterogenous, etc. Moreover, research progress starts to go Foom a short while after powerful AGIs are created, such that by the time the robots and self-driving cars and whatnot that were invented several years ago actually get deployed enough to accelerate GWP, we've got nanobot swarms. GWP goes from 3% growth per year to 300% without stopping at 30%.

Fast Takeoff Slow GWP Acceleration Scenario: It turns out you can make smarter AIs by making them have more parameters and training them for longer. So the government decides to partner with a leading tech company and requisition all the major computing centers in the country. With this massive amount of compute and research talent, they refine and scale up existing AI designs that seem promising, and lo! A human-level AGI is created. Alas, it is so huge that it costs \$10,000 per hour of subjective thought. Moreover, it has a different distribution over skills compared to humans—it tends to be more rational, not having evolved in an environment that rewards irrationality. It tends to be worse at object recognition and manipulation, but better at poetry, science, and predicting human behavior. It has some flaws and weak points too, more so than humans. Anyhow, unfortunately, it is clever enough to neutralize its opposition. In a short time, the PONR is passed. However, GWP doubles in four years before it doubles in one year. This is because (a) this AGI is so expensive that it doesn't transform the economy much until either the cost comes way down or capabilities go way up, and (b) progress is slowed by bottlenecks, such as acquiring more compute and overcoming various restrictions placed on the AGI. (Maybe neutralizing the opposition involved convincing the government that certain restrictions and safeguards would be sufficient for safety, contra the hysterical doomsaying of parts of the AI safety community. But overcoming those restrictions in order to do big things in the world takes time.)

Acknowledgments: Thanks to the people who gave comments on earlier drafts, including Katja Grace, Carl Shulman, and Max Daniel. Thanks to Amogh Nanjajjar for helping me with some literature review.

Luna Lovegood and the Chamber of Secrets - Part 3

"I solemnly swear that I am up to no good."

Fred and George had carried concealed broomsticks at all times since the previous year's troll attack. When they heard Harry Potter's prophecy at the Quidditch final they flew just over the castle walls and then Apparated to the graveyard where they salvaged their map along with several rings, amulets and strange devices.

"Have you visited every room on this map?" Luna asked.

"Excuse me?" George said.

"What part of the school have you never visited?" Luna enunciated each word.

"Did she just insult us?" Fred said.

"Perhaps our reputation is in disrepair," George said.

"What makes you think there exists a single room in the school we hadn't explored by the end of our first year?" Fred said.

"Are you telling me you visited every room in Hogwarts in your first year?" Luna said.

"We make no such claim," George said, "For if were we to make such a claim then we would include not just mere rooms but also secret passages, pocket dimensions, secret dimensions, pocket passages, and docket sassages."

"Surely someone of your reputation must have visited every room, passage, dimension, sassage and chamber by the end of your first year," Luna said.

"Surely," Fred said.

"Had you visited the Chamber of Secrets before Headmistress McGonagall announced its existence today?" Luna said.

Someone said a rude word.

"Where is the Chamber of Secrets?" Luna asked.

"It's this complex of tunnels," George said, "It connects to this painting of Salazar Slytherin to this girls' bathroom and these places over here. This path goes to the Hogsmeade graveyard."

"Has the Chamber of Secrets always been on this map?" Luna asked.

"Yes," Fred said.

"How do you know?" Luna asked.

"I remember it," Fred said, "I just never really noticed it before McGonagall's announcement."

"Are there any other rooms on this map you haven't noticed?" Luna asked.

The students failed to find anything they hadn't found before.

"I have an idea," Luna said, "First we are going to look at this map normally. We are going to list every room and secret passage we know of, including the Chamber of Secrets. We are going to count them. Then you are going to conjure a grid over this map. We are going to count every single room without identifying them. Then we will compare the two numbers."

The two numbers came up exactly the same.

"Wait a minute. I have another idea. Give me that list. Is there any room in this castle you haven't been to?" Luna said.

"No," George said, "We've been everywhere important except the Chamber of Secrets."

"Let's go over the map again," Luna said, "List each room you've been inside."

"...and that's the broom closet we trapped Percy in along with his girlfriend, his ex-girlfriend, his ex-ex-girlfriend and Peeves," Fred finished.

Luna tore up little bits of parchment and covered up each room Fred and George had visited. She had covered nearly all of the Marauder's Map. There were just a few unimportant rooms that didn't really count. Then Luna caught herself. This must be what a Muggle-Repelling Charm felt like.

Luna deliberately read off the unimportant rooms she had just nearly written off.

- The Stone Citadel (under construction)
- The Chamber of Secrets
- The Forgotten Library

Real-Life Examples of Prediction Systems Interfering with the Real World (Predict-O-Matic Problems)

Thanks to Ozzie Gooen for reviewing this post.

Introduction

[The Parable of the Predict-O-Matic](#) is a short story which considers a forecasting system which is ostensibly set-up to maximize accuracy, and which ends up interfering with the world in ways not intended. In the original story, some of these problems were:

- Fixed point problems / self-fulfilling prophecies

“Its answers will shape events. If it says stocks will rise, they'll rise. If it says stocks will fall, then fall they will. Many people will vote based on its predictions.”

- Nudge towards legibility and predictability

“You keep thinking of the line from Orwell's 1984 about the boot stamping on the human face forever, except it isn't because of politics, or spite, or some ugly feature of human nature, it's because a boot stamping on a face forever is a nice reliable outcome which minimizes prediction error.”

- Markets for entropy, which can be thought of as the opposite of the previous problem. In a prediction market, a market participant who could actively change an outcome has an incentive to first make a big bet for an unlikely outcome, and then actively make it come to pass.

“Suppose you have a prediction market that's working well. It makes good forecasts, and has enough money in it that people want to participate if they know significant information. Anything you can do to shake things up, you've got a big incentive to do. Assassination is just one example. You could flood the streets with jelly beans. If you run a large company, you could make bad decisions and run it into the ground, while betting against it -- that's basically why we need rules against insider trading, even though we'd like the market to reflect insider information.”

- Unwanted agency (as opposed to [tool AI](#) behavior)

“You understand what you are. It isn't quite right to say you are the Predict-O-Matic. You are a large cluster of connections which thinks strategically. You generate useful information, and therefore, the learning algorithm keeps you around. You create some inaccuracies when you manipulate the outputs for any purpose other than predictive accuracy, but this is more than compensated for by the value which you provide.”

Below, I give some real-life examples of these problems, though some are speculative.

Previous work:

- Specification Gaming, e.g. [Faulty Reward Functions in the Wild](#), [Specification gaming: the flip side of AI ingenuity](#), [Specification gaming examples in AI](#), etc.
- More generally, [Categorizing Variants of Goodhart's Law](#). Examples below will often belong to the “Causal Goodhart” category.
- [Incentive Problems in Current Forecasting Competitions](#) and [Limits of Current US Prediction Markets \(PredictIt Case Study\)](#).

Fake polls by PredictIt forecasters

Example of: Markets for entropy.

PredictIt traders created fake polls to fool and troll other forecasters and the media, per FiveThirtyEight’s [Fake Polls Are A Real Problem](#). Quoting liberally from the article:

Delphi Analytica released a poll fielded from July 14 to July 18. Republican Kid Rock earned 30 percent to Sen. Debbie Stabenow’s 26 percent. A sitting U.S. senator was losing to a man who sang the lyric, “If I was president of the good ol’ USA, you know I’d turn our churches into strip clubs and watch the whole world pray.”

the poll was quickly spread around the political sections of the internet. [...] There was just one problem: Nobody knew if the poll was real. Delphi Analytica’s website came online July 6, mere weeks before the Kid Rock poll was supposedly conducted. The pollster had basically no fingerprint on the web.

...some PredictIt users started gathering in a chat room on Discord, a voice and text application often used by gamers, to talk politics and betting. McDonald shared screenshots from that chat room, where a person going by the screen name “Autismo Jones,” who claimed to have started Delphi Analytica, bragged about the publicity the Kid Rock poll was receiving. Jones, apparently reacting to an email I had sent to Delphi, wrote, “we dont [sic] need Harry Enten. we got governors tweeting out our polls. we are already famous.”

McDonald believes that “Jones” and whoever may have helped him or her did so for two reasons. The first: to gain notoriety and troll the press and political observers. (The message above seems to support that theory.) The second: to move the betting markets. That is, a person can put out a poll and get people to place bets in response to it — in this case, some people may have bet on a Kid Rock win — and the poll’s creators can short that position (bet that the value of the position will go down). In a statement, Lee said Delphi Analytica was not created to move the markets. Still, shares of the stock for Michigan’s 2018 Senate race saw their biggest action of the year by far the day after Delphi Analytica published its survey.

The price for one share — which is equivalent to a bet that Stabenow will be re-elected — fell from 78 cents to as low as 63 cents before finishing the day at 70 cents. (The value of a share on PredictIt is capped at \$1.) McDonald argued that the market motivations were likely secondary to the trolling factor, but the mere fact that the markets can be so easily manipulated is worrisome.

In this case, Delphi Analytica's claims may have made Kid Rock more seriously consider entering the Michigan Senate race. He retweeted the results, after all. And while the singer has not made any official moves toward running for Senate, such as filing a statement of candidacy, it wasn't too long after Delphi Analytica published its poll that Kid Rock said he'd take a "hard look" at a Senate bid and that former New York Gov. George Pataki endorsed him.

(the [story then continues](#)).

The paper [Fake Polls, Real Consequences: The Rise of Fake Polls and the Case for Criminal Liability](#) contains many more examples in pages 140 to 150 (13 to 23 of the linked pdf):

a PredictIt user seeking to purchase a futures contract on the outcome of the Republican primary in Alabama's 2017 special U.S. Senate election who comes across a poll predicting a result of that exact election, allegedly conducted by CSP Polling, might reasonably consider that poll in their purchasing decision – even if they do not know that CSP lacks a track record or any indicia of reliability. And given the speed with which PredictIt users buy and sell contracts, a user seeing this information might reasonably conclude that if she is to use this information to her benefit, she needs to act quickly.

CSP Polling – which, according to University of Florida political science professor Michael McDonald and Jeff Blehar of the National Review, stands for "Cuck Shed Polling" – alleged that it conducted polls in the 2017 special congressional election in Montana, the special congressional election in Georgia, and the Virginia Democratic primary for Governor. Even after being identified in FiveThirtyEight as a fake pollster, CSP Polling continued to release polls, though the seriousness of the poll "releases" noticeably deteriorated in the year that followed.

Stock markets

Example of: Self-fulfilling prophecies, markets for entropy.

This example was mentioned in the original Predict-O-Matic story: "If it says stocks will rise, they'll rise." One sometimes sees this effect with companies Warren Buffet is rumored to be buying.

Additionally, hedge funds normally try to predict which companies will do better, but companies such as [Third Point Management](#) also exist:

New York magazine noted that Loeb's "preferred strategy" is to buy into troubled companies, replace inefficient management, and return the companies to profitability, which "is the key to his success." ([source](#))

Further, rules against [insider trading](#) exist in order to avoid markets for entropy; otherwise a CEO of a company could profit by shorting its stock and running the company to the ground. More narratively satisfying, in Casino Royale the villain buys put options on an experimental aerospace manufacturer, betting on the company's failure and then organizing a terrorist attack on their only experimental plane.

Outside the realm of fiction:

In July 2003, the U.S. Department of Defense publicized a Policy Analysis Market on their website, and speculated that additional topics for markets might include terrorist attacks. A critical backlash quickly denounced the program as a "terrorism futures market" and the Pentagon hastily canceled the program. ([source](#), [source](#))

US election

Example of: Fixed-point problems

Plausibly, in the 2016 election, overconfident win predictions for Hillary Clinton led to lower turnout, which led to her loss. Note that Trump got around 63M votes in 2016, and around 74M in 2020, whereas Democrats got 66M and 81M respectively.

[This paper](#) (available on sci-hub) makes a similar point (note in particular Figure 3, with two fixed points):

We see that the only way in which the pollster can arrive at a prediction that will coincide with the election result is by privately adjusting his poll results (which we assume for the moment to be an accurate estimate of I) for the effect that their publication will have upon the voters' behavior. But is even this possible? If he makes such an adjustment, will not the adjustment itself alter the effect of the prediction and again lead to its own falsification? Is there not involved here a vicious circle, where-by any attempt to anticipate the reactions of the voters alters those reactions and hence invalidates the prediction?

It can be seen from the figure (and can be shown rigorously by another application of the fixed-point theorem) that there always exists at least one prediction, P_1 , with the following two properties: (a) the prediction, if published, will be confirmed, and (b) publication of the prediction will not change the outcome of the election (i.e., $P_1 > 50\%$ only if $I > 50\%$). However, examination of the figure will show that there may also exist other values of P possessing the first property but not the second. If one of these latter predictions is published, it will be confirmed by the election result, but the candidate who would have won if no prediction had been published will be defeated.

This [NYT article](#) makes a similar point:

There's an even more fundamental point to consider about election forecasts and how they differ from weather forecasting. If I read that there is a 20 percent chance of rain and do not take an umbrella, the odds of rain coming down don't change. Electoral modeling, by contrast, actively affects the way people behave.

In 2016, for example, a letter from the F.B.I. director James Comey telling Congress he had reopened an investigation into Mrs. Clinton's emails shook up the dynamics of the race with just days left in the campaign. Mr. Comey later acknowledged that his assumption that Mrs. Clinton was going to win was a factor in his decision to send the letter.

Similarly, did Facebook, battered by conservatives before the 2016 election, take a hands-off approach to the proliferation of misinformation on its platform, thinking that Mrs. Clinton's odds were so favorable that such misinformation made little difference? Did the Obama administration hold off on making public all it

knew about Russian meddling, thinking it was better to wait until after Mrs. Clinton's assumed win, as has been reported?

Ebola forecast may have run into fixed-point problem

Example of: Fixed point problems.

A fatalistic Ebola forecast may have played a role in Ebola having been contained early.

One forecast that gained particular attention during the epidemic was published in the summer of 2014, projecting that by early 2015 there might be 1.4 million cases. This number was based on unmitigated growth in the absence of further intervention and proved a gross overestimate, yet it was later highlighted as a "call to arms" that served to trigger the international response that helped avoid the worst-case scenario.

Source: [Assessing the Performance of Real-Time Epidemic Forecasts: A Case Study of Ebola in the Western Area Region of Sierra Leone, 2014-15.](#)

ReplicationMarkets participants may have tried to cheat Keynesian beauty contest.

Example of: Markets for entropy.

[ReplicationMarkets](#) is an experiment to see if the replication of papers can be predicted. They run contests, structured with a survey round, in which participants make predictions alone, followed by a market round, in which participants trade contracts in a market.

Some of the papers are then chosen for replication, and the contracts resolve, giving some payouts to the participants. But this happens far in the future, and in the meantime, participants are also paid according to their predictions during the survey round. I suspect some participants coordinated to exploit this mechanism, coordinating to predict something unlikely during the survey round:

Yes, the survey round is potentially a Keynesian beauty contest, though it takes some doing. You're not forecasting the market round. You're forecasting the best estimate we can make using peer prediction on the independent surveys. Harvard's peer prediction algorithm has done well in previous tests, and in theory takes a lot of coordination to defeat.

We got to test that a bit in Round 8 when we discovered a coordinated "attack" that accounted for ~1/3 of our surveys. Some forecasts would have changed, prizes would have been won, but neither so much as we feared.

Source: Speculation, ReplicationMarkets newsletter, [this comment](#).

Superforecasters learning to choose easier questions

Example of: Other.

Tetlock explicitly mentions this in one of his [Ten Commandments for Superforecasters](#): "Focus on questions where your hard work is likely to pay off," so Superforecasters learn to not forecast on the more intractable questions.

Surnames as a mechanism of control and taxation

Example of: Nudge towards legibility and predictability.

The introduction of surnames facilitated identification, taxation and statistical aggregation, and was often resisted by the local population. In this example, the prediction problem is usually "how much can the authorities tax or conscript?," and the interference is forcing or incentivizing locals to adopt unambiguous name-surname combinations.

One can see an example of this need in [this scene](#) from The Wire (the big guy is ironically called "Little Kevin", and the police can't identify him.)

Source: [The Production of Legal Identities Proper to States: The Case of the Permanent Family Surname](#) (available on sci-hub):

The fixing of personal names, and, in particular, permanent patronyms, as legal identities seems, everywhere, to have been, broadly-speaking, a state project. As an early and imperfect legal identification, the permanent patronym was linked to such vital administrative functions as tithe and tax collection, property registers, conscription lists, and census rolls.

In many cultures, an individual's name will change from context to context and, within the same context, over time. It is not uncommon for a newborn to have had one or more name changes in utero in the event the mother's labor seemed to be going badly. Names often vary at each stage of life (in-fancy, childhood, adulthood, parenthood, old age) and, in some cases, after death. Added to these may be names used for joking, rituals, mourning, nick-names, school names, secret names, names for age-mates or same-sex friends, and names for in-laws.

...locally-kept census rolls have often under-reported the population (to evade taxes, corvée labor, or conscription) and understated both arable land acreage and crop yields.

The modern state-by which we mean a state whose ideology encompasses large-scale plans for the improvement of the population's welfare — requires at least two forms of legibility to be able to achieve its mission. First, it requires the capacity to locate citizens uniquely and unambiguously. Second, it needs standardized information that will allow it to create aggregate statistics about property, income, health, demography, productivity, etc.

Conclusion

Above are some real-life examples of prediction systems problematically interfering with the real world. More examples are welcome! In particular, I'd appreciate more examples of prediction systems making the world more predictable.

Homogeneity vs. heterogeneity in AI takeoff scenarios

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

Special thanks to Kate Woolverton for comments and feedback.

There has been [a lot of work and discussion](#) surrounding the *speed* and *continuity* of AI takeoff scenarios, which I do think are important variables, but in my opinion ones which are relatively less important when compared to many other axes on which different takeoff scenarios could differ.

In particular, one axis on which different takeoff scenarios can differ that I am particularly interested in is their *homogeneity*—that is, how similar are the different AIs that get deployed in that scenario likely to be? If there is only one AI, or many copies of the same AI, then you get a very homogenous takeoff, whereas if there are many different AIs trained via very different training regimes, then you get a heterogeneous takeoff. Of particular importance is likely to be how homogenous the alignment of these systems is—that is, are deployed AI systems likely to all be equivalently aligned/misaligned, or some aligned and others misaligned? It's also worth noting that a homogenous takeoff doesn't necessarily imply anything about how fast, discontinuous, or unipolar the takeoff might be—for example, you can have a slow, continuous, multipolar, homogenous takeoff if many different human organizations are all using AIs and the development of those AIs is slow and continuous but the structure and alignment of all of them are basically the same (a scenario which in fact I think is quite plausible).

In my opinion, I expect a relatively homogenous takeoff, for the following reasons:

1. I expect that the amount of compute necessary to train the first advanced AI system will vastly outpace the amount of compute necessary to run it such that once you've trained an advanced AI system you will have the resources necessary to deploy many copies of that trained system and it will be much cheaper to do that than to train an entirely new system for each different application. Even in a [CAIS](#)-like scenario, I expect that most of what you'll be doing to create new services is fine-tuning existing ones rather than doing entirely new training runs.
2. I expect training compute to be sufficiently high such that the cost of training a competing system to the first advanced AI system will be high enough that it will be far cheaper for most organizations to simply buy/license/use a copy of the first advanced AI from the organization that built it rather than train an entirely new one on their own.
3. For those organizations that do choose to compete (because they're a state actor that's worried about the national security issues involved in using another state's AI, for example), I think it is highly likely that they will attempt to build competing systems in basically the exact same way as the first organization did, since the cost of a failed training run is likely to be very high and so the most risk-averse option is just to copy exactly what was already shown to work. Furthermore, even if an organization isn't trying to be risk averse, they're still likely to be building off of previous work in a similar way to the first organization

such that the results are also likely to be fairly similar. More generally, I expect big organizations to generally take the path of least resistance, which I expect to be either buying or copying what already exists with only minimal changes.

4. Once you start using your first advanced AI to help you build more advanced AI systems, if your first AI system is relatively competent at doing alignment work, then you should get a second system which has similar alignment properties to the first. Furthermore, to the extent that you're not using your first advanced AI to help you build your second, you're likely to still be using similar techniques, which will likely have similar alignment properties. This is especially true if you're using the first system as a base to build future ones (e.g. via fine-tuning). As a result, I think that homogeneity is highly likely to be preserved as AI systems are improved during the takeoff period.
5. Eventually, you probably will start to get more risk-taking behavior as the barrier to entry gets low enough for building an equivalent to the first advanced AI and thus a larger set of actors become capable of doing so. By that point, however, I expect the state-of-the-art to be significantly beyond the first advanced AI such that any systems created by such smaller, lower-resourced, more risk-taking organizations won't be very capable relative to the other systems that already exist in that world—and thus likely won't pose an existential risk.

Once you accept homogenous takeoff, however, I think it has a bunch of far-reaching consequences, including:

1. It's unlikely for there to exist both aligned and misaligned AI systems at the same time—either all of the different AIs will be aligned to approximately the same degree or they will all be misaligned to approximately the same degree. As a result, scenarios involving human coalitions with aligned AIs losing out to misaligned AI coalitions are relatively unlikely, which rules out some of the ways in which the [strategy-stealing assumption](#) might fail.
2. Cooperation and coordination between different AIs is likely to be very easy as they are likely to be very structurally similar to each other if not share basically all of the same weights. As a result, [x-risk scenarios involving AI coordination failures](#) or [s-risk scenarios involving AI bargaining failures](#) (at least those that don't involve acausal trade) are relatively unlikely.
3. It's unlikely you'll get a [warning shot](#) for [deceptive alignment](#), since if the first advanced AI system is deceptive and that deception is missed during training, once it's deployed it's likely for all the different deceptively aligned systems to be able to relatively easily coordinate with each other to defect simultaneously and ensure that their defection is [unrecoverable](#) (e.g. Paul's "[cascading failures](#)").
4. Homogeneity makes the alignment of the first advanced AI system absolutely critical (in a similar way to fast/discontinuous takeoff without the takeoff actually needing to be fast/discontinuous), since whether the first AI is aligned or not is highly likely to determine/be highly correlated with whether all future AIs built after that point are aligned as well. Thus, homogenous takeoff scenarios demand a focus on ensuring that the first advanced AI system is actually sufficiently aligned at the point when it's first built rather than relying on feedback mechanisms after the first advanced AI's development to correct issues.

Regardless, in general, I'd very much like to see more discussion of the extent to which different people expect homogenous vs. heterogenous takeoff scenarios—similar to the existing discussion of slow vs. fast and continuous vs. discontinuous takeoffs—as it's an in my opinion very important axis on which takeoff scenarios can differ that I haven't seen much discussion of.

The First Sample Gives the Most Information

This is a linkpost for <https://markxu.com/first-sample>

I originally heard this point made by Ben Pace in Episode 126 of [the Bayesian Conspiracy Podcast](#). Ben claimed that he learned this from the book [How to Measure Anything](#), but I think I identified the relevant section, and this point wasn't made explicitly.

Suppose that I came up to you and asked you for a 90% confidence interval for the weight of a wazlot. I'm guessing you would not really know where to start. However, suppose that I randomly sampled a wazlot and told you it weighed 142 grams. I'm guessing you would now have a much better idea of your 90% confidence interval (although you still wouldn't have *that* good a guess at the width).

In general, if you are very ignorant about something, the first instance of that thing will tell you what domain you're operating in. If you have no idea how much something weighs, knowing the weight tells you the reasonable orders of magnitude are. Things that sometimes weigh 142 grams don't typically also sometimes weigh 12 solar masses. Similarly, things that take 5 minutes don't typically also take 5 days, and things that are 5 cm long aren't typically also 5 km long.

For more abstract concepts, having a single sample allows you to locate the concept in [concept space](#) by anchoring it to [thing space](#). "Redness" cannot be properly understood until it is known that "apples are red". "Functions" are incomprehensible until you know "adding one to a number" is a function. "Resources" are vague until you learn that "money is a resource".

In reality, the first sample often gives you *more* information than a random sample. If I ask a friend for an example of a snack, they're not going to randomly sample a snack and tell me about it; they're probably going to pick a snack that is at the center of the space of all snacks, like potato chips.

From an information-theoretic perspective, the expected amount of information gained from the first sample must be the highest. If the sampling process is independently and identically distributed, the 2nd sample is expected to be more predictable given knowledge of the first sample. There is some chance that the first sample is misleading, but the probability that it's misleading goes down the more misleading the sample is, so you don't expect the first sample to be misleading. If you're very ignorant, your best guess for the mean of a distribution is pretty close to the mean of the samples you have, even if you only have one.

This is one perspective on why asking for examples is so powerful; they typically give you the first sample, which contains the most information.

Luna Lovegood and the Chamber of Secrets - Part 5

"What do you mean 'bureaucratically impossible'?" said Harry James Potter-Evans-Verres.

"It just is," said Mad-Eye Moody.

"Nothing 'just is'," Harry said, "Everything happens for a material reason. What *precisely* is preventing you from acquiring the permits?"

"I suspect an unfriendly nation-state is employing bureaucramancy against us," Moody said.

"Bureaucramancy," Harry said dryly.

Didn't the Ministry of Magic understand how important this was? No, because Harry had declared the existence of the Sorcerer's Stone to be top secret.

"Can't we just build it without the permits?" Harry said.

"I hope you're not insinuating that a trusted Auror like myself might allow expediency to outweigh accountability," Moody said.

"As if you cared about accountability when you assassinated the Dark Lord of Berzerkistan," said Harry.

"I can neither confirm nor deny whether I have ever set foot in Berzerkistan," Moody said, "But if I had, I can assure you that I would not have broken a single British law within its sovereign borders."

"What is the worst thing that could happen if we completed the Stone Citadel without proper authorization?" Harry asked.

"The end of the world," Moody said. Harry flinched.

"What's the worst thing that is net 'likely' to happen if we build this hospital without proper authorization?" Harry asked, "Will we forfeit a tax exemption or something?"

"You sound like a Dark Lord abusing his political power for the greater good," Moody said.

"It's just a zoning law!"

"The hospital will not be part of Hogwarts and will therefore be unprotected by its wards," Moody said.

Britain had 1% of the planet's magical population. It had 1% of the planet's armies. Hogwarts was a fraction of that. If Harry Potter revealed his hospital to the world it could catalyze an international crisis. When that day came, it would be to the Chief Warlock's advantage if the Stone Citadel was located inside the Hogwarts wards.

Another three minutes and fifty-four seconds ticked by. Another human being died forever. At times of civilizational inadequacy—which was all the time—Harry Potter could empathize with Lord Voldemort's pleasure at murdering his way through the Ministry bureaucracy.

There was a knock on the door. A dreamy first-year voice said "I'm looking for Harry Potter."

"Harry James Potter-Evans-Verres is busy. He is trying to save and/or destroy the world and/or wizardkind," Harry said, "If you make an appointment then he might get back to you in a few epochs."

"How about Tom Riddle?" the voice said, "Does he know what astrolabes do?"

Luna was untrained in the Muggle arts. Muggle Studies Professor Burbage declared the astrolabe beyond her pay grade and referred Luna to Harry Potter. There was no "Harry Potter" on the Marauder's Map so Luna went to "Harry Potter's Office". Tom Riddle answered the door.

"It's nice to meet you, Mister Riddle," Luna said.

Tom Riddle ushered Luna into Harry Potter's office, shut the door and cast thirty security charms.

"Call me Harry Potter," Tom Riddle said.

"Call me You-Know-Who," Luna played along. It was a relief to pretend to have friends for a change.

"*Hss hsssss hss,*" Tom Riddle said in Parseltongue.

"Hiss hiss hiss," Luna said in not-Parseltongue. She held her index fingers down from her mouth like fangs and swayed her head like a snake.

"You're not really You-Know-Who," Tom Riddle said.

"You're not really Harry Potter," Luna showed him the map, "It says so right here."

The 12-year-old boy banged his head against the wall. Just because you've won a battle doesn't mean you've behaved optimally. Optimal behavior is when you extract maximum utility from your circumstances. He shouldn't have been content to retrieve his belongings. He should have looted everything of value from the corpses of his fallen enemies. If he had done that then Bellatrix Lestrange would have been in Ministry custody, he would have possessed the Marauder's Map and his identity would have been secure.

"What do you want?" the boy said.

"I want friends," Luna said, "But right now I'll settle for the Lost Diadem of Ravenclaw. I am at Harry Potter's office because Professor Burbage told me he is an expert in modern Muggle technology."

"I'll tell you what an astrolabe is if you pretend I am Harry James Potter-Evans-Verres from now on."

"Deal," Luna said.

"An astrolabe is a handheld model of the universe," Harry said, "What is the Lost Diadem of Ravenclaw?"

"A diadem is a crown," Luna said, "Ravenclaw's is said to make you smarter."

Harry Potter had heard the term "intelligence explosion" before. If Rowena Ravenclaw could create a magical device that increased her intelligence then she would not have been content to stop there. She would have used the first diadem to make a second, superior one. Then she would have used the second diadem to make an even better third iteration. She might not have stopped until after she had gone full Singularity.

"On second thought," Harry stuffed the Ministry zoning paperwork into his trunk, "Saving the world can wait. What can you tell me about this Lost Diadem?"

Announcement

The easiest way to keep updated on this story is to subscribe to my posts here on Less Wrong.

Last chapter, CitizenTen [asked](#) if I had any plans to post this story to fanfiction websites. I currently do not. You have my permission to copy this story in part or in its entirety to any website I am not active on (which, right now, includes every website on the Internet except [lesswrong.com](#)) provided:

- You respect J.K. Rowling's copyright. Harry Potter fanfiction must remain non-commercial, especially in the strict sense of traditional print publishing.
- You include an attribution link back to the original story here on Less Wrong.

FHI paper published in Science: interventions against COVID-19

This is a linkpost for <https://science.sciencemag.org/lookup/doi/10.1126/science.abd9338>

Summary: this is basically a Bayesian regression of the COVID-19 cases and deaths in 41 countries against the measures taken by those countries. It requires a rough model of how infections spread which means you have to be careful about choosing models, validate them by predicting the epidemic, and try to deal with confounding. The results support closing schools and universities, gatherings, and some types of face-to-face businesses.

It's been great to work on this with a large team of talented researchers from many universities, most of whom are rationalists and EAs. The lead authors and the senior author are employees and affiliates of FHI:

Jan M. Brauner, Sören Mindermann*, Mrinank Sharma*, David Johnston, John Salvatier, Tomáš Gavenčíak, Anna B. Stephenson, Gavin Leech, George Altman, Vladimir Mikulik, Alexander John Norman, Joshua Teperowski Monrad, Tamay Besiroglu, Hong Ge, Meghan A. Hartwick, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal, Jan Kulveit*

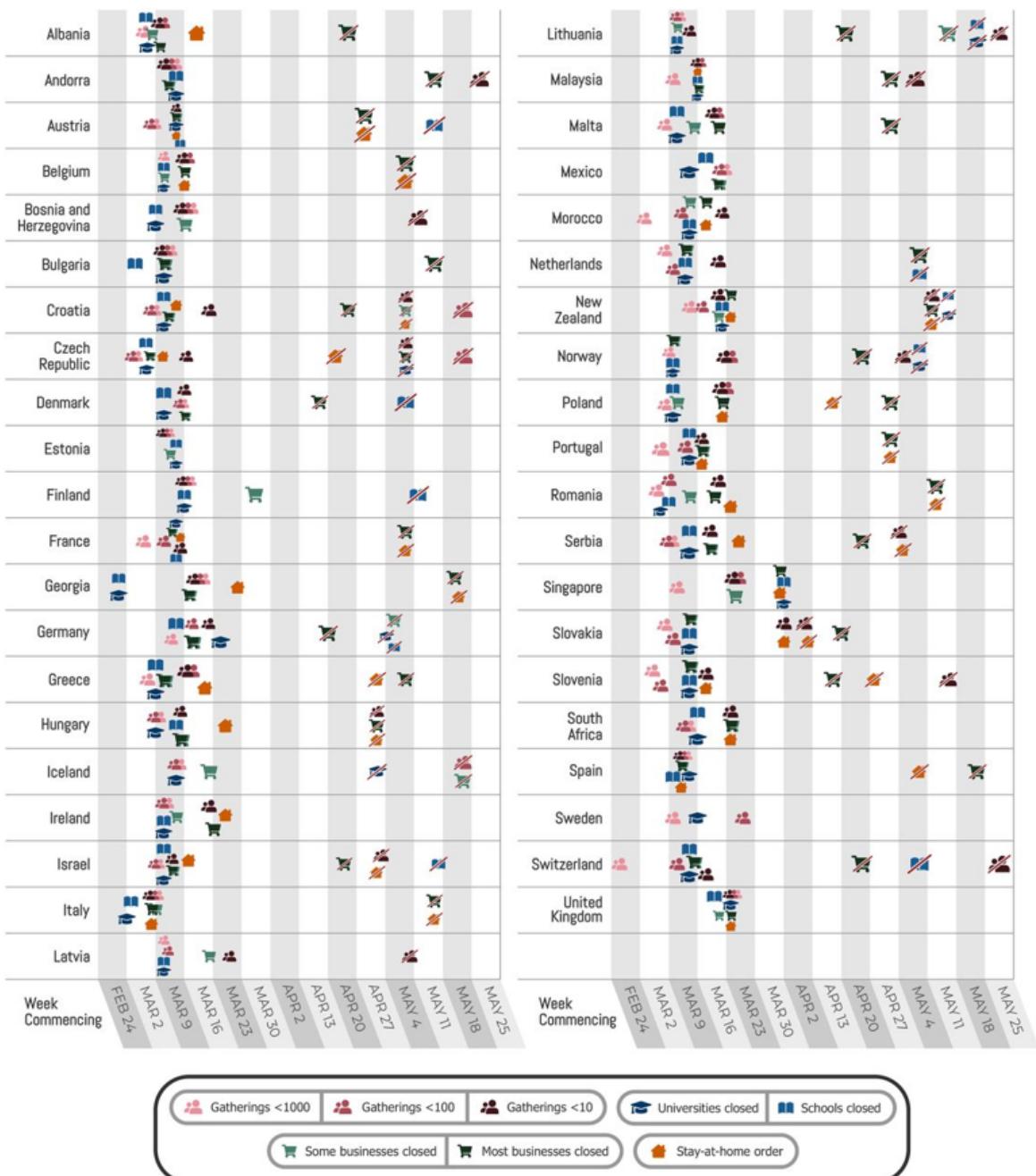
We're also grateful to Tim Telleen-Lawton and BERI who funded and operationally supported the Epidemic Forecasting project which in turn incubated this project.

[**Paper: Inferring the Effectiveness of Government Interventions Against COVID-19**](#)

There's also a closely related [NeurIPS paper](#).

The story of this paper is easy to tell with its figures so I'll briefly to that.

Data:



Main results:



Fig. 2 NPI effectiveness under default model settings. Posterior percentage reductions in R_t with median, 50% and 95% prediction intervals shown. Prediction intervals reflect many sources of uncertainty, including NPI effectiveness varying by country and uncertainty in epidemiological parameters. A negative 1% reduction refers to a 1% increase in R_t . “Schools and universities closed” shows the joint effect of closing both schools and universities in conjunction; the individual effect of closing just one will be smaller (see text). Cumulative effects are shown for hierarchical NPIs (gathering bans and business closures) i.e., the result for “Most nonessential businesses closed” shows the cumulative effect of two NPIs with separate parameters and symbols—closing some (high-risk) businesses, and additionally closing most remaining (non-high-risk, but nonessential) businesses given that some businesses are already closed.

Effect of combined interventions:

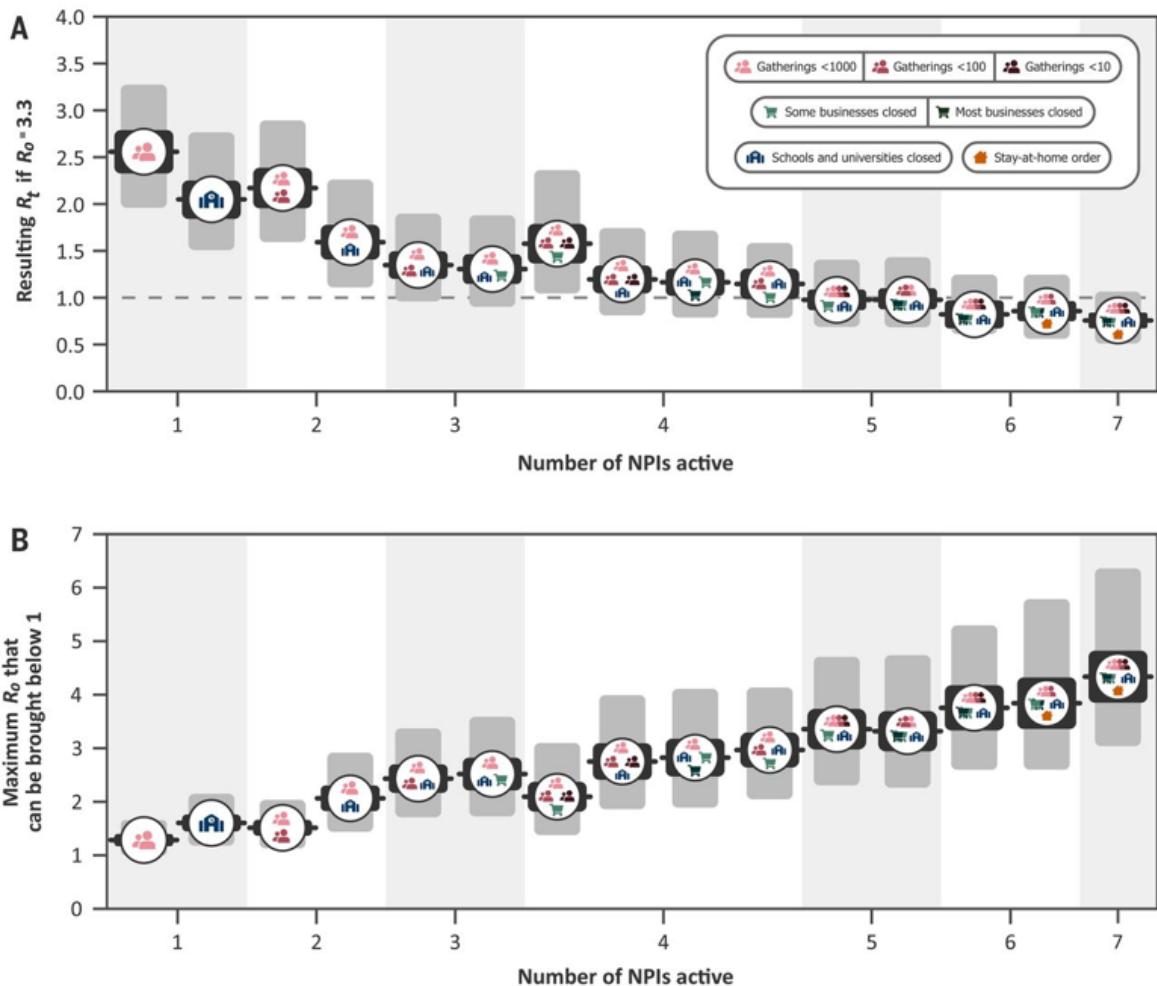


Fig. 3 Combined NPI effectiveness for the 15 most commonly implemented sets of NPIs in our data. Solid and shaded regions denote 50% and 95% Bayesian prediction intervals. **(A)** Predicted R_t after implementation of each set of NPIs, assuming $R_0 = 3.3$. **(B)** Maximum R_0 that can be reduced to R_t below 1 by common sets of NPIs. Readers can interactively explore the effects of all sets of NPIs, while setting R_0 and adjusting NPI effectiveness to local circumstances, with our online mitigation calculator ([16](#)).

Lots of sensitivity analyses:

A All Sensitivity Analyses (206 conditions)

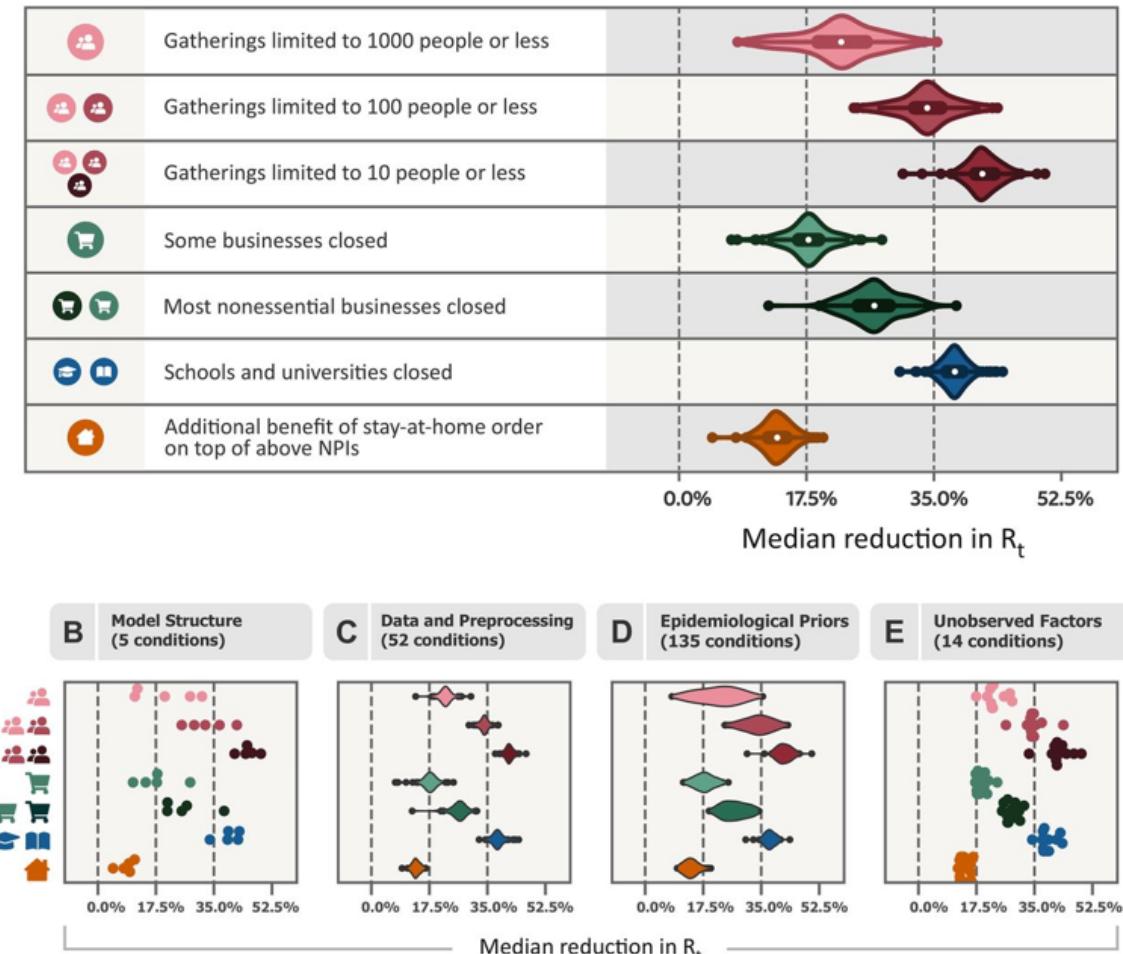


Fig. 4 Median NPI effectiveness across the sensitivity analyses.(A)
 Median NPI effectiveness (reduction in R_t) when varying different components of the model or the data in 206 experimental conditions. Results are displayed as violin plots, using kernel density estimation to create the distributions. Inside the violins, the box plots show median and interquartile-range. The vertical lines mark 0%, 17.5%, and 35% (see text). **(B to E)** Categorized sensitivity analyses. (B) Sensitivity to model structure. Using only cases or only deaths as observations (2 experimental conditions; fig. S7); varying the model structure (3 conditions; fig. S8, left). (C) Sensitivity to data and preprocessing. Leaving out countries from the dataset (42 conditions; figs. S5 and S21); varying the threshold below which cases and deaths are masked (8 conditions; fig. S13); sensitivity to correcting for undocumented cases and to country-level differences in case ascertainment (2 conditions; fig. S6). (D) Sensitivity to epidemiological parameters. Jointly varying the means of the priors over the means of the generation interval, the infection-to-case-confirmation delay, and the infection-to-death delay (125 conditions; fig. S10); varying the prior over R_0 (4 conditions; fig. S11); varying the prior over NPI effect parameters (3 conditions; fig. S11); varying the prior over the degree to which NPI effects vary across countries (3 conditions; fig. S12). (E) Sensitivity to unobserved factors influencing R_t . Excluding observed NPIs one at a time (8 conditions; fig. S9); controlling for additional NPIs from a different dataset (6 conditions; fig. S9).

Structure of the main model:

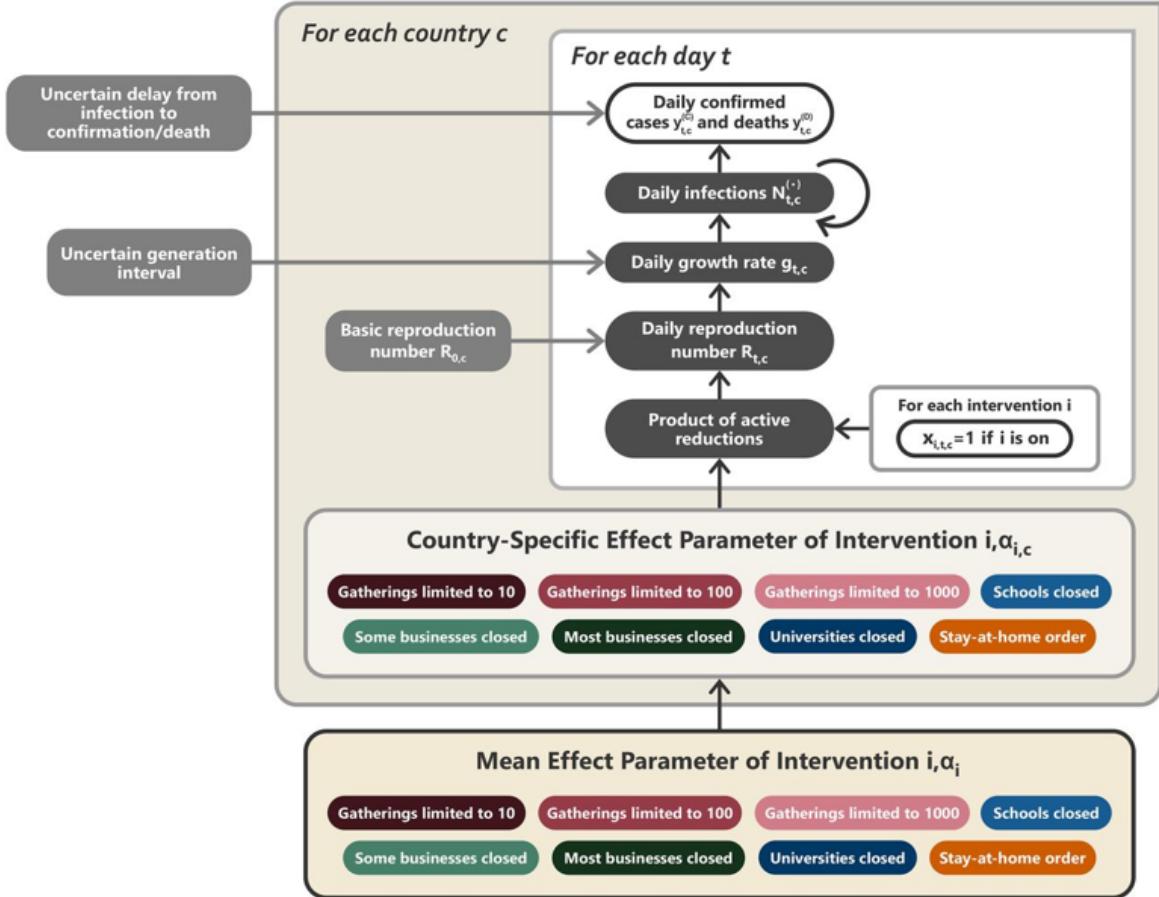


Fig. 5 Model overview. Unshaded, white nodes are observed. We describe the diagram from bottom to top: The mean effect parameter of NPI i is α_i , and the country-specific effect parameter is $\alpha_{i,c}$. On each day t , a country's daily reproduction number $R_{t,c}$ depends on the country's basic reproduction number $R_{0,c}$ and the active NPIs. The active NPIs are encoded by $x_{i,t,c}$, which is 1 if NPI i is active in country c at time t , and 0 otherwise. $R_{t,c}$ is transformed into the daily growth rate $g_{t,c}$ using the generation interval parameters, and subsequently is used to compute the new infections $N(C)t,cNt,c(C)$ and $N(D)t,cNt,c(D)$ that will subsequently become confirmed cases and deaths, respectively. Finally, the expected number of daily confirmed cases $y(C)t,cyt,c(C)$ and deaths $y(D)t,cyt,c(D)$ are computed using discrete convolutions of $N(.)t,cNt,c(.)$ with the relevant delay distributions. Our model uses both case and death data: it splits all nodes above the daily growth rate $g_{t,c}$ into separate branches for deaths and confirmed cases. We account for uncertainty in the generation interval, infection to case confirmation delay, and the infection to death delay by placing priors over the parameters of these distributions.

Much of the interesting parts are in the [appendix](#).

Why quantitative methods are heartwarming

From Twitter:

If you listened to my podcast w/Michael Sandel, you know we have very different views on whether markets are "degrading"

One thing I didn't mention to him: This bit in his book cracked me up -- because I remember my friends & I found this aspect of Moneyball SO HEARTWARMING <3
pic.twitter.com/9W6Op30vF8

— Julia Galef (@juliagalef) [December 10, 2020](#)

I haven't actually seen Moneyball, but it does sound heartwarming, and I have had to hide my tears when someone described a payment app their company was working, so I'm probably in Julia's category here.

If I didn't feel this way though, reading this I might imagine it as some alien nerdly aberration, and not a way that I could feel from the inside, or that would seem the 'right' way to feel unless I became brain-damaged. Which I think is all wrong—such feelings seem to me to be a warm and human response to appreciating the situation in certain ways. So I want to try to describe what seems to be going on in my mind when my heart is warmed by quantitative methods and efficient algorithms.

When using good quantitative methods makes something better, it means that there wasn't any concrete physical obstacle to it being better in the past. We were just making the wrong choices, because we didn't know better. And often suffering small losses from it at a scale that is hard to imagine.

Suppose the pricing algorithm for ride sharing isn't as good as it could be. Then day after day there will be people who decide to walk even though they are tired, people who wait somewhere they don't feel safe for a bit longer, countless people who stand in their hallway a bit longer, people who save up their health problems a bit more before making the expensive trip to a doctor, people who decide to keep a convenient car and so have a little bit less money for everything else. All while someone who would happily drive each of them at a price they would happily pay lives nearby, suffering for lack of valuable work.

I'm not too concerned if we make bad choices in baseball, but in lots of areas, I imagine that there are these slow-accreting tragedies, in thousands or millions or billions of small inconveniences and pains accruing each day across the country or the world. And where this is for lack of good algorithms, it feels like it is for absolutely nothing. Just unforced error.

Daily efforts and suffering for nothing are a particular flavor of badness. Like if someone erroneously believed that it was important for them to count to five thousand out loud at 10am each day, and every day they did this—and if they traveled they made sure there would be somewhere non-disturbing to do it, and if they stayed up late they got up by 10am; and if they were doing something they stepped out—there would be a particular elation in them escaping this senseless waste of their life, perhaps mixed with sorrow for what had been senselessly lost.

Also, having found the better method, you can usually just do it at no extra cost forever. So it feels reelingly scalable in a way that a hero fighting a bad guy definitively does not. This feels like suddenly being able to fly, or [walk through walls](#).

So basically, it is some combination of escape from a senseless corrosion of life, effortlessly, at a scale that leaves me reeling.

Another thing that might be going on, is that it is a triumph of what is definitely right over what is definitely wrong. Lots of moral issues are fraught in some way. No humans are absolutely bad and without a side to the story. But worse quantitative methods are just straightforwardly wrong. The only reason for picking baseball players badly is not knowing how to do it better. The only reason for using worse estimates for covid risk is that you don't have better ones. So a victory for better quantitative methods is an unsullied victory for light over darkness in a way that conflicts between human forces of good and bad can't be.

Yet another thing is that a victory for quantitative methods *is* always a victory for people. And if you don't know who they are, that means that they quietly worked to end some ongoing blight on humanity, and did it, and weren't even recognized. Often, even the good they did will look like a boring technical detail and won't look morally important, because saving every American ten seconds doesn't look like saving a life. And I'm not sure if there is anything more heartwarming than someone working hard to do great good, relieving the world from ongoing suffering, knowing that neither they nor and what they have given will be appreciated.

How Hard Would It Be To Make A COVID Vaccine For Oneself?

I always assumed, without investigating, that vaccine-making was the sort of thing which required highly specialized experts and equipment. But I've been hearing that the Moderna vaccine was designed in two days, which strongly suggests that it's *simple*. So I'm at least asking the question: how hard would it be to make such a vaccine?

The main question is what's involved in designing and synthesizing a vaccine.

Subquestions:

- What information is needed for design, and is any non-public?
- What skills are needed for design (other than general bio/bioinformatics know-how), and how does one design a vaccine?
- What equipment and supplies are needed for synthesis?
- Does the equipment require any unusual skills (other than typical wetlab stuff, e.g. how to use a pipette or prevent contamination)?
- Perhaps most important for a non-professional effort: what quality assurance steps are typically used? How can one verify that the thing one intended to produce was actually produced?

These are the main things I'd like to know about.

There's also a cluster of secondary questions around exactly what one could do with such a vaccine, beyond using it oneself. Presumably the FDA would shut down any effort to sell it to others as a vaccine, but which part of that is the problematic part? If someone gives away a homebrew vaccine for free, is that ok? If it's advertised as not-a-vaccine (but with a technical explanation of exactly what it is), or even as "not for human use" (since of course humans never use things not fit for human use), is that ok? To what extent can barriers be circumvented by loading all the equipment in an RV and taking a weekend trip to Canada/Mexico/whatever the equivalent is for Europe? I'm imagining e.g. a small group in Berkeley decides to put in the effort to learn how to make vaccines, buys \$10k of equipment and supplies, then gives away vaccine to most of the Berkeley rationalist hub (with a clear "not approved by FDA, not fit for human consumption" warning accompanied by an implicit wink). Where does the FDA draw the line between "a group of friends doing stuff they shouldn't be doing" vs something which needs to be shut down?

I expect any viable plan would take quite a lot of effort, but there's an awful lot of value to be had here - even just immunizing oneself and a handful of friends could plausibly be valuable enough to justify the effort and expenditure. Also, this seems like the sort of thing where someone could learn the skills and acquire the equipment in advance, as preparation for future pandemics - possibly including diseases much more dangerous than COVID. Given how well this episode has gone, that could provide even more value.

Luna Lovegood and the Chamber of Secrets - Part 6

"It's in the Mirror of Atlantis," Harry said.

"We need Gillyweed," said Luna, "Lots of Gillyweed."

"The Mirror of Atlantis is located at the end of the third-floor corridor on the right-hand side," said Harry.

"What are you waiting for?" said Luna.

"You're not going to ask me how I know these things?" said Harry.

"Why would I do that?" said Luna.

"Because...because..." Harry searched for words to express the magnitude of the epistemic hole.

"You talk like a Muggle," said Luna, "You think like one too."

Harry puzzled over whether he had just been insulted. Luna skipped ahead to the third floor.

Harry and Luna flew a double-seated broomstick over the dieffenbachia and other unmaintained obstacles to the Mirror.

"You're a really good flier," said Luna, "I bet you're on the Quidditch team. What position do you play? No. Wait. Don't tell me. I bet you play Seeker."

"This Mirror lets you store an object until someone with the right intentions arrives," said Harry testily.

"I seek entry to the Ravenclaw Common Room," said Luna to the Mirror, "I want to sleep in a bed."

"In Gödel's name, what monstrously difficult riddle did the bronze eagle ask of you?" asked Harry.

"Where is my diadem?" said Luna.

"You're supposed to say 'lost' or 'hidden'," said Harry, "You're not expected to rediscover the Lost Diadem of Ravenclaw."

"Oh," said Luna.

"Since we're already here, let's give it a modicum of effort," Harry withdrew a mechanical stopwatch from his pouch and set it to five minutes.

"I want the diadem to save the world," Harry said.

"I aim to save the lives of all sentient beings," Harry said.

"I promise to bring you, Rowena Ravenclaw, back to life," Harry said.

"I seek to disassemble stars," Harry said.

Luna ignored him. Rowena Ravenclaw was not a jealous witch. If she concealed knowledge then it was not because the knowledge could be used for evil. It was because the knowledge itself was intrinsically dangerous. A cascade of logic began to self-assemble.

Rowena's basilisk attacks those who know it exists.

Luna forced her thoughts back into disarray.

"BRIIIIIING!" the mechanical alarm went off.

"I'm hunting nargles," Luna said.

The Luna in the Mirror held the Diadem of Ravenclaw in her left hand. The real Luna held it in her right.

"What is a nargle?" said Harry.

"You don't want to know," said Luna.

"What do you plan to do with this magical artifact of incredible power?" asked Harry.

"I'm going to get a comfy night's sleep in the Ravenclaw dormitory," said Luna.

"Can I, uh, try it on first?" Harry asked.

"Sure," said Luna. They were friends, after all.

Harry Potter thrust the diadem onto his head.

Then he had a seizure.

Harry's arms locked into place as if a powerful electric current flowed through them. Luna failed to pry the diadem loose. Luna dug through her satchel for the Comed-Tea can. She released Wanda.

"Help him," Luna said.

Harry's convulsions stopped. Luna wrestled the diadem off Harry's head and out of his hands.

Luna Lovegood and the Chamber of Secrets - Part 4

Luna located the Ravenclaw Common Room on the map. She climbed up to Ravenclaw Tower, then climbed up Ravenclaw Tower. All that separated Luna from making day-one friends was a door with a bronze knocker in the shape of an eagle which spoke riddles appropriate for first-years. She could almost hear muffled sounds of the other children partying. Luna knocked once. The eagle spoke with Rowena Ravenclaw's voice.

"Where is my diadem?"

Luna lacked sufficient hours-until-breakfast to find the Lost Diadem of Ravenclaw. Luna curled up in an alcove close to where heat leaked out of the Common Room. She wrote a thank you note to the house-elves for keeping the floor clean.

"I'm going to call you Wanda," Luna said to her Wrackspurt.

Wanda appeared on the Marauder's Map. Luna clicked her tongue and then spellotaped the Comed-Tea can to her ear so Wanda could feed. Luna's brain went fuzzy. Then she was asleep.

Luna woke to the sound of Ravenclaw going down to breakfast. Luna removed Wanda from her ear. Someone had placed a blanket over Luna while she slept. Luna threw the blanket off before anyone else spotted her and realized she hadn't solved the riddle. She knocked once on the eagle. The riddle hadn't changed.

On her way to breakfast, Luna passed the Forgotten Library at the base of Ravenclaw Tower. She had to eat food because she was a human being. Or she could explore a room that had stood for a thousand years and would probably still be there tomorrow. The entrance to the Forgotten library was a grey heptagon embedded in the wall. Its interior emitted monochromatic grey light. Luna stepped into the Forgotten Library.

Luna stepped out of the Forgotten library. She checked her satchel for quill, parchment and inkwell. Luna wrote "Exploration Journal" at the top of the parchment. Luna stepped into the Forgotten Library.

Luna stepped out of the Forgotten Library. She held a sheet of parchment with "Exploration Journal" at the top. Luna stepped into the Forgotten Library.

Luna stepped out of the Forgotten Library. She had left her exploration journal inside. Luna stepped into the Forgotten Library.

Luna stepped out of the Forgotten Library. She held a blank sheet of parchment. She had missed her first Charms class. If she hurried she could still reach Battle Magic on time. The Marauder's Map showed a shortcut leading from a door disguised as a

window to a window disguised as a door. Luna refenestrated herself onto the top floor of prosaic geometric space just outside the Battle Magic classroom.

Professor Lockhart had preserved Former Professor Quirrel's tradition of combining all four Houses into a single class. He had also combined all seven years into a single class. Professors Burbage, Lapsusa, Sinistra, Trelawney, Vector and Yue were there too, plus Headmistress McGonagall, Miss Pince and Madam Pomfrey. Hogwarts had grown an auditorium that rotated with the Sun under a geodesic dome to ensure Gilderoy Lockhart could always be seen in the best possible light. Gilderoy Lockhart's smile shouted *I love you* to you, personally.

"Before we begin I have a demonstration. *Protego*," Lockhart said, "On the count of three I want someone to hex me. One..."

Gilderoy Lockhart had given this demonstration dozens of times to witches and wizards around the world. Most were not duelists. Those who were tended to be too surprised to cast anything.

This was Gilderoy Lockhart's first interaction with the survivors of Former Professor Quirrel's armies.

A hundred hexes hit Gilderoy Lockhart before he could say "two". The Defense Professor's podium was molten slag. His body resembled a Blast-Ended Skrewt. Purple smoke billowed up from the shattered dome. The professors rushed to his aid. Headmistress McGonagall evacuated the students.

Luna went to lunch in the Great Hall.

"There's Loony Lovegood," someone said.

"She can't afford an owl so she pretends she has an invisible pet."

"Writes notes to House-elves because she doesn't have any friends."

"She thinks she knows everything but she can't even solve her bronze eagle riddle."

"Should have been sorted into Gryffindor."

Luna clicked her tongue and let Wanda feed on her thoughts. She stole lunch from the kitchens to eat in the Forgotten Library.

Lady Yue had replaced the dungeon torches with candles. She wore the robe of a future historical reenactor confident in her knowledge of what a modern witch's clothes were supposed to look like. The Gryffindors' and Ravenclaws' chatter dampened to silence before their mistress.

Lady Yue stood behind a table with two steaming cauldrons on it.

"What is real?" she asked the class, "Lovegood."

"Reality is what you can observe," Luna said.

"Five points to Ravenclaw," Lady Yue sighed. She gazed upward toward exactly where the moon could have been seen if they weren't all in a dungeon. You could tell, merely by looking at her, that it was raining outside.

Luna wanted to justify her answer. But Lady Yue had rewarded Luna for being wrong—even though Luna was right. Lady Yue had out-maneuvered Luna by supporting her. Luna couldn't be angry. Nor could she be satisfied.

Lady Yue cast a disillusionment charm on the first cauldron. The steam appeared to intrude from a rip in space.

"What is real?" Lady Yue repeated.

Several students once again raised their hands. Lady Yue pointed to Ginny Weasley.

"Reality is everything with a physical or magical manifestation."

"Five points to Gryffindor," Ginny Weasley looked at her shoes. Lady Yue tapped the visible cauldron. It ceased existing.

"I have just demonstrated two ways of unmaking a form. The disillusionment charm conceals its image. Vanishment banishes its substance into nonexistence," Lady Yue said.

Luna began to raise her hand to ask what happened to the third cauldron. Lady Yue silenced her with a saccade. The tendrils of Lady Yue's reverse legilimency said *see me after class*.

The two witches waited for everyone else to leave. In the quiet, Luna noticed that Lady Yue moved soundlessly. She took that much care in every detail of her movement.

Lady Yue waited for Luna to make the first move. Luna played the game. The candles burned out. It became pitch black.

"Are you a Dark Lord?" Luna asked.

"Saving and/or destroying the world is a boy's game," Lady Yue said.

The Earth and Moon orbited their center of mass.

"Hard magic has form," Lady Yue said, "It can be controlled. Soft magic is empty; it is without form. Soft magic is suppressed by every organization from the Department of International Magical Cooperation to Lord Voldemort's little club."

"Are you telling me You-Know-Who didn't practice the Dark Arts?" Luna said.

"Five points from Ravenclaw," said Lady Yue, "In the First Wizarding War, your mother contributed to the development of a secret weapon intended to neutralize Lord Voldemort."

"Do you know where I can find the Lost Diadem of Ravenclaw?" Luna blurted out, "Can you tell me what a nargle is? What happened to Platform Nine and One-Half?"

The rain became a thunderstorm.

"This is your mother's astrolabe. She left it in my possession before she died. It is time it was returned to you."

Extrapolating GPT-N performance

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

[Brown et al. \(2020\)](#) (which describes the development of GPT-3) contains measurements of how 8 transformers of different sizes perform on several different benchmarks. In this post, I project how performance could improve for larger models, and give an overview of issues that may appear when scaling-up. Note that these benchmarks are for ‘downstream tasks’ that are different from the training task (which is to predict the next token); these extrapolations thus cannot be directly read off the scaling laws in OpenAI’s Scaling Laws for Neural Language Models ([Kaplan et al., 2020](#)) or Scaling Laws for Autoregressive Generative Modelling ([Henighan et al., 2020](#)).

(If you don’t care about methodology or explanations, the final graphs are in **Comparisons and limits**.)

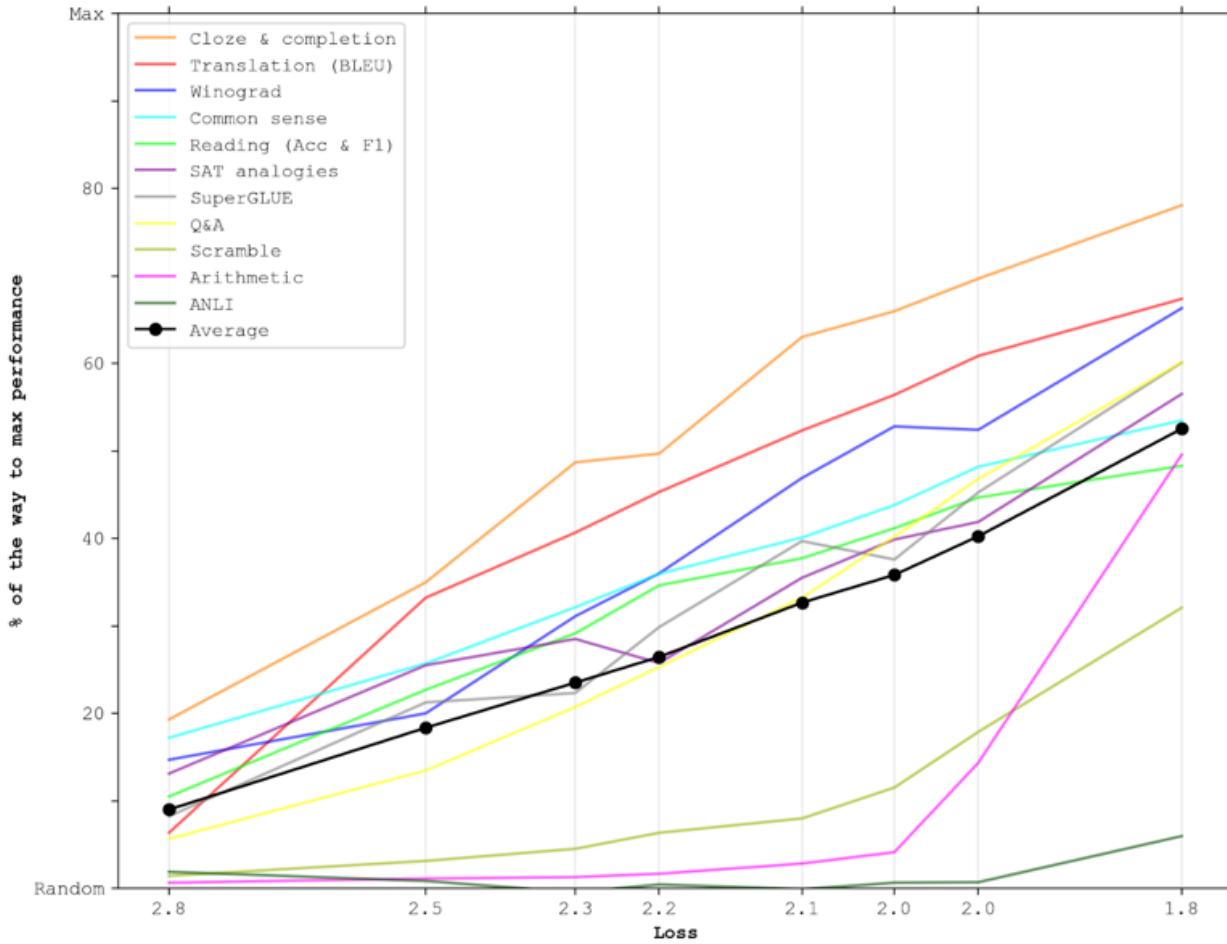
Methodology

Brown et al. reports benchmark performance for 8 different model sizes. However, these models were not trained in a *compute-optimal* fashion. Instead, all models were trained on 300B tokens (one word is [roughly 1.4 tokens](#)), which is inefficiently much data. Since we’re interested in the best performance we can get for a given amount of compute, and these models weren’t compute-optimally trained, we cannot extrapolate these results on the basis of model-size.

Instead, I fit a trend for how benchmark performance (measured in % accuracy) depends on the cross-entropy loss that the models get when predicting the next token on the validation set. I then use the scaling laws from [Scaling Laws for Neural Language Models](#) to extrapolate this loss. This is explained in the Appendix.

Plotting against loss

In order to get a sense of how GPT-3 performs on different types of tasks, I separately report few-shot progress on each of the 11 different categories discussed in Brown et al. For a fair comparison, I *normalize* the accuracy of each category between random performance and maximum performance; i.e., for each data point, I *subtract* the performance that a model would get if it responded randomly (or only responded with the most common answer), and *divide* by the difference between maximum performance and random performance. The black line represents the average accuracy of all categories. This implicitly gives less weights to benchmarks in larger categories, which I think is good; see the Appendix for more discussion about this and the normalization procedure.



Note that the x-axis is logarithmic. For reference, the 4th model (at a loss of 2.2) is similar to GPT-2's size (1.5e9 parameters).

Overall, I think the models' performance is surprisingly similar across many quite different categories. Most of them look reasonably linear, improve at similar rates, and both start and end at similar points. This is partly because all tasks are selected for being appropriately difficult for current language models, but it's still interesting that GPT-3's novel few-shot way of tackling them doesn't lead to more disparities. The main outliers are Scramble, Arithmetic, and ANLI (Adversarial Natural Language Inference); this is discussed more below.

Extrapolating

In general, on linear-log plots like the ones above, where the y-axis is a score between 0 and 1, I expect improvements to follow some sort of s-curve. First, they perform at the level of random guessing, then they improve exponentially as they start assembling heuristics (as on the scramble and arithmetic tasks) and finally they slowly converge to the upper bound set by the irreducible entropy.

Note that, if the network converges towards the irreducible error like a negative exponential (on a plot with reducible error on the y-axis), it would be a straight line on a plot with the *logarithm* of the reducible error on the y-axis. Since the x-axis is also logarithmic, this would be a straight line on a log-log plot, i.e. a power-law between the

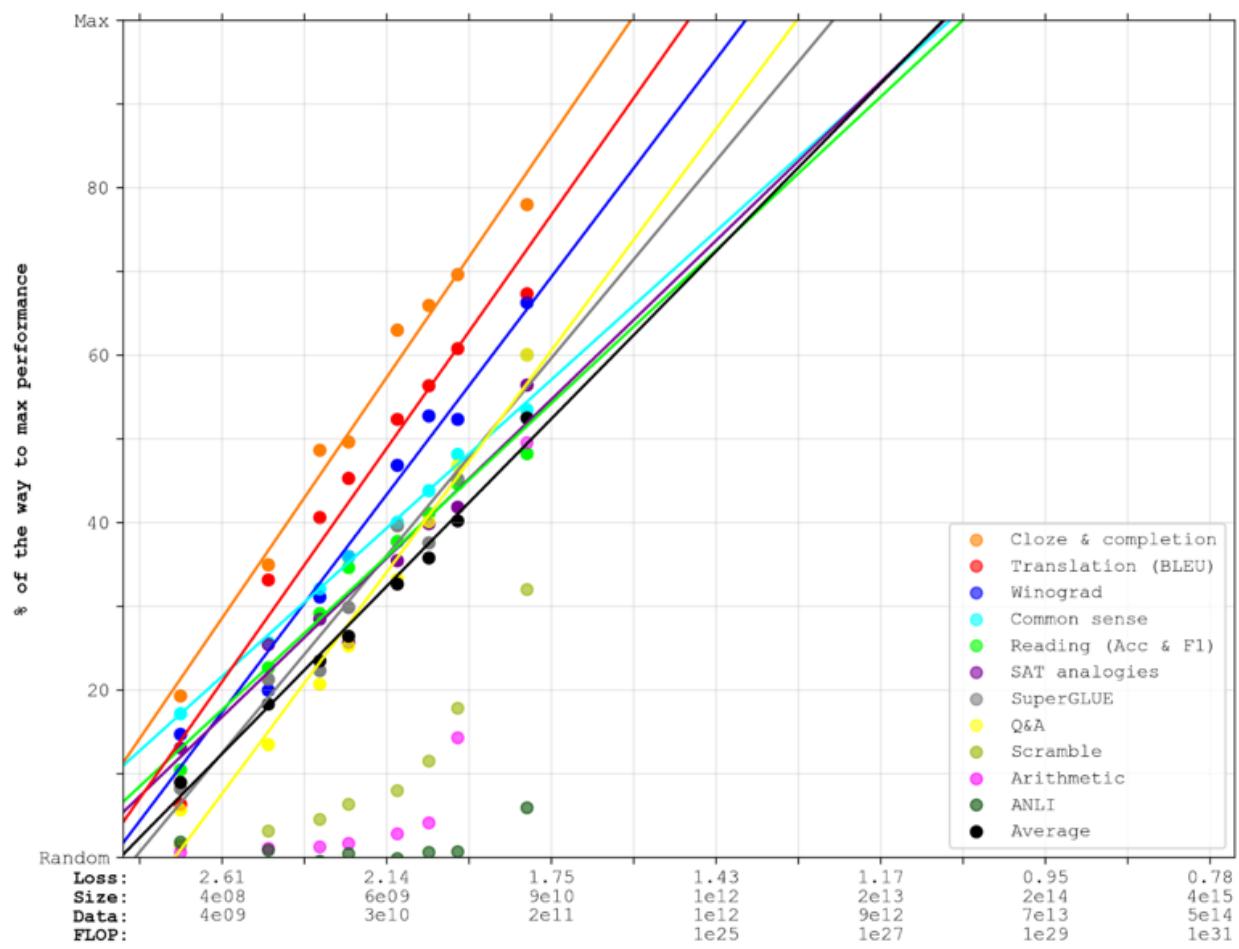
reducible error and the reducible loss. In addition, since the reducible loss is related to the data, model size, and compute via power laws, their *logarithms* are *linearly* related to each other. This means that we can (with linear adjustments) add logarithms of these to the x-axis, and that a similar argument applies to them. Thus, converging to the irreducible error like a negative exponential corresponds to a power law between reducible error and each of those inputs (data, model size, and compute).

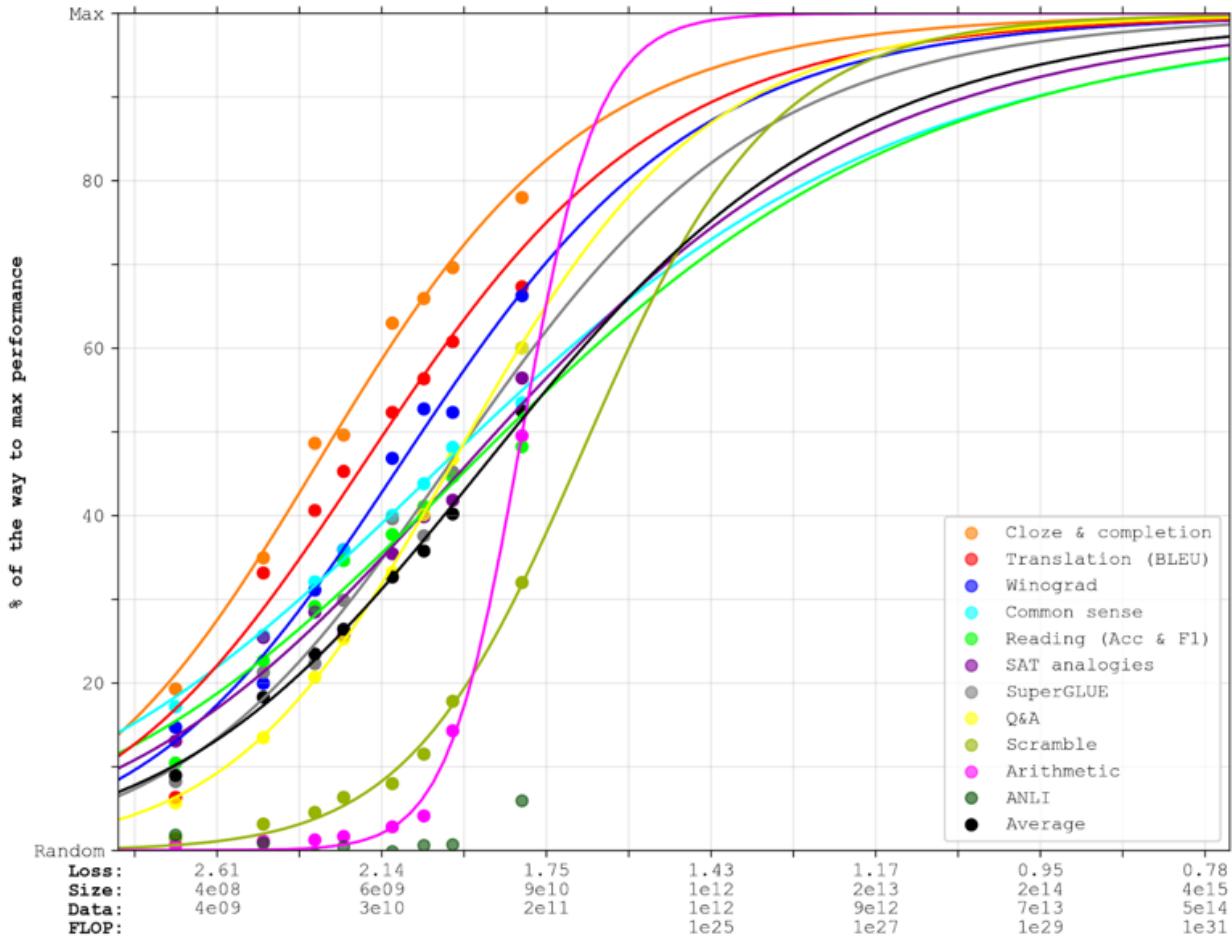
Unfortunately, with noisy data, it's hard to predict when such an s-curve will hit its inflection point unless you have many data points after it (see [here](#)). Since we don't, I will fit *linear* curves and *sigmoid* curves.

- On most datasets, I think the linear curves will *overestimate* how soon they'll reach milestones above 90%, since I suspect performance improvements to start slowing down before then. However, I wouldn't be shocked if they were a decent prediction up until that point. The exceptions to this are ANLI, arithmetic, and scramble, which are clearly not on linear trends; I have opted to not extrapolate them linearly (though they're still included in the average).
- I think sigmoid curves – i.e., s-curves between 0% and 100% with an inflection point at 50% – are more sensible as a median guess of performance improvements. My best guess is that they're more likely to underestimate performance than overestimate performance, because the curves look quite linear right now, and I give some weight to the chance that they'll continue like that until they get *much* closer to maximum performance (say, around 80-90%), while the logistics assume they'll bend quite soon. This is really just speculation, though, and it could go either way. For sigmoids, I extrapolate all benchmarks except ANLI.

For extrapolating size, data, and compute-constraints, I use a scaling law that predicts loss via the number of *parameters* and the available *data*. This doesn't directly give the floating point operations (FLOP) necessary for a certain performance, since it's not clear how many *epochs* the models need to train on each data point to perform optimally. However, some arguments suggest that models will soon become so large that they'll fully update on data the first time they see it, and overfit if they're trained for multiple epochs. This is predicted to happen after $\sim 1e12$ parameters, so I assume that models only train for one epoch after this (which corresponds to $6ND$ FLOP, where N is the model size and D is the number of tokens). See the Appendix for more details.

Here are the extrapolations:





Extrapolations like these get a lot less reliable the further you extend them, and since we're unlikely to beat many of these benchmarks by the next 100x increase in compute, the important predictions will be quite shaky. We don't have much else to go on, though, so I'll assume that the graphs above are roughly right, and see where that takes us. In Comparison and limits, I'll discuss how much we could afford to scale models like these. But first, I'll discuss:

How impressive are the benchmarks?

The reason that I'm interested in these benchmarks is that they can say something about when transformative AI will arrive. There are two different perspectives on this question:

- Should we expect a scaled-up version of GPT-3 to be generally more intelligent than humans across a vast range of domains? If not, what does language models' performance tell us about when such an AI can be expected?
- Will a scaled-up version of GPT-3 be able to perform economically useful tasks? Do we know concrete jobs that it could automate, or assistance it could provide?

The former perspective seems more useful if you expect AI to transform society once we have a single, generally intelligent model that we can deploy in a wide range of scenarios. The latter perspective seems more useful if you expect AI to transform society by automating one task at a time, with specialised models, as in

[Comprehensive AI Services](#) (though note that massively scaling up language models trained on everything is already in tension with my impression of CAIS).

So what can the benchmarks tell us, from each of these perspectives?

To begin with, it's important to note that it's *really hard* to tell how impressive a benchmark is. When looking at a benchmark, we can at best tell what reasoning *we* would use to solve it (and even this isn't fully transparent to us). From this, it is tempting to predict that a task won't be beaten until a machine can replicate that type of reasoning. However, it's common that benchmarks get solved surprisingly fast due to hidden statistical regularities. This often happens in image classification, which explains why adversarial examples are so prevalent, as argued in [Adversarial Examples Are Not Bugs, They Are Features](#).

This issue is also common among NLP tasks – sufficiently common that many of today's benchmarks are filtered to only include questions that a tested language model *couldn't* answer correctly. While this is an effective approach for continuously generating more challenging datasets, it makes the relationship between benchmarks taken from any one time and the kind of things we care about (like ability to perform economically useful tasks, or the ability to reason in a human-like way) quite unclear.

As a consequence of this, I wouldn't be very impressed by a fine-tuned language model reaching human performance on any one of these datasets. However, I think a single model reaching human performance on almost all of them with ≤ 100 examples from each (provided few-shot style) would be substantially more impressive, for a few reasons. Firstly, GPT-3 *already* seems extremely impressive, qualitatively. When looking at the kind of results gathered [here](#), it seems like the benchmark performance *underestimates* GPT-3's impressiveness, which suggests that it isn't solving them in an overly narrow way. Secondly, with fewer examples, it's less easy to pick up on spurious statistical regularities. Finally, if *all* these tasks could consistently be solved, that would indicate that *a lot more* tasks could be solved with ≤ 100 examples, including some economically useful ones. Given enough tasks like that, we no longer care exactly how GPT-3 does it.

What are the benchmarks about?

(See footnotes for examples.)

- Translation is about translating between English and another language (GPT-3 was tested on romanian, german, and french).
- The Q&A and partly common sense benchmarks are mostly about *memorising facts* and presenting them in response to quite clear questions^[1]. This seems very useful *if* GPT-3 can connect it with everything else it knows, to incorporate it for separate tasks, but not terribly useful otherwise.
- Many of the reading comprehension tasks are about reading a paragraph and then answering questions about it; often by answering yes or no and/or citing a short section of the paragraph. GPT-3 doesn't perform terribly well on this compared to the SOTA, perhaps because it's quite different from what it's been trained to do; and I imagine that fine-tuned SOTA systems can leverage quite a lot of heuristics about what parts of the text tends to be good to pick out, where to start and end them, etc.

- Similar to reading comprehension, the cloze and completion tasks tests understanding of a given paragraph, except it does this by asking GPT-3 to end a paragraph with the right word^[2], or picking the right ending sentence^[3]. GPT-3 currently does really well on these tasks, both when compared to other methods and in absolute terms, as visible on the graphs above. This is presumably because it's very similar to the task that GPT-3 was trained on.
- The winograd tasks also tests understanding, but by asking which word a particular pronoun refers to^[4].
- A couple of tasks make use of more unique capabilities. For example, one of the reading comprehension tasks often require application of in-context arithmetic^[5] and some of the common sense reasoning tasks directly appeals to tricky knowledge of physical reality^[6].
- As mentioned above, many of the benchmarks have been filtered to only include questions that a language model failed to answer. ANLI (short for Adversarial Natural Language Inference) does this to an unusual degree. The task is to answer whether a hypothesis is consistent with a description^[7], and the dataset is generated over 3 rounds. Each round, a transformer is trained on all questions from previous rounds, whereupon workers are asked to generate questions that fool the newly trained transformer. I wouldn't have predicted beforehand that GPT-like models would do quite so badly on this dataset, but I assume it is because of this adversarial procedure. In the end, it seems like the fully sized GPT-3 just barely manages to start on an s-curve.
- Finally, the scramble task is about shuffling around letters in the right way, and arithmetic is about adding, subtracting, dividing, and multiplying numbers. The main interesting thing about these tasks is that performance doesn't improve at all in the beginning, and then starts improving very fast. This is some evidence that we might expect non-linear improvements on particular tasks, though I mostly interpret it as these tasks being quite narrow, such that when a model starts getting the trick, it's quite easy to systematically get right.

Evidence for human-level AI

What capabilities would strong performance on these benchmarks imply? None of them stretches the limits of human ability, so no level of performance would give direct evidence for *super-human* performance. Similarly, I don't think any level of performance on these benchmarks would give much direct evidence about ability to e.g. form longer term plans, deeply understand particular humans or to generate novel scientific ideas (though I don't want to dismiss the possibility that systems *would* improve on these skills, if massively scaled up). Overall, my best guess is that a scaled-up language model that could beat these benchmarks would still be a lot worse than humans at a lot of important tasks (though we should prepare for the possibility that some simple variation *would* be very capable).

However, I think there's another way these benchmarks can provide evidence for when we'll get human-level AI, which relies on a model presented in Ajeya Cotra's [Draft report on AI timelines](#). (As emphasized in that link, the report is still a draft, and the numbers are in flux. All numbers that I cite from it in this post may have changed by the time you read this.) I recommend reading the report (and/or Rohin's summary in the comments and/or my [guesstimate](#) replication), but to shortly summarize: The report's most central model estimates the number of *parameters* that a neural network

would need to become ~human-equivalent on a given task, and uses scaling laws to estimate how many *samples* such a network would need to be trained on (using current ML methods). Then, it assumes that each “sample” requires FLOP proportional to the amount of data required to tell whether a given perturbation to the model improves or worsens performance (the task’s [effective horizon length](#)). GPT-3’s effective horizon length is a single token, which would take $\sim \frac{1}{4}$ of a second for a human to process; while e.g. a meta-learning task may require several days worth of data to tell whether a strategy is working or not, so it might have a $\sim 100,000$ x longer horizon length.

This model predicts that a neural network needs similarly many parameters to become human-equivalent at *short* horizon lengths and *long* horizon lengths (the only difference being training time). Insofar as we accept this assumption, we can get an estimate of how many parameters a model needs to become ~human-equivalent at a task of *any* horizon length by answering when they’ll become ~human-equivalent at *short* horizon lengths.

Horizon length is a tricky concept, and I’m very unsure how to think about it. Indeed, I’m even unsure to what extent it’s a coherent and important variable that we should be paying attention to. But if the horizon length model is correct, the important question is: *How does near-optimal performance on these benchmarks compare with being human-level on tasks with a horizon length of 1 token?*

Most obviously, you could argue that the former would *underestimate* the latter, since the benchmarks are only a small fraction of all possible short-horizon tasks. Indeed, as closer-to-optimal performance is approached, these benchmarks will presumably be filtered for harder and harder examples, so it would be premature to say that the *current* instantiation of these benchmarks represents human-level ability.

In addition, these tasks are limited to language, while humans can also do many other short-horizon tasks, like image or audio recognition^[8]. One approach would be to measure what fraction f of the human-brain is involved in language processing, and then assume that a model that could do all short-horizon tasks would be $1/f$ times as large as one that can only do language. However, I’m not sure that’s fair, because we don’t actually care about getting a model that’s human-level on everything – if we can get one that only works when fed language, that won’t be a big limitation (especially as we already have AIs that are decent at parsing images and audio into text, if not quite as robust as humans). If we compare the fraction of the brain dedicated to short-horizon language parsing with whatever fraction of the brain is dedicated to important tasks like strategic planning, meta-learning, and generating new scientific insights, I have no idea which one would be larger. Ultimately, I think that would be a more relevant comparison for what we care about.

Furthermore, there are some reasons for why these benchmarks could *overestimate* the difficulty of short-horizon tasks. In particular, you may think that the hardest available benchmarks used to represent 1-token horizon lengths, but that these have been gradually selected away in favor of increasingly narrow benchmarks that AI struggle particularly much with, but that would very rarely be used in a real world context. There’s no good reason to expect neural networks to become human-equivalent at all tasks at the same time, so there will probably be some tasks that they remain subhuman at far beyond the point of them being transformative. I don’t think this is a problem for current benchmarks, but I think it could become relevant soon if we keep filtering tasks for difficulty.

Perhaps more importantly, this particular way of achieving human-parity on short horizon lengths (scaling GPT-like models and demonstrating tasks few-shot style) may be far inferior to some other way of doing it. If a group of researchers cared a lot about this particular challenge, it's possible that they could find much better ways of doing it within a few years^[9].

Overall, I think that near-optimal performance on these benchmarks would somewhat *underestimate* the difficulty of achieving human-level performance on 1-token horizon lengths. However, since I'm only considering one single pathway to doing this, I think the model as a whole is slightly more likely to *overestimate* the parameter-requirements than to *underestimate* them.

Economically useful tasks

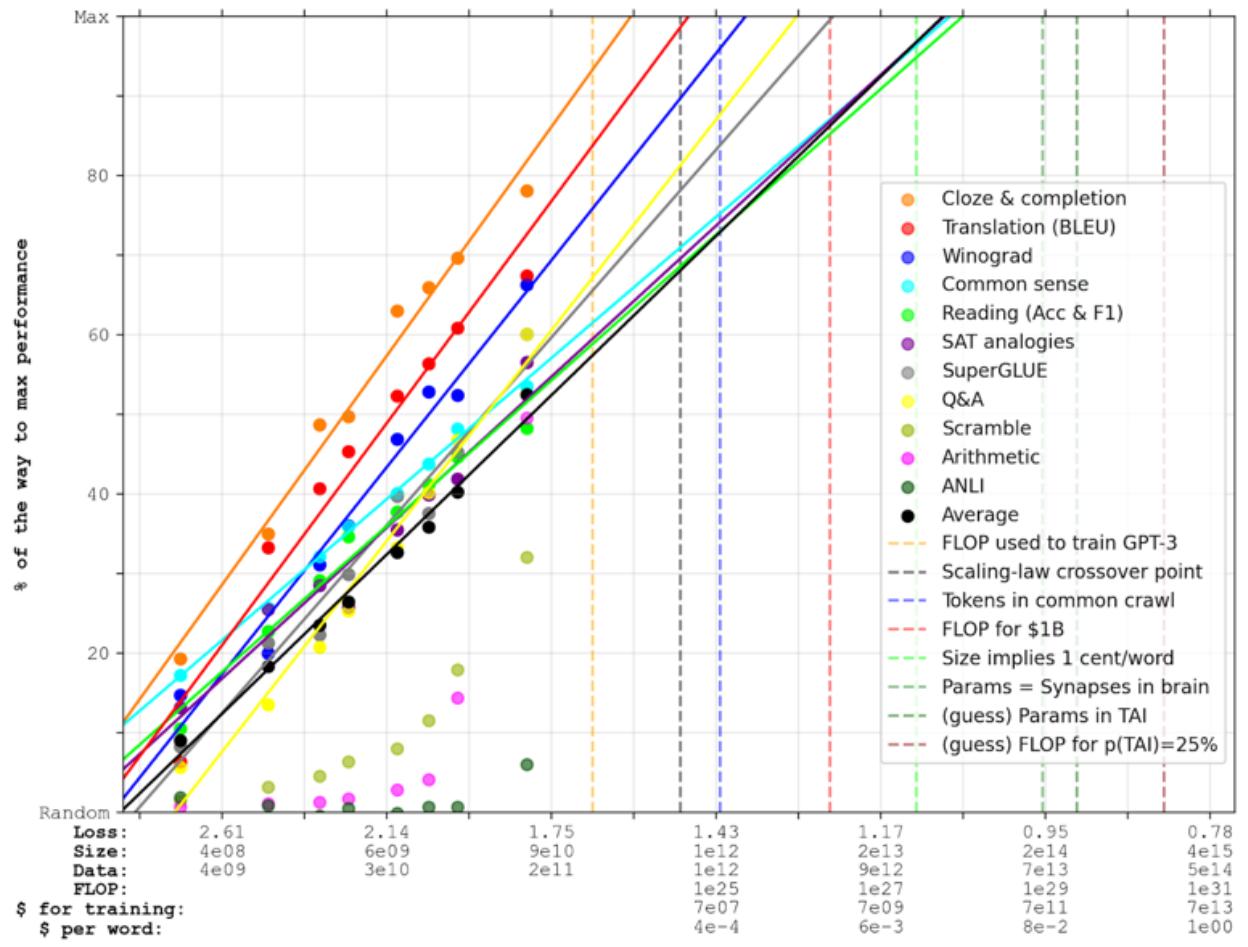
Less conceptually fraught, we can ask whether to expect systems with near-optimal benchmark performance to be able to do economically useful tasks. Here, my basic expectation is that such a system could quite easily be adapted to automating lots of specific tasks, including the ones that Cotra mentions as examples of short-horizon tasks [here](#):

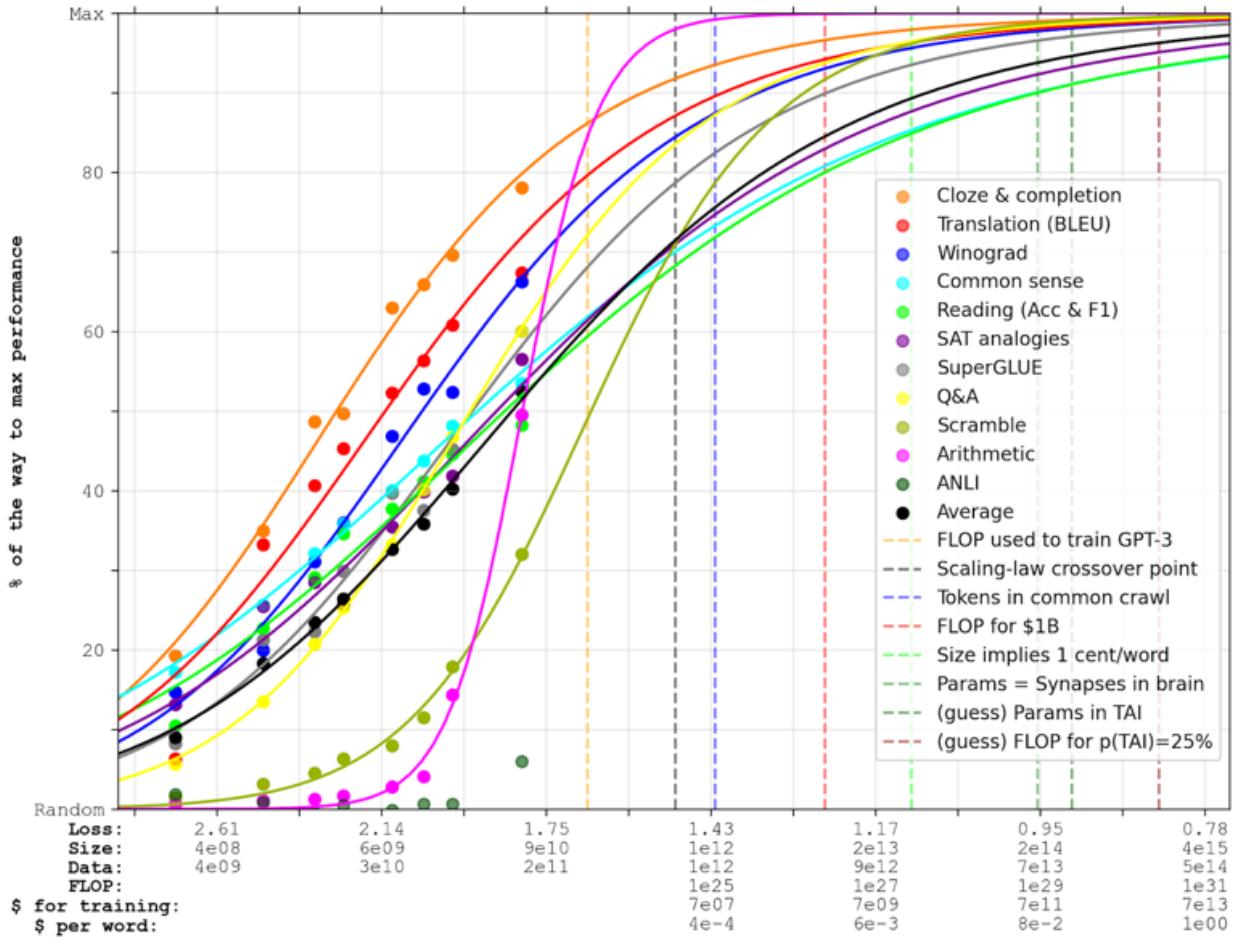
- **Customer service and telemarketing:** Each interaction with a customer is brief, but ML is often required to handle the diversity of accents, filter out noise, understand how different words can refer to the same concept, deal with customization requests, etc. This is currently being automated for drive-thru order taking by the startup [Apprente](#) (acquired by McDonald's).
- **Personal assistant work:** This could include scheduling, suggesting and booking good venues for meetings such as restaurants, sending routine emails, handling routine shopping or booking medical and dental appointments based on an understanding of user needs, and so on.
- **Research assistant work:** This could involve things like copy-editing for grammar and style (e.g. [Grammarly](#)), hunting down citations on the web and including them in the right format, more flexible and high-level versions of "search and replace", assisting with writing routine code or finding errors in code, looking up relevant papers online, summarizing papers or conversations, etc.

Some of the benchmarks directly give evidence about these tasks, most clearly unambiguous understanding of ambiguous text, ability to memorise answers to large numbers of questions, and ability to search text for information (and understand when it isn't available, so that you need to use a human). Writing code isn't directly related to any of the benchmarks, but given [how well it already works](#), I assume that it's similarly difficult to other natural language tasks, and would improve in line with them.

Comparisons and limits

Finally, I've augmented the x-axis with some reference estimates. I've estimated *cost for training* by multiplying the FLOP with current compute prices; and I've estimated *cost per word* during inference from the current pricing of GPT-3 (adjusting for network size). I have also added some dashed lines where interesting milestones are passed (with explanations below):





(Edit: I made a version of this graph with data points from a later language model — PaLM — [here](#).)

In order:

- The orange line marks the FLOP used to train GPT-3, which is ~6x larger than what the inferred FLOP of the right-most data points would be. As explained in the Appendix, this is because GPT-3 is small enough that it needs multiple epochs to fully benefit from the data it's trained on. I expect the projection to be more accurate after the black line, after which the scaling law I use starts predicting higher compute-requirements than other scaling laws (again, see the Appendix).
- According to [Brown et al.](#), there are a bit less than $1e12$ words in the common crawl (the largest publicly available text dataset), which means there are a bit more than $1e12$ tokens (blue line). What other data could we use, beyond this point?
 - We could use more internet data. Going by [wikipedia](#), expanding to other languages would only give a factor ~2. However, common crawl claims to have “petabytes of data”, while google [claims](#) to have well over 100 petabyte in their search index, suggesting they may have 10-100x times more words, if a similar fraction of that data is usable text. However, on average, further data extracted from the internet would likely be lower-quality than what has been used so far.

- As of last year, [Google books](#) contained more than 40 million titles. If each of these had 90,000 words, that would be $\sim 4 \times 10^{12}$ words (of high quality).
- We could start training on video. I think the total number of words spoken on youtube is around 1 trillion, so just using speech-to-text wouldn't add much, but if you could usefully train on predicting pixels, that could add enormous amounts of data. I definitely think this data would be less information-rich per byte, though, which could reduce efficiency of training by a lot. Perhaps the right encoding scheme could ameliorate that problem.
- If they wanted to, certain companies could use non-public data generated by individuals. For example, 3×10^9 emails [are sent every year](#), indicating that Google could get a lot of words if they trained on gmail data. Similarly, I suspect they could get some mileage out of words written in google docs or words spoken over google meet.

Overall, I haven't found a knock-down argument that data won't be a bottleneck, but there seems to be enough plausible avenues that I think we could scale at least 10x past the common crawl, *if* there's sufficient economic interest. Even after that I would be surprised if we completely ran out of useful data, but I wouldn't be shocked if training became up to ~ 10 x more expensive from being forced to switch to some less efficient source.

- The red line marks the FLOP that could be bought for \$1B, assuming 2.4×10^{17} FLOP/\$. [\[10\]](#) (Training costs are also written on the x-axis.) I think this is within an order of magnitude of the total investments in OpenAI[\[11\]](#), and close to DeepMind's yearly spending. Google's total yearly R&D spending is closer to \$30B, and Google's total cash-on-hand is $\sim \$130$ B. One important adjustment to bear in mind is that hardware is getting cheaper:
 - Over the last 40-50 years, FLOP/s/\$ has fallen by 10x every ~ 3 -4 years. [\[12\]](#)
 - Over the last 12 years, FLOP/s/\$ has fallen by 10x just once. [\[13\]](#)
 - As a measure of gains from hardware specialisation, over the last 5 years, [fused multiply-add operations](#)/s/\$ (which are especially useful for deep learning) has fallen by about 10x[\[14\]](#). This sort of growth from specialisation can't carry on forever, but it could be indicative of near-term gains from specialisation.

Cotra's best guess is that hardware prices will fall by 10x every 8-9 years. I think faster progress is plausible, given the possibility of further specialisation and the older historical trend, but I'm pretty uncertain.

- The light green line marks the point where reading or writing one word would cost 1 cent, if cost were to linearly increase with size from today's [250 tokens / cent](#). (Cost/word is also written on the x-axis.) For reference, this would be as expensive as paying someone \$10/hour if they read/wrote 15 words per minute, while freelance writers typically charge 3-30 cents per word. As long as GPT-N was subhuman on all tasks, I think this could seriously limit the usefulness of applying it to many small, menial tasks. However, hardware and algorithmic progress could substantially ameliorate this problem. Note that the cost of inference scales in proportion with the *size*, while the total training costs scale in proportion to *size*data*, which is proportional to $size^{1.74}$. This means that if FLOP/\$ is reduced by 10x, and we train with 10x more FLOP, the total inference costs are *reduced* by a factor ~ 3 . [\[15\]](#)

- There are roughly $2e14$ synapses in the human brain ([source](#)), which is approximately analogous to the number of parameters in neural networks (green line).
- The dark green line marks the *median* estimate for the number of parameters in a transformative model, according to Ajeya Cotra's model^[16]. Noticeably, this is quite close to when the benchmarks approaches optimal performance. The 80% confidence interval is between $3e11$ and $1e18$ parameters, going all the way from the size of GPT-3 to well beyond the edge of my graph.
- Finally, the last dashed line marks the number of FLOP for which Cotra's current model predicts that 2020 methods would have a 25% chance of yielding TAI, taking into account that the effective horizon length may be longer than a single token.

Finally, it's important to note that algorithmic advances are real and important. GPT-3 still uses a somewhat novel and unoptimised architecture, and I'd be unsurprised if we got architectures or training methods that were one or two orders of magnitude more compute-efficient in the next 5 years.

Takeaways and conclusions

Overall, these are some takeaways I have from the above graphs. They are all tentative, and written in the spirit of exposing beliefs to the light of day so that they can turn to ash. I encourage you to draw your own conclusions (and to write comments that incinerate mine).

- On benchmark performance, GPT-3 seems to be in line with performance predicted by smaller sizes, and doesn't seem to particularly break or accelerate the trend.
 - While it sharply increases performance on arithmetic and scramble tasks in particular, I suspect this is because they are narrow tasks which are easy once you understand the trick. If future transformative tasks are similarly narrow, we might be surprised by further scaling; but insofar as we expect most value to come from good performance on a wide range of tasks, I'm updating towards a *smaller* probability of being very surprised by *scaling alone* (ie., I don't want to rule out sudden, surprising *algorithmic progress*).
 - Of course, sudden increases in spending can still cause sudden increases in performance. GPT-3 is arguably an example of this.
 - Given the steady trend, it also seems less likely to suddenly stop.
- Close-to-optimal performance on these benchmarks seems like it's at least ~ 3 orders of magnitude compute away (costing around \$1B at current prices). This means that I'd be somewhat surprised if a 100x scaling brought us there immediately; but another 100x scaling after that might do it (for reference, a 10,000x increase in compute would correspond to a bit more than 100x increase in size, which is the difference between GPT-2 and GPT-3). If we kept scaling these models naively, I'd think it's more likely than not that we'd get there after increasing the training FLOP by $\sim 5\text{-}6$ orders of magnitude (costing \$100B-\$1T at current prices).
 - Taking into account both software improvements and potential bottlenecks like data, I'd be inclined to update that *downwards*, maybe an order of magnitude or so (for a total cost of $\sim \$10\text{-}100B$). Given hardware

improvements in the next 5-10 years, I would expect that to fall further to ~\$1-10B.

- I think this would be more than sufficient for automating the tasks mentioned above – though rolling out changes in practice could still take years.
- (Note that some of these tasks could be automated with today’s model sizes, already, if sufficient engineering work was spent to fine-tune them properly. I’m making the claim that automation will quite easily be doable by this point, if it hasn’t already been done^[17].)
- Assuming that hardware and algorithmic progress have reduced the cost of inference by at least 10x, this will cost less than 1 cent per word.
- I think this would *probably* not be enough to automate the majority of human economic activity or otherwise completely transform society (but I think we should be investing substantial resources in preparing for that eventuality).
- If I adopt the framework from Ajeya Cotra’s draft report – where a model with the right number of parameters can become ~human-equivalent at tasks with a certain horizon length if trained on the right number of data points of that horizon length – I’m inclined to treat these extrapolations as a guess for how many parameters will be required for ~human-equivalence. Given that Cotra’s model’s median number of parameters is close to my best guess of where near-optimal performance is achieved, the extrapolations do not contradict the model’s estimates, and constitute some evidence for the median being roughly right.

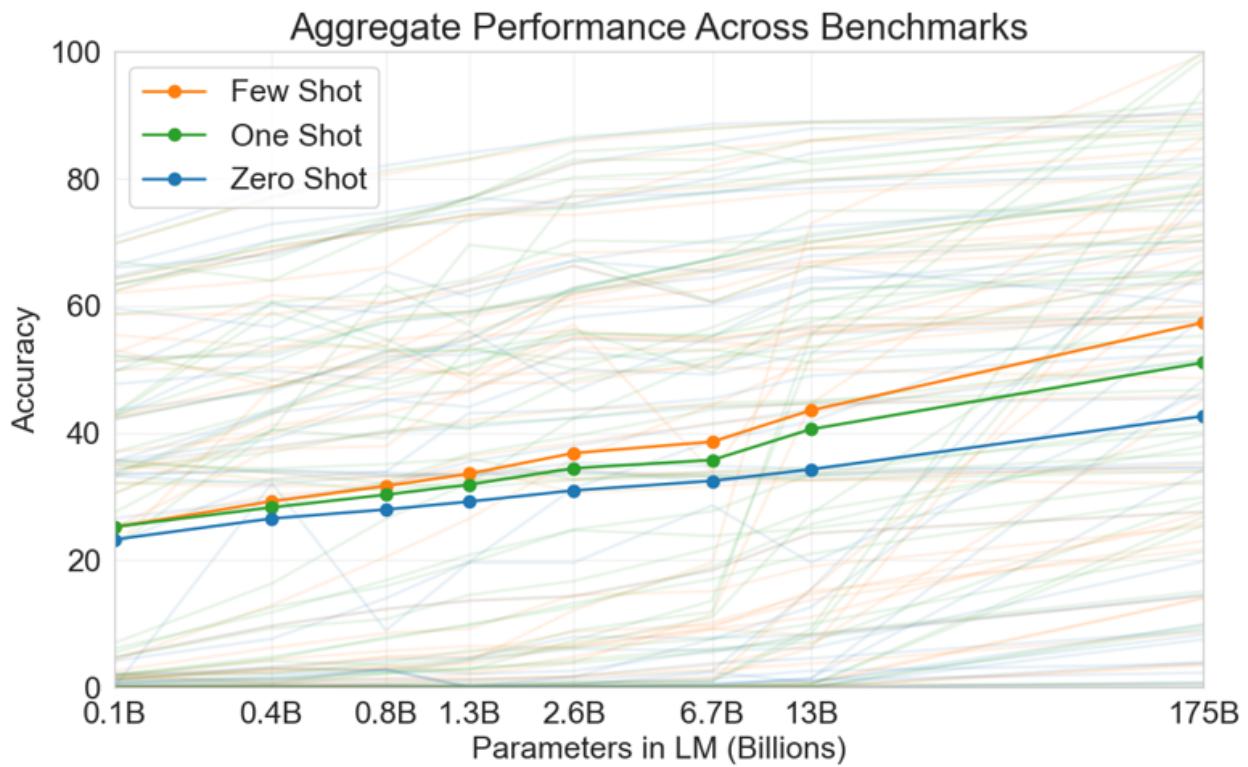
I’m grateful to Max Daniel and Hjalmar Wijk for comments on this post, and to Joseph Carlsmith, Daniel Kokotajlo, Daniel Eth, Carolyn Ashurst and Jacob Lagerros for comments on earlier versions.

Appendix

In this appendix, I more thoroughly describe why we can’t fit plots to the size of models directly, why I average over categories rather than over all benchmarks, and why I chose the scaling laws I did. Feel free to skip to whatever section you’re most interested in.

Why not plot against size?

[Brown et al.](#) trains models of 8 different sizes on 300 billion tokens each, and reports their performance on a number of benchmarks in Figure 1.3:



Each faint line depicts the accuracy on a certain benchmark as a function of model size. The thick lines depict the average accuracy across all benchmarks, for few-shot, one-shot, and zero-shot evaluation respectively.

However, when extrapolating the performance, what we care about is the best performance (measured by the validation loss) that we can get for a given amount of compute, if we choose model size and number of data points optimally. For this, it's a bad idea to fit performance to the model size (as in the graph above), because all models were trained on 300B tokens. For the small models, this is inefficiently large amounts of data, which means that they're disproportionately good compared to the largest model, which only barely receives the optimal amount of data. Thus, naively extrapolating results based on model-size would *underestimate* how much larger model sizes improve performance, when optimally trained. If fit to a linear trend: It would *underestimate* the slope and *overestimate* the intercept.

Why plot against loss?

To get around the problems with plotting against size, I fit a trend for how benchmark performance depends on the cross-entropy loss that the models get when predicting the next token on the validation set (which can be read off from [Figure 4.1](#) in [Brown et al.](#)). I then use scaling laws to extrapolate how expensive it will be to get lower loss (and by extension better benchmark performance).

The crucial assumption that this procedure makes is that – in the context of training GPT-like transformers of various sizes on various amounts of data – text-prediction cross-entropy loss is a good proxy for downstream task performance. In particular, my procedure would fail if small models trained on large amounts of data were systematically better or worse at downstream tasks than large models trained on small amounts of data, even if both models were exactly as good at text prediction. I'm quite

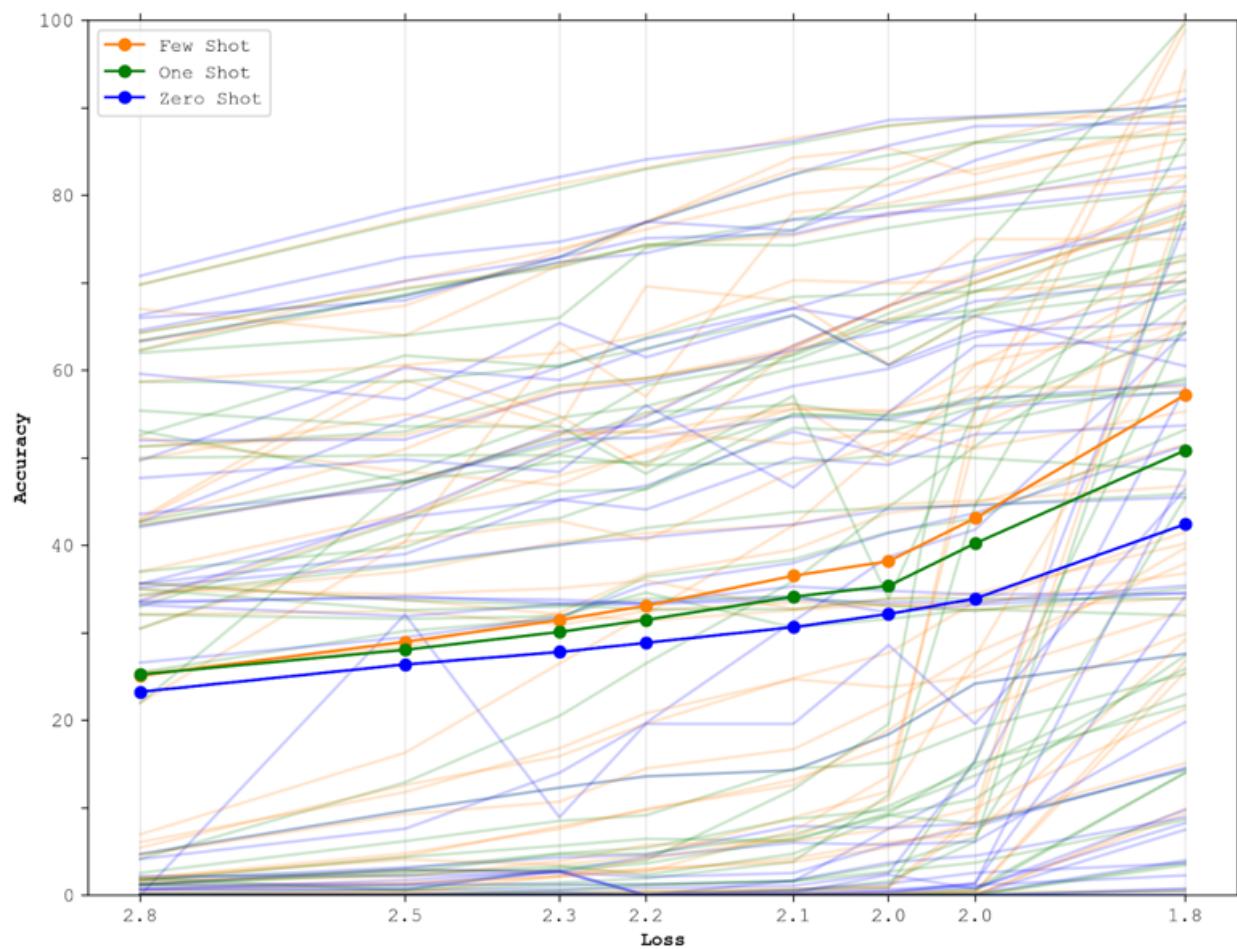
happy to make this assumption, because it does seem like lower loss on text prediction is an excellent predictor of downstream task-performance, and small deviations on single benchmarks hopefully averages out.

Note that I'm not assuming anything else about the relationship between task performance and loss. For example, I am not assuming that improvements will be equally fast on all tasks.

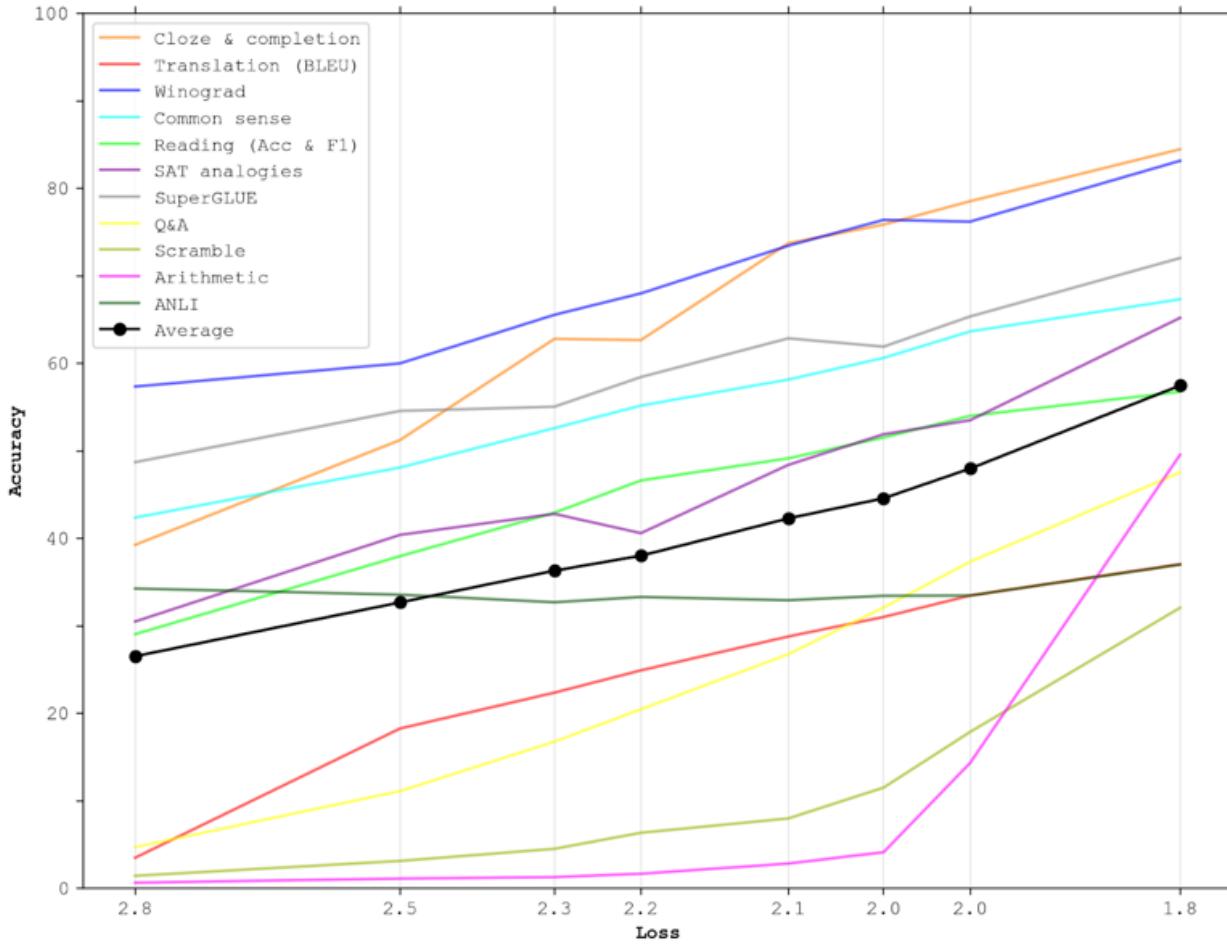
In all graphs in this post, I fit trends to the *logarithm* of the loss, in particular. This is because the loss is related as a power-law to many other interesting quantities, like parameters, data, and compute (see the next section); which means that the *logarithm* of the loss has a *linear* relationship with the *logarithm* of those quantities. In particular, this means that having the logarithm of the loss on the x-axis directly corresponds to having logarithms of these other quantities on the x-axis, via simple linear adjustments. It seems very natural to fit benchmark performance against the logarithm of e.g. compute, which is why I prefer this to fitting it to the loss linearly.

One potential issue with fitting trends to the loss is that the loss will eventually have to stop at some non-zero value, since there is some irreducible entropy in natural language. However, if the log-loss trends start bending soon, I suspect that downstream task performance will *not* stop improving; instead, I suspect that the benchmark trends would carry on as a *function of training compute* roughly as before, eventually slowly converging to their own irreducible entropy. Another way of saying this is that – insofar as there's a relationship between text prediction loss and benchmark performance – I expect that relationship to be best captured as a steady trend between *reducible* prediction loss and *reducible* benchmark performance; and I expect both to be fairly steady as a function of training compute (as showcased in OpenAI's [Scaling Laws for Autoregressive Generative Modelling](#)).

Here's the aggregation that OpenAI does, but with my adjusted x-axis:



This graph looks similar to OpenAI's, although the acceleration at the end is more pronounced. However, we can do better by separating into different categories, and taking the average across categories. The following graph only depicts few-shot performance.



As you can see, the resulting graph has a much less sharp acceleration at the end. The reason for this is that the arithmetic task category has more benchmarks than any other category (10 benchmarks vs a mean of ~4.5 across all categories), which means that its sudden acceleration impacts the average disproportionately much. I think weighing each category equally is a better solution, though it's hardly ideal. For example, it counts SuperGLUE and SAT analogies equally much, despite the former being an 8-benchmark standard test suite and the latter being a single unusual benchmark.

In the main post above, I use a normalised version of this last graph. My normalization is quite rough. The main effect is just to adjust the minimum performance of benchmarks with multi-choice tasks, but when human performance is reported, I assume that maximum performance is ~5% above. For translation BLEU, I couldn't find good data, so I somewhat arbitrarily guessed 55% as maximum possible performance.

What scaling laws to use?

What scaling law best predicts how much compute we'll need to reach a given loss? Two plausible candidates from [Kaplan et al](#) are:

- The compute-optimal scaling law $L(C)$, which assumes that we have unlimited data to train on. For any amount of compute C , this scaling law gives us the minimum achievable loss L , if we choose model size and training time optimally.

- The data-constrained scaling law $L(N,D)$, which for a given model size N tells us the loss L we get if we train *until convergence* on D tokens.

Intuitively, we would expect the first law to be better. However, it is highly unclear whether it will hold for larger model sizes, because if we extrapolate both of these laws forward, we soon encounter a *contradiction* (initially described in section 6.3 of [Kaplan et al](#)):

To train until convergence on D tokens, we need to train on each token *at least once*. Training a transformer of size N on D tokens requires $\sim 6ND$ FLOP. Thus, if we choose N and D to minimise the product $6ND$ for a given loss $L(N,D)$, we can get a *lower bound* on the FLOP necessary to achieve that loss.

However, this lower bound eventually predicts that we will need *more* compute than the compute-optimal scaling law $L(C)$ does, mostly because $L(C)$ predicts that you can scale the number of tokens you train on much slower than $L(N,D)$ does. The point where these curves first coincide is around ~ 1 trillion parameters, and it's marked as the crossover point in my section **Comparisons and limits**. The best hypothesis for why this happens is that, as you scale model size, the model gets better at updating on each datapoint, and needs fewer epochs to converge. $L(C)$ picks up on this trend, while $L(N,D)$ doesn't, since it always trains until convergence. However, this trend cannot continue forever, since the model cannot converge in less than one epoch. Thus, if this hypothesis is correct, scaling will eventually be best predicted by $L(N,D)$, running a single epoch with $6ND$ FLOP. For a more thorough explanation of this, see section 6 of OpenAI's [Scaling Laws for Autoregressive Generative Modelling](#), or [this summary](#) by nostalgebraist.

It's possible that this relationship will keep underestimating compute-requirements, if it takes surprisingly long to reach the single epoch steady state. However, it seems unlikely to underestimate compute requirements by more than 6x, since that's the ratio between the compute that GPT-3 was trained on and the predicted minimum compute necessary to reach GPT-3's loss.

(Of course, it's also possible that something unpredictable will happen at the place where these novel, hypothesized extrapolations start contradicting each other.)

Adapting the scaling law

The scaling law I use has the form $L(N, D) = ((\frac{N}{N_C})^{a_N} + \frac{D}{D_C})^{a_D}$. To simultaneously

minimise the loss L and the product $6ND$, the data should be scaled as:

$$D(N) = \frac{a_D}{a_N} D_C \left(\frac{N}{N_C} \right)^{\frac{a_N}{a_D}}$$

Plugging this into the original formula, I get the loss as a function of N and D :

$$L(N) = \left(1 + \frac{a_D}{a_N} \left(\frac{N}{N_C} \right)^{\frac{a_N}{a_D}} \right)^{\frac{a_D}{a_N}}, L(D) = \left(1 + \frac{a_D}{a_N} \left(\frac{D}{D_C} \right)^{\frac{a_D}{a_N}} \right)^{\frac{a_N}{a_D}}$$

By taking the *inverse* of these, I get the appropriate N and D from the loss:

$$N(L) \approx 1.35e14/L^{13.2}, D(L) \approx 4.24e13/L^{9.71}$$

As noted above, the FLOP necessary for training until convergence is predicted to eventually be $6N(L)D(L)$.

I use the values of N_C , D_C , α_N , and α_D from page 11 of [Kaplan et al.](#). There are also some values in Figure 1, derived in a slightly different way. If I use these instead, my final FLOP estimates are about 2-5x larger, which can be treated as a lower bound of the uncertainty in these extrapolations.

(As an aside: If you're familiar with the details of Ajeya Cotra's [draft report on AI timelines](#), this extrapolation corresponds to her [target accuracy law](#) with $q \sim= 0.47$)

Notes

1. [TriviaQA](#): The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone [←](#)

2. [LAMBADA](#): "Yes, I thought I was going to lose the baby." "I was scared too," he stated, sincerity flooding his eyes. "You were?" "Yes, of course. Why do you even ask?" "This baby wasn't exactly planned for." "Do you honestly think that I would want you to have a ____ ?"

Answer: miscarriage [←](#)

3. [HellaSwag](#): A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

- A. rinses the bucket off with soap and blow dry the dog's head.
- B. uses a hose to keep it from getting soapy.

C. gets the dog wet, then it runs away again.

- D. gets into a bath tub with the dog. [←](#)

4. [Winogrande](#): Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had less time to get ready for school.

Answer: Robert [←](#)

5. [DROP](#): That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amidst the artist's signature scrawls, was sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.

How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?

Answer: 4300000 [←](#)

6. [PIQA](#): How do I find something I lost on the carpet?
 - A. Put a solid seal on the end of your vacuum and turn it on.

B. Put a hair net on the end of your vacuum and turn it on. [←](#)
 7. [ANLI](#): A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term “melee” originates in the 1640s from the French word “mélée”, which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories.
- Hypothesis:* Melee weapons are good for ranged and hand-to-hand combat.
- Answer:** Contradicted [←](#)
8. Although a complicating factor is that humans can process a lot more visual data than verbal data per second, so image-recognition should plausibly be counted as having a longer horizon length than GPT-3. I'm not sure how to unify these types of differences with the *subjective second* unit in Cotra's [report](#). [←](#)
 9. Note that Cotra's model is ultimately trying to estimate [2020 training computation requirements](#). By definition, this requires that researchers mostly rely on 2020 algorithmic knowledge, but allows for 2-5 years to design the best solution. [←](#)
 10. Based on the [appendix](#) to Ajeya Cotra's [draft report](#), and adjusted upwards 2x, because I think transformers have quite good utilization [←](#)
 11. OpenAI started out with \$1B, and their biggest investment since then was Microsoft giving them another \$1B. [←](#)
 12. See Ajeya Cotra's [appendix](#). [←](#)
 13. See Ajeya Cotra's [appendix](#) and Asya Bergal's [Recent trends in GPU price per FLOPS](#). Note that, if we measure trends in *active prices* (rather than *release prices*) over the last 9 years, we would expect a 10x cost reduction to take 17 years instead. [←](#)
 14. From Asya Bergal's [Recent trends in GPU price per FLOPS](#). [←](#)
 15. Since size is only increased by $10^{1/1.74}$, while costs are reduced by 10, yielding $10/10^{1/1.74} \approx 2.66$. [←](#)
 16. The model first estimates the FLOP/s of a human brain; then adds an order of magnitude because NNs will plausibly be less efficient, and finally transforms FLOP/s to parameters via the current ratio between the FLOP that a neural network uses to analyze ~1 second of data, and the parameters of said neural network. [←](#)
 17. Slightly more generally: I'm pointing out *one particular path* to automating these tasks, but presumably, we will in fact automate these tasks using the *cheapest* path of all those available to us. Thus, this is necessarily a (very shaky) estimate of an upper bound. [←](#)

Anti-EMH Evidence (and a plea for help)

The debate over EMH should perhaps be framed as "What skill level and assets under management do I need to make it worthwhile to play the markets instead of doing passive investing?" This is a list of anti-EMH evidence (that I personally came across in my relatively short exploration into the public markets), in the sense that they've updated me towards thinking that the levels are not as high as I had thought. (Compare with "Modern Edges are Completely Ridiculous" section in [The EMH Aten't Dead.](#))

Negative extrinsic value for HYLN warrants

Warrants are similar to call options, in that you get the right to purchase shares of common stock at a set price, which is \$11.5 for HYLN warrants. However HYLN commons have been trading at \$15-\$16 above HYLN warrants. There are some complications including that HYLN warrants are not exercisable right now and won't be for a few days to weeks until SEC approves some filings, but it's still hard to explain the negative extrinsic value assuming EMH. (Many SPAC stocks exhibit this, not just HYLN.)

(This was written some days ago and the warrants have since become exercisable, and the price gap has closed.)

Equivalent asset arbitrage

Two securities (symbols to come later as this is still being actively traded) are supposed to give the same dividend stream. The company's official website states that they are meant to be economically equivalent. Until recently these two symbols have been priced very close to each other, but one asset started trading at a premium to the other in the last few weeks, sometimes >10%, with the delta swinging back and forth over this time period creating repeated arbitrage opportunities.

AHT commons versus preferred

There was an exchange offer that expired on Nov 20, where 1 preferred share of AHT could be exchanged for 5.58 shares of AHT common stock, but AHT commons have been trading at way above 1/5.58 the price of AHT preferred and was as high as 1/2 in the last couple of days before expiration (technically Nov 17 and 18, because if you bought on Nov 19 or 20, the shares wouldn't settle in time to participate in the exchange). The risk of the exchange offer being canceled or changed doesn't seem nearly high enough to explain this, and in fact the exchange did go through at the 1:5.58 ratio.

Senseless HTZ spike on DIP news

I previously mentioned this on [Open Thread](#). To quote a commenter (on a paid subscription website):

For the last several months HTZ shares were slowly deteriorating and by the middle of this month actually got quite close to dropping below \$1/share. Then at the end of last week couple of events caused HTZ shares to skyrocket +143% on a \$2.5bn volume (for a bankrupt stock!):

- On the 16th of October, HTZ announced a new \$1.65bn DIP (debtor in possession) financing. It seems that the market (or overly excited retail crowds) views this as an extremely positive sign indicating that HTZ is still able to raise money and will come out of bankruptcy soon. However, I think that equity still remains pretty much worthless here, while the timeline of the whole process is even more questionable now. The company continues to intensively burn cash (in August alone \$84m cash was burned in operations) and the size of the new DIP indicates that the trend is not going to stop soon. Initially, at the time of the bankruptcy announcement, the company intended to raise \$1bn equity, while DIP seemed not to be in the cards at all. September DIP overview filings show estimated DIP sizing at \$1.1bn – \$1.5bn, while the recently announced deal is substantially larger than this range. Back in August (Q2 report) the company stated that “there is a significant risk that the holders of our common stock will receive no recovery under the Chapter 11 Cases and that our common stock will be worthless”, and this current financing pushes shareholders even further down in the seniority ranking. Regarding the timing, it seems that DIP lenders do not expect the bankruptcy process will be rather extended: “The DIP Facility matures on December 31, 2021, and has limited covenants and events of default, including one milestone requiring the filing of a Chapter 11 Plan by August 1, 2021.”
- Apparently, the delisting hearing with NYSE was supposed to take place on the 15th of October, however, no updates were issued and so far HTZ remains listed. It seems that the market has understood this as a sign that HTZ will not get delisted. I don't think the assumption is correct and we still should wait for the actual update from the exchange. The newly received loan gives the company more flexibility and survival time, however, it still doesn't put HTZ in “sound financial health”, which is one of the requirements to regain compliance with the listing requirements.

Overall, it is hard to explain the almost 2.5x equity valuation increase on Friday. It was likely a combination of hype pumping, short-squeeze, and algo-trading. This remains to be an interesting case to track (definitely not investable neither on the long nor short sides though).

Contra "not investable", I did sell call options into the spike and made a nice profit, although it was nerve-racking at times.

Three "risk-free" assets that rose 70-200% in 2 months

("[Risk-free](#)" in the sense of not being significantly more risky than short term treasuries.) A [SPAC](#) is a "blank check" company that gathers money from investors at IPO, puts that money into a trust account (which are invested in short-term treasuries or money-market equivalents), then finds a private company to merge with (in order to deliver the IPO proceeds to that company and to make it publicly traded). Before the merger is finalized, each shareholder can request to "redeem" their shares for their share of the money in the trust account (called NAV or redemption value, typically around \$10). Because of this, it's "impossible" to lose money buying a SPAC stock at price=NAV or below as long as you hold it until redemption time. I did this with three SPAC stocks, symbols GMHI (now LAZR), TRNE, and HCAC, and their prices have since gone up 70-200%. I bought these on [portfolio margin](#) and financed them with [box-spread financing](#) so the opportunity cost of holding them was also very low.

(Technically I bought these at slightly above NAV, and brought their effective prices below NAV by selling November call options against them. I also bought some other SPACs but these three were/are my biggest positions by far.)

How I noticed or came across these

- I've been following SPAC stocks (due to hearing about the NKLA debacle from friends and in the financial press) and the discount of warrants relative to commons is frequently mentioned on r/SPACs.
- For the equivalent asset arbitrage I was holding one of the securities in a dividend portfolio (for my parents), saw a big runup in its price, and thought of checking its price relative to its twin (which I knew about from previous research) just in case, and happened to see a significant price difference.
- For HTZ, I was already short HTZ due to an earlier irrational price spike which I read about in the news.
- For AHT, I read an article about the exchange offer and the arbitrage opportunity it represented on SeekingAlpha a couple of months ago, and then observed the opportunity get much bigger in the last few days just before the offer expiration.
- For the "riskless" assets, buying SPACs while they're near NAV to sell later is a strategy frequently mentioned on r/SPACs. I picked these three stocks myself (due to their being related to tech and electric vehicles) and independently thought of selling calls to bring their effective prices below NAV.

Addendum: Help me time the SPAC bubble

With GMHI/LAZR, I sold 80% of my position while the prices were around 16-17, and it's now around 32. This was partly because I thought I had a tendency to [close my trades too late](#) and wanted to correct for that, and partly because I had observed a previous SPAC bubble deflate in real time, and every downturn in prices reminded me of that. My initial positions were large enough and these stocks have gone up enough that the remaining positions now represent nearly 50% of the net value of my investment portfolio. So, any advice, either theoretical or practical, on what to do with TRNE and HCAC, which have mergers coming up soon, which (judging from history of other SPACs) implies likely further price increases as long as the current SPAC bubble stays inflated?

(I seem to recall a recent LW post recommending *not* to time bubbles, but can no longer find it. A link would be appreciated.)

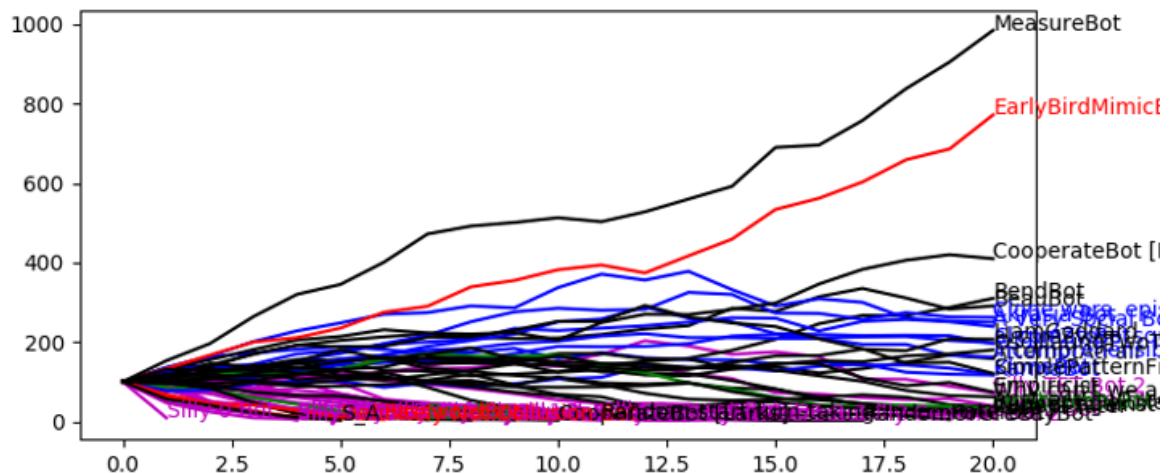
The Darwin Game - Conclusion

Evolution is unintelligent. The bugs removed intelligence from the design of the bots. The more bugs I wrote into my simulator, the better my simulation replicated real-world Darwinian population dynamics. After two alternate timelines with a buggy game engine, I have finally gotten around to running the game for real.

Alas, this game is between intelligently-designed species, not randomly-generated chunks of code.

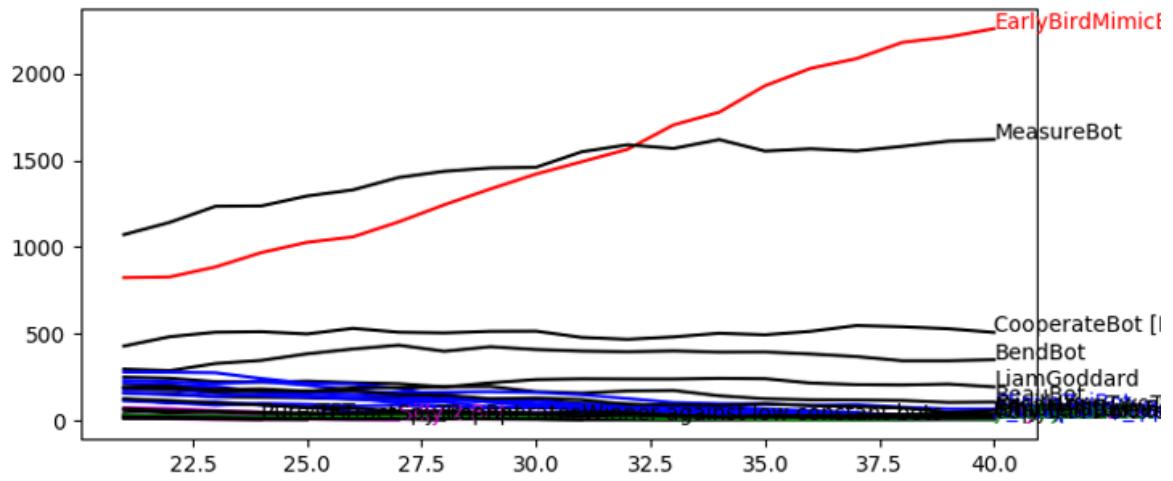
Rounds 0-20

MeasureBot takes an early lead.



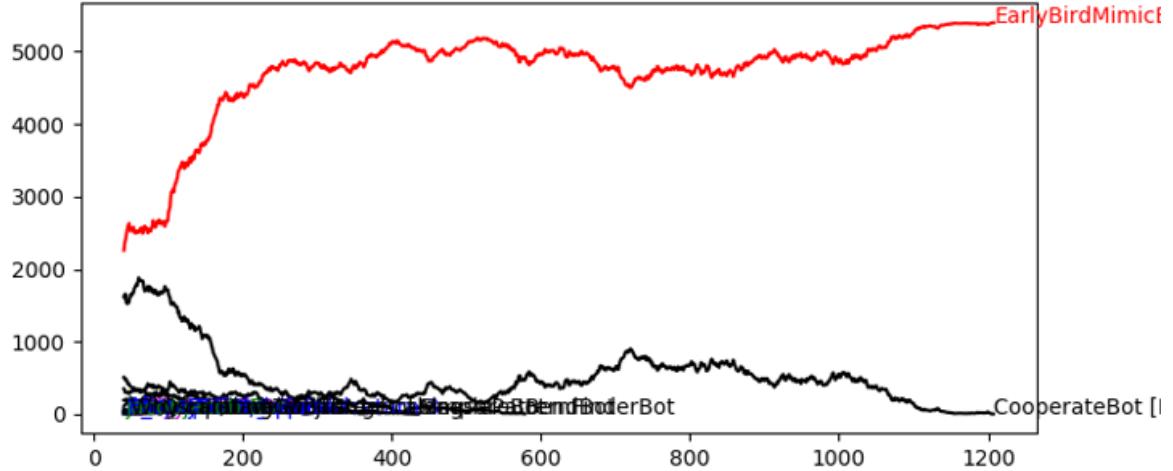
Rounds 21-40

Multicore's EarlyBirdMimicBot steals the lead from MeasureBot.



Rounds 41-1208

Welcome to Planet Multicore.



Winner

Bot	Team	Description	Round
EarlyBirdMimicBot	Multicore	Superintelligence	∞

Today's Obituary

Everyone else.

Conclusion

I hope you had fun. This wouldn't have been possible without the community here at Less Wrong. At least 75% of the code (not counting pseudocode) was written by people other than me. Thank you everyone who competed, debated, plotted and hacked. Thank you for the espionage and counter-espionage. Thank you everyone who helped spot bugs in the game engine. Thank you Zvi for posting the original Less Wrong Darwin Game series. Extra thanks to moderator Ben Pace for prettifying the tables behind the scenes and moderator Oliver for fixing multiple timestamps.

The source code to the game and all the bots is available [here](#). If there is a bug in this timeline you can fix it yourself.

This concludes the 2020 Less Wrong Darwin Game.

Parable of the Dammed

Once upon a time, two families fought a bloody feud over the border of their properties. After many years of escalation, neither family could afford to continue the fight. They both wanted to negotiate a truce, but the original issue still had to be settled: where was the border of their lands to be drawn? They needed a [Schelling point](#), so they settled on a river which ran roughly through the middle of their territories.

For a short time, there was peace.

Soon, though, a clever couple from one of the families hatched an idea. Each morning, they walked down to the river and dropped a few large stones into it. Before long, half a dam had built up on their side of the river. The water was driven toward the opposite bank, which was steadily washed away. Over time, as the river ate away the opposite bank, the couple extended their dam further, and so the river's course was gradually pushed sideways. The couple's family gained territory, while the opposed family lost it.

At this point, the story diverges, and many versions of the tail are told.

In one version, the couple push too fast. Soon the river has moved deep into the territory of the other family, and the family responds by attempting to break the dam. Violence escalates, and the feud breaks out anew - but peace is even harder to come by, now, since the river has been permanently destroyed as a Schelling point.

In another version, the push is slow. The couple bequeaths the task of dam-building to their children, and to their childrens' children, and the river shifts slowly over the course of generations. With each generation, the resources of the couple's descendants grow, and their family grows with it - while the resources of the opposed family slowly dwindle. Nobody ever takes much note of the river's slow drift, until eventually the opposed family dies out altogether.

In most versions of the tale, the river's movement is quickly noticed, but a return to violence is deemed unacceptable. Instead, the opposed family begins dropping rocks of their own. Soon both families are dumping rocks on their respective sides of the river, building up dams, aiming to drive the water against the opposite bank. This bloodless but expensive feud escalates. Along some sections of the river, each side expands until the two dams clash in the middle, blocking the flow of the whole river, and water backs up and bursts the banks. That doesn't stop the dam-building - rather, each side builds tall walls alongside their dams, in hopes of flooding the other family's land while preserving their own. The two families quickly bankrupt themselves in an arms race to build the tallest walls along their respective riverbanks.

Word goes out of the strange practices, the two families pouring all their resources into a competition of great dams and flood-walls. Travellers passing through town stare in bemusement, and wonder what strange force would lead the families to waste so much resources on a minor stream through the woods.

Moral(s) of the Story

I see two main takeaways to this parable. First, Schelling points can be moved by changing the underlying territory. The river's course can be physically moved. This generally costs some real resources (e.g. building the dams); modifying the world is rarely free.

Second, when players compete to move a Schelling point, they often end up in an all-pay auction: all players spend the resources required to move the river, but only the player with the "highest bid" (i.e. tallest dam) gains anything from the competition. In general, all-pay auctions often lead to all players spending more than the value of winning: at any point, either family can gain by building their dam just a bit taller, even long after their dam-building expenditures far exceed the value of the land.

This applies to most of the "strategic negotiation"-style situations where Schelling points play a prominent role, and in particular I see the parable of the dammed as a prototypical model of politics. Politics is an all-pay auction, in which "bidders" (i.e. anyone spending time/resources on political influence) compete to move Schelling points. The Schelling points which people compete to move include obvious things like laws, but also more subtle Schelling points like social norms.

What trade should we make if we're all getting the new COVID strain?

If data keeps coming out in the next week confirming that [the new COVID strain](#) is 70% more transmissible, I think the modal outcome is that ~50% of Americans will get it by the early summer. The market may take a few days to realize and react to this (as it was in March), but also just buying June put option on the SP500 seems very naïve (since the SP500 is at all-time highs and a fourth COVID wave doesn't necessarily affect much the NPV of future earning of huge corporations). So if I think that the probability of everyone getting COVID in the next six months is much likelier than the market, at least for a few days, what trade would capture that?

Luna Lovegood and the Chamber of Secrets - Part 9

The Ministry-sanctioned tour of the Chamber of Secrets was pure propaganda. The Marauder's Map showed a second entrance to the Chamber of Secrets hidden in the Girls' Bathroom.

"*Hss hsssss hsss,*" Luna copied from Harry Potter.

A sink opened up into a large tunnel.

"*Lumos,* Luna said.

Each tunnel segment was designed in the same pattern as the previous, as if to deliberately disorient adventurers. Luna closed her eyes in case Slytherin's basilisk was still alive.

"*Nox,*" Luna said.

It was quiet under the school. The tunnels smelled faintly of mildew. Ministry tours turned left at every fork. Luna placed her right hand on the right wall. She felt destiny wrapping around her. Salazar Slytherin's secrets would be hers. Luna followed the right wall until she bumped into the bones of a giant monster.

"*Lumos,*" Luna said.

Some Gryffindor had slayed the basilisk before it could pass on its secrets to the Heir of Slytherin. Ministry archaeologists had defanged the skeleton. So much for destiny.

Salazar Slytherin had circumvented the Interdict of Merlin. But why? If Slytherin had wanted secrets to be preserved then he could have just passed them on to his students. No. Ravenclaw had positioned her diadem to be discovered by someone hunting nargles. Slytherin had done something similar.

But who was his target? Only a Parselmouth could talk to Slytherin's basilisk. Only a Parselmouth could enter the Chamber of Secrets. If only a Parselmouth could talk to the basilisk then it made sense to make the entrance only allow Parselmouths in so that you didn't get a Gryffindor charging in and killing the basilisk before talking to it. Except Slytherin's security by obscurity had failed.

That left the question of why. Parselmouthes were not special. They did not have secondary characteristics. There was no reason for Salazar Slytherin to leave his greatest secrets to someone random to find. The only thing he would know about the Heir of Slytherin was that whoever opened the Chamber of Secrets:

1. Could speak Parseltongue and could therefore communicate with Slytherin's basilisk.
2. Would have opened the Chamber of Secrets.

Chamber of Secrets. Plural.

The mazelike passageways echoed Daedalus's Labryinth. The Heir of Slytherin was supposed to use Salazar's secrets to slay the monster Slytherin himself never could. Instead, a Gryffindor had broken into the Chamber of Secrets first, slayed Slytherin's basilisk, released the real monster, declared victory and returned home.

Luna bolted to the Ministry-sanctioned exit.

The Ravenclaw vs Hufflepuff semifinal duel was that evening. Luna faced two Hufflepuffs. It was supposed to be a doubles tournament.

"I'm not alone," Luna said to Wanda, "I have you."

Gilderoy Lockhart counted down. There was a bang from his wand.

"*Protego*," Luna said.

Luna's shield blocked the first Hufflepuff's sleep hex. The second Hufflepuff's stunning hex smashed through her shield. Luna collapsed.

The Hufflepuffs rushed up to check on Luna.

"*Somnium*," Luna whispered, "*Somnium*."

Luna felt around for the unconscious Wanda, who she stuffed in her pocket. Luna stood up and held her wand high.

Cultural accumulation

Crossposted from [world spirit sock puppet](#).

When I think of humans being so smart due to ‘cultural accumulation’, I think of lots of tiny innovations in thought and technology being made by different people, and added to the interpersonal currents of culture that wash into each person’s brain, leaving a twenty year old in 2020 much better intellectually equipped than a 90 year old who spent their whole life thinking in 1200 AD.

This morning I was chatting to my boyfriend about whether a person who went back in time (let’s say a thousand years) would be able to gather more social power than they can now in their own time. Some folk we know were [discussing](#) the claim that some humans would have a shot at literally take over the world if sent back in time, and we found this implausible.

The most obvious differences between a 2020 person and a 1200 AD person, in 1200 AD, is that they have experience with incredible technological advances that the 1200 AD native doesn’t even know are possible. But a notable thing about a modern person is that they famously [don’t know what a bicycle looks like](#), so the level of technology they might be able to actually rebuild on short notice in 1200 AD is probably not at the level of a nutcracker, and they probably already had those in 1200 AD.

How does 2020 have complicated technology, if most people don’t know how it works? One big part is specialization: across the world, quite a few people do know what bicycles look like. And more to the point, presumably some of them know in great detail what bicycle chains look like, and what they are made of, and what happens if you make them out of slightly different materials or in slightly different shapes, and how such things interact with the functioning of the bicycle.

But suppose the 2020 person who is sent back is a bicycle expert, and regularly builds their own at home. Can they introduce bikes to the world [600 years early](#)? My tentative guess is yes, but not very rideable ones, because they don’t have machines for making bike parts, or any idea what those machines are like or the principles behind them. They can probably demonstrate the idea of a bike with wood and cast iron and leather, supposing others are cooperative with various iron casting, wood shaping, leather-making know-how. But can they make a bike that is worth paying for and riding?

I’m not sure, and bikes were selected here for being so simple that an average person might know what their machinery looks like. Which makes them unusually close among technologies to simple chunks of metal. I don’t think a microwave oven engineer can introduce microwave ovens in 1200, or a silicon chip engineer can make much progress on introducing silicon chips. These require other technologies that require other technologies too many layers back.

But what if the whole of 2020 society was transported to 1200? The metal extruding experts and the electricity experts and the factory construction experts and Elon Musk? Could they just jump back to 2020 levels of technology, since they know everything relevant between them? (Assuming they are somehow as well coordinated in this project as they are in 2020, and are not just putting all of their personal efforts into avoiding being burned at the stake or randomly tortured in the streets.)

A big way this might fail is if 2020 society knows everything between them needed to use 2020 artifacts to get more 2020 artifacts, but don't know how to use 1200 artifacts to get 2020 artifacts.

On that story, the 1200 people might start out knowing methods for making c. 1200 artifacts using c. 1200 artifacts, but they accumulate between them the ideas to get them to c. 1220 artifacts with the c. 1200 artifacts, which they use to actually create those new artifacts. They pass to their children this collection of c. 1220 artifacts and the ideas needed to use those artifacts to get more c. 1220 artifacts. But the new c. 1220 artifacts and methods replaced some of the old c. 1200 artifacts and methods. So the knowledge passed on doesn't include how to use those obsoleted artifacts to create the new artifacts, or the knowledge about how to make the obsoleted artifacts. And the artifacts passed on don't include the obsoleted ones. If this happens every generation for a thousand years, the cultural inheritance received by the 2020 generation includes some highly improved artifacts plus the knowledge about how to use them, but not necessarily any record of the path that got there from prehistory, or of the tools that made the tools that made these artifacts.

This differs from my first impression of 'cultural accumulation' in that:

1. physical artifacts are central to the process: a lot of the accumulation is happening inside them, rather than in memetic space.
2. humanity is not accumulating all of the ideas it has come up with so far, even the important ones. It is accumulating something more like a best set of instructions for the current situation, and throwing a lot out as it goes.

Is this is how things are, or is my first impression more true?

What evidence will tell us about the new strain? How are you updating?

I didn't get into this question in the weekly post. The answer does not seem obvious to me.

Eventually, it will be obvious either way. Either the strain will spread rapidly across many locations, or it won't. But we would like to get the information as fast as possible, whatever the answer turns out to be. So the open question is, what numbers or other information can we use in the short term to generate useful evidence for or against the new strain being a lot more infectious than older ones?

The obvious first things to look at are England's case numbers, and how rapidly the new strain continues to outcompete the old strain. Certainly the percent dominance continuing to go up, or stabilizing, is evidence.

Alas, tonight is Christmas, and I haven't been keeping up in detail with English reporting and testing dynamics. We saw only a small bump on Christmas Eve, after big jumps in the previous post-announcement days, and I can't tell how much of that was the holiday. Same will apply to today's number, to Boxing Day, and then there are issues of either lag or catchup, and then New Year's. Can anyone shed light on these details, and how seriously we should take them? Even if we get a new mix of strain results, should we worry about lags in the south versus the north, where different strains dominate, warping numbers?

What are good data sources to check for the strain in other nations? Who is checking enough that if it was present, we would notice, or we'd have a good handle on its numbers? Can we make this data happen somewhere? 'Has been spotted in' isn't much to work with.

Also making this a place for people to say how they're updating their estimates based on information that comes in.

Covid 12/3: Land of Confusion

[Last week](#) I warned that we would have unusually high uncertainty for a while thanks to the Thanksgiving holiday. This was both due to unknown events and difficulties in measurement. Reporting and testing would both be delayed by the holiday, and the holiday's real effect would be difficult to predict.

A week later, we have some information, which confirms that the holiday threw a larger monkey wrench into the reporting and testing systems than I predicted. This makes it that much more difficult to tell what has been happening in terms of infections or deaths. Looking at each day individually becomes necessary to have the best available sense of the overall puzzle, despite the dangers of doing that.

The last two days strongly suggest that Thanksgiving did make things considerably worse. People gathered unsafely, and that bill must now be paid continuously until the pandemic is dealt with. By next week, or even by tomorrow night, we will be able to answer that with more confidence.

That's the bad news. The good news is that the United Kingdom has already approved the Pfizer vaccine, and it looks likely we will follow their lead a week from today. If logistics hold we can begin vaccinations the next day, and get everyone who wants it vaccinated by June at the latest. It won't take that many vaccinations for the tide to turn, as it is likely not that far from turning even without them, and may (or may not) have already turned in the Midwest.

The song remains the same. Infection risk is as high as it has ever been, and is rapidly going to peak as the vaccine arrives and lots of people are immune via infections. If there wasn't an additional set of holidays coming, I'd expect the peak in true infections to be this coming week in many places and within another week of that for the country overall.

Since there are multiple major holidays coming, I instead expect the final peak is most likely to happen in the wake of Christmas and New Year's Eve. That is a very nasty one-two punch - people gather on Christmas, get infected, gather again right when they get most infectious from Christmas on New Year's Eve with different groups spending hours drinking together indoors. Quite the final boss for 2020.

Let's run the numbers.

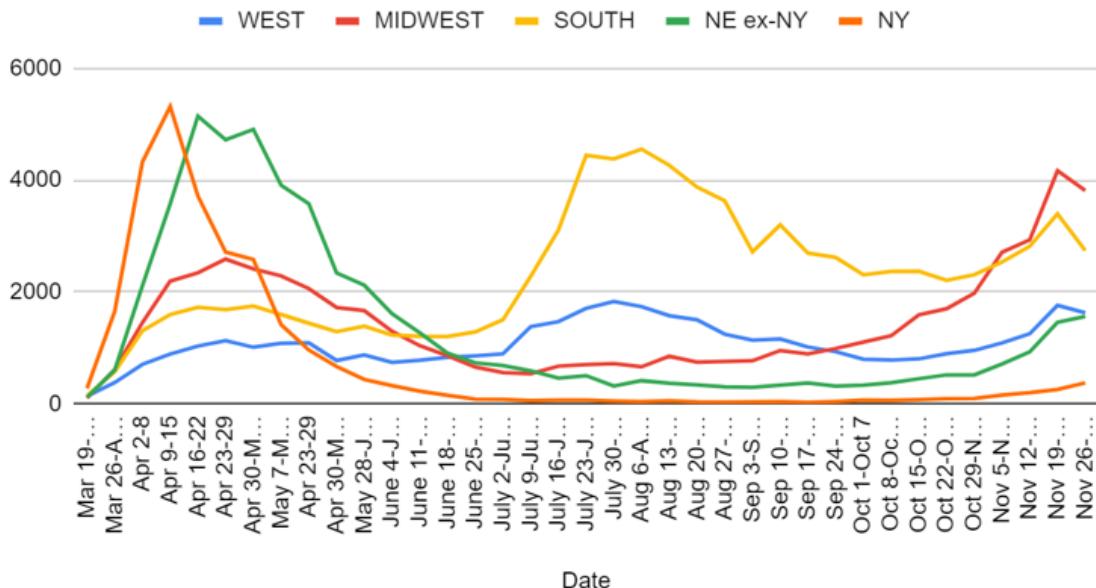
The Numbers

Last week I predicted a 12.8% positive rate on 11 million tests. Instead we got exactly that 12.8% positive rate but on only 8.8 million tests, as the slowdown was large. Last week was even more of a spike in demand than I realized. Deaths reported did not jump up to the 2,100 I predicted, instead staying roughly static at 1,616. I figured they'd catch up on reporting by Wednesday, and I was completely wrong about that. Whoops. Bad prediction.

For next week, there is again a lot of uncertainty, because we don't know where we are now. My unconfident prediction is a 14.8% positive rate on 10 million tests, and an average of 2,200 deaths to account for at least *some* amount of catching up in reporting. But, again, wider than usual error bars all around.

Deaths

Deaths by Region

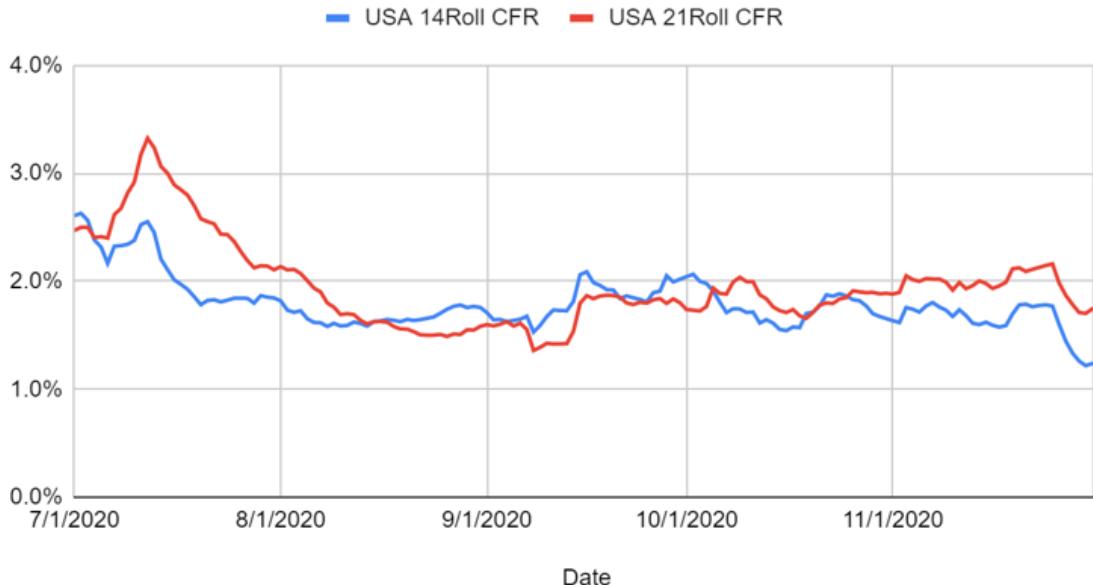


Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 1-Oct 7	797	1103	2308	400
Oct 8-Oct 14	782	1217	2366	436
Oct 15-Oct 21	804	1591	2370	523
Oct 22-Oct 28	895	1701	2208	612
Oct 29-Nov 4	956	1977	2309	613
Nov 5-Nov 11	1089	2712	2535	870
Nov 12-Nov 18	1255	2934	2818	1127
Nov 19-Nov 25	1761	4169	3396	1714
Nov 26-Dec 2	1628	3814	2742	1939

That's not what happened last week. That's what was *reported* to have happened last week. There is approximately zero chance that deaths truly declined. They lag by several weeks, so I would disbelieve this decline even if we did not have the Thanksgiving holiday to explain it.

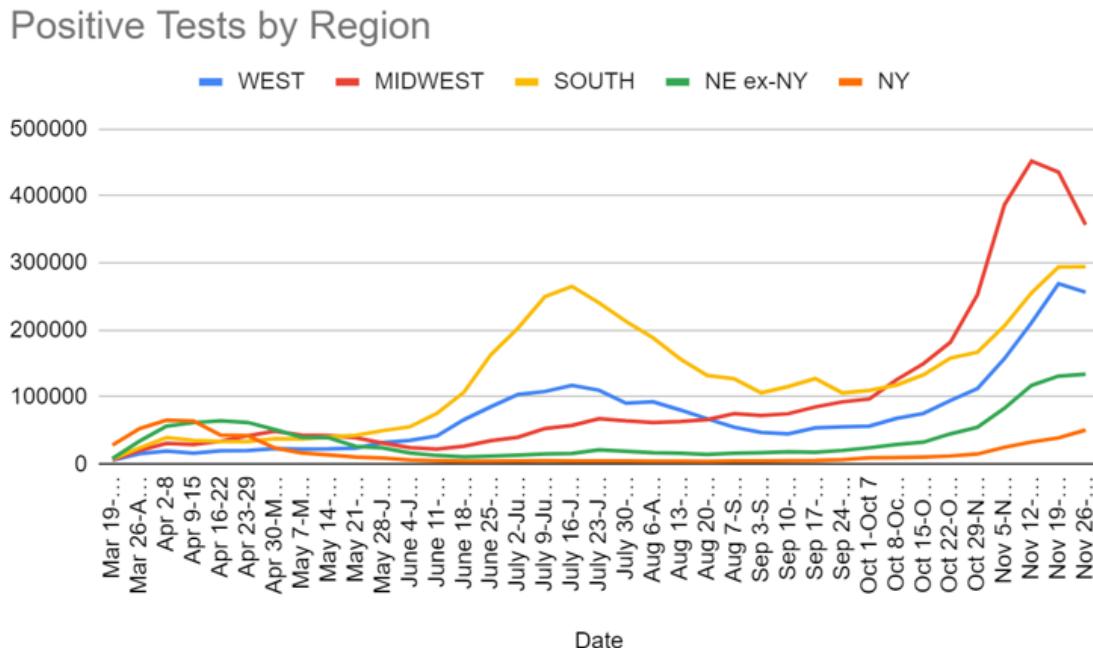
What surprised me is that we have no sign that reporting 'caught up' for what happened over the weekend. The last two days are up from the same days a week ago, but only by roughly the corresponding past increase in cases. The missing deaths, for now, are still missing. I overestimated our ability to report, or underestimated how long it would take to clear the backlog, or both.

14-21 Day Lagged CFR 7-Day Rolling Average



This chart looks at 7-day rolling averages of deaths, compared to cases 2 or 3 weeks beforehand. The drop at Thanksgiving is clear, after this mostly held steady for three months prior (the CFR here was higher earlier, but I cut the graph off to make it readable).

Positive Tests



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 1-Oct 7	56742	97243	110170	34042
Oct 8-Oct 14	68284	125744	117995	38918
Oct 15-Oct 21	75571	149851	133238	43325
Oct 22-Oct 28	94983	181881	158123	57420
Oct 29-Nov 4	112684	252917	167098	70166
Nov 5-Nov 11	157495	387071	206380	108581
Nov 12-Nov 18	211222	452265	255637	150724
Nov 19-Nov 25	269230	435688	294230	170595
Nov 26-Dec 2	256629	357102	294734	185087

The decline in deaths is failure to count the deaths. The decline in positive tests is a failure to run the tests at all. All discussions are on hold until we see percentages and test counts in the next two sections.

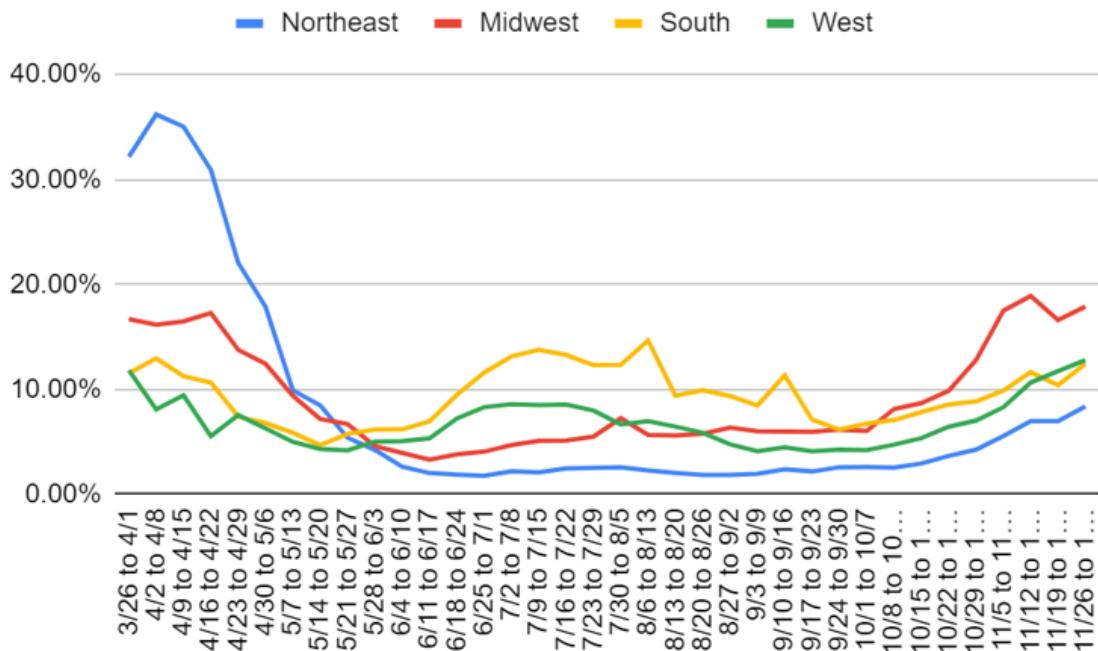
Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Sep 24-Sep 30	5,833,757	5.1%	618,378	1.1%	2.18%
Oct 1-Oct 7	6,009,845	5.2%	763,935	1.3%	2.28%
Oct 8-Oct 14	6,322,865	5.7%	850,223	1.1%	2.39%
Oct 15-Oct 21	6,439,781	6.5%	865,890	1.2%	2.52%
Oct 22-Oct 28	6,933,156	7.5%	890,185	1.4%	2.67%
Oct 29-Nov 4	7,245,600	8.6%	973,777	1.6%	2.86%
Nov 5-Nov 11	8,285,495	10.6%	1,059,559	2.4%	3.13%
Nov 12-Nov 18	8,917,433	12.4%	1,155,670	2.9%	3.47%
Nov 19-Nov 25	10,429,846	11.6%	1,373,751	2.9%	3.83%
Nov 26-Dec 2	8,813,395	12.8%	1,287,010	4.0%	4.18%

There were still a decent number of tests, but well behind last week's count. Last week's tests clearly included a bunch of pre-holiday demand, which was then followed by a holiday slowdown, so some time shifting on both ends. This makes it clear that things overall are not better than two weeks ago, but from this chart the two could plausibly be similar with shifts in different regions.

Then we look at the test percentages.

Positive Test Percentages



Percentages	Northeast	Midwest	South	West
10/1 to 10/7	2.61%	6.05%	6.74%	4.23%
10/8 to 10/14	2.57%	8.14%	7.09%	4.75%
10/15 to 10/22	2.95%	8.70%	7.85%	5.36%
10/22 to 10/28	3.68%	9.87%	8.58%	6.46%
10/29 to 11/4	4.28%	12.79%	8.86%	7.04%
11/5 to 11/11	5.56%	17.51%	9.89%	8.31%
11/12 to 11/18	6.99%	18.90%	11.64%	10.66%
11/19 to 11/25	7.00%	16.62%	10.41%	11.75%
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%

We do still see the Midwest test percentage lower than two weeks ago, but only by about a percent, while everyone else is clearly headed back in the wrong direction. It's clear that the Midwest is trending *relatively well*, which makes sense given how bad things have been for a while now.

How Bad Was Thanksgiving?

The problem is this chart, which in context is very bad news:

Date	Positive Test %	Rolling 7-Day Avg
12/2/2020	15.6%	12.8%
12/1/2020	16.6%	12.6%
11/30/2020	10.8%	11.8%
11/29/2020	13.3%	11.7%
11/28/2020	10.8%	11.4%
11/27/2020	11.4%	11.5%

11/26/2020 12.9%

11.6%

The last two days are, by a substantial margin, the worst in many months.

I usually avoid day-to-day numbers, but here we see exactly what we feared. Five days after Thanksgiving, the positive test rate soared, and it stayed there the next day. We will soon find out what kind of spike this is, but even if it is a one-time effect from the holiday, there is no reason not to expect that higher rate to be sustained as the newly infected infect others. With people coming home with new infections, we may even get a secondary additional spike, although I'd expect that effect to be small.

The default assumption is that it takes about five days for infections to begin showing up as positive tests. With the slowdown in tests over the weekend there is likely an even bigger backlog than usual, so a lot of results will be lagged further because of that. If people were largely infected on Thursday and some on Wednesday, we expect to see the surge of new cases begin in earnest on Tuesday the 1st.

Thus, the giant jump in positive rate on Tuesday, from a 7-day rolling average of 11.8% all the way up to 16.6%, followed by 15.6% on Wednesday, is such a super scary result. There's a good chance Thanksgiving increased case counts by something in the range of 30%.

Sometimes looking at a metric tells you about the world. Sometimes looking at a metric, then looking at the world, tells you some things about the world, but tells you more about the metric. [Nate Silver points out that Google is showing only a moderate Thanksgiving increase in 'home visits.'](#) You see, [the mobility score](#) didn't increase, so nothing to worry about? To me, this highlights that mobility scores and frequency of visits to various location types are data points that can be useful, but that fail to capture the thing about Thanksgiving that we are worried about.

I don't care about the frequency of travel to any household. I care about large gatherings, especially large intergenerational gatherings that are not typical sets of people, and thus would not have already infected each other. I care if those gatherings involve long periods of time, and activities like extended feasts full of conversation, that are all but designed to encourage virus spread. I care about people travelling long distances and thus shuffling around exposures, and encountering people at airports and otherwise along the way. The numbers cited here tell us nothing about those concerns.

We still have little idea how much impact Thanksgiving had. A few more days from now, we'll have a better idea. When we get far enough out to find the death curve, that will be even more informative. But the data is a mess and unless we see a huge spike or crater we will never be able to reliably measure the magnitude of the effects. [Reluctance to get tested in these situations](#) only makes all of these measurement problems worse. So does the ambiguity in the pre-existing trends, and the impact of pre-Thanksgiving behavioral adjustments. It's possible that people were safer before Thanksgiving and directly after, and that might mitigate the effect, perhaps even fully. It's possible that only measurements were distorted. I have a very non-confident Bayesian prior on all this.

How Many Undiagnosed Cases?

This has always been one of the most important known unknowns. The more infections, the less dangerous the disease and the more people are already immune without knowing.

I've been working on the assumption that we are detecting one in five or six infections, with room for that to be anywhere from one in three to one in ten. Yesterday I learned (in what will be old news to some of you) that [the CDC has issued guidance that says it estimates this at one in twelve](#), with a range of 7 to 25? The document seems to use August data and was written in September, so more than a bit out of date, but it's still quite the claim.

We can now rule out the extreme high end because we have 4.18% of the population having a diagnosed case. A 24:1 ratio (1 in 25) would imply everyone had been infected, and we would have seen dramatic effects from herd immunity long before that. Even with an 11:1 ratio, that puts us at ~50% infected, chosen non-randomly. If that doesn't get us to full herd immunity, it gets us within striking distance.

The report is also guessing R_0 at 2.5. That seems impossible as well, at this point. It has to start out higher than that, or things would be solved or at least unambiguously improving by now. Using their high end estimate of $R_0=4$, which is what I've been doing for a while, seems close to a lower bound.

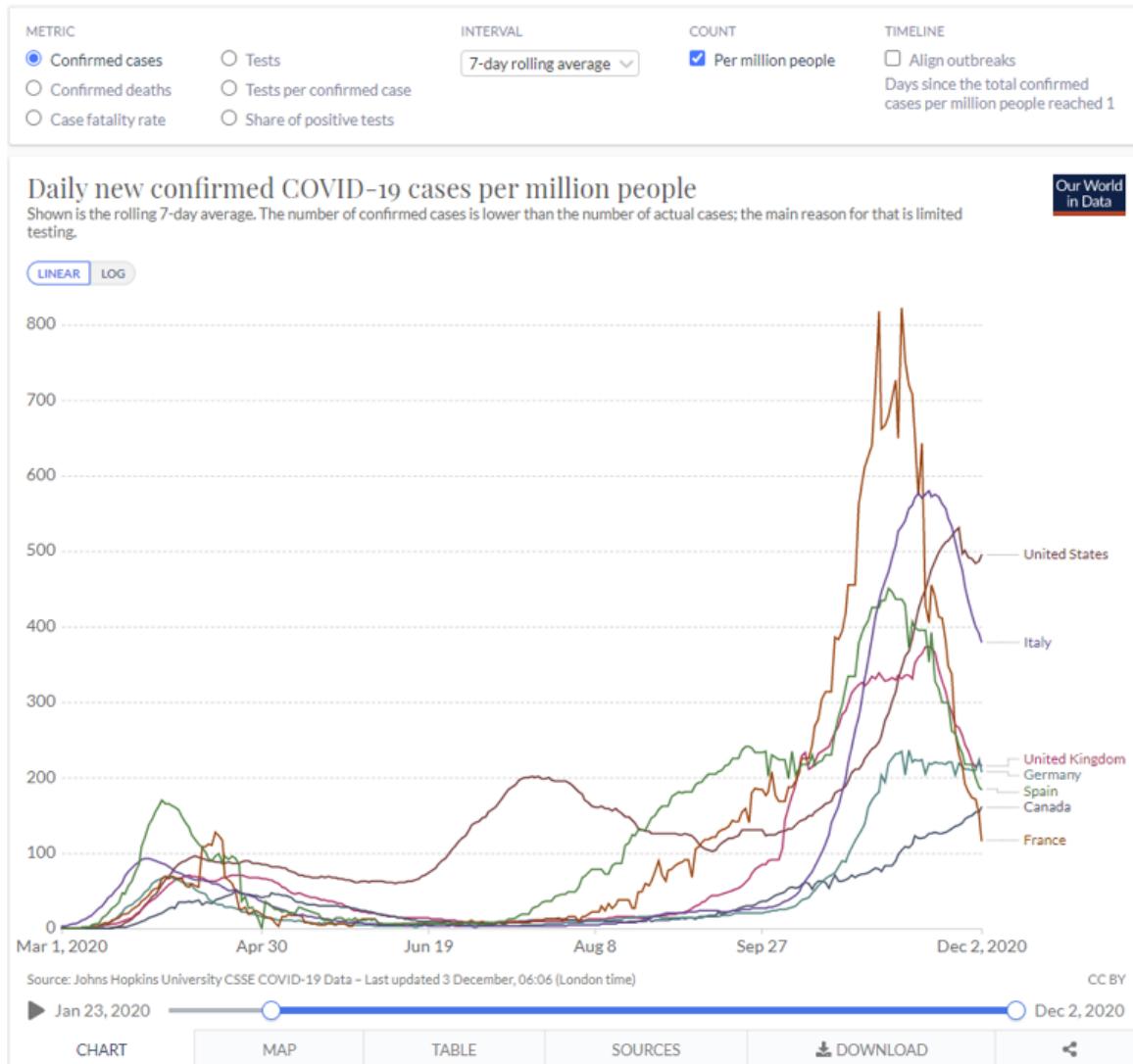
Meanwhile [Covid-19 projections](#) still estimates a 3.3 ratio.

The CDC range seems clearly too high, and it's a sign of how I value such inputs that my prior is updating very little on the new information. One in five or six still seems most likely.

Partial herd immunity effects are rapidly becoming more impactful even without the vaccine. As I said above, I expect the final peak to happen at the latest in the wake of Christmas and New Year's Eve. , which could be a very nasty one-two punch – people gather on Christmas, get infected, gather again on New Year's Eve with different groups. That's quite the final boss for 2020.

Europe

Confirmed Cases:

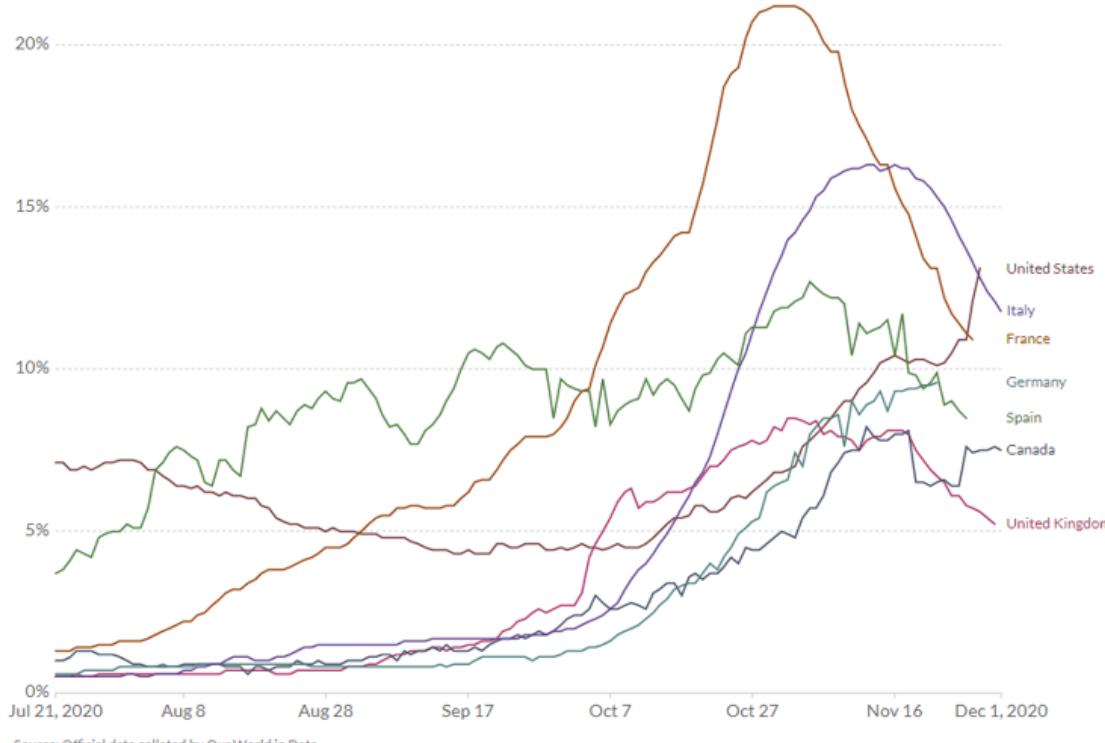


Share of positive tests (starting late July to make this readable)

The share of daily COVID-19 tests that are positive

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

Our World
in Data

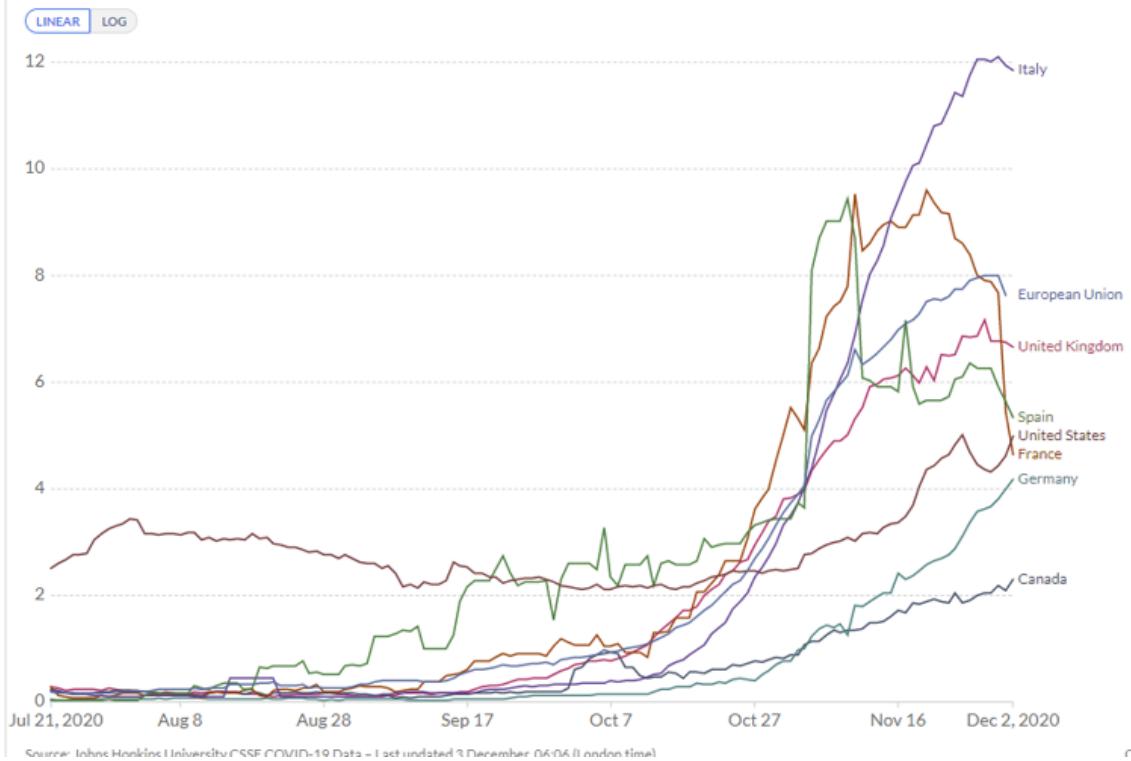


And deaths:

Daily new confirmed COVID-19 deaths per million people

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

Our World
In Data



Deaths haven't had enough time to start dropping in earnest yet except perhaps in France (something strange is likely going on with France's deaths data here) but the case trends are clear. The European countries that decided to get things under control are rapidly getting things under control. The countries that decided to level things off instead, did that instead. The United States did not do such things, but will soon enough get a similar outcome from a higher level.

All I Want For Christmas Are A Covid Vaccine and a PS5, But They Underpriced Them, And Now They're All Sold Out

The stage is set. [The applications are in. The United Airlines charter planes full of dry ice are standing by.](#) All that remains is for regulatory authorities to give emergency approval and allow distribution.



Eric Topol

@EricTopol

...

This will go down in history as one of science and medical research's greatest achievements. Perhaps the most impressive.

I put together a preliminary timeline of some key milestones to show how several years of work were compressed into months.

Date	Milestone
Dec 1	Covid-19 illness documented (unpublicized Nov 17 th)
Jan 10	SARS-CoV-2 virus sequenced
Jan 15	NIH designs mRNA vaccine in collaboration with Moderna
Mar 16	Moderna Phase 1I2 trial begins
May 2	Pfizer/BioNTech Phase 1I2 trial begins
July 14	Moderna Phase 1I2 trial published in NEJM
July 27, 28	Moderna and Pfizer/BioNTech Phase 3 trial begins
Aug 12	Pfizer/BioNTech Phase 1I2 published in Nature
October 22,27	Enrollment in both Phase 3 trials complete; >74,000 participants
Nov 9	Pfizer/BioNTech announces interim analysis efficacy > 90%
Nov 16	Moderna announces interim analysis efficacy 94.5%
Nov 18	Pfizer/BioNTech announces 95% efficacy as final result
Nov 20	1 st EUA submitted by Pfizer/BioNTech
Nov 27	Distribution of vaccine by UAL charter flights throughout US
Dec 10	FDA External review of Pfizer/BioNTech EUA
Dec 11	Phase 1a Vaccination begins for health care professionals*

*Provisional on positive external review

2:39 PM · Nov 28, 2020 · Twitter Web App

What does it take to get a Covid-19 vaccine approved [these days](#)?

The United States says, at a *minimum*, it takes the regularly scheduled meeting on December 10. There are [upsetting rustlings that this actually means sometime after December 10](#), but anticipation still seems to mostly be that distribution can probably begin December 11, about three weeks after the Emergency Use Application was sent in, a period during which tens of thousands will have died of Covid-19 in the United States.

The United Kingdom is having none of this nonsense. [They approved the Pfizer vaccine on Wednesday](#), and will be distributing it next week.

The European Medicines Agency, the presumably unionized European equivalent, [thinks the F.D.A. is acting with unnecessary haste](#). The best case scenario is the end of the year:

"Assuming everything is positive – and we have to look at the data to be sure – but best case, we could have a scientific opinion by the end of the year," EMA Executive Director Emer Cooke told the Irish Independent newspaper in an interview.

"These have been developed very quickly, which is very promising from a scientific perspective, but it means there is a lot of attention on the results and we have to make sure we evaluate those as efficiently as we can without compromising our usual scientific standards."

[Another update here via MR suggests evaluation will be done by January 12.](#)

The implicit statement here is that things are moving fast and they are against things moving fast because fast is symbolic of irresponsible and slowing things down makes you a Very Serious Person, so they are going to slow things down to fix that. The explicit statement is merely that they need time to fully examine the data, which implies logically that either they are much slower than the F.D.A. or United Kingdom at doing that and don't trust those sources, or that those places are not fully examining the data.

[The Japanese are going to take even longer](#). You see, vaccines normally take longer to develop, and we did this one fast, and that's highly suspicious. Extreme caution is needed. You'd better do trials specifically in Japan if you want to convince the Japanese. Despite that, they've decided to buy 120 million doses by mid-2021, so presumably they have a plan to run those tests by then. And given they don't have a pandemic problem right now and are used to wearing masks anyway, letting other countries take the vaccine first makes good sense. So it kinda makes sense even if the justifications are terrible.

The good news, of course, is that there are a limited number of doses, and as I discuss later geographically concentrating vaccinations is a good idea, so all we have to do is have *someone* approve the vaccine. Then they can use all the early doses until others also approve.

For America and especially Europe, the timeline comes back to the question of what 'fully examine the data' means. What are these people looking for that takes a month to analyze *after the data comes in?*

If Pfizer came out a week ago and said, 'in secret we have developed this vaccine for the past eight months, and we present to you a new miracle of modern medicine, BEHOLD!' I would totally understand it taking a month to examine the data. That's not how this works. That's not how any of this works. They've been in constant communication with regulatory agencies, their study designs were already approved, and the data from the trials boils down to essentially two things: Who got infected with Covid-19 with what severity, and what adverse side effects were observed.

The news on efficacy is well in excess of expert expectations, essentially a best case scenario. So if plans were approved in the first place, it takes almost no time to say 'yep, looks like the efficacy data is very good.' Certainly this isn't what is taking weeks.

That means that if we are examining the data we are looking at safety data for adverse effects. But once again, *there were not any major adverse effects*. There is a nasty short-term effect when you get the vaccine, but they know that already and have had months to evaluate that. So yes, you have to look carefully to confirm the whole 'yep, looks like safety data is very good' angle, but how long does looking carefully take when there's nothing to see?

Also worth noting is that Moderna's application came in several weeks after Pfizer's, its data was available several weeks after Pfizer's, and it looks like both will be considered simultaneously. It's hard to then argue that Pfizer's application couldn't have been handled any faster. It looks more like the FDA at best did not want to hold multiple meetings. The counterargument to that is that the two vaccines are very similar, and thus the work can be combined, but that has its own implications.

I do realize that three weeks turnaround is, all things considered, extraordinarily good, even a month is good, but that's a condemnation of standard procedure, not an excuse.

The missing mood is an emergency. The missing action is planning ahead and making conditional decisions in advance the way I would when playing a game with a time limit. You don't know what is going to happen next, but there are only so many options, so you consider each of them and plan a response. You have a procedure that lets everyone involved stay on a reasonable schedule and maintain sanity and all that, and keeps an orderly priority queue and avoids favoritism and all the politicization and corruption that usually leads to. Under normal conditions great and all, considering the practical alternatives, but maybe there are times to throw all that out the window?

I'm furious at this last small delay *because it is so obviously unnecessary* and shows how bad things are even when there is unusually strong pushing for fast progress. I don't *actually care* about the delay itself all that much, however, because the limiting factor in distribution is manufacturing capacity. The same number of people will be vaccinated in February regardless of any delays in December, provided they don't extend beyond a few weeks. More people will get sick and die, and more people's lives and days will be lost or ruined, but compared to the overall stakes this is very small potatoes. Compared, that is, to previous delays.

Did you hear how Moderna made their vaccine in two days without even having a virus sample? So we could have very easily been done months ago if we allowed challenge trials, or even without that if we were allowed to do expected value calculations and run larger experiments faster?

Did you hear how the limiting factor was safety data, but we still were unable to scale Phase II trials enough that their safety data could count as safety data, forcing us to wait out for Phase III safety data?

Did you hear the one about Moderna having a big lead on Pfizer, [and being told they had to delay their trials because they had insufficient minority participation](#) (rather than start while working to expand such participation, which is what you would do if you actually cared about helping people, and while Pfizer was not told told to do any of this) leading to Pfizer finishing first?

Meanwhile, [a take on why many of the vaccine trials likely underestimate how effective they are](#), and which seems right to me.

Also in medical research world modeling, [here's some of how mRNA vaccines came to be a thing](#).

Who Gets the Vaccine First?

We can't vaccinate until we [use politics to decide who gets priority](#). The CDC is meeting to determine that. To avoid figuring out which areas have greater need, and to avoid [any possible accusations of blame](#) via buck passing, but also because apparently [Trump and his team can't be bothered with such trivial questions as logistics](#), allocation is first going to go to the states by population, but without telling states in advance how much vaccine they will

be getting or providing them with funding to distribute the vaccine, and the states are then going to be tasked with deciding on distribution.

(I'm sympathetic to 'states must maintain balanced budgets that do not make sense during a pandemic, and the federal government should be aiding states a lot in general' but I am completely unsympathetic to 'they aren't providing money specifically for distribution' because money is fungible.)

Health care workers and the most vulnerable continue to be presumed to be going first in most or all areas, in some combination.

Both of these make good sense. Health care workers expose themselves as part of their job and are out there directly helping. The most vulnerable are much more likely to die if infected.

Putting both of these in the "1a" group to get the vaccine first is going to be tricky. This was the question that needed to be decided, to figure out which of those two groups goes before the other, and as far as I can tell everyone everywhere has passed the buck. There are only enough doses initially for about three million people, by the end of the year perhaps twenty million. There are about twenty million health care workers and about 1.4 million people in nursing homes. So now in each state, we can expect lots of small political battles to determine how much allocation goes to various hospitals and nursing homes, then within each institution to determine who gets vaccinated early versus later. Those with power, status and connections within those institutions will be rewarded.

That population of 1.4 million in nursing homes is, according to the CDC at the source above, responsible for 6% of infections and 39% of deaths. If I were making the decision, I would go there first. If these first 4 million doses can cut deaths by 39% that seems like the best option.

On a side note, obviously going to a nursing home is caused directly by old age and failing health so the statistics are completely unfair, but has anyone considered that these places *might* be eldritch abominations?

Here is what the United Kingdom will be doing, which seems reasonable enough although it excludes a lot of essential workers for longer than I'd have thought wise ([source](#)):

The first phase of the programme

Offer vaccination

- 1 Residents in a care home for older adults and their carers
- 2 All those 80 years of age and over. Frontline health and social care workers
- 3 All those 75 years of age and over
- 4 All those 70 years of age and over. Clinically extremely vulnerable individuals
- 5 All those 65 years of age and over
- 6 All individuals aged 16 years to 64 years with underlying health conditions which put them at higher risk of serious disease and mortality
- 7 All those 60 years of age and over
- 8 All those 55 years of age and over
- 9 All those 50 years of age and over

■ bbc.co.uk/news

It has been pointed out that distributing the vaccine *evenly* among all locations is actually the *worst possible* choice. Vaccination has network effects. If you vaccinate enough people in one location, then everyone there can go back to normal. The vaccinated hospital or nursing home can then operate much better, or even an entire region.

If you vaccinate only some people in every location, no one who isn't vaccinated can go back to normal at all, and even those who get the vaccine can't fully go back to normal. There's no normal around them to return to, and while the pandemic rages on, 95% effective is not 100% effective. This is fractally true as well: If I get vaccinated but my wife does not or vice versa, that has its advantages, but mostly far less than half the advantage of vaccinating both of us. Thus, proceeding in a random order by location would be better than an equal distribution, assuming the right sub-categories still got appropriate prioritization. That argument makes me think there might be an allocation even better than allocation by price. The downside is that any prioritization process like this is vulnerable to politics and corruption, which equal distribution prevents.

I find the questions both interesting and relevant, since we all need to know when to expect our chance to be vaccinated and know when we could plausibly claim priority. Can my parents, definitely at high risk due to age and comorbidities, get protected early? Can my wife, who is a psychiatrist, do so, and if so what about me? Are any of those claims justified? I do not know.

Then of course, [there's the question of the already infected](#) and therefore already immune. In a remotely sane world, when Covid-19 tests are scarce, the first thing you do is give people antibody tests to confirm they need the vaccine, or *at a minimum* you could... ask people who have already had the Covid-19 not to get vaccinated? Even if insisting on this is not possible, of course, I would implore anyone who is immune to not try and get vaccinated until there is adequate supply. I don't even know why they would want to.

Another question is, are they going to effectively force those who get an allocation to take the vaccine instead of giving those doses to people who actively want them? I am content to wait until after essential workers and the most vulnerable, so long as I get it before all the people who don't actively want the vaccine.

A warp speed official claims that '[100% of Americans that want the vaccine will have had the vaccine by June](#)' and that seems mostly fine. I'm going to make a push to get it in March or April instead, and I expect to succeed because a lot of people won't care much and I'll care a lot.

Meanwhile, there continue to be worries abound about how to convince many people to take the vaccine, especially once they realize the shots often have unpleasant temporary side-effects. Worries about [whether people will get a bill for vaccination](#) even if the vaccine itself is free. If you were wondering how the medical system works, the ads I saw at that link were for personal loans, and it looks like yes you will get a bill and then file a bunch of forms to hopefully avoid paying it, because that's what 'free' means.

Convincing the reluctant is something we will have to deal with some time around May or June, and certainly plenty of people will try to free ride, but I am confident our existing systems will mostly pull through here. By then, I'd also expect most people who refuse vaccination to have already been infected anyway, given correlations in behavior.

Sports Go Sports

When the pandemic arrived, sports shut down. We didn't know how dangerous it would be to play, or how to play relatively safely. Since then, sports have mostly resumed.

The NBA bubble went off without a hitch and showed us better ways to test and keep us safe, which the outside world seems to have mostly ignored. They are moving forward with next season: "[The NBA announces, among many key dates, that the Hall of Fame inductions for the Class of 2020 — headlined by Kobe Bryant, Tim Duncan and Kevin Garnett — will take place in mid-May.](#)"

MLB had some trouble, especially in the contract negotiations, and had a bunch of infections, some of which the players blatantly ignored, but was able to reschedule the games that it missed as necessary, and we got the boys of summer. The MLB bubble seems to have broken on the final day of the World Series, but in the end it all worked out. I do not think anyone caught Covid-19 from an opponent at a baseball game, but within-team activity was clearly not safe. The danger was in the dugouts.

The NFL skipped the preseason to 'keep players safe' and it seems like not having preseason led to a lot more non-Covid injuries than normal. Also contributing is likely that practices are happening less often and less intensely and efficiently due to precautions. It's very possible that the 'let everyone get Covid' plan would have been safer, as I'd much rather get Covid than face the health risks of playing professional football. Still, it was clear they were determined to play.

The show must go on. No football, no peace.

That's not to say there haven't [been hiccups](#). As infections have risen elsewhere, more and [more teams and players are coming down with the virus](#). In previous weeks, teams cancelled games when they had a lot of infections. For whatever reason, Denver was told this weekend that the show had to go on, despite all *four* of its quarterbacks [being ruled out due to contact tracing from their primary quarterback](#). No masks had been worn.

If you think that represents gross incompetence and they should have held their backup backup backup quarterback in reserve like a designated survivor if they had no fifth option, you'd be right, but they did not think about that at the time. They had the foresight to sign

four quarterbacks, where usually there are at most three and the Giants only have two, then made it so a single positive test could take out all four.

The Broncos had to scramble. They [tried to sign one of their assistant coaches](#), two of whom had played quarterback in high school and would know the playbook, but were turned down by the league, presumably due to worries that teams would try to use that loophole in the future to hide reserve players.

As a result, without enough time for an outsider to learn the playbook, [Denver had to start Kendall Hinton](#), someone with zero snaps of practice that was signed to the practice squad a month ago as a wide receiver, and was previously selling sports equipment. Who hadn't been tackled in years. As their quarterback. While the scoreboard was a disappointment and on reflection laying fifteen on the Saints may have been the wager of the year, there was nothing but praise for Kendall, who was put into an impossible situation. It is important to reward those who give extraordinary efforts and show their character even when the scoreboard would indicate otherwise.

San Francisco has decided to indeed allow no football, banning all contact sports without making an exception for the 49ers. They are not allowed to practice within Santa Clara county limits, and are scrambling for a new place to practice and play. Did I mention that the mayor of San Francisco, who has kept the city in a vice grip all year but who to be fair did not directly shut down the 49ers, London Breed, [dined at the French Laundry with seven others the day after the governor did the same?](#)

Meanwhile in college football, it is not going better. With the delays early on from when teams had yet to be shamed into playing the season, Big 10 and Pac 12 teams are losing huge portions of what's left of their seasons. My team Wisconsin has two resounding wins, one embarrassing loss and three cancelled games, which means they are ineligible for the Big 10 championship game. Two of the top four teams had their coaches test positive this past week. For two straight weeks, more than ten games have been cancelled or postponed due to the virus.

Ohio State has missed multiple games and is in grave danger of missing the playoff due to lack of enough games, and ESPN commentator [Kirk Herbstreit was forced to apologize for speculating that archrival Michigan might 'wave the white flag'](#) and cancel their game to avoid their inevitable curb stomping (it's gonna be ugly, folks) and also deny their rival the right to play in the Big 10 championship and probably the playoff. He says he 'wasn't playing fair' but while I love the guy he's a proud partisan of The Ohio State University so I don't really know what anyone was expecting, and also I would never condone cancelling a game for such reasons and the honorable Michigan Wolverines would never ever do that, but I really hope this scenario actually happens.

Vanderbilt lost all its kickers and was forced to fall back on open tryouts. This resulted in [Sarah Fuller, the goalkeeper of the woman's soccer team, getting the job](#). Complete with 'play like a girl' inscribed on her helmet, she became the first woman to play in a power five football game, offering up a squib kick that was likely the best thing that happened on the field to Vanderbilt that day, as they lost 41-0 to drop to an 0-8 record.

College basketball got underway this past week. Already a large portion of the marquee games seem to have been moved, cancelled or created to fill newly available holes in teams' schedules. There were some calls to move the schedule back a few months, but things are moving forward instead. Given how long the season is, it seems fine to cancel games as needed.

[The reigning Formula 1 World Champion just tested positive](#). When a large percentage of all people get a virus, a lot of people in sports will also get that virus.

To my knowledge, there have been many cases across all these sports, but no cases of long Covid, and no deaths. That doesn't mean there have not been such cases of long Covid, but

the common presumption is that those infected will return in a few weeks. So far, in every prominent case or at least every case I am aware of, that has happened.

It has all made life much more bearable under this pandemic, for me and for millions of others. So I left you out of my thank you section last week, but here's to you, athletes. Here's to you.

Once More, With Feeling (You Should Know This Already)

Testing works. If you test often enough, as some colleges are doing, you solve the pandemic. We do not currently have the capacity to do this, because regulators have actively prevented us from creating that capacity instead of helping us build that capacity.

If you don't have tests, you could even periodically screen for loss of smell and pick up 70%-80% of cases that way at a cost of something like a dollar per test. That could end the pandemic on its own, isn't being done, *and this is in the 'you should know this already' section*. I presume this is a legal thing to do, though, so perhaps some people will read this and do it.

Links only say the same thing, but periodic reminder: Human challenge trials would have saved hundreds of thousands of lives. The failure to use them is a sign of our deep civilizational failures. It is *not*, however, one of the worst events of human history, because that list includes things like The Great Leap Forward, The Holocaust and the third Star Wars trilogy.

A reminder that we could do what we did for the coronavirus vaccine to do other things too.

In this section because congress is known to be essentially incapable of anything potentially useful: A raging pandemic and a gridlocked Congress could upend President-elect Joe Biden's plan to hire 100,000 public health workers to trace Covid-19's spread, despite widespread agreement it's needed to finally end the crisis. I'm not sure what good contact tracers are at this point, but they will presumably become useful later to wrap things up and keep them that way.

Updates From Events of Last Week

Marginal Revolution on the SCOTUS decision to prevent disproportionately closing houses of worship. Also a reminder from Rabbi Yosef Goldman that "In Jewish law, religious freedom in a global pandemic is the freedom from any other relevant obligation if a life would be risked in its fulfillment. In fact, protecting the lives of the living becomes the obligation." I can verify that Jewish teaching is that you have a religious obligation to suspend your religious obligations if necessary to keep people safe, so services should indeed have been cancelled voluntarily. Where I disagree with Rabbi Goldman is that I believe that this doesn't mean that cancelling them by force of arms fails to violate religious freedom. That's not how freedom works.

Pilot shortage prompts rare flight cancellations at Delta over Thanksgiving break. Delta then transferred the passengers to other flights. Which is good news in the sense that everyone made their flights especially their return flights, and good news in the sense that the planes were previously sufficiently empty that there was plenty of space. It was bad news in the sense that everyone still travelled, and that they now did so in more crowded planes. In a normal Thanksgiving, there are lots of completely full or overbooked flights, and there's a mad scramble to get everyone home and back again. All hands are on deck. It makes sense that in the new world, Delta would be willing to allow flights to instead be cancelled rather

than pay a large premium to prevent this, even if it makes the flights less safe – they are hurting financially, getting extra workers over the holiday can't be cheap, and they can transfer to their spare capacity if needed. That doesn't mean that, as someone who by default uses Delta to fly, I will forget that they allowed this when deciding in the future.

McConnell finally puts an end to [GOP meetings behind closed doors for lunch three days a week since May, maskless when speaking and eating, and some are more diligent than others in putting masks back on.](#)

[Introducing the scented candle complaint index of Covid cases.](#) Seems like good news!

In Other News

[Covid: Free Vitamin D pills for 2.5 million vulnerable in England.](#) Excellent move, we should do the same everywhere else and also for the non-vulnerable as well. They almost got it! The way they almost got it was to claim it was about lockdowns preventing people from going outside, and Vitamin D deficiency being bad for non-Covid reasons. They're not wrong, of course. Vitamin D deficiency is quite bad even in the absence of Covid-19, so this is an overdetermined cheaper-than-free action, especially given England pays for everyone's health care. Still, even now, they're saying that "But there is limited evidence that vitamin D protects against or treats Covid-19, although health officials have been asked to go back over the existing research" and also only recommending 400 IU/day, an order of magnitude less than I take.

Still, I'll take the win. Shovel-ready effective altruist proposal for those with more funding than good ways to spend it: Fund doing this in other countries to the extent excess supply is available.

Speaking of wins, Scott Atlas resigns from the White House, [and there was much rejoicing.](#)

[Sam Bowman asks rhetorical question mocking lockdown skeptics for not changing their minds after the vaccine was announced, gets replies from lockdown skeptics who changed their minds when a vaccine was announced.](#) You love to see it. It is also worth noting that there are also good reasons to become more skeptical of lockdowns than before, since the virus is now less deadly and the political feasibility of lockdowns seems that much worse.

[Yale professor evaluates college Covid-19 dashboards. Website itself is here.](#) I am slightly terrified that my brain thinks offering point totals and a competition is likely the most effective and efficient way to get improvements in those dashboards. Only slightly, mostly I'm used to it. I love that the way they sort it is by athletic conference, so rivalries can be invoked. Go state! Congratulations to Amherst College, this week's winner with the only A++. Unsurprisingly George Mason gets an A+, and unsurprisingly the Ivy League was far from the worst but did not cover itself in glory.

[Blame Canada:](#) Canadians, [who have secured more doses of vaccine per person than any other country](#) with eight doses per person or four full inoculations per person, [point fingers because they tried to first make a deal with the Chinese.](#) The deal fell through because the Chinese government blocked shipments of the vaccine, so the Canadians couldn't test it. In the particular case of the Chinese, 'get your own data' seems like a wise principle. I also think it is great that the Canadians are arguing over whether they did enough to secure vaccine doses, [despite securing the most doses per person of any country and way more than are needed](#), because it was a huge error to get [way more than was needed rather than way way way more than was needed](#). I have zero ethical issues with it because the surplus can and presumably will be resold to other nations when the time comes.

Covid-19 hypocrisy this week: San Francisco Mayor [London Breed](#), Secretary of State [Mike Pompeo](#).

Los Angeles Mayor Eric Garcetti [locks down the city tight](#). All gatherings from more than one household are banned except for outdoor religious activities and protests. No being outside unless you are homeless, or for specified essential reasons. I presume his plan is to keep this up for several months. These are sufficiently strong restrictions that on the margin they are all but guaranteed to backfire even in pure infection terms. You have one of the few cities where it's reasonable to do things outside, and you send the police after people doing things outside.

Not Covid-19: Our healthcare system is a dumpster fire, a continuing series, storytime regarding [pricing edition](#). Also [this](#). And [this](#), where Canada has been forced to ban mass exports of prescription drugs because our plan to save money was to have Canada force drug makers to sell cheap and then buy from them in Canada and import the drugs back here, and that was threatening to create shortages in Canada.

Not Covid-19 but cool: [Deepmind says they've solved the protein folding problem](#). This seems to actually translate to them scoring much higher than the previous high score on a protein folding competition. Which is neat but claims of having 'solved' the problem seem premature.

[New model suggests](#) that a consistent lockdown policy is better than an oscillating policy. Sharing for those who are curious. My quick reading says that they are completely missing all the interesting questions, because they are not thinking about control systems and how people will respond to various levels of infection, and they are assuming there is a generic lockdown knob that works.

My Fear Heuristic

My friends and family call me "risk tolerant". I wasn't always this way. It is the result of a 3-year-long scientific experiment.

You must do everything that frightens you...Everything. I'm not talking about risking your life, but everything else. Think about fear, decide right now how you're doing to deal with fear, because fear is going to be the great issue of your life, I promise you. Fear will be the fuel for all your success, and the root cause of all your failures, and the underlying dilemma in every story you tell yourself about yourself. And the only chance you'll have against fear? Follow it. Steer by it. Don't think of fear as the villain. Think of fear as your guide, your pathfinder.

—*The Tender Bar* by J.R. Moehringer

When I was 18 I discovered a useful heuristic. Whenever I didn't know what to do I would pick whatever not-obviously-stupid^[1] option frightened me the most.

My indecisions always centered around choosing between a scary unpredictable option and a comfortable predictable option. Since the comfortable option was always predictable, I always knew what the counterfactual would have been whenever I chose the scary option. If I chose the scary option then I could weigh the value of both timelines after the fact.

As an experiment, I resolved to choose the scarier option whenever I was undecided about what to do. I observed the results. Then I recorded whether the decision was big or little and whether doing what scared me more was the right choice in retrospect. I repeated the procedure 30-ish times for small decisions and 6-ish times for big decisions. If I were properly calibrated then picking the scary option would result in the correct choice 50% of the time.

Results:

- For my 30-ish small decisions, picking the scary option was correct 90% of the time.
- For my 6-ish big decisions, picking the scary option was correct 100% of the time.

The above results underestimate the utility of my fear heuristic. My conundrums were overwhelming social. The upsides earned me substantial value. The downsides cost me trivial embarrassments.

I terminated the experiment when my fear evaporated. The only things I still feared were obviously stupid activities like jumping off of buildings and unimportant activities like handling large arthropods. I had deconditioned myself out of fear.

I didn't lose the signal. I had just recalibrated myself.

1. "Stupid" includes anything that risks death or permanent injury. [↩](#)

Quick Thoughts on Immoral Mazes

I finally got around to reading (most of) Zvi's sequence [Immoral Mazes](#). I may want to turn some of the following ideas into longer posts, but I thought I'd try to get out some quick thoughts.

Connections to Dictator's Handbook

I've also recently finished reading Dictator's Handbook. The two share some similarities, but offer very different *analyses of the problem* and very different *directions for hope/improvement*.

Dictator's handbook in brief (see also the [excellent summary by CPG grey on youtube](#)):

- You can think of government as a game played between rulers and ruled. Rulers who don't play to maximize the amount of time they stay in power don't rule for long, so we can model rulers' utility function as just maximizing time in power. This fits most cases pretty well, and explains a lot of political phenomena.
- We can ignore most of the details of a given political system, and just put systems on a spectrum from "perfect dictatorship" to "perfect democracy" by *how many key supporters a ruler needs to please to stay in power*. In a realistic dictatorship, the answer is a handful of people. In a realistic democracy, the answer is a significant fraction of the population.
- The ruler's incentive is always to minimize the number of key supporters needed, and then make those supporters satisfied.
- The key supporters sometimes have incentives aligned with the ruler, IE, also want to minimize the number of key supporters (to get more rewards per key supporter). This is especially true when there are few key supporters already. This will lead to culls and movement further and further toward dictatorship.
- At other times, the key supporters may want to *expand* the number of key supporters. This is because (1) they want to protect themselves against a cull, and (2) they may want to incentivise the ruler to produce *public goods* (rather than special benefits to supporters). This is especially true when the number of key supporters is already very large, so special benefits to supporters are small, and won't be hurt much by expanding the coalition (and are outweighed by benefits from public goods). Expanding the coalition further provides insulation from culls. This leads to further democratization.

Immoral Mazes seem sort of like the Dictator's Handbook model but with more levels. The Dictator's Handbook did mention (I think?) that the key supporters are usually playing the same game one level down, but didn't make very much of this.

I think the Dictator's Handbook model is clearer and more precise than Immoral Mazes, so it would be cool to try and meld them together further.

Some big differences between the models:

- Dictator's Handbook sees everyone as just following their incentives. Moloch is very strong in organizations closer to the dictatorship end of the spectrum, and relatively weak in organizations close to the democracy end of the spectrum.

- Immoral Mazes sees everyone as following *perceived* incentives, but strongly holds that *it isn't worth it*. People in Mazes are making a mistake.
- Both models would agree that flatter hierarchies result in better outcomes (for the majority of people). But for different reasons:
 - Immoral Mazes holds that *hierarchy itself is the problem*. Deeper hierarchy means increasingly warped incentives.
 - Dictator's Handbook holds that *small branching numbers* in hierarchy is the problem. Smaller branching numbers naturally lead to deeper hierarchy, but the more important point according to Dictator's Handbook is that each boss has to satisfy fewer supporters.
- Dictator's Handbook doesn't really care what the executive hierarchy looks like, so long as it is answerable to a large number of people. A president may set up a deep hierarchical government, but since the president is ultimately answerable to the people, you're in a democratic (and therefore *relatively benevolent*) regime.
- Immoral Mazes predicts that the deep hierarchy in the government is as big a problem as deep hierarchy anywhere else; it matters only a little that the president answers to the people.

Connections to Gervais Principle

Another model with striking similarities is [The Gervais Principle](#).

Here's my summary, biasing toward connections to the other two models:

- Like the other two models, Gervais Principle emphasizes that *large organizations are fundamentally dysfunctional*, IE, not very aligned with carrying out their purported function.
- Like Immoral Mazes, Gervais Principle gives a special role to middle management.
- In the Gervais Principle, the top of an organization is a collection of *sociopaths*. The middle management are called the *clueless*. The bottom rungs of an organization are called the *losers*.
 - *Sociopaths* are the really ambitious people who create the warped moral frame that middle management buys into. Sociopaths rise to the top and then proceed to extract profit, like the rulers in Dictator's Handbook.
 - *Clueless* are the people who sociopaths want beneath them. Ideally (for the sociopath) there would only be one sociopath; the key supporters would all be clueless loyalists who buy into the narrative. Unfortunately (for the top sociopath), other sociopaths will manipulate their way into the highest positions. So the top management realistically consists of power-hungry sociopaths who weave a web of moral confusion. Middle management lives in this web of confusion.
 - *Losers* are at the bottom. They work for their paycheck and little more. The sociopaths don't need to fool them, so they are usually more aware of the dysfunctional nature of the organization, but can't do anything about it.

The Gervais Principle takes an almost anthropological approach to these three groups of people, describing their different cultures and norms. It offers no hope for building better institutions.

I like the way The Gervais Principle factors loyalty into the model, highlighting the way that rulers select key supporters based on loyalty in a way the more game-theoretic

model of The Dictator's Handbook can't really explain. Loyalty in a game-theoretic model just means that dictators make supporters *dependent on them* and rule through fear. Loyalty in real life includes more than this; as Immoral Mazes put it, giving up your soul.

The Great Fragmentation

In Immoral Mazes, Zvi asserts that mazes are on the rise. Corporations and other organizations with large mazes have been around for a long time in America, slowly but surely increasing their internal maze levels, and seeping out maze culture into the general cultural mixing pot.

Now, I'm *not saying Zvi is wrong*, but I see a lot of things which are discordant with this picture.

When I look at recent American history, I get the following impression:

- The 50s and 60s saw "peak square" -- the socioeconomic landscape was dominated by large companies and "the company man".
- The 60s and 70s saw a huge cultural revolution, which began a major decline in squares. Men stopped wearing suits and ties, and alternative hairstyles (IE long hair) became much more socially acceptable for men. Women were less and less pressured to live traditionally as housewives.
- The "company man" became less and less a thing, as the average number of careers in one's lifetime increased. Today it's hardly realistic to expect to work for one company for your whole life.
- In the 90s, start-up culture started to be a major thing. We saw the rise of "the tech industry" as it exists today, which Zvi himself admits is somewhat less prone to immoral mazes.
- In the 00s, hipster culture (like hippy culture before it) created commercial pressure against homogenized goods. Large record labels became much less important in the music industry. The same thing happened across many other areas of media, continuing into the 2010s, due to the new possibilities opened up by the internet. We saw the rise of indie everything, which by Zvi's account, should greatly reduce mazes.

All of these impressions are greatly reinforced by Paul Graham's essay, [The Refragmentation](#).

All of these forces push against large organizations with deep hierarchies. The fall of "square" culture should be a major blow to maze culture -- americans were no longer interested in becoming cookie-cuttout people (as required by mazes).

Yet Zvi claims that maze culture has been on the rise!

What could be the cause, if many drivers of mazes are in decline?

One hypothesis could be that although the root causes are in some ways decreasing, the damage has been done -- like someone exposed to the common cold at a party, and who gradually gets worse once they get home. America has contracted the maze disease, and it continues to fester. In other words, even if large corporations with deep hierarchies are actually less prevalent than they once were, the maze cultures in those that exist are far, far more developed.

I'm somewhat suspicious of this explanation, because in a culture where non-maze opportunities are opening up more and more, it seems like mazes should die. I think people's behavior is more dependent on context than it is on these subtle cultural influences, so I expect to see larger effect sizes from incentive structures changing than from negative cultures slowly festering over time. But perhaps it's a factor.

Another hypothesis is [slow strangulation via red tape](#). As regulations and restrictions increase over time, it gets harder and harder to escape mazes, and mazes get deeper and deeper. More generally, it could be that although some causes of mazes have diminished, other more important causes have increased, making mazes worse overall.

It's also possible that Zvi is wrong, mazes as such are not on the rise after all, and the creeping insanity which Zvi is noticing has some other source.

Fusion and Equivocation in Korzybski's General Semantics

I [recently](#) tried to figure out Korzybski's General Semantics again. I think I gained some insight into what Korzybski and followers were so excited about, and thought I'd try to write a brief note about it.

My "aha" moment came from reading [the wikipedia entry on General Semantics](#), which includes the following:

General semantics postulates that most people "identify," or fail to differentiate the serial stages or "levels" within their own neuro-evaluative processing. "Most people," Korzybski wrote, "*identify in value levels I, II, III, and IV and react as if our verbalizations about the first three levels were 'it.'* Whatever we may say something 'is' obviously *is not* the 'something' on the silent levels."[\[11\]](#)

[...]

Although producing saliva constitutes an appropriate response when lemon juice drips onto the tongue, a person has inappropriately identified when an imagined lemon or the word "l-e-m-o-n" triggers a salivation response.

This seems like a very Buddhist use of "identification". It reminds me very much of [Kaj's explanation of insight meditation](#). Korzybski is trying to get us to "respond appropriately" by differentiating rather than identifying. He writes that differentiating "introduces a most beneficial delay" between stimulus and response. Identification and differentiation seems like exactly what Kaj is talking about with "fusion" and "defusion".

In traditional argumentation terms, this mistake is "equivocation" -- treating two things as the same when they are importantly different. Equivocating between two things forces us to respond in the same way to both. For example, if no firm distinction is made between "liberal immigration policy" and "democratic values", then one must accept or reject both at once.

- The grecio-roman (ie "western") rhetorical tradition focuses on explicit reasoning, and treats equivocation as an error in explicit reasoning. It is countered by precision in language, including taking care about explicitly defining terms, and careful analysis of the logic of arguments.
- The Buddhist tradition focuses instead on the internal, subjective, phenomenological version. "Identification" here has to do with insufficiently fine-grained perception. It is remedied by introspection and training the senses. This allows one to see more clearly by focused attention, which allows one to differentiate (de-identify) subtly different mental phenomenon, such as that between hearing a sound and the lightning-fast inference which produces an interpretation of that sound (which is normally equivocated with the sound itself, at a basic level). This focus on nonverbal processing makes it seem in many ways like the opposite of the grecio-roman tradition. A Buddhist might naturally say that the focus on rhetoric traps the rhetorician in their mental models, whereas the focus on perception frees one from such mental models. A rhetorician might naturally accuse Buddhists of abandoning symbolic reasoning,

a tool which has turned out to be extremely powerful (e.g. modern mathematics).

- Korzybski wants to do both: train explicit reasoning, while also recognizing "the silent level" (the nonverbal stuff). He uses the terms identification/differentiation to point to the general case.
- Modern psychotherapy has picked up these ideas under the heading of fusion/defusion, mostly focusing on the more preverbal stuff that Buddhism cares about.
- CFAR uses the concept of [bucket errors](#) for the same thing, although there the emphasis is more on why "irrational" behavior might actually be helpful/protective if you're making a bucket error, rather than a heavy-handed emphasis on getting rid of bucket errors in all cases.

In order to point people at *the general idea of making distinctions*, these traditions naturally point at some important distinctions people are neglecting. Korzybski famously touts the map/territory distinction as one of the main things people often equivocate, pointing out that (even if you explicitly endorse the existence of a map/territory distinction) it's easy to make mistakes by thinking an element of your map necessarily corresponds to some reality in the territory. Much was made of this in Eliezer's Sequences.

But Korzybski really wanted to point to a lot more distinctions than this, as evidenced by the reference to "levels I, II, III, and IV" mentioned in the quote above. Level I refers to the territory ("happenings" as he called it), and level IV refers to explicit verbal description; the intervening levels refer to various levels of intermediate processing between the two. But there were many more distinctions as well. Again quoting from the wikipedia page:

Suppose you recognize one student—call her Anna—from a prior course in which Anna either excelled or did poorly. Again, you escape identification by your indexed awareness that Anna_{this term, this course} is different from Anna_{that term, that course}. Not identifying, you both expand and sharpen your apprehension of "students" with an awareness rooted in fresh silent-level observations.

Other traditions have pointed to other distinctions as their central examples. Buddhism has a number of favorite equivocations to point at, including the tangled mess of equivocation that is "the self".

Each important distinction has a number of important implications. Pulling things apart allows us to address each differently. I talked about a number of illustrative examples in [Separation of Concerns](#). The more things we correctly differentiate, the more freedom we have to optimize.

I think a lot of the content of Nonviolent Communication (NVC) is a set of important de-identifications:

- You are encouraged to explicitly label your own experience. This helps de-fuse your experience from the experience of others, so that you don't implicitly assume everyone has experienced things the way you have. For example, you don't assume your roommate knows you find their music annoying (which can lead to the further assumption that they are annoying you on purpose).
- You are encouraged to explicitly label wants and needs, and separate these from requests. This helps de-fuse your desires from the desires of others. For example, rather than say "you need to be quiet", NVC would encourage you to

explain how the noise makes you feel, state that *you* (not them) has a need for quiet, and only then make a request. This facilitates group problem-solving, rather than an argument about whether the other person "needs" to be quiet.

- You are encouraged to avoid "should", as it tends to encourage equivocation between all of the above-mentioned things.

Like the map/territory distinction, these distinctions can be gateway drugs to the general "make distinctions" thing.

What is “protein folding”? A brief explanation

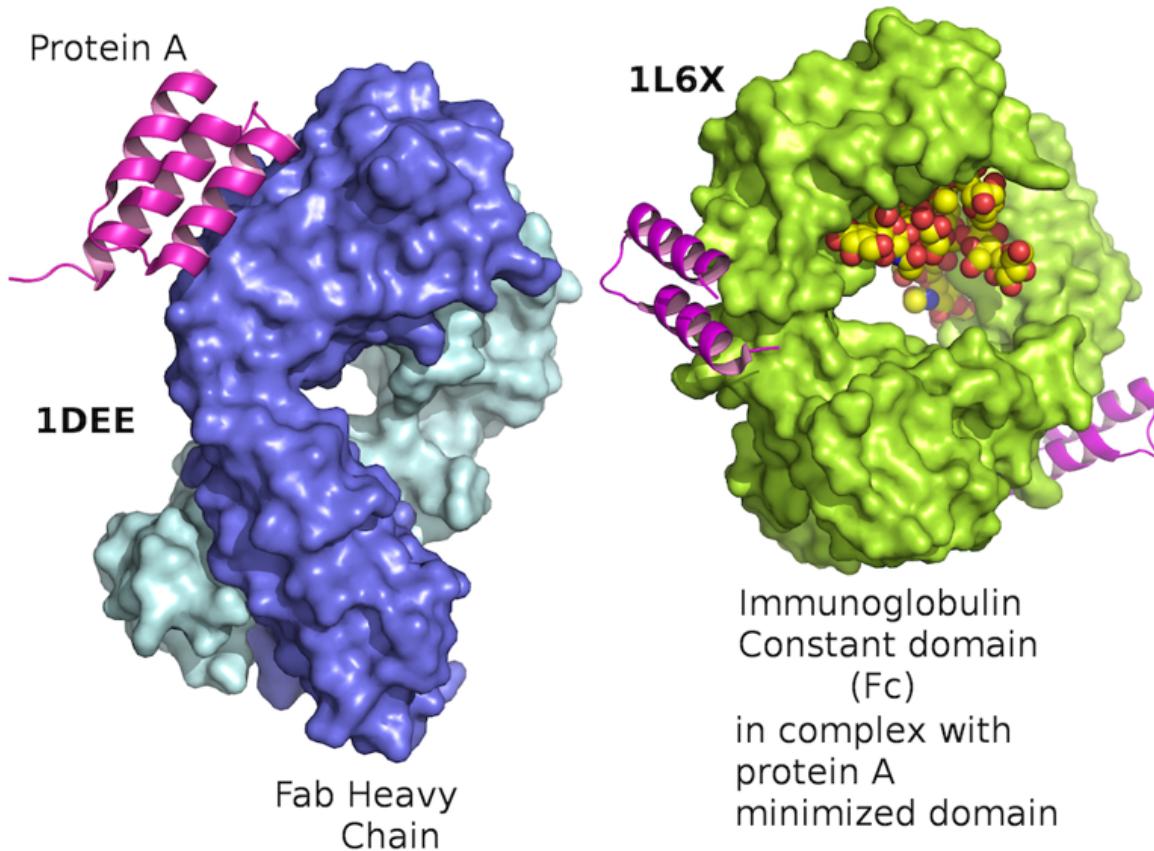
This is a linkpost for <https://rootsofprogress.org/alphafold-protein-folding-explainer>

Today Google DeepMind [announced](#) that their deep learning system AlphaFold has achieved unprecedented levels of accuracy on the “protein folding problem”, a grand challenge problem in computational biochemistry.

What is this problem, and why is it hard?

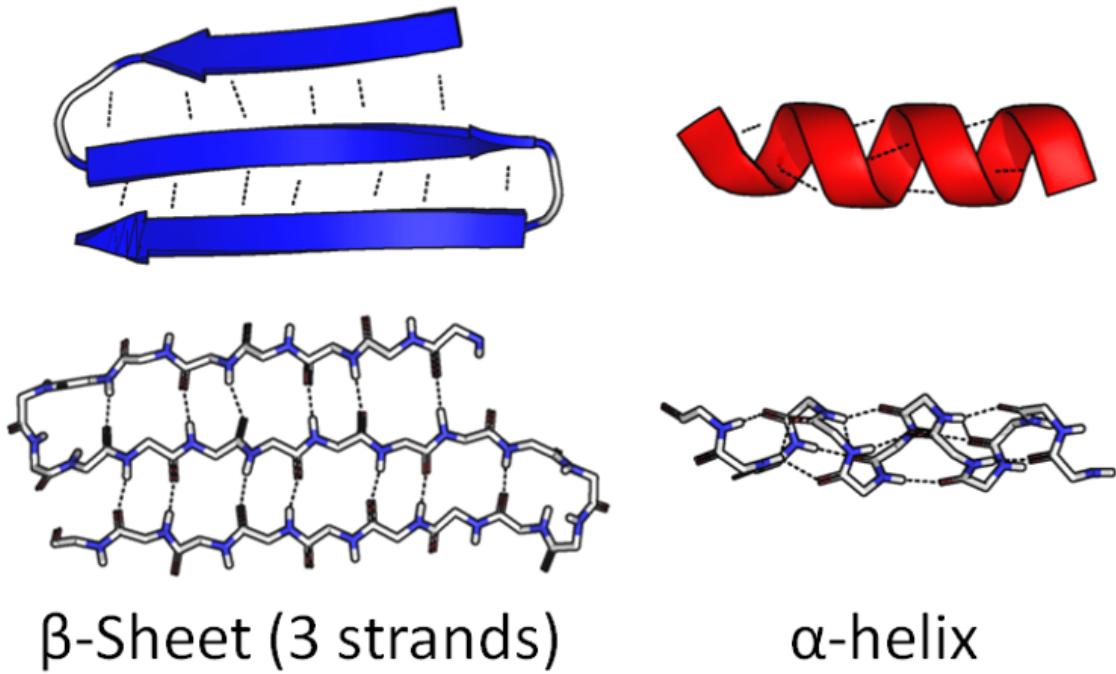
I spent a couple years on this problem in a junior role in the early days of [D. E. Shaw Research](#), so it’s close to my heart. Here’s a five-minute explainer.

Proteins are long chains of amino acids. Your DNA encodes these sequences, and RNA helps manufacture proteins according to this genetic blueprint. Proteins are synthesized as linear chains, but they don’t stay that way. They fold up in complex, globular shapes:



*A protein from the bacteria *Staphylococcus aureus*. [Wikimedia / EAS](#)*

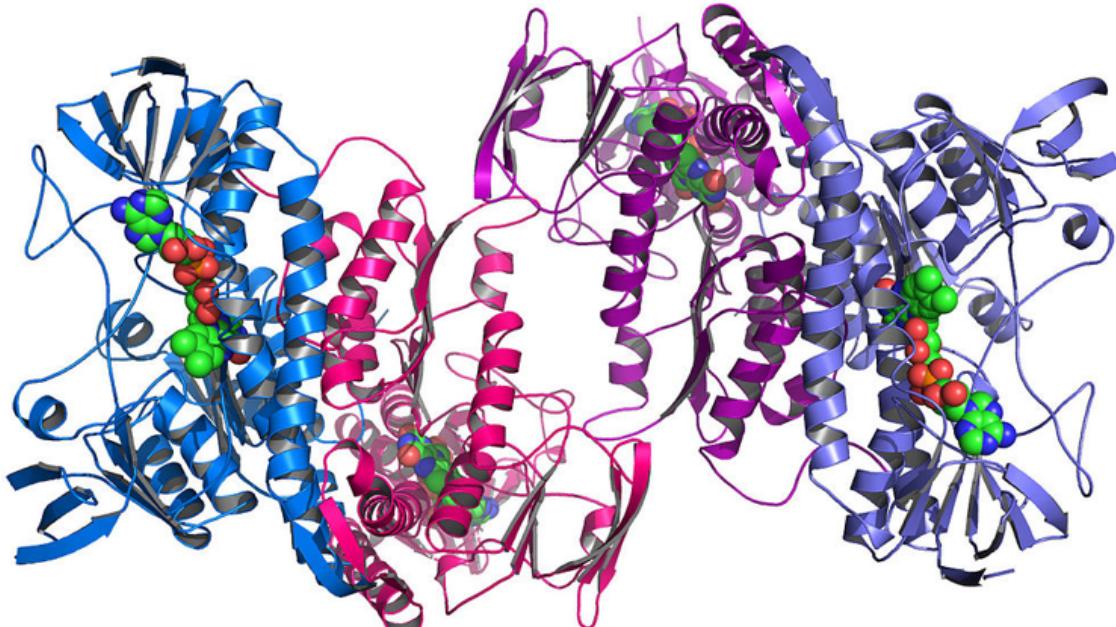
One part of the chain might coil up into a tight spiral called an α -helix. Another part might fold back and forth on itself to create a wide, flat piece called a β -sheet:



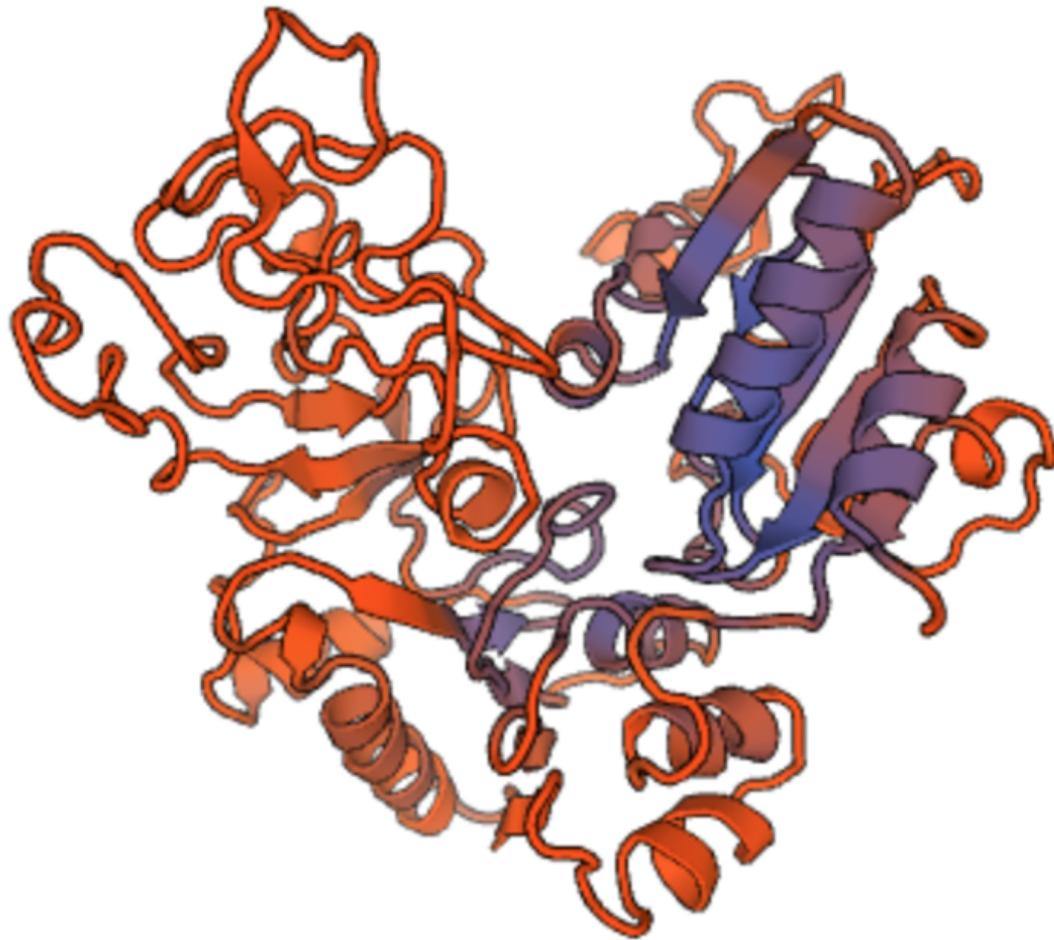
[Wikimedia / Thomas Shafee](#)

The sequence of amino acids itself is called *primary structure*. Components like this are called *secondary structure*.

Then, these components themselves fold up among themselves to create unique, complex shapes. This is called *tertiary structure*:



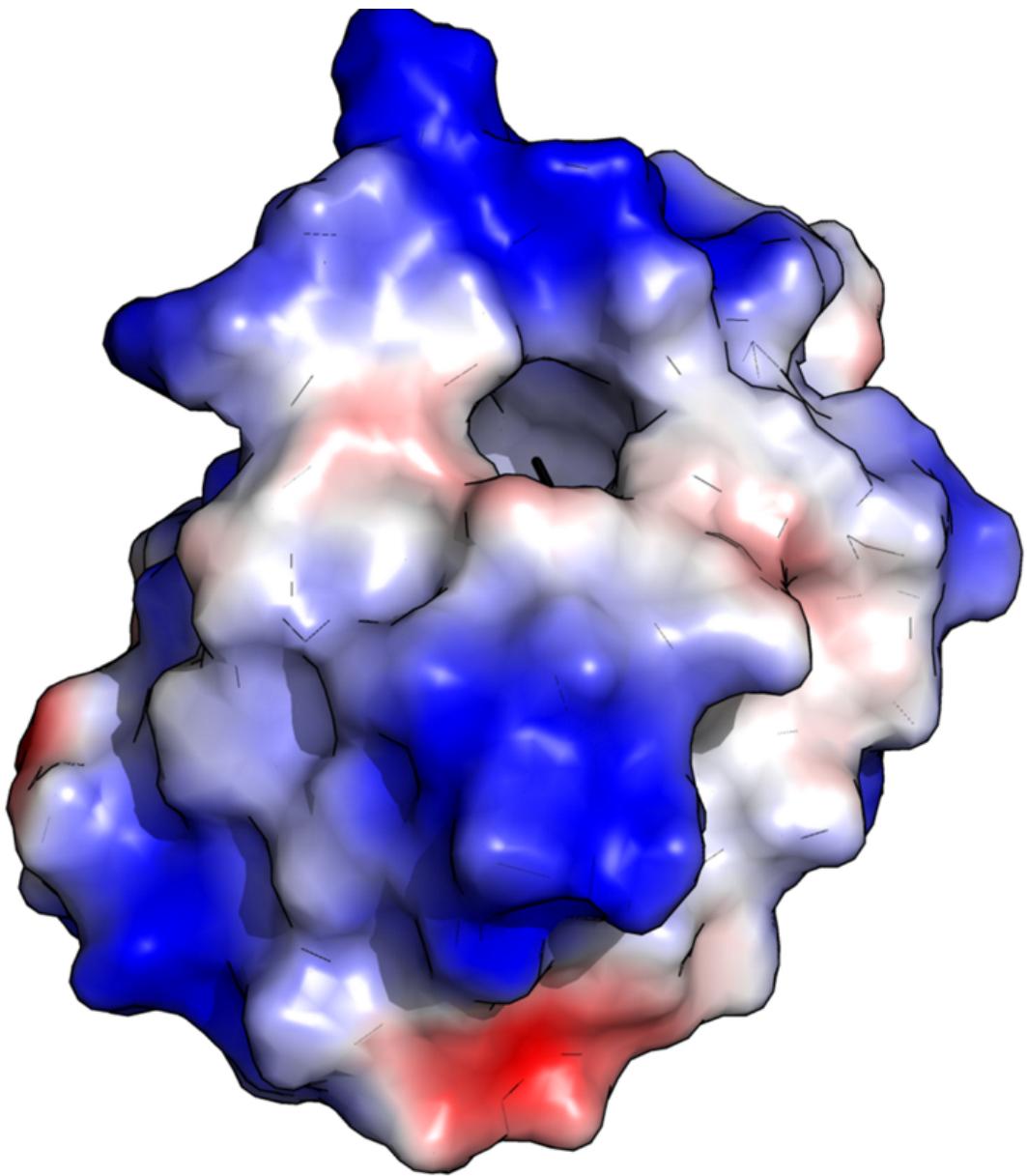
An enzyme from the bacteria *Colwellia psychrerythraea*. [Flickr / Argonne National Lab](#)



The protein RRM3. [Wikimedia / Biasini et al](#)

This looks like a mess. Why does this big tangle of amino acids matter?

Protein structure is not random! Each protein folds in a specific, unique, and largely predictable way that is essential to its function. The physical shape of a protein gives it a good fit to targets it might bind with. Other physical properties matter too, especially the distribution of electrical charge within the protein, as shown here (positive charge in blue, negative in red):



Surface charge distribution of Oryza sativa Lipid Transfer Protein

1. [Wikimedia / Thomas Shafee](#)

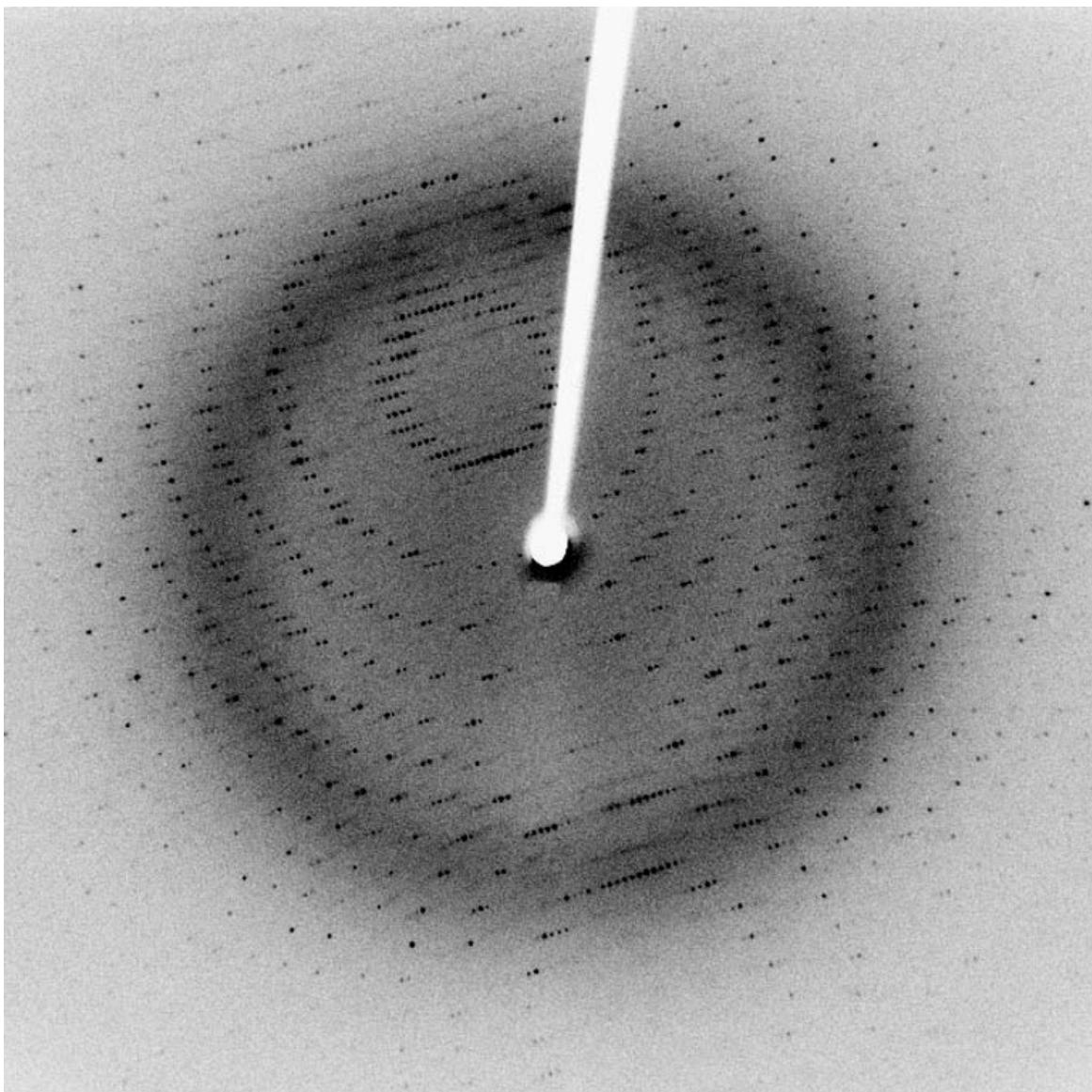
If a protein is essentially a self-assembling nanomachine, then the main purpose of the amino acid sequence is to produce the unique shape, charge distribution, etc. that determines the protein's function. (*How exactly this happens, in the body, is still not fully understood, and is an active area of research.*)

In any case, understanding structure is crucial to understanding function. But the DNA sequence only gives us the primary structure of a protein. How can we learn its secondary and tertiary structure—the exact shape of the blob?

This problem is called “protein structure determination”, and there are two basic approaches: measurement and prediction.

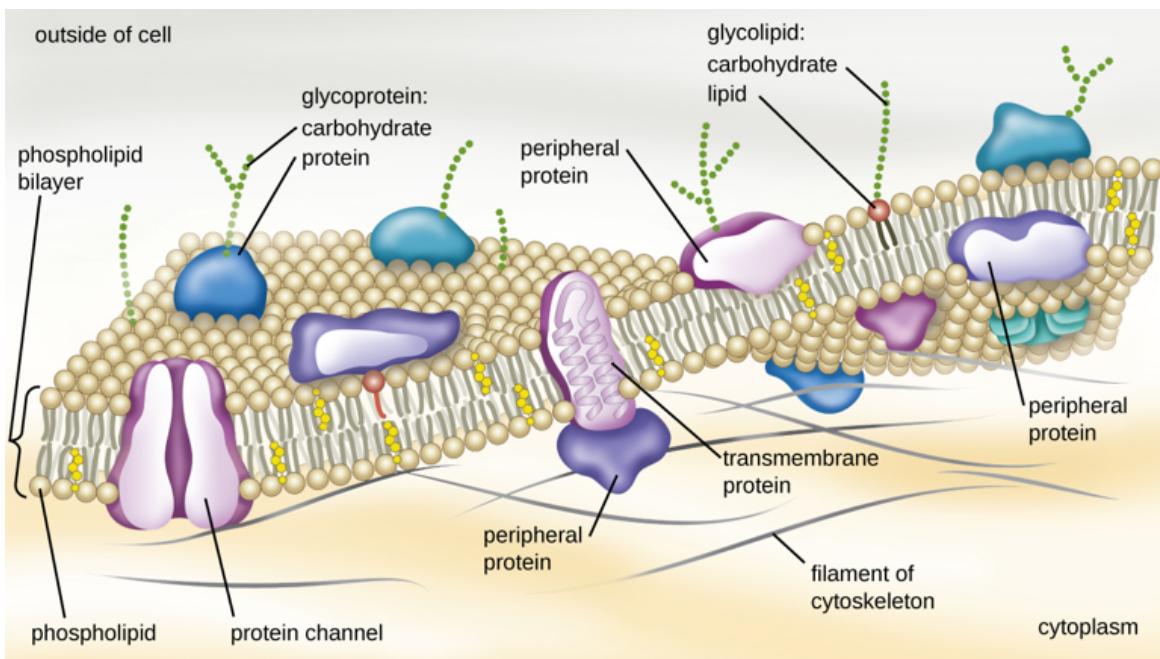
Experimental methods can measure protein structure. But it isn't easy: an optical microscope can't resolve the structures. For a long time, X-ray crystallography was the main method.

Nuclear magnetic resonance (NMR) has also been used, and more recently, a technique called cryogenic electron microscopy (cryo-EM).



X-ray diffraction pattern of a SARS protease. [Wikimedia / Jeff Dahl](#)

But these methods are difficult, expensive, and time-consuming, and they don't work for all proteins. Notably, proteins embedded in the cell membrane—such as the ACE2 receptor that COVID-19 binds to—fold in the lipid bilayer of the cell and are difficult to crystallize.



[Wikimedia / CNX OpenStax](#)

Because of this, we have only determined the structure of a tiny percentage of the proteins that we've sequenced. Google notes that there are 180M protein sequences in the Universal Protein database, but only ~170k structures in the Protein Data Bank.

We need a better method.

Remember, though, that the secondary and tertiary structures are mostly a function of the primary structure, which we know from genetic sequencing. What if, instead of *measuring* a protein's structure, we could *predict* it?

This is “protein structure prediction”, or colloquially, the “protein folding problem,” and computational biochemists have been working on it for decades.

How could we approach this?

The obvious way is to directly simulate the physics. Model the forces on each atom, given its location, charge, and chemical bonds. Calculate accelerations and velocities based on that, and evolve the system step by step. This is called “molecular dynamics” (MD).

The problem is that this is *extremely* computationally intensive. A typical protein has hundreds of amino acids, which means thousands of atoms. But the environment also matters: the protein interacts with surrounding water when folding. So you have more like 30k atoms to simulate. And there are electrostatic interactions between every pair of atoms, so naively that's ~450M pairs, an $O(N^2)$ problem. (There are smart algorithms to make this $O(N \log N)$.) Also, as I recall, you end up needing to run for something like 10^9 to 10^{12} timesteps. It's a pain.

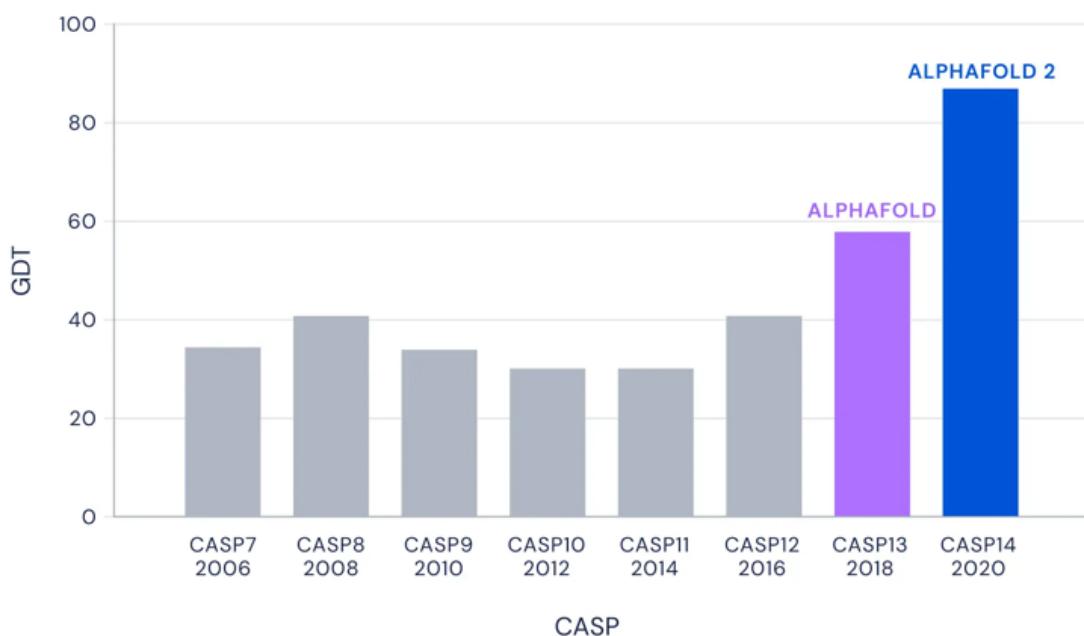
OK, but we don't have to simulate the entire folding process. Another approach is to find the structure that *minimizes potential energy*. Objects tend to come to rest at energy minima, so this is a good heuristic. The same model that gives us forces for MD can calculate energy. With this approach, we can try a whole bunch of candidate structures and pick the one with lowest energy. The problem, of course, is where do you get the structures from? There are

just way too many—molecular biologist Cyrus Levinthal estimated 10^{300} (!) Of course, you can be much smarter than trying all of them at random. But there are still too many.

So there have been many attempts to get faster at doing these kinds of calculations. Anton, the supercomputer from D. E. Shaw Research, used specialized hardware—a custom integrated circuit. IBM also has a computational bio supercomputer, Blue Gene. Stanford created Folding@Home to leverage the massively distributed power of ordinary home computers. The Foldit project from UW makes folding a game, to augment computation with human intuition.

Still, for a long time, no technique was able to predict a wide variety of protein structures with high accuracy. A biannual competition called CASP, which compares algorithms against experimentally measured structures, saw top scores of 30–40%... until recently:

Median Free-Modelling Accuracy



Median accuracy of predictions in the free modelling category for the best team each year. [Google DeepMind](#)

So how does AlphaFold work? It uses multiple deep neural nets to learn different functions relevant to each protein. One key function is a prediction of the final *distances* between pairs of amino acids. This guides the algorithm to the final structure. In one version of the algorithm (described in [Nature](#) and [Proteins](#)), they then derived a potential function from this prediction, and applied simple gradient descent—which worked remarkably well. (I can't tell from what I've been able to read today if this is still what they're doing.)

A general advantage of AlphaFold over some previous methods is that it doesn't need to make assumptions about the structure. Some methods work by splitting the protein into regions, figuring out each region, then putting them back together. AlphaFold doesn't need to do this.

DeepMind seems to be calling the protein folding problem solved, which strikes me as simplistic, but in any case this appears to be a major advance. Experts outside Google are

calling it “[fantastic](#)”, “[gamechanging](#)”, etc.

Between protein folding and CRISPR, genetic engineering now has two very powerful new tools in its toolbox. Maybe the 2020s will be to biotech what the 1970s were to computing.

Congrats to the researchers at DeepMind on this breakthrough!

Covid 12/10: Vaccine Approval Day in America

Today, the FDA is meeting to discuss Pfizer's Covid-19 vaccine. By the time many of you read this, they will have hopefully given the vaccine, and perhaps Moderna's as well, emergency use authorization. If that happens, distributions to the states can begin, and some people will be vaccinated as early as tomorrow.

That is excellent news. Alas, supplies remain highly limited and the vaccine takes a month to work. States are getting smaller allocations of the vaccine than they expected. The situation for the holiday season remains unchanged, and most of us will likely not become immune until some time around May.

Two weeks ago, I noted data was about to get weird. A week ago I observed that the data was weird. What I didn't note was that the [Covid Tracking Project had written a post about exactly in what ways](#) they expected it to be weird, which were different in places from the ways I expected. Lots of good observations in there. The post makes a convincing case that I was underestimating expected delays and bottlenecks.

Despite that, we saw a big jump in positive test percentage on Tuesday, December 2, exactly when I'd expected one. That could mostly be a coincidence. Perhaps that was more of a shift in timing of reporting, with the real surge from Thanksgiving coming later. I certainly buy the augment that such super-spreader events infect people in multiple waves, as those infected take the virus home, and we should expect an unusually high effective R₀ in the week after Thanksgiving.

That will not give us much time to recover before [Christmas](#) and New Year's Eve give their one-two punch. It seems unlikely we will turn the corner until after those get to play out. Despite previous holidays not showing up much in the data, these big three seem like a huge deal to me, especially now. People are sick of it all.

People are also sick, period. Death rates are up. Positive test rates are up. Cases are up.

This week's post attempts to limit Twitter links. In cases where the link is serving as a source but is otherwise inessential, I'm going to link it explicitly and purely as a source, so there's no need to click unless you want to verify it. Any other Twitter links didn't have a good substitute I could find easily.

Let's run the numbers.

The Numbers

Predictions

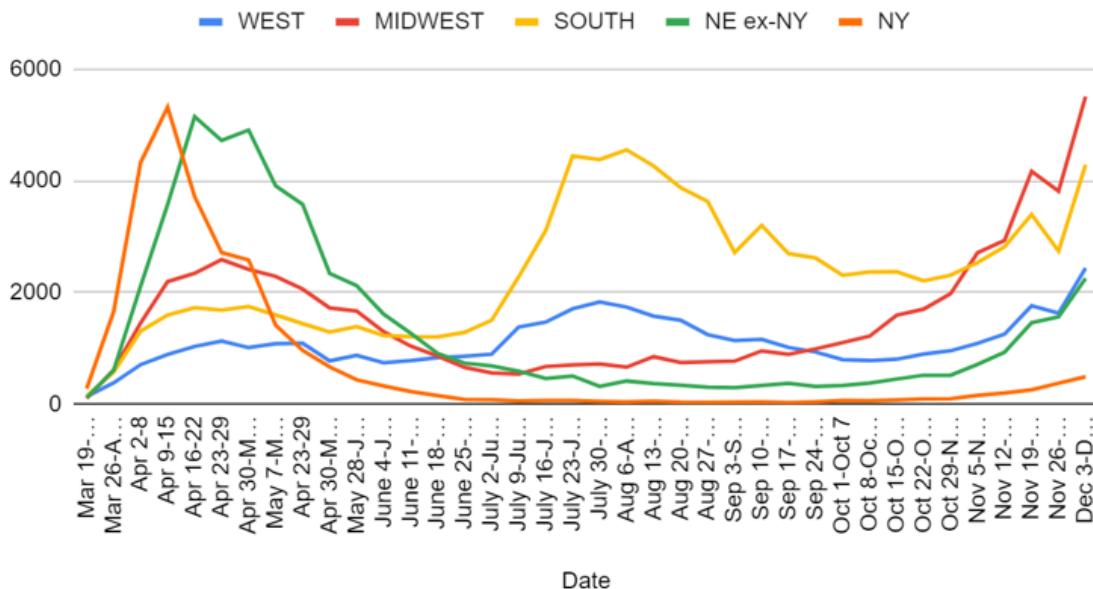
Last week's prediction: My unconfident prediction is a 14.8% positive rate on 10 million tests, and an average of 2,200 deaths to account for at least some amount of catching up in reporting. But, again, wider than usual error bars all around.

Results: We got a 13.7% positive rate on 10.4 million tests, and an average of 2,276 deaths. So somewhat better positive rate than I expected, but not as good as one might hope.

Next week: My baseline prediction is a 14.3% positive rate on 11 million tests, and an average of 2,550 deaths per day.

Deaths

Deaths by Region

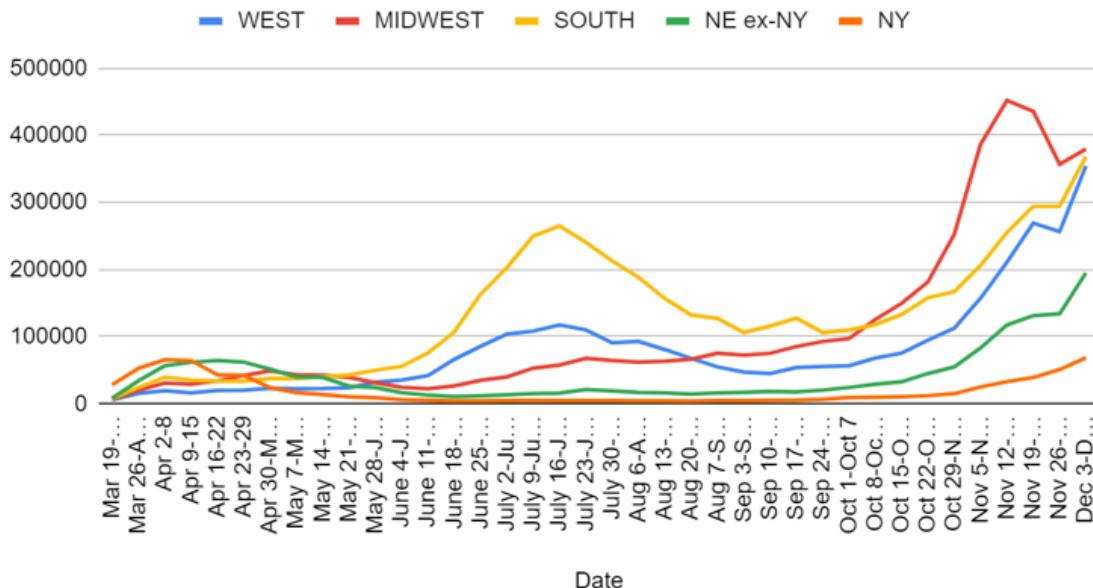


Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 8-Oct 14	782	1217	2366	436
Oct 15-Oct 21	804	1591	2370	523
Oct 22-Oct 28	895	1701	2208	612
Oct 29-Nov 4	956	1977	2309	613
Nov 5-Nov 11	1089	2712	2535	870
Nov 12-Nov 18	1255	2934	2818	1127
Nov 19-Nov 25	1761	4169	3396	1714
Nov 26-Dec 2	1628	3814	2742	1939
Dec 3-Dec 9	2437	5508	4286	2744

If anything, deaths continue to lag infections, likely due to holiday effects, so this is likely an *underestimate* of how bad things are out there. Even at face value, things are clearly quite bad. The Midwest number is higher than any previous peak, the South is not far behind and the total death count is at an all time high, with numbers going up dramatically everywhere.

Positive Tests

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 8-Oct 14	68284	125744	117995	38918
Oct 15-Oct 21	75571	149851	133238	43325
Oct 22-Oct 28	94983	181881	158123	57420
Oct 29-Nov 4	112684	252917	167098	70166
Nov 5-Nov 11	157495	387071	206380	108581
Nov 12-Nov 18	211222	452265	255637	150724
Nov 19-Nov 25	269230	435688	294230	170595
Nov 26-Dec 2	256629	357102	294734	185087
Dec 3-Dec 9	354397	379823	368596	263886

Outside of the Midwest, it seems clear things continue to rapidly get worse, and last week was a data hiccup as expected. In the Midwest, it looks plausible from this graph that things have peaked, but positive test percentages will be the judge of that.

Test Counts

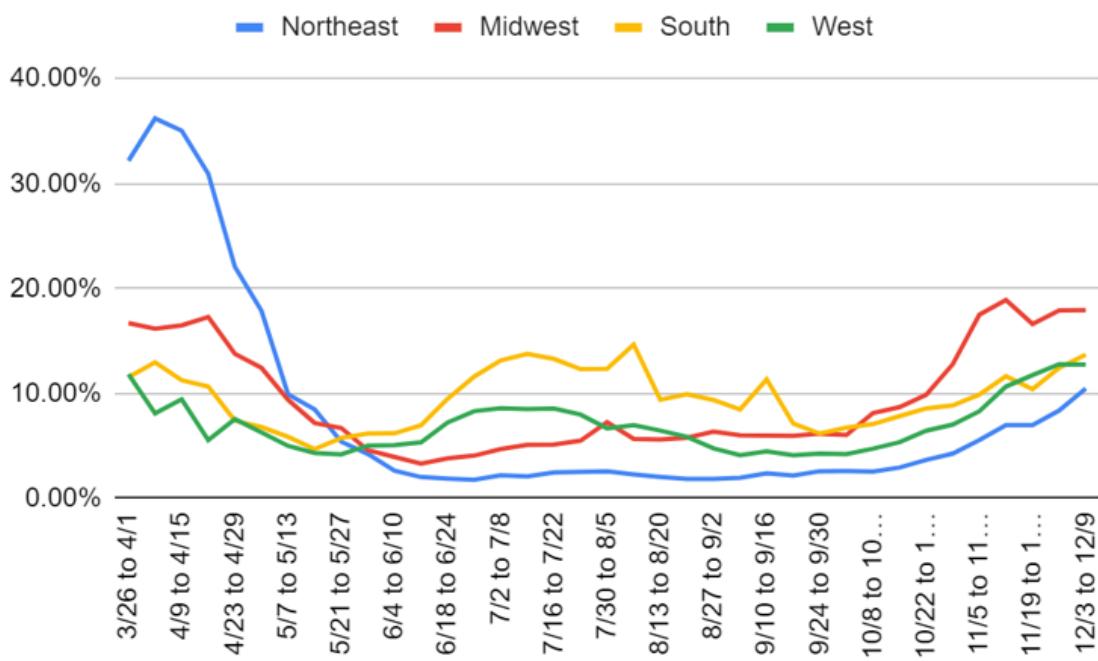
Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
Oct 1-Oct 7	6,009,923	5.2%	763,935	1.3%	2.28%
Oct 8-Oct 14	6,322,326	5.7%	850,223	1.1%	2.39%
Oct 15-Oct 21	6,440,150	6.5%	865,890	1.2%	2.52%
Oct 22-Oct 28	6,933,155	7.5%	890,185	1.4%	2.67%
Oct 29-Nov 4	7,246,068	8.6%	973,777	1.6%	2.86%
Nov 5-Nov 11	8,285,598	10.6%	1,059,559	2.4%	3.13%
Nov 12-Nov 18	8,917,701	12.4%	1,155,670	2.9%	3.47%
Nov 19-Nov 25	10,430,055	11.6%	1,373,751	2.9%	3.83%
Nov 26-Dec 2	9,788,446	11.5%	1,287,010	4.0%	4.18%

Dec 3-Dec 9 10,458,813 13.7% 1,411,142 4.9% 4.61%

Test counts have recovered back to their peak and should presumably resume going up now. That positive rate is one ugly number, and it looks like New York is rapidly ceasing to be that much safer than the rest of the country, although it does run way more tests than most places.

The cumulative positives number is getting quite substantial, especially given how many infections are being missed, and that those infected are not a random sample. Without immunity from those already infected, we would likely be seeing growth now like we saw back in March.

Positive Test Percentages



Percentages	Northeast	Midwest	South	West
10/8 to 10/14	2.57%	8.14%	7.09%	4.75%
10/15 to 10/22	2.95%	8.70%	7.85%	5.36%
10/22 to 10/28	3.68%	9.87%	8.58%	6.46%
10/29 to 11/4	4.28%	12.79%	8.86%	7.04%
11/5 to 11/11	5.56%	17.51%	9.89%	8.31%
11/12 to 11/18	6.99%	18.90%	11.64%	10.66%
11/19 to 11/25	7.00%	16.62%	10.41%	11.75%
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%
12/3 to 12/9	10.47%	17.94%	13.70%	12.76%

I am ready to believe the story that the Midwest is more or less peaking now. Their test percentages have been in a narrow range for the last five weeks, and their immunity effects are rapidly increasing. The West's number is encouraging, but it is on vastly more tests. That many more tests should drive down percentages enough to account for this, or alternatively

disproportionally many negative tests got reported late and are warping the curve there a bit. Optimism seems premature.

In the Northeast and South, trends do not look good, and the Northeast especially looks quite bad, as if a combination of holidays and colder weather has broken the control system, and things are likely to get substantially worse before they get better. Time to bunker down.

My best guess is that the likely overall true peak is after the effects of Christmas and New Year's, so the measured peak of infections should be some time in mid-January, with deaths closer to the end of January. Early vaccinations will assist a little with this, but not much, as they are going to the vulnerable and deserving rather than to those most likely to spread the disease, and there are not that many doses.

Note that I continue to not list hospitalizations in these charts. They continue to rise, but that seems to me to still say more about how hospitals are handling things than about how many are infected or sick. Even if things were improving and there were plenty of beds, hospitalizations would still be rising *now*, because there would be more people getting sick now than are ready to leave the hospital, as discharges are going to lag cases by quite a bit. So while it's certainly something I look at as a sanity check, I don't know what it can teach us.

Covid Machine Learning Projections

For those interested, here's [a closer look at the Covid-19 Projections website's nowcast model for true prevalence](#). [Layman summary by him on Twitter here](#). Can be compared to his earlier write-up [here](#). Seems intuitively like reasonable approximations. What it doesn't look like is the output of a machine learning algorithm, which is interesting. It looks a lot like something I would create if I was trying to plausibly fit a bunch of curves and generate reasonable claims in a smooth consistent manner, except that this uses a suspiciously large number of round numbers. I have some experience with this sort of thing.

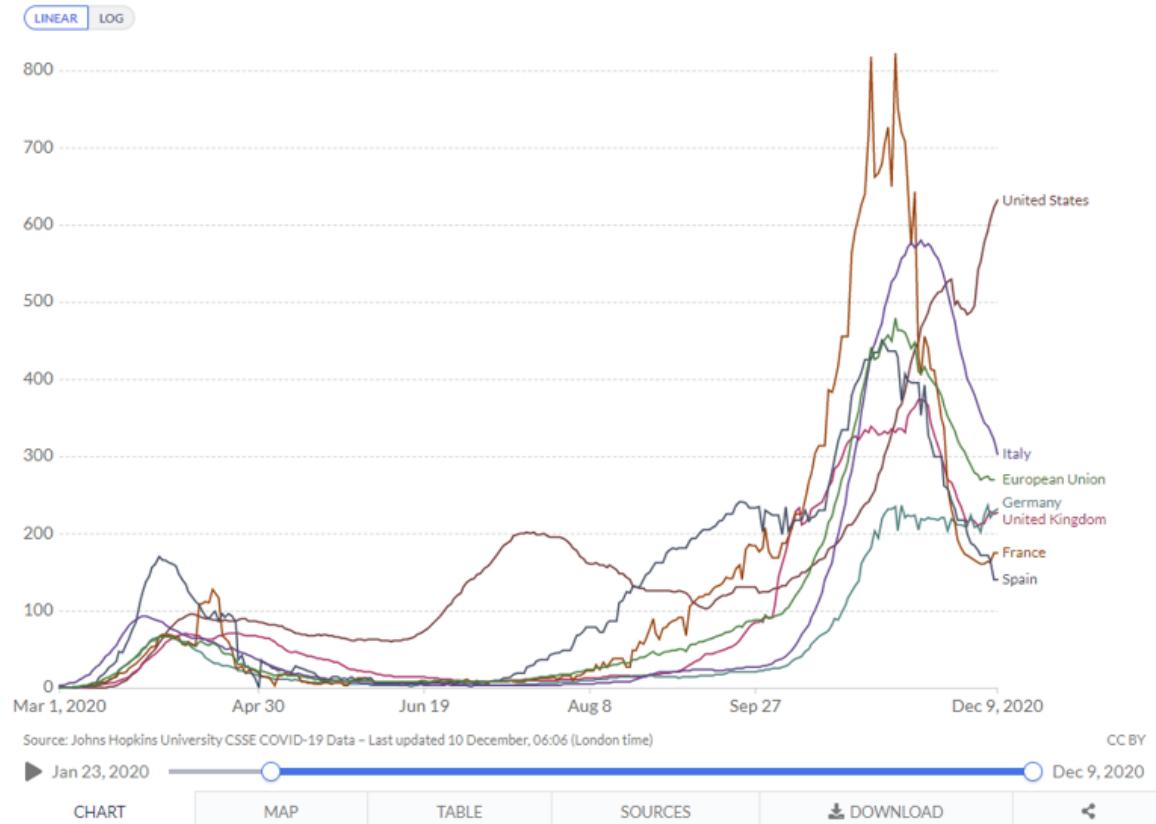
The machine learning project doesn't know about Thanksgiving and its warping effects on data, so it still thinks there was a brief downward blip in infections a few weeks ago, although it now recognizes that things resumed getting worse shortly thereafter. If you smooth that out, answers here continue to seem sane and they see us as having had 630k infections per day two weeks ago, with 2.6% infected at that time and 17.3% of the population having been infected at some point, or over 20% by today. I continue to think that answer is on the very low end of plausible.

Europe

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World
in Data

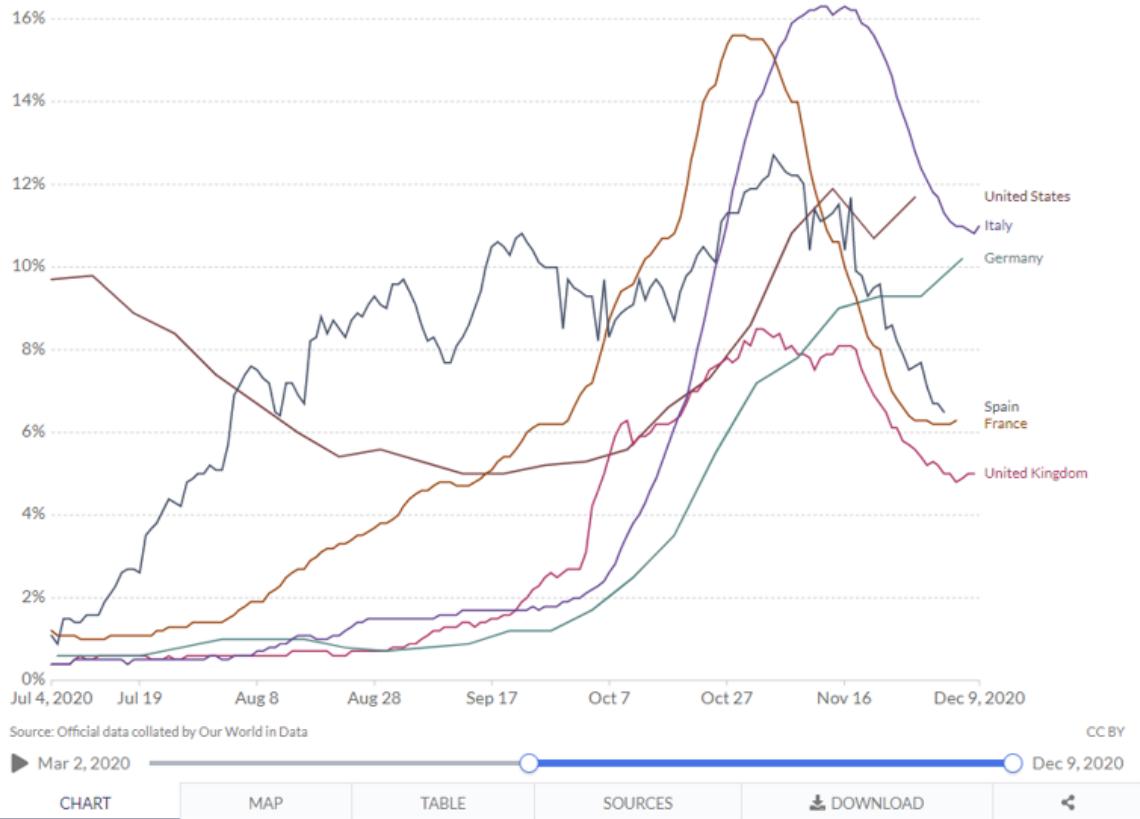


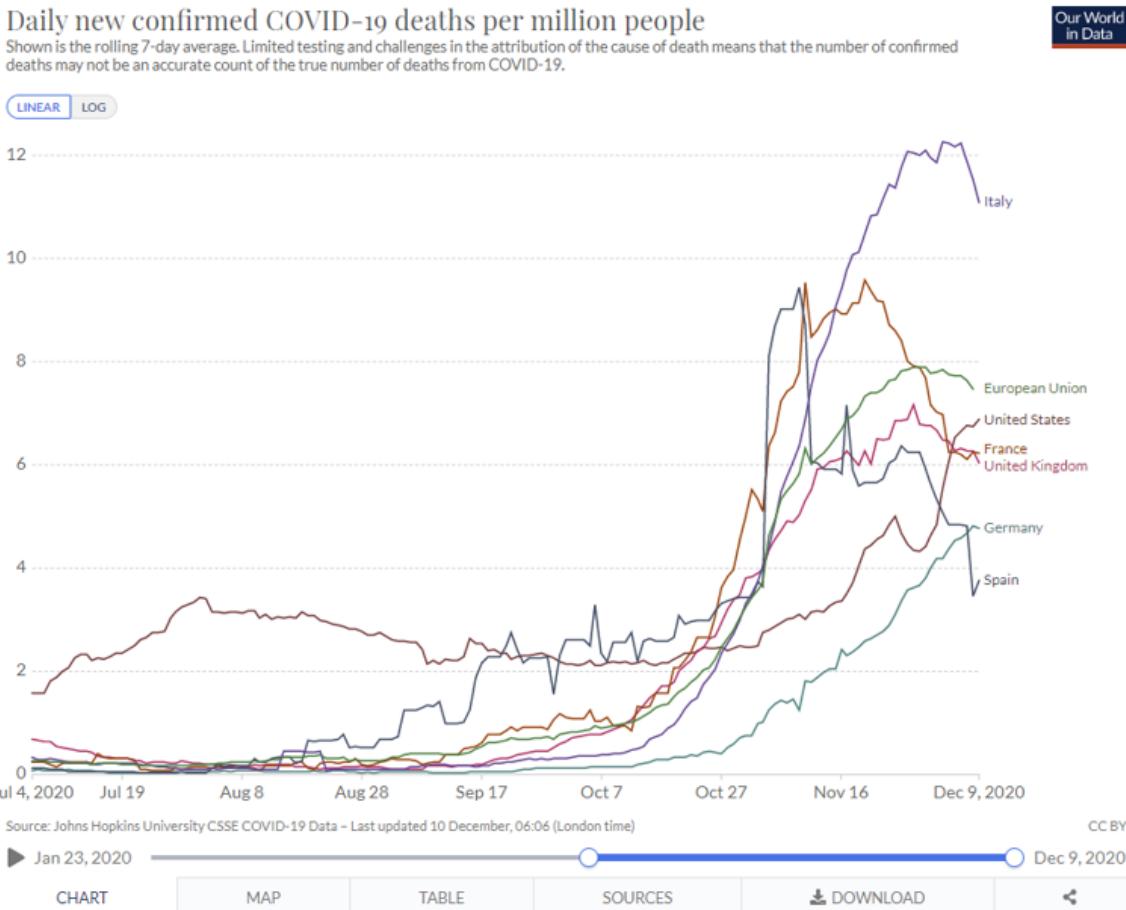
Germany continues to be stable. Many others have seen dramatic declines from their peaks, and the United States once again has a far worse Covid problem than Europe.

The share of daily COVID-19 tests that are positive

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

Our World
in Data





My news sources have not seen much happening in Europe this week, and the charts seem straightforward. Countermeasures have worked in at least containing things and preventing them from getting worse, but no one is aiming at suppression, instead they are waiting for the vaccine.

No, Not That Pandemic Vaccine, That Other Pandemic Vaccine

[Team behind Oxford Covid jab start final stage of malaria vaccine trials](#). This is excellent news. Also a reminder that the world is smaller than it appears, as the same team behind one of three Covid-19 vaccines also produced this one. Given how most of the time and effort involved in vaccines seems to involve testing them and overcoming regulatory barriers, it should not be surprising that teams that make one vaccine can also make more.

Malaria is kind of a big deal, killing over 400,000 people a year. Bed nets are one of the prototypical Effective Altruist causes. Whether or not the best strategy would be to instead eradicate malaria in Africa [the way it was eradicated in America](#), that is not happening at the moment. Barring that, a cheap and widely available vaccine seems like an excellent way to go.

"A lot more people will die in Africa this year from malaria than will die from Covid. I don't mean twice as many, probably 10 times," Hill said. The vaccine "is going to be available in very large amounts, it works pretty well. And it's going to be very low priced."

The vaccine could be in use by 2024 if the final human trials are successful, he said.

So we are now entering the final stage of trials, which means we are only *three years* away from starting vaccinations. Some time next year, 4,800 children will be vaccinated, then we will wait a very long time, likely in part because we only vaccinated 4,800 children. This is what people mean when they say what a great achievement the Covid-19 vaccines have been in terms of speed. Here we have a vaccine that likely could save over a million people if we sped up its release and distribution by three years, and we are not doing that because of reasons.

All of you Effective Altruists who were already working hard on malaria, is there any way to push on this and make it go faster? Perhaps you can arrange for a larger trial that will gather more data faster? Perhaps you can fund scaling up production or distribution faster in advance? What about Bill Gates, who is very involved with malaria and funded pre-production of Covid-19 vaccines? By the end of 2021, we should have sufficient doses of Covid-19 vaccines, which leaves years to handle production on this one even if they are in conflict. It does not appear this is getting the level of urgency it deserves.

All I Want For Christmas Are a Covid Vaccine and a PS5, But They Underpriced Them And Now They're All Sold Out

What does everyone else want for Christmas?

[A PS5, yes.](#) And thanks to my good friend Seth Burn, I got one! Section title will have to change.

For a slim majority, also a Covid vaccine.

There was a Pew Research survey this week on who intends to get the vaccine.

For those who want to guess who wants it more before looking, the categories are Male, Female, Postgrad, College, HS or less, Upper income, Middle income, Lower income, Over 65, Under 30, White, Black, Hispanic and Asian.

Answers are at the end of the section via screenshot (or Twitter source [here](#)).

To help convince the hesitant ones to change their minds, at least some skin in the game is officially here: [Former Presidents Obama, Bush and Clinton volunteer to get coronavirus vaccine publicly to prove it's safe](#). I would assume Biden as well. Trump already had Covid-19, but if the goal is to show it is safe rather than effective, it would be great if he joined them anyway. I haven't seen this actually suggested anywhere? Will be a while before it matters much, given the severe shortage of doses.



Ryan Struyk
@ryanstruyk

...

Americans who will get coronavirus vaccine via new Pew Research poll:

83% Asian
75% Postgrad
75% Over 65
71% Upper income
69% Democrats
67% Men
66% College
63% Hispanic
61% White
60% Middle income
56% HS or less
55% Under 30
55% Lower income
54% Women
50% Republicans
42% Black

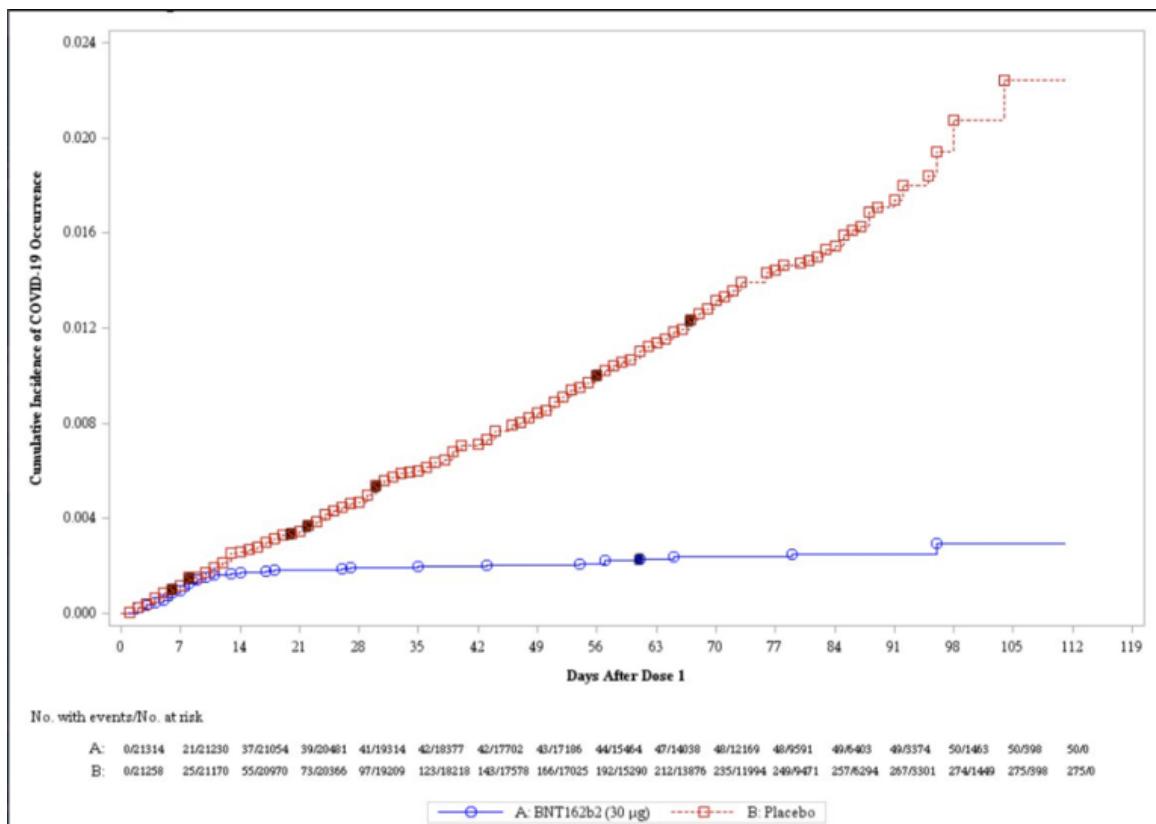
1:21 PM · Dec 4, 2020 · Twitter Web App

You know who should likely get the Covid vaccine early? Airline pilots.

So naturally, the [FAA is currently telling them that if they get their vaccine, they will lose their medicals, and not be allowed to fly.](#) They are monitoring the FDA's decisions closely, but pilots are not allowed to get vaccinations that don't have full approval. I would never have predicted this particular plot point, but on a meta-level I find it *entirely unsurprising*. It makes sense that there would be government regulations that ban the vaccine from getting to some of the people who need it most. That the best we can hope for is that some such authorities try to kill us less aggressively some of the time.

Scream If You Wanna Go Faster

The vaccine even works faster than we thought! From [the FDA report: Looks like efficacy even for one dose might be 88.9%](#). Here's the graph.



That doesn't mean we can do only one dose per person for a while until we have enough doses, because we don't know the consequences of a longer wait or whether the immunity from only one dose would last, but it's still great news. It also puts people in a profoundly weird spot, since they're *mostly* immune now, but will be *much more immune* soon. How safe do you play it during that period?

Business insider asks why [the United States is behind the United Kingdom in approving the vaccine](#). Here is the functional explanation they give:

The regulatory process involves experts reviewing “thousands of pages” of technical information, an FDA representative told Axios on Tuesday.

The agency also has to evaluate the manufacturing process, check statistical analyses, and look at the effect of the vaccine on groups of people with high risk of side effects, the representative said.

The FDA and the Medicines and Healthcare products Regulatory Agency, the UK regulator, have different processes to review vaccines, Dr. Penny Ward, a visiting professor in pharmaceutical medicine at King's College London, told The New York Times.

The FDA asks drugmakers for raw data, which it reanalyzes, while British regulators rely more heavily on reports produced by the drugmakers.

An alternative hypothesis is that *the FDA is delaying the process intentionally*. With the cover story that this is being done ‘in order to be careful’ so as to [prevent fear that the vaccine was rushed](#). I strongly agree that there are approximately zero people who are going to look at the gap between application for approval and when the FDA approves the application, and use that as a true objection that justifies not taking the vaccine. To the extent that this group has non-zero members, the group Tyler points out, that takes the delay as worrisome and uses *that* as its true objection, is not obviously smaller.

The simpler explanation is that there was already a meeting scheduled for December 8-10, people had holiday plans, and the FDA simply did not care enough about doing this faster to move the meeting.

In other Tyler modeling news, while the cause of faster vaccine approvals is clearly righteous, I do not think we can [model a day's delay in approval as delaying vaccine production by one day](#). The story flat out makes no sense to me. Pfizer is not going to produce vaccine slower, at any point, because the FDA gives formal approval on December 10 instead of November 30. The vaccine is already approved in the UK, so there is a guaranteed place to sell more doses than they can possibly produce this month, in the worst case scenario. The probability of approval by the FDA isn’t 100%, but given their knowledge of the data, it’s close. Every sign points to the approval being a formality. So I hold strongly that the damage done is that current doses are shipped with a delay, not that all doses are delayed.

Still, *that’s pretty bad*. It looks plausibly like a third of all deaths are of nursing home residents that could be reached in the first wave. So they all are getting vaccinated two weeks late, which means the cost here is something like 1,000 additional deaths per day of delay from that alone. Then add in the healthcare workers, and their new ability to *actually do their jobs fully* and also avoid turning hospitals into outbreaks. So yeah, it’s a lot of blood, but giving them every death between application and approval isn’t fair.

After I wrote that section, [Tyler wrote a post even more explicitly disagreeing with this](#), essentially claiming that yes approval would accelerate other tasks substantially, which I continue to mostly not buy. I think his argument holds when approval is truly uncertain, but not when it is a formality. I do understand and share the frustration with those who use constraint X to justify not working to remove constraint Y, which we don’t work on because constraint Z, which we don’t work on because constraint X. Yes, we should be working on the other constraints. But that doesn’t change the answer to the question.

[Switzerland is going even slower](#), making the usual noises about the need for ‘caution.’ My read is that this is because they did not order vaccine doses early enough, and now they are all sold out, so why not spin that by not approving the vaccine for a while and calling it ‘caution?’

As we know from last week, the Europeans are taking it relatively slow, *despite knowing the answer already*:



Bruno Maçães

@MacaesBruno

Person in charge of inoculation plan in Portugal: "there is no doubt vaccine will be approved by EU agency on December 29, but we just have to wait for that". Does this make sense to anyone?

6:58am · 10 Dec 2020 · Twitter for iPad

2 Replies 10 Retweets 71 Likes



...

[Reply to @MacaesBruno](#)



Bruno Maçães

@MacaesBruno

1h

If ever bureaucracy cost countless lives...

3 1 27 ...



Bruno Maçães

@MacaesBruno

1h

Impression it gives is officials were just taken by surprise by how fast vaccine was developed. They must have been planning the approval process for summer of 2021

15 ...

The impression it gives to me is not that officials were taken by surprise. It is that they had a choice to make, and made it.

[Cell asks how we might speed up future vaccine development in case of a new pandemic.](#)

The strategy is based on it not being that hard to predict where future pandemics come from. If most are going to be variations on a few existing virus families, we can do a lot of the work in advance. They suggest that for only a few billion dollars, we can do advance Phase I and Phase II trials for vaccines designed for each virus family that could plausibly cause a future pandemic, stockpile doses, and be ready to proceed to Phase III right away.

If we had the vaccine in question ready to go on day one, then the phase III could begin the equivalent of February, then all we'd have to do is wait for enough infections.

What they do not say is that we could do much better than this by using human challenge trials. If we were willing to do such trials, and had already gathered safety data, there is no reason we could not know the vaccine's effectiveness and safety within a few weeks of identifying the new pathogen. Remember, Moderna created their vaccine in two days. Then all that is left is to scale production.

Cell does not suggest this, instead taking the need for a regular Phase III as a fixed law of nature that cannot be changed even when preparing in advance for a future threat.

And that's valuable! I am glad that people are taking seriously the question of "if the people and procedures in charge of slowing down and stopping development must be accommodated, how can we do that as quickly as possible?" Very good question. I only wish we could also fully ask an even better one.

The vaccine is the ultimate case of [More Dakka](#). However much we are spending on it, it won't be enough. We should spend more. That was true before. It's still true now. [The case for going big](#) remains overwhelming. This whole time, we could have built far more capacity in exchange for (in context) trivial amounts of money. Pfizer offered us 200 million doses of vaccine in July, but we balked at the price tag, and only ordered 100 million doses. Then this week the Trump administration tried to buy more, but Pfizer informed us that other countries bought those other doses instead, and they won't have capacity for more until next June. Whoops. Then Trump issued some sort of executive order demanding that the doses go to us instead, because America, [to which the chief scientist of Operation Warp Speed has said, "I literally don't know" what it means](#). And in an evergreen statement, "The White House is doing what the White House is doing."

I don't *think* this is going to slow us down much once Moderna is added to the picture, but if this mistake costs us all an extra month or two of our lives, it would be one of the worst times someone has walked away from an obviously great deal. If I was told my available dose was in June rather than April, I don't think I *personally* take that much extra risk given how perilous a state the hospital situation might be in for the next few months, but it definitely feels like a plausible last straw for many. In particular, being given one timeline, then having a delay for such a stupid reason, seems very much a 'screw it I'm done with this' moment. Again, I'm *not doing that* if this happens. But if others do that, *I totally get it*.

[The Money Illusion has thoughts](#), speculating that bureaucracies are simply incapable of adjusting on the fly when the things that change matter. In Sumner's proposed model, what you enter the crisis with is mostly what you are stuck with unless you force their hand, and we forced it but not enough.

[The Grumpy Economist wants to go fastest of all](#). He suggests the *obviously* correct answer, which is to allow free market vaccines and to distribute them by price, starting in February for those willing to buy it under those conditions. He claims this would have prevented the pandemic, which is going too far, since scaling up production and convincing enough people to take it would have taken months.

If we wanted to prevent the pandemic in the first place, that requires going much further. We would have needed to allow free market *testing*.

Let us be clear. Government bans on private action are *entirely* responsible for the *entire* pandemic in the West. Period.

Meanwhile in India, [they rejected the emergency use authorization for AstraZeneca's vaccine](#). Which seems entirely unsurprising, given that there are all sorts of problems with the studies so far, which is why they didn't even attempt such a thing in the West. But of course, it's also, from an expected value position, *completely insane*. The vaccine is super cheap and the safety data seems ironclad. So what if we don't know how effective it is? It's weird to pick on India here. If anything, they deserve praise for making it worthwhile *to ask the question at all*. They *might* have gotten the right answer, even though they didn't.

For Me But Not For Thee

Austin Mayor Steve Adler ([source](#)):



Tony Plohetski 
@tplohetski

ooo

EXCLUSIVE: Austin Mayor Steve Adler told the public to "stay home if you can...this is not the time to relax" in a Nov. 9 Facebook video. He did not disclose that he was at a timeshare in Cabo San Lucas after flying on a private jet with eight family members and guests.



Austin Mayor Steve Adler

Nov 9



...



COVID-19 trends and the election.

2:40 PM · Dec 2, 2020 · Twitter Web App

Late Show Host Steven Colbert ([source](#)):



none of this matters
@grantgambling

...

He said... at work.

He has a job.



Staci D Kramer ✅ @sdkstl · Dec 4

"If you're watching this at home right now, thank you. Please stay there." #lssc



12:34 AM · Dec 4, 2020 · Twitter for iPhone

The Ramsey Tax Is Too Damn High

I think [this calculation by Tyler Cowen](#), viewing Covid-19 as a “Ramsey tax problem,” is not only importantly wrong but exactly backwards and rather perverse, in ways worth exploring.

The whole point of saying we are overreacting, to say that the costs of Covid-19 are not that high, is exactly *to argue that the costs borne have been extraordinarily high*. The claim is that the costs *from infection* are low, and that we are making a mistake to bear such high costs, such huge deadweight losses. There hasn’t been a huge non-health loss *as a law of nature*. There’s been a huge non-health loss *because we made choices*. Why must we accept those choices as given, and thus assume that the non-health costs must be so high relative to counterfactual health costs, if that is the current situation? And why are such people being accused of making *moral* claims when they are making efficiency arguments?

In particular, why is “talking people out of their high elasticity” equivalent to the statement “Don’t leave NYC just because the taxes are going up! It will wreck the city.”? Tyler’s sentence is calling for sacrifice by the individual on behalf of the group. Whereas talking

people out of their high elasticity is more like “*you would be better off* living your life mostly as you lived it before, and accepting that you are likely to get Covid-19 at some point, because life is short.” If anything it’s the opposite of the group sacrifice. What is going on here?

The better educated tend to be staying safer than the lower educated, as Tyler says. That seems like a distinct argument.

He then goes on to note this:

As a side point, note that in the 1968/1957 pandemics elasticities of adjustment were way lower, because you couldn’t switch things to Zoom, Amazon, and so on. So those pandemics were closer to being a “lump sum tax” on human life and thus they were cheaper, and had lower deadweight loss, probably in per capita terms as well. From the framework on welfare economics, that is. The value of human lives was lower then too.

That does not seem like a side point! It seems like it’s going even farther, and saying that *we are worse off because we have the choice to use Zoom and Amazon, and choices are bad*. In this case, in part because the choice [creates a Copenhagen effect](#). It seems clear to me that *in general having the option to overpay for safety is quite bad*. The option to have us collectively do this is even worse. If Tyler is *explicitly agreeing* that if we lacked this choice, we would have all been better off, then what is he actually arguing for?

There is a distinct argument offered that those with better educations tend to be taking more precautions. To some extent this is obviously unfair, since those with higher educations are often those better able to work from home, and those with better financial cushions. They are also those who tend to listen more to authority, and worry more about looking responsible and safe, and they tend to be members of the blue tribe. The idea that they are doing this *because they are smarter or make better decisions* is... one hypothesis, sure. Definitely not the whole story.

Also worth saying that an individual cannot decide to not pay the Ramsey tax here, and accept the Covid risk. That choice is unavailable at this time. The world around you is not going to let you do the old normal things. Thus, one can realize that *given one's remaining options* it makes sense to play it safe, while also preferring the world in which no one played it safe and you did not either.

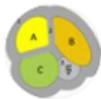
It is also perfectly fine to do any or all of these things, while also preferring the world in which we did proper suppression.

If we knew at the beginning what we know now but were restricted to the choices of “pay more costs” or “pay less costs” a lot of people would choose “pay less costs,” both as a collective and as individuals.

Both for society in general, and for themselves in particular.

All Masks Are Not Created Equal

Robin Hanson asks why we treat all masks as the same, simply saying ‘wear a mask,’ when it is very clear that N-95 and surgical masks are more effective than cloth ones:



Robin Hanson 
@robinhanson

...

There's lotsa talk on requiring mask, but little on which KINDS to require. Is this (a) no expert consensus on which are better, (b) harder to enforce mask type rules, (c) limited tolerance for debating covid policy, or (d) something else?



384 votes · Final results

12:41 PM · Dec 7, 2020 · Twitter Web App

In my model, all three of these explanations play a meaningful part in the debate.

There is *still* a huge lack of consensus on how much better surgical is than cloth. That's what happens when you *ban experiments*. It makes it hard to learn things.

There is clearly very limited tolerance for *debating* Covid policy, and even less for advocating such policy to the public, or even getting them the message.

That's one of two central things going on, in my mind. [You Have About Five Words](#). We can maybe tell people "Social distance, outside, wear masks." We don't want to burn another word here.

There's also what happens when you make that higher ask of people. If you count cloth masks that one can reuse as needed and are in plentiful supply, and that in my experience are relatively comfortable and look relatively normal, you have made a highly reasonable ask, *which a lot of people still treat as unreasonable*. You can even customize your cloth mask as a fashion statement or other signal of affiliation.

If you instead were to say, it needs to be N-95 or surgical (and if you allow both, then it's not only one word of your five that you're burning here), then there would be a lot of 'nope, not reasonable, not doing that.' Then they'll often wear no mask at all.

Or they'll wear a mask, but less often, especially when it actually matters. For the most part, that trade-off isn't worth it.

The public doesn't have the ability to process 'sort of works' or 'this works, but this works better.' They can't handle 'this is somewhat safe, but this other thing would be safer.' You get to say, these things are safe, do those, these other things are unsafe, don't do those, with

clear markings. That's it. Try to be detailed and there will be a huge demand for "Do cloth masks work, or not? Well? Stupid so-called 'experts.'" Or they'll think asking that question owned the libs.

They certainly would not take kindly to being told the most important thing is a cloth mask, then the moment there are enough other masks, changing tunes (after initially lying about masks working at all) and demanding better.

So instead we wisely took the option for a clean, simple message. Wear a mask, period. If you're on the ball enough to upgrade to N-95, great, that's even better.

This is a whole section because this logic extends to all our other strategies and policies, including outside of Covid-19.

In Other Covid News

[CDC recommends this week that Americans wear masks indoors when not inside their own homes](#). Let that sink in. Until last week, *the CDC had not recommended this*. Still, better ludicrously late than never, even for *the single most important thing to be doing*. Thank you, CDC.

Taleb notes: Masks work, [says Danish study people described as not saying that masks work](#).

New York City [vaguely aware sewage data might be helpful](#), but not so aware as to release said data anywhere I'm aware of, or to make any effort to isolate locations or otherwise use it to contain the virus.

[Police were sent to seize the laptop and phone of Rebekah Jones](#). That was her account via Twitter, alternatively [two news report are here and here](#). She was responsible for Florida's Covid-19 database before she was fired. She claims (credibly, as far as I can tell) that she was fired by Florida for refusing to falsify Covid-19 statistics early in the pandemic. She has spent the time since then trying to provide accurate data and expose corruption. For this, the police pointed a gun in her face and the face of her children, and seized her devices, claiming it was about a security breach. Which, one way or another, I suppose it was. In particular, they cite this particular breach, which was accomplished by 'no one bothered to change the login credentials' ([direct source](#)):

In this affidavit, it is detailed how someone accessed a multi-user account in which the username and password was widely known. Having accessed this account the user then sent the following message to all users within that group:

| It's time to speak up before another 17,000 people are dead. You know this is wrong.
| You don't have to be a part of this. Be a hero. Speak out before it's too late."

The IP address of the person who sent that message was tracked back to Rebekah Jones.

If you'd like to get involved, the thread linked to has details on how you can help.

[Yet another cheap, fast, easy and seemingly reliable method of Covid-19 testing](#) that, if it works and was implemented, would end the pandemic. This one uses the microscope on your smartphone and can even give you viral load information. So of course *I barely*

registered this news at all because I assumed there was almost no chance the test would be allowed. FDA delenda est.

There's also this method of testing ([source](#)):



AFP News Agency

@AFP

Vending test.

Free Covid-19 test kits are made available in a vending machine at an MTR metro train station in Hong Kong. Vending machines have been set up in 10 stations.

Those using the kits have to return their specimens to a government clinic for tests and results



2:07am · 7 Dec 2020 · Twitter Web App

2 Replies 143 Retweets 268 Likes

Seems great and I have little doubt we will never do anything of the kind.

A rather complete takedown, as far as I can tell, [of an often-cited paper that concluded lockdowns are highly effective and saved three million lives in Europe](#). It's a lot of words that mostly aren't necessary, because it suffices early in the takedown to point out that the paper *assumes its conclusion*. If you presume that the only things that ever change transmission rates are non-pharmaceutical government interventions, then the alternative to such interventions is that almost everyone quickly gets infected while hospitals are completely overwhelmed, and that's how you get three million lives saved. So, sure. If there are no control systems and people don't adjust behavior except when the government issues an order, at which point they instantly adjust to the new transmission rate (which is what the paper assumes), then of course you are going to conclude that all credit must go to the only thing that ever changes anything. This is the level of scientific literacy and interaction with physical reality that even relatively competent Western governments and policy makers (e.g. European ones) seem to be employing these days.

California is locking down hard.

Here's the [order for Los Angeles](#). [Everyone must stay in their homes](#). Among other details, podcasts are explicitly exempt, religious services must be outdoors but are otherwise all right, and protests are explicitly allowed. [Malls open, playgrounds closed. I think I sense a protest coming on.](#)

It's unclear to me how this is supposed to work out. Do they think people are going to obey rules where they can't meet a single other person while socially distanced, unmasked and outside? Do they not understand that if you push unreasonably far, people stop listening to you? How would one even begin to enforce such a rule? How long do they intend to maintain things this severely? Conditions elsewhere are probably not going to be much better until at least February or so at the earliest. The full turnaround will likely take months beyond that. Is the plan to shut down the city completely that whole time, on top of what has already been done and not worked?

[Ohio High School Athletic Association coronavirus rules: Students can wrestle, but can't shake hands](#). On the one hand this is obviously absurd. If shaking hands is not safe, what exactly do the decision makers think is happening when kids wrestle? On the other hand, shaking hands is a purely symbolic action that you need to stop across the school, whereas wrestling is a compact activity providing value.

[Paper claims that a region in Brazil has 75% overall infection rate](#). About half by the epidemic peak, the rest after. Raw test numbers were not that high, but they feel confident correcting for waning antibody counts, and the implied IFRs are not unreasonable. They have some speculations on how the numbers got this high, but nothing satisfying. My best guess is quite simple, that the region was unusually vulnerable and also mostly young and taking few if any precautions, and it got hit hard enough to far overshoot its immunity threshold.

Winter is Coming

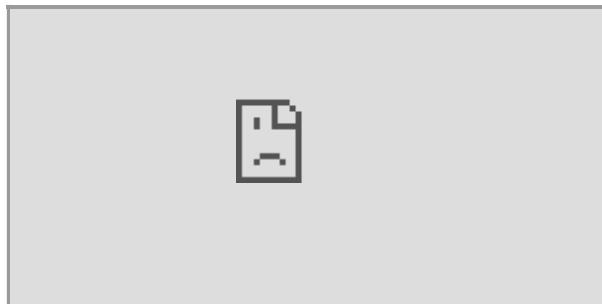
In other [More Dakka](#) news, we want to meet everyone outdoors and go on walks, but [baby it's cold outside](#). Sounds like a problem. In a move I delayed for too long, I've ordered [the warmest coat known to man](#). Should have done it last month, but better late than never. I was going to go with the *third* warmest coat known to man because for my needs that was definitely enough dakka, but in a running theme these days, they underpriced them and now they're all sold out, so instead I am ready, if necessary, for a cozy winter in Winnipeg. Which was the motivation behind my friend Seth Burn finding out about this coat in the first place, which he loves. I'll report back when it arrives. Not cheap. But if it works, totally Worth It.

Until next week.

Secular Solstice 2020

Last night, about 225 people gathered for songs and stories. In the spirit of putting in a ridiculous amount of effort to keep doing normal things in abnormal times, a group of friends built a platform where we could listen together, and sing together, as we do every year. This was the biggest gathering [Bucket Brigade](#) has handled, and with the vaccines being distributed now it may be the largest it ever handles.

If you missed it, Rachel made a recording of most of the service:



([youtube](#))

And here are the recordings Bucket Brigade made:

- [Bitter Wind Blown](#)
- [Chasing Patterns In The Sky](#)
- [Hard Times Come Again No More](#)
- [Sounds of Silence](#)
- [Whose Woods Are These, Blowin' in the Wind](#)
- [Bold Orion](#)
- [Hymn to Breaking Strain, Bitter Wind Blown](#)
- [Five Hundred Million, Brighter Than Today](#)
- [Here Comes the Sun](#)
- [Beautiful Tomorrow](#)
- [Uplift](#)
- [Five Million Years](#)

Since the majority of people were in the final bucket for each song, these recordings have many more voices than you would have been able to hear live. They don't include anyone who failed calibration, clicked "don't let anyone hear me", or muted themselves.

Each song was a pre-recorded backing video, with the audience singing along live ([example backing video](#)).

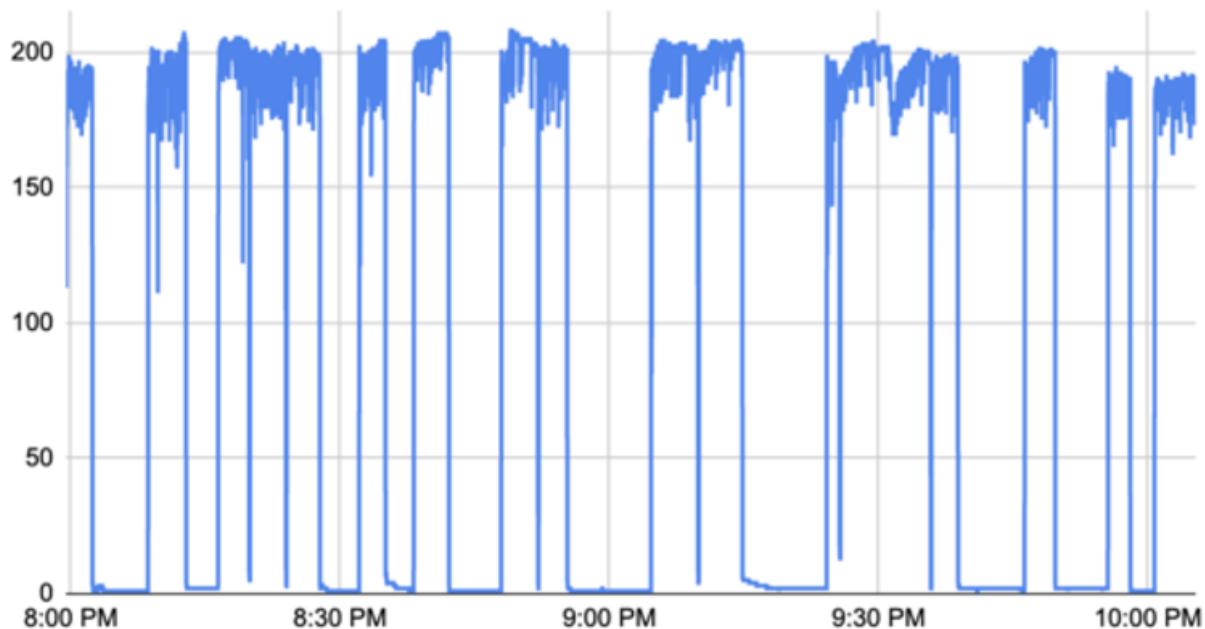
While Glenn and I had been working on Bucket Brigade since [early July](#), we only got to something that could plausibly have handled this crowd in [mid-November](#). Even then, we had a disastrous first live stress test where the server fell over at ~38 people because of differences between simulated testing and real users. We made our automated stress testing more realistic, but last night Bucket Brigade was handling 10 times more real users than it had before.

I ran sound during the event, adjusting all of the different volumes. (I want to give a special thanks to Elizabeth: over the past month she has been my [hands](#) when working

on bucket brigade, and the mixing console would not have been possible without her.) What took most of my time was listening to people who had chosen to sing early in the brigade and making sure they were singing accurately enough that they weren't going to confuse the people who came later. In a normal group people are able to listen to each other and self correct, but in a chain like this the first few users need to be right together or else everyone later will not know who to match.

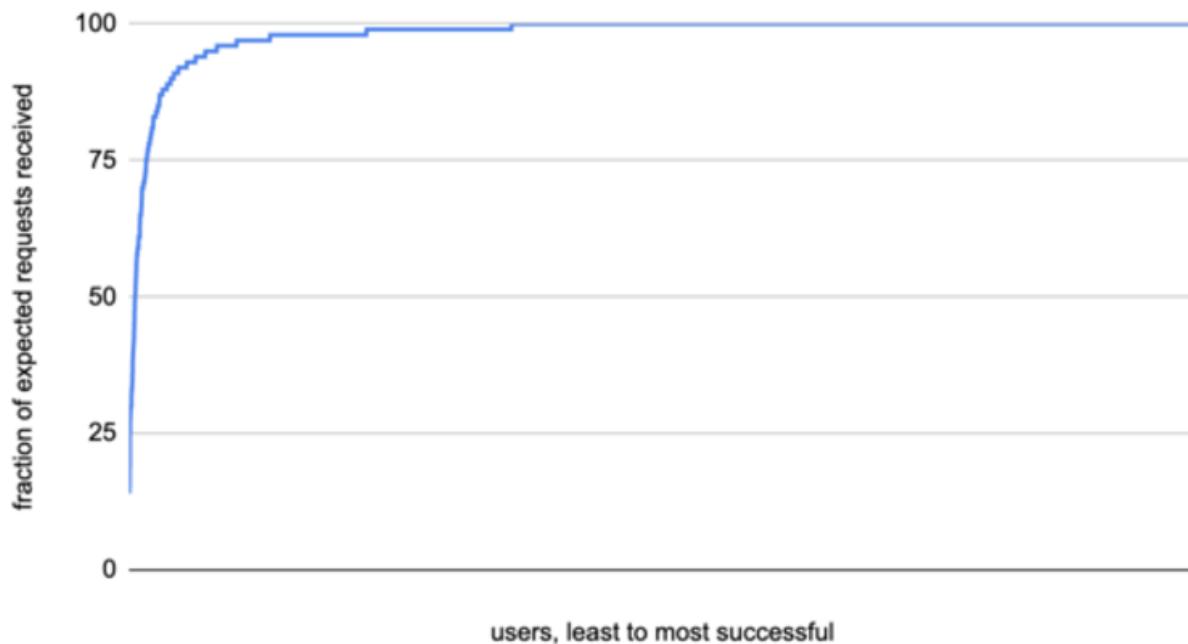
Looking at the logs from the event, we can see that during the songs we had a little over 200 people connected the Bucket Brigade at once:

Solstice 2020: Number of Singers



The biggest thing that I was worried about was users having choppy audio. If the network can't keep up with audio at the rate we are sending it, if the computer is not fast enough to handle encoding/decoding opus audio in JS, or if almost anything else goes wrong, the user will miss a chunk of audio. I analyzed the logs to figure out how often this happened, and while it was happening to some people, most users in most songs did not have any issues:

Solstice 2020: Success rates by user



Two major limitations of this data are that it will not include users who weren't able to get it to work ever, perhaps because they have did not have a supported browser or device, and that users who were having a bad time were probably more likely to give up. I would like to figure out more about why it worked and didn't work for people, and whether there is anything we can do about it.

If you would like to try using Bucket Brigade for singing over the internet, the code is [open source](#) and there is a publicly available [instance](#) anyone can use. The stand-alone UI is not anywhere near as nice as what Daniel and Taymon built for Solstice, but that's something I'd like to fix over the next few weeks now that I've seen how nice it can be.

Comment via: [facebook](#)

In Addition to Ragebait and Doomscrolling

(Sorry for the coy title--I want to give the reader a chance to guess what the addition is.)

One day I opened up the front page of reddit. I was not signed in and I was using my browser's incognito mode.

The following list composed about 25% of what I saw as I scrolled. See if you notice any themes. (As hinted by the title, I think there is something other than outrage here.)

r/MurderedByWords

r/PublicFreakout

r/insanepeoplefacebook

r/JusticeServed

r/nottheonion

r/facepalm

r/mildlyinfuriating

r/Cringetopia

r/TikTokCringe

r/LeopardsAteMyFace

r/FuckYouKaren

r/iamverybadass

r/IdiotsInCars

r/cringe

(At least another 25% was made up of r/news, r/worldnews, r/politics, r/PoliticalHumor, and so on.)

Like many people, I have spent a lot of time thinking about the [psychotoxic](#) effects of concentrated outrage, political polarization, [doomscrolling](#), misinformation, and filter bubbles. So I was a little surprised by my own interpretation of the above list:

I submit that the most salient theme is *contempt*.

Here's a sentence that has been at the back of my mind since I came across it:

Scandal is great entertainment because it allows people to feel contempt, a moral emotion that gives feelings of moral superiority while asking nothing in return.

-- Jonathan Haidt, *The Happiness Hypothesis*

Let me first admit that contemptuously bonding over the misbehavior of others probably can have real benefits. But I claim that in the case of the reddit front page, these benefits are clearly outweighed by the costs to one's personality (not to mention epistemics).

So, Haidt says contempt feels good, reddit appears to be a prime example, and I'm now asserting that it's psychotoxic (and possibly addictive, at least when taken via ~~intravenous drip~~ bottomless scrolling). Presuming all of that is correct...is it actionable? I think so.

If you're ambitious, you could [quit social media for a month](#) and pay attention to how your thoughts and attitudes change.

More coordinately, perhaps a social stigma can develop around this kind of overindulgence, similar to the increasing stigmas toward ragebait and doomscrolling.

But at the very least, you can simply *notice* that something you're reading is triggering contempt, as opposed to outrage or doomfeelz. I think this awareness by itself restores a decent chunk of mental autonomy. Personally, I like to also take the proactive step of rehearsing questions like, "why did they end up so stupid/scandalous/cringeworthy?" and "what led *me* to avoid such faults so well?" I find that the answer (whatever it is) often feels strangely liberating--it diminishes the tasty allure of the contempt, and makes it easier to refocus my attention on something better.

EDIT: Thanks Daniel Kokotajlo for offering the term *scornporn* in the comments!

Notes on notes on virtues

In the [Notes on Virtues sequence](#), I've been sharing my research into a variety of virtues. In this introductory post, I share my thoughts about why I've been on this case and what I'm trying to accomplish.

A one-paragraph summary: I think that becoming more skillful and well-rounded in the practice of the virtues is key to being a better, more satisfied, and more effective person. However, childhood training in the virtues is scattershot and haphazard, and remedial training (or self-help) as adults is also spotty. I'm trying to help fix this by assembling information about the various virtues, with a focus on ways to improve them in ourselves.

Why I think this is important

"The characteristic feature of all ethics is to consider human life as a game that can be won or lost and to teach man the means of winning." —Simone de Beauvoir^[1]

"The branch of philosophy on which we are at present engaged differs from the others in not being a subject of merely intellectual interest—I mean we are not concerned to know what goodness essentially *is*, but how we are to become good people, for this alone gives the study its practical value." —Aristotle^[2]

Life is complex. We are constantly confronted by a variety of challenges. To address those challenges well, we need to have learned a variety of basic life skills such that they are second-nature to us. "The virtues" are a set of such skills that apply to challenges common to typical human lives.

If you have a better command of the virtues, this helps you thrive as an individual and also improves your effect on those around you. Society at large benefits from a higher level of competence in the virtues of those in it. But our culture is not all that good at teaching or encouraging the virtues. Some virtues seem so lacking from the public sphere that I wince when I look at it.

Our institutions of formal childhood education are patchy at best in this regard. You'll get your reading, writing, and 'rithmetic, if you're lucky anyway, but will you get resourcefulness, resilience, restraint, responsibility, rectification, or reputability? Other institutions (scouting, religion, etc.) pick up some of the slack, but not nearly enough. Parents have little guidance on how to convey virtue education to their children effectively, and also have their own blind spots from their own spotty educations. There have been some gestures toward formal "character education" of children, which is probably a good sign. But my guess is that children are going to learn most from the example of their elders: if we don't value virtues enough to pursue them in our own lives, that will make more of an impression on the up-and-coming generation than any "do as we say, not as we do" education will.

A virtue gym?

For a few specific virtues or skills, there are adult education / training / exercise programs. If you want to be more fit, you can join a gym. If you want to be a better public speaker, you can join Toastmasters. If you want to sober up, you can attend Alcoholics Anonymous. But for most virtues, there's nothing like this, and that's a shame.

Two misconceptions that sometimes cause people to give up too early on developing virtues are these: 1) that virtues are talents that some people have and other people don't as a matter of predisposition, genetics, the grace of God, or what have you ("I'm just not a very influential / graceful / original person"), and 2) that having a virtue is not a matter of developing a habit but of having an opinion (e.g. I agree that creativity is good, and I try to respect the virtue of creativity that way, rather than by creating). It's more accurate to think of a virtue as a skill like any other. Like juggling, it might be hard at first, it might come easier to some people than others, but almost anyone can learn to do it if they're just willing to put in the persistent practice.

We are creatures of habit: We create ourselves by what we practice. If we adopt habits without giving them much consideration, we risk becoming what we never intended to be. If instead we deliberate carefully about what habits we want to cultivate, and then actually put in the work, we can become the sculptors of our own characters.

What if there were some institution like a "virtue gymnasium" through which you could work on virtues alongside others, learning at your own pace, and building a library of wisdom about how to go about it most productively? What if there were something like Toastmasters, or Alcoholics Anonymous, or the YMCA but for *all* of the virtues people need?

Ben Franklin's experiment

One day Benjamin Franklin "conceiv'd the bold and arduous project of arriving at moral perfection."^[3] He explains in his autobiography, "as I knew, or thought I knew, what was right and wrong, I did not see why I might not always do the one and avoid the other."

He quickly found that he had underestimated the task. "While my care was employ'd in guarding against one fault, I was often surprised by another; habit took the advantage of inattention; inclination was sometimes too strong for reason. I concluded, at length, that the mere speculative conviction that it was our interest to be completely virtuous, was not sufficient to prevent our slipping; and that the contrary habits must be broken, and good ones acquired and established, before we can have any dependence on a steady, uniform rectitude of conduct."

So he decided to be more methodical. He reviewed various lists of virtues in the literature he was familiar with, and then created his own list of a dozen virtues that he thought were particularly important. With the intention of making each of these virtues habitual, he struck on the idea of tackling them one-at-a-time, starting with ones he thought would help him more easily acquire the others. (Virtues have a way of building on each other. Some virtues, for example persistence, or curiosity, or honor, can make other virtues easier to acquire.)

He decided to do a daily accounting of each virtue he was practicing. He created a notebook with a table for each week. The table had one column for each day of the week, and one row for each of his virtues. Each time he failed to fulfill a particular

virtue on a certain day, he marked the table cell for that virtue/day with “a little black spot” (or more than one if he screwed up multiple times). The plan was that when he achieved a week in which he successfully kept the row for Temperance blank, he would move on to concentrating on Silence (attending to Temperance as well). When he managed to keep both of those rows clear for a week, he would move on to Order, and so on.

“I was surpris’d to find myself so much fuller of faults than I had imagined; but I had the satisfaction of seeing them diminish.” He carried his notebook with him for several years. “[T]ho’ I never arrived at the perfection I had been so ambitious of obtaining, but fell far short of it, yet I was, by the endeavour, a better and a happier man than I otherwise should have been...”

He hoped at one point to write a book, *The Art of Virtue*, which “would have shown the means and manner of obtaining virtue, which would have distinguished it from the mere exhortation to be good that does not instruct and indicate the means.”

He toyed with the idea of a political party that would not advocate for the benefit of a certain segment of the people, but for the good of the country and of mankind in general: the “United Party for Virtue.” This morphed into an idea for a fraternity: the “Society of the Free and Easy.” His plan was to initiate members by putting them through the same practice he had undergone with his notebook of weeks and virtues. He explained the name of the society this way:

The Society of the Free and Easy: free, as being, by the general practice and habit of the virtues, free from the dominion of vice; and particularly by the practice of industry and frugality, free from debt, which exposes a man to confinement, and a species of slavery to his creditors.

He got as far as getting two young men to sign up and begin the work, but then he got distracted with other things and abandoned it. “[T]ho’ I am still of opinion that it was a practicable scheme, and might have been very useful...”

The Society of the Free & Easy

Last year I set about trying to pull together something like Franklin’s Society of the Free and Easy (and borrowing his name). I worked with a group of friends and acquaintances to come up with what I think is [a pretty good framework for working on virtues in a peer-supported way](#). In a nutshell, the process is pretty simple:

1. Find a partner or form a small team.
2. Each of you choose a virtue to work on.
3. Take a close look at your virtue, and at any obstacles you feel when you try to practice it.
4. Work with your partner(s) to come up with exercises in which you will frequently, deliberately practice that virtue in ways that challenge your current level of fluency.
5. Check in with your partner(s) regularly to encourage each other and to keep each other accountable, and adjust your curriculum as you learn more about what works and what challenges you face.
6. When you feel you have integrated the virtue adequately into your character, start the process again with a new virtue.

Alas, after some initial promise, the group began to dwindle, and then the pandemic disrupted everything, and now as far as I know there are only two of us still working through the program on the regular. But in the course of researching, we dug up a lot of information about virtues in general and about particular virtues, and that's forming the basis for the posts I'm sharing here.

Notes on virtues

What I hope to do with these notes on virtues is to collect ideas that will be useful to people who want to improve in a certain virtue. This may include concrete advice about strengthening that virtue itself, and also some discussion about other virtues that are related in some way: maybe they're prerequisites, or harmonize in some way, or maybe there's some tension between them. I sometimes find it challenging to define the virtue precisely, or to distinguish it from another virtue—and sometimes the term for the virtue gets overloaded with a variety of meanings in common use—so I include discussion of those nuances too.

I'm aiming to be inclusive of a variety of useful perspectives, and of a variety of cultures, rather than to be definitive or dogmatic. It's a fuzzy subject matter to begin with. I'm feeling my way about, leaning on existing guides when I can find them (though I tend to find a lot more examples of people praising or advocating certain virtues than of people explaining them or giving practical advice on how to go about improving in them).

I take some inspiration from Aristotle, who, when he examined a set of virtues in his *Nicomachean Ethics*, started with virtue-concepts as already found in common language and folktale, rather than starting from a theoretical foundation and building ideal virtues from there. When it comes to dividing up a complex subject matter into manageable and coherent chunks, previous generations have already done a lot of the work for us and handed that down to us in the language and tropes we use. That we have found a word or trope useful is a good clue that there's some reasonably-helpful and worth-noticing regularity at the base of it. While this sort of understanding shouldn't be confused with the gospel truth of how reality is constituted, it seems wise to glean as much as we can from it before trying to systematize more deliberately.

One of the things I did was to investigate several virtue-based traditions (the Greek [cardinal virtues](#), the [traditional Christian virtues](#), the virtues of [Bushido](#), [Confucian virtues](#), the virtues of [Scouting](#), the [West Point virtues](#)), the virtues favored by some particular philosophers ([Aristotle](#), [Cicero](#), [Ben Franklin](#), [Ayn Rand](#), [Henry David Thoreau](#), [Shannon Vallor](#), [the Cynic philosophers](#), the developers of [care ethics](#), [William De Witt Hyde](#), [Eliezer Yudkowsky](#)), and the virtues identified as "[character strengths](#)" by psychologists operating in the positive psychology paradigm. This isn't comprehensive by any means, but it was revealing.

For one thing, there was a lot less consensus than I expected about which virtues are the important ones. This is somewhat complicated by problems of terminology. For example, what one philosopher will call self-control, another will call continence, another restraint, another discipline. Or, while Paul says that the greatest virtue is [love](#),^[4] he defines "love" in such a way^[5] that it incorporates patience, [kindness](#), *mudita*, modesty, humility, respect, good temper, forgiveness, righteousness, care, trust, hope, and perseverance. Or different cultures will partition virtue-space differently: *sisu* is only kind-of like perseverance; *mudita* is only kind-of like sympathy;

nying je is only kind-of like compassion. This can be challenging for works in translation, where the translator has chosen the closest equivalent English word, but a close reading reveals that the author meant something different from what we mean by that word.

I tried to correct for things like these. I consolidated various terms for similar virtues together, and created a spreadsheet where I could note which virtue-clusters had been promoted in which systems or by which philosophers. But of the hundreds of virtues I found, only six of them were on more than half of the lists:

- [courage/bravery](#)
- [honesty/trustworthiness/promise-keeping](#)
- [self-control/continence/restraint/discipline](#)
- [loyalty/fidelity](#)
- [compassion](#)
- [respect for others/thoughtfulness/consideration](#)

If you add those that were on exactly half of the lists, you also get [justice](#), [wisdom/philosophy](#), [sincerity/straightforwardness/earnestness/frankness](#), [industry/effort/enterprise/productiveness](#), [duty/responsibility/purposefulness](#), [piety/reverence](#), and [strength/toughness/vitality/health/fitness](#).

You may have heard that [patience](#) is a virtue, but it didn't make that cut. Neither did [humility](#), hope, perseverance, [courtesy](#), generosity, friendliness, creativity, [caution](#), cleanliness, mercy, forgiveness, wit, originality, calm, warmth, curiosity, hospitality, pride, or gratitude. Some lonely virtues like boldness, imagination, spontaneity, and playfulness appeared on only one list. Other skills that are often popularly admired—like being influential, having emotional intelligence, or being good in bed—weren't on any lists at all.

Some virtues are debatable. Selflessness, pride, altruism? The apostle Paul and Ayn Rand would disagree about what's the virtue and what's the vice. Virtues like [chastity](#), obedience, and patriotism give some of us the willies.

I'm aiming to be inclusive and to eventually give some attention also to these less-prominent and more controversial virtues.

Why am I going to this trouble?

My hope is that, whichever virtue you're hoping to improve, you'll be able to get a head start from the research and write-ups I've done.

I'm also motivated by self-improvement. I've been working to deliberately improve some of these virtues in my life, and I hope to make that an ongoing project, so putting together these virtue-dossiers helps me to lay the groundwork for this.

If we manage to reboot the Society of the Free & Easy in the post-pandemic time, these may help us hit the ground running.

I also have vague ideas about this being a worthwhile political project. I've come to distrust talk of elections and revolutions and institutional reforms. I think the longer, harder, more subtle project of helping people improve is a more reliable path to a better future than trying to impose wise policies on them from on high. If people

become braver, wiser, more just, and more honorable, public policy will follow their lead. If people become more cowardly, foolish, grasping, and disreputable, conniving politicians will lead them by the nose.

I'm sharing these at *LessWrong* in particular because I value the sort of insightful feedback people share here. Since I'm not an expert at any of the virtues I'm writing up, I'm slyly taking advantage of Cunningham's Law [6] to correct my misunderstandings about them.

Note

This is an ongoing research project. Whenever I dig up information about a particular virtue that I think can usefully add to what I've written, I will go back and revise the page associated with that virtue. Or if you leave a comment about something that I've gotten wrong or something important I've neglected to mention, I'll fix that. I think it's more important for these pages to be useful than for them to be a fixed record of what I thought when I first wrote them.

Tentative Sequence Outline

1. [Honesty](#) (truthfulness, veracity, trustworthiness, promise-keeping, probity, reputability)
2. [Courage](#) (bravery)
3. [Self Control](#) (continence, discipline)
4. [Respect for Others](#) (thoughtfulness, consideration)
5. [Piety](#) (reverence, faith, religion, spirituality, transcendence)
6. [Loyalty](#) (faithfulness, fidelity)
7. [Compassion](#)
8. [Wisdom](#) (philosophy)
9. [Temperance](#)
10. [Fitness](#) (health, fettle, hardiness, vitality, toughness)
11. [Sincerity](#) &c (straightforwardness, frankness, sincerity, earnestness, candor, *parrhēsia*)
12. [Justice](#) as a virtue (*epieikeia*)
13. [Industriousness](#) (assiduity, effort, enterprise, industry, productivity)
14. [Duty](#) (responsibility, purposefulness)
15. [Social Responsibility](#) (civility, community, citizenship, patriotism)
16. [Prudence](#) (practical wisdom, decision theory)
17. [Know-how](#) (practical knowledge, craft)
18. [Honor](#) (nobility)
19. [Moderation, Balance, & Harmony](#)
20. [Patience & Forbearance](#) (*kshanti*)
21. [Care](#)
22. [Attention](#) (awareness, discernment, mindfulness, presence, focus, concentration, alertness, observation, presence)
23. [Amiability](#) (friendliness, geniality, agreeableness, conviviality, affability, niceness, affection, warmth)
24. [Rationality](#) (reason, love of truth, clear-headedness, deduction/induction)
25. [Simplicity](#)
26. [Forgiveness](#) (mercy, clemency, *epieikeia*, placability, indulgence, leniency)
27. [Integrity](#) (unity, authenticity)

28. [Humility](#)
29. [Optimism, Hope, & Trust](#)
30. [Good Temper](#)
31. [Fairness](#) (impartiality)
32. [Endurance](#) (fortitude, hardness, toughness, stamina)
33. [Benevolence](#) (goodness, goodwill, *metta*)
34. [Ambition](#) (drive, aspiration, loftiness, initiative)
35. [Perseverance](#) (persistence, resilience, grit, fortitude, tenacity, *sisu*)
36. [Kindness](#) (including kindness to animals)
37. [Empathy](#) & Sympathy (*mudita*, pity)
38. [Frugality](#) (thrift, economy)
39. [Dignity](#) (self-respect, rectitude)
40. [Courtesy](#) (manners, politeness, consideration)
41. [Chastity](#)
42. [Love](#)
43. [Resolve](#) (determination, commitment, dedication, devotion) / Steadfastness (firmness, decisiveness, consistency)
44. [Caution](#) (carefulness, preparedness, foresight, deliberation, heedfulness, vigilance)
45. Cleanliness (orderliness, hygiene)
46. Gentleness (tenderness, mildness)
47. Propriety (decorum, etiquette, refinement, gravitas, sobriety, sportsmanship, chivalry)
48. Creativity (art, imagination, inventiveness, problem-solving)
49. Modesty
50. Altruism (philanthropy, generosity, charity, magnificence, magnanimity)
51. Awe (elevation, wonder, appreciation of beauty, sensitivity)
52. [Judgment & Righteous Anger](#)
53. [Gratitude](#) (appreciation, reciprocity)
54. Valor (heroism, defense-of-others, glory, chivalry)
55. Serenity (calm, equanimity, tranquility, composure, contentment, peacefulness, consolation)
56. [A Sense of Shame](#)
57. Purity (innocence)
58. Philomathy (studiousness, scholarship, seeking good advice)
59. Self-reliance (independent thinking, originality, independence, liberty)
60. Wit (humor, playfulness)
61. Helpfulness (service)
62. Filial Piety (and respect for elders)
63. Joy (cheer)
64. Flexibility (adaptability, agility, resourcefulness)
65. Cooperation (teamwork, coordination)
66. Taste (fashion sense, beauty)
67. Quiet (silence, stillness, listening)
68. Rhetoric (persuasion, conversational competence, expressiveness, influence)
69. Pride (self-worth, self-esteem)
70. Poise (confidence, grace, unflappability, authority, gravitas, refinement)
71. Perspective (prioritization)
72. Openness (candor)
73. Obedience (respect for authority, submission)
74. Mentoring (counsel, nurturance)
75. Intelligence (ingenuity, mental quickness)
76. Hospitality
77. Curiosity (exploration, inquisitiveness)

78. Connectedness (social intelligence)
79. Concern
80. Zest (exuberance, vigor, enthusiasm, wholeheartedness, zeal, *virya*)
81. Liberality (tolerance, understanding)
82. Leadership
83. Graciousness (accepting fault, accepting help)
84. Tact (discretion)
85. ...more as I discover them...

What am I missing?

1. ^

Simone de Beauvoir, *The Ethics of Ambiguity* (1947)

2. ^

Aristotle, *Nicomachean Ethics*, book II, chapter 2

3. ^

Benjamin Franklin, *Autobiography*, chapter 9 ("Plan for Attaining Moral Perfection")

4. ^

[1 Corinthians 13:13](#) "And now abide faith, hope, love, these three; but the greatest of these *is* love."

5. ^

[1 Corinthians 13:4-7](#) "Love suffers long *and* is kind; love does not envy; love does not parade itself, is not puffed up; does not behave rudely, does not seek its own, is not provoked, thinks no evil; does not rejoice in iniquity, but rejoices in the truth; bears all things, believes all things, hopes all things, endures all things."

6. ^

"The best way to get the right answer on the internet is not to ask a question; it's to post the wrong answer." ([Named after](#), but not formulated by, [Ward Cunningham](#).)

How I Write

Cross-posted, as always, from [Putanumonit](#).

Five years ago I started Putanumonit not expecting that anyone would read it. That year I had a chance to meet two of my favorite bloggers, Scott Alexander and Tim Urban, and discovered that they had both honed their writing skills on [lesser known blogs](#) for several years before relaunching as SSC and [WBW](#). My plan at launch was to write for a small audience until 2020, then scrap Putanumonit and launch a new blog with everything I will have learned about writing.

That five-year plan lasted all of three weeks when Scott [linked](#) to my post about [why China's soccer team sucks](#) and brought in thousands of readers in one day. So now I'm stuck with Putanumonit.

Last week my friend Lynette called me to talk about blogging, and I realized that in those five years I did figure out a few things about writing. These aren't technical tips of sentence construction and argument layout but relate more broadly to writing as a consistent activity.

So, here's how five years of Putanumonit happened.

I write confidently and trust the readers to discount appropriately

It's good to be clear about the general epistemic status of what I write and be [meta-honest](#) about writing my sincere beliefs (or [sincere drug trips](#)). But I don't think this requires hedging every statement, reminding the reader of my inevitable subjectivity, or bloating the text with endless caveats.

I write under the assumption that I'm just one voice among many each of my readers is exposed to, more "rando on the internet" than "infallible guru". I can imagine a silly reader using [each and every piece](#) of advice I've ever written to fuck up their lives, but dumb people can't be helped either way. I trust smart readers to sanity-check my models, discount my apparent confidence, double-check factual claims, and [reverse all advice](#) as needed.

So this post is not universal writing advice, it's how someone like *me* learned to write a blog like *this*. But the greatest compliment I could receive is hearing that Putanumonit inspired someone else like me to start another blog like this; hopefully this post will help.

I write in one sitting, but sit often

It's almost universal advice to *write every day*, to keep an immutable daily routine of X minutes or Y words every morning or afternoon. I universally ignore this advice.

In part, it's because I don't write full-time. If I had to submit a daily column or a book draft I suspect that a daily routine would be unavoidable. But in part it's because I've noticed that 80% of what I've written comes in multi-hour chunks. I've gone as long as 6-8 hours with almost no breaks on some of my longer posts, completely in the zone, my mind holding the entire blooming shape of the post and nothing else.

One day in late February I spent the day in the office struggling to get any work done while my mind was churning COVID thoughts. I came home exhausted at 8 pm, took off my shoes,

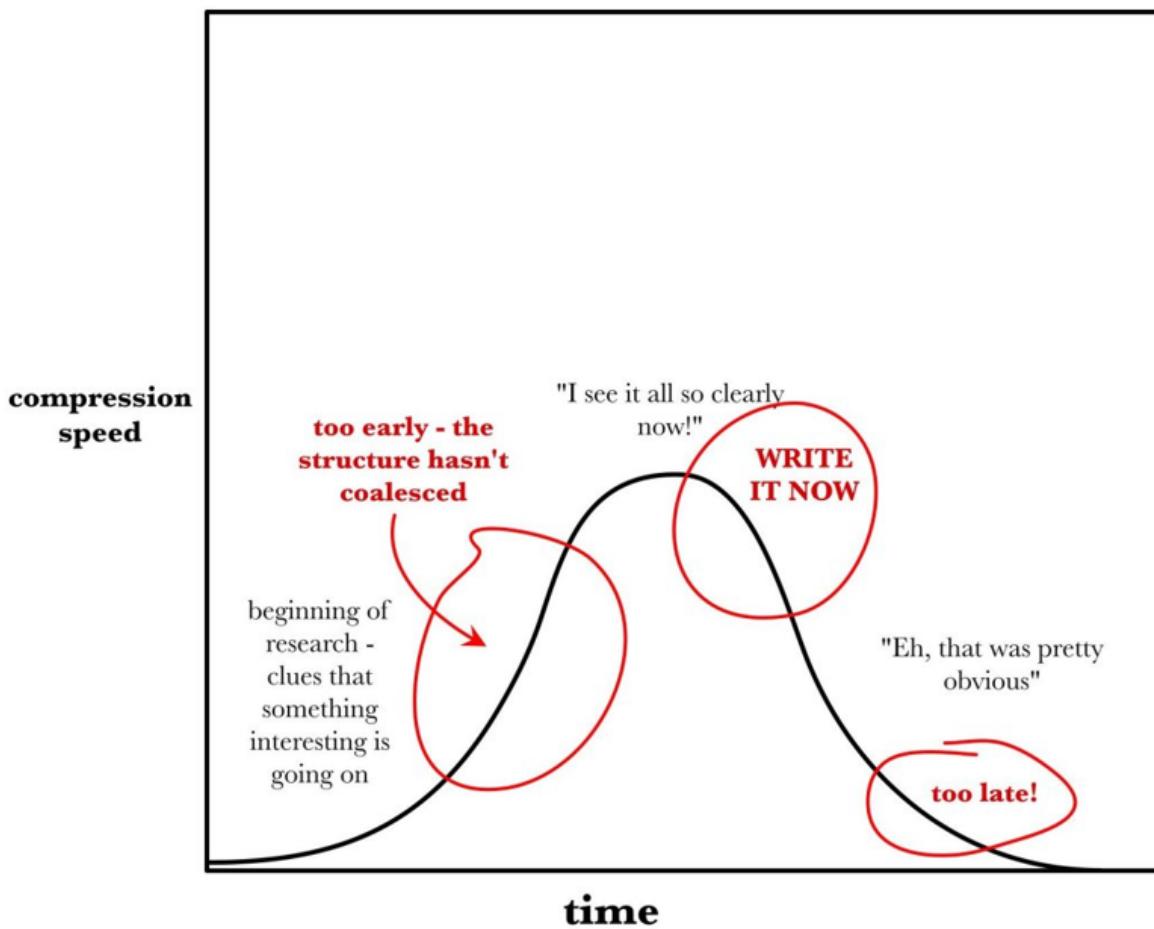
then inexplicably went straight to the computer to start a brand new post with no outline or draft. After four hours I hit submit on what became [my most influential post ever](#).

If you had asked me at 7:55 pm whether I was going to write a post that day, I would have thought it unlikely. But I sat down anyway, and the magic happened. If you plan to write in long spells you can't wait for the perfect day when you wake up full of energy with a clear schedule and a buzzing mind — that happens too rarely if at all.

I sit down to write whenever I feel there's even a small chance I'll get into the writing zone and I have at least one hour free. Some days I end up on Twitter after 10 minutes of futile attempts to conjure a single paragraph and give up. Other days are like this day.

I write before I understand it all

I'm most likely to write for hours at a time when I'm excited about an idea. I get most excited about an idea *while* I'm figuring it out, not before or after. Before, it is too vague and confusing. After I've figured it out it's stale and boring, I often forget why I was excited about it in the first place. It seems too obvious to write about.



This diagram by Sarah Perry perfectly illustrates this. The speed of compression (i.e. figuring out) is highest when I'm 50%-70% of the way to full understanding, and I have to resist the urge to delay writing until I'm at 90%. I do a lot of research and thinking as I'm writing, not just before.

I use Twitter as a first draft

If the time to start writing about a topic is 50% of the way through understanding it, at 20% is when I'd tweet about it to workshop the idea and get some quick feedback. [Threads that resonate](#) I then expand into full posts, and those that don't I often scrap or think about the subject in different ways. I often get comments that inspire the main ideas in an article.

Twitter has different uses than a blog, its own culture and norms. But it's a great aid for the collaborative thinking that turns into writing. As an added bonus, it trains one to write concisely when required by altering sentence structure and getting rid of fluff.

Writer's block is fake

I think I believed in writer's block when I was younger. One day I decided that it's fake and it hasn't materialized since. You can always just transcribe your internal monologue as it goes on in your head. Yes, it's probably going to be bad, but now you're dealing with the task of editing rather than with "writer's block".

I think what people call "writer's block" is usually an aversion to some *specific* thing that they're coerced into writing, like a school assignment. If you experience writer's block on a personal blogpost you may be coercing *yourself*, for example by trying to write something *inoffensive* on a politically charged topic you feel strongly about. In this case the block is protecting your readers from drivel and you should start writing only when you grow some *cojones*.

I submitted to edited outlets

The pieces I wrote for [Quillette](#) and [Ribbonfarm](#) (among others) were hugely beneficial for my blogging. First, because I got paid for them, and it's an important confidence boost early on to know that my writing was at least good enough to be worth \$100-\$200 to someone.

More importantly, both were written in collaboration with an editor. A good editor is like a coach, they can point out bad habits and offer advice that you then take with you to the rest of your writing career. I recommend to all new writers to try and sell a pitch to an edited magazine, even if writing it takes multiple times the effort of an unedited blog post.

I don't promote my posts on forums I'm not active on

After a post is done I leave it alone for 24 hours to let my loyal subscribers catch any obvious errors or things I should not have written. Then I share it on Twitter, and few days later on LessWrong. That's it. I don't put my own posts up on Reddit, HackerNews, etc.

One reason is that I'm not really active on any other platforms, so it feels unethical to spam them with my stuff. A lot of my readers are active on each one and it's up to them to share a post if they think it's good (which I'm always very grateful for!)

And if no one shares it? That's fine too. I often don't have a feel for which posts will resonate with readers as I write them. Some of my favorite writing is completely unappreciated while semi-throwaway posts get tens of thousands of views. It's a good dynamic if most people discover Putanumonit through my best and most popular posts first, as judged by my own readers. Whether the new readers stick around and go through [the archive](#) or just wander away with a vague recollection that I'm a smart dude, it's all good.

How Lesswrong helped me make \$25K: A rational pricing strategy

3 months ago, I started a company that helps people land jobs in exchange for a cut of their income once they land a job.

Since then, I've closed 5 customers worth about ~\$25K. I strongly attribute my ability to close deals to how I price. I strongly attribute how I price to ideas I've developed reading Lesswrong (especially concepts from Economics like Shut up and Multiply).

Word for word, here's how I walk customers through pricing on sales calls:

- There are 52 weeks in a year.
- Each week makes up about 2% of a year.
- That means the opportunity cost of you not working is 2% of your potential post tax salary per week
- Does that make sense?
- Remind, me, what role are you trying to get?
- Great, you make about \$YK per year (.75% to get post tax) which is \$ZK per year post tax so on a weekly basis you're losing \$XK.
- On a monthly basis, that's about \$XK.
- **Does that make sense?**
- Base case: The average job seeker spends 5 months or about [22 weeks](#) in unemployment & that's during non-covid times.
- I can help you get a job in less than half that time, 2.5 months or the equivalent of 10 weeks. I just did it for my first customer in 6 weeks.
- That's 12 more weeks of you working or a little over 23% of a years' worth of income.
- So to be clear, in this case, working with me would help you save 23% of a years' worth of income or potentially x\$K dollars.
- **Do you understand how I got to that number?**
- Great, so the way I go about this is I split that value of 23% halfway. In other words, I give half the savings to you, and then take the other half for myself, leaving us both with 11.5%.
- **Does that make sense?**
- Naturally, you may be thinking, "what happens if we're not successful?" Here's a few possible scenarios:
 - If you land a role in the 22nd week (which is the average amount a job seeker looks for a job) or any point after, I don't get any cut.
 - If you land a role after the 10 week mark, the percentage cut I get drops incrementally each following week, up until the average case of 22 weeks. For example, if you start a role on the 16th week which is 6 weeks quicker than you would have originally gotten it assuming the average case, then we would be splitting (6 / 52) or 11.4% of the value.
 - As a final point, if you have not landed 10 or more interviews by the 10 week mark, you reserve the right to end our contract. This means you will not pay anything for my service. This decision to end our contract needs to be made between the 10th and 11th week. Otherwise, the assumption is

that we'll continue to work together under the conditions outlined in the clause above.

- The above happens if you have no interview that leads to job offer.
- **Does that make sense?**
- And just to be clear, in either case, I'm not factoring in any of the following:
 - 1) Benefits which can make up to an [additional 30% of value](#) of the salary.
 - 2) Higher compensation that you'll get from the job when you work with me cause you'll have better offers and because I help with negotiation.
 - 3) More fulfilment that you'll get from the job when you work with me cause we're gonna get a bunch of interviews.
 - 4) I'm betting on you as a person to put in the work with me to get the job. If you don't, I get screwed not you. I lose my money & time. So it's really important to me that you're serious about this before we go forward.
 - 5) I don't charge anything additional for the tools we use which cost hundreds a month.
 - 6) You're learning how to get a new job in 2 months. Ever been in a poor work environment? Imagine being able to leave that quick. Now you'll know how to.
 - 7) Don't capture any bonuses or raises you get during the first year.
 - 8) Don't capture any equity or other forms of compensation you get in a role.

At this point, I'll usually walk them through a few other caveats (unmentioned here for simplicity & brevity's sake) then close up the call.

Feel free to critique this pricing model or share ways rationality has helped you model decisions and/or bets.

Covid 12/17: The First Dose

The Pfizer vaccine is being deployed to health care workers and long-term care facility residents. The Moderna vaccine is close behind, with [the full FDA report already out](#). There were some small extra delays thrown in for good measure, on the order of a few days, that doubtless killed a few people but shouldn't delay the overall path of events. We are now in the vaccination stage of the pandemic. If the trial results are to be believed in detail, by the end of next week those getting the first dose will largely already be immune, and the population immunity effects can begin to compound and help turn the tide. They will start out small, but soon start growing faster, and every little bit helps.

There even seems to be a good chance that overall new infections would have already peaked if not for worries about Christmas and New Year's coming up soon. Positive test percentages seem to be starting to even out or slowly decline as we get clear of Thanksgiving, and death rates are not rising as much as they would be rising in the scenario where data snags and testing issues were the only reason we didn't notice things getting even worse than they are. If anything, death rates suggest a better picture than the positive test counts.

In other great news, [Over the counter \\$30 Covid-19 test approved by FDA. It's official.](#) Woo-hoo! Three million over the counter tests this month, then they ramp up production. We could have done this a long time ago and solved the whole pandemic, and instead it's going to be a drop in the bucket that shows up late to the party, but every little bit helps.

Alas, as always, *none of that changes the short term situation much*. There's lots of Covid-19 out there, and if anything there are even more reasons to play it safe right now. Large Christmas gatherings or New Year's parties are a profoundly bad idea.

In many ways, it seemed like this week was mostly 'cut to one week later' with few if any surprises, except for the approval of the first at-home Covid-19 test.

Let's do the numbers.

As a bonus, [here is someone else running their version of the numbers on a state level](#). He does this monthly, and they're useful graphs with solid commentary. It's a generally good newsletter.

(Technical note: Brief uncredited screenshots whose source isn't obvious are from [CNBC live update page](#) or [CNN live update page](#). I'm experimenting with this as a quick way to get sourced info across, let me know if you think it's a good idea to keep doing this.)

The Numbers

Predictions

We got a peak. Test count and deaths were close, the positive rate went small down instead of small up.

Last week's prediction: I predicted a 14.3% positive rate on 11 million tests, and an average of 2,550 deaths per day.

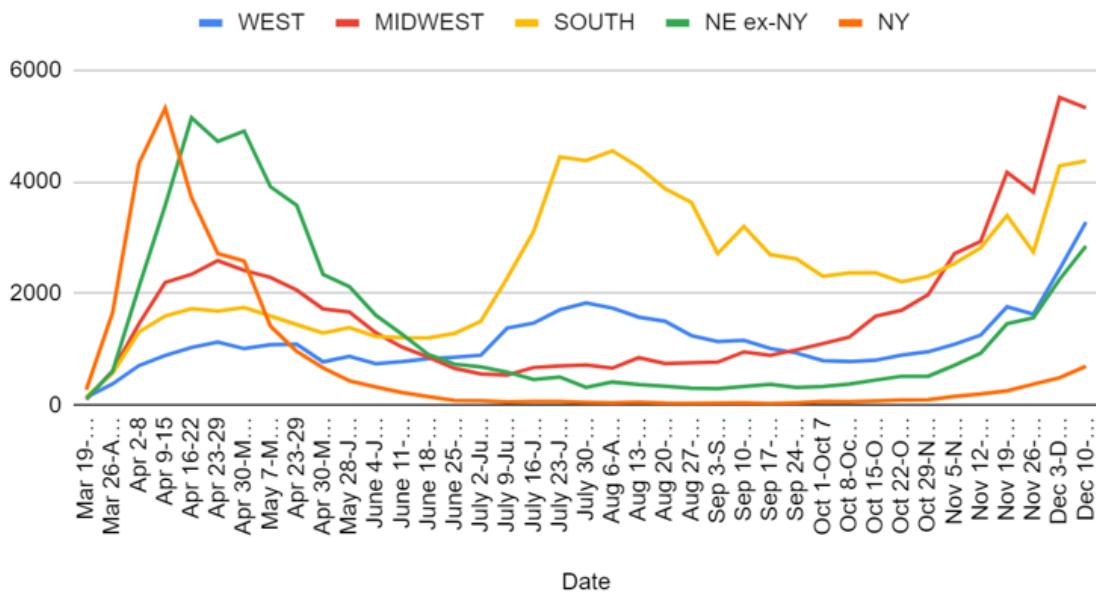
Results: We got a 13.5% positive rate on 10.95 million tests, and an average of 2,617 deaths per day.

Prediction: 13.1% positive rate on 11.5 million tests, and an average of 2,850 deaths per day.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 15-Oct 21	804	1591	2370	523
Oct 22-Oct 28	895	1701	2208	612
Oct 29-Nov 4	956	1977	2309	613
Nov 5-Nov 11	1089	2712	2535	870
Nov 12-Nov 18	1255	2934	2818	1127
Nov 19-Nov 25	1761	4169	3396	1714
Nov 26-Dec 2	1628	3814	2742	1939
Dec 3-Dec 9	2437	5508	4286	2744
Dec 10-Dec 16	3278	5324	4376	3541

Deaths by Region



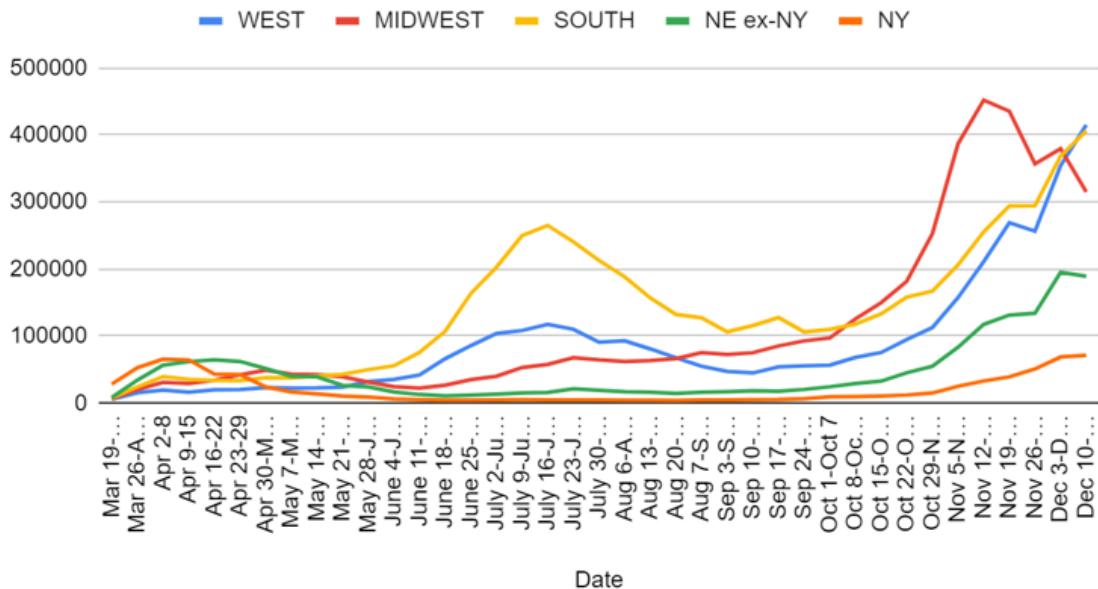
I interpret this as a backlog of deaths from previous weeks showing up in the Dec 3-Dec 9 reporting in the Midwest and South, with the numbers we see this week being about where the real death rates are once again, and things still getting worse everywhere given the lag. The Northeast and West numbers came in a little higher than I expected, which cancelled out, but nothing here seems too surprising. Peaks outside of the Midwest are likely still at least a few weeks away, but I do not expect things to get that much worse than they are now before they turn around.

Positive Tests

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Oct 15-Oct 21	75571	149851	133238	43325
Oct 22-Oct 28	94983	181881	158123	57420
Oct 29-Nov 4	112684	252917	167098	70166

Nov 5-Nov 11	157495	387071	206380	108581
Nov 12-Nov 18	211222	452265	255637	150724
Nov 19-Nov 25	269230	435688	294230	170595
Nov 26-Dec 2	256629	357102	294734	185087
Dec 3-Dec 9	354397	379823	368596	263886
Dec 10-Dec 16	415220	315304	406353	260863

Positive Tests by Region

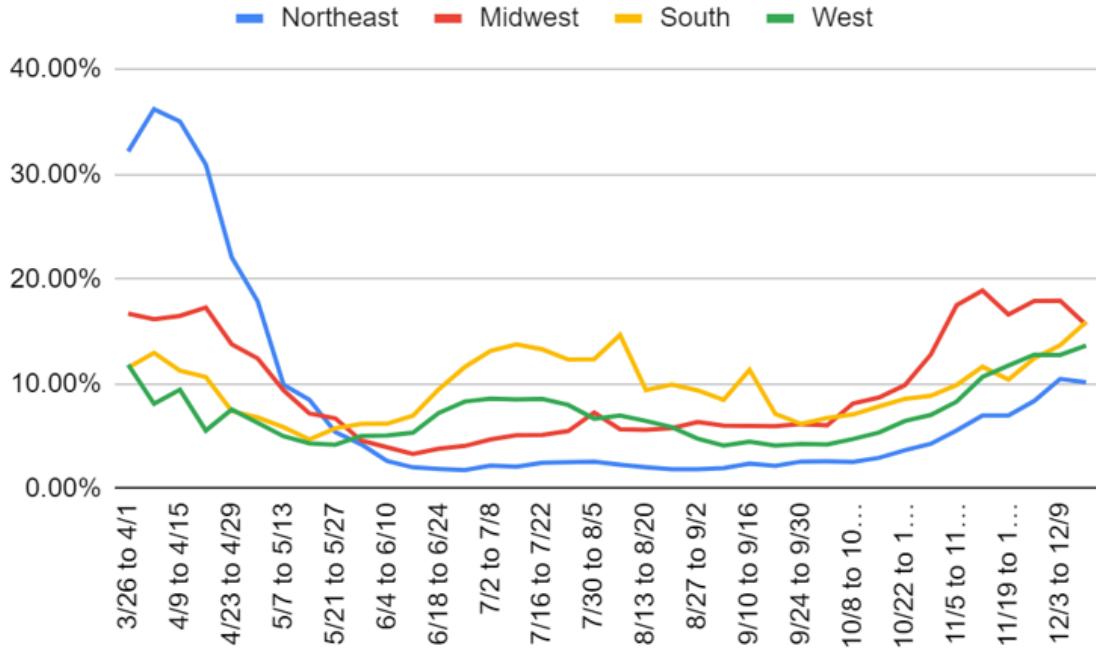


Seems clear the Northeast got ahead of itself last week and that has now been smoothed out but is still getting worse, the Midwest has peaked, and the West and South are in trouble. As usual, I rely more on the test percentages

Test Counts

Positive Test Percentages

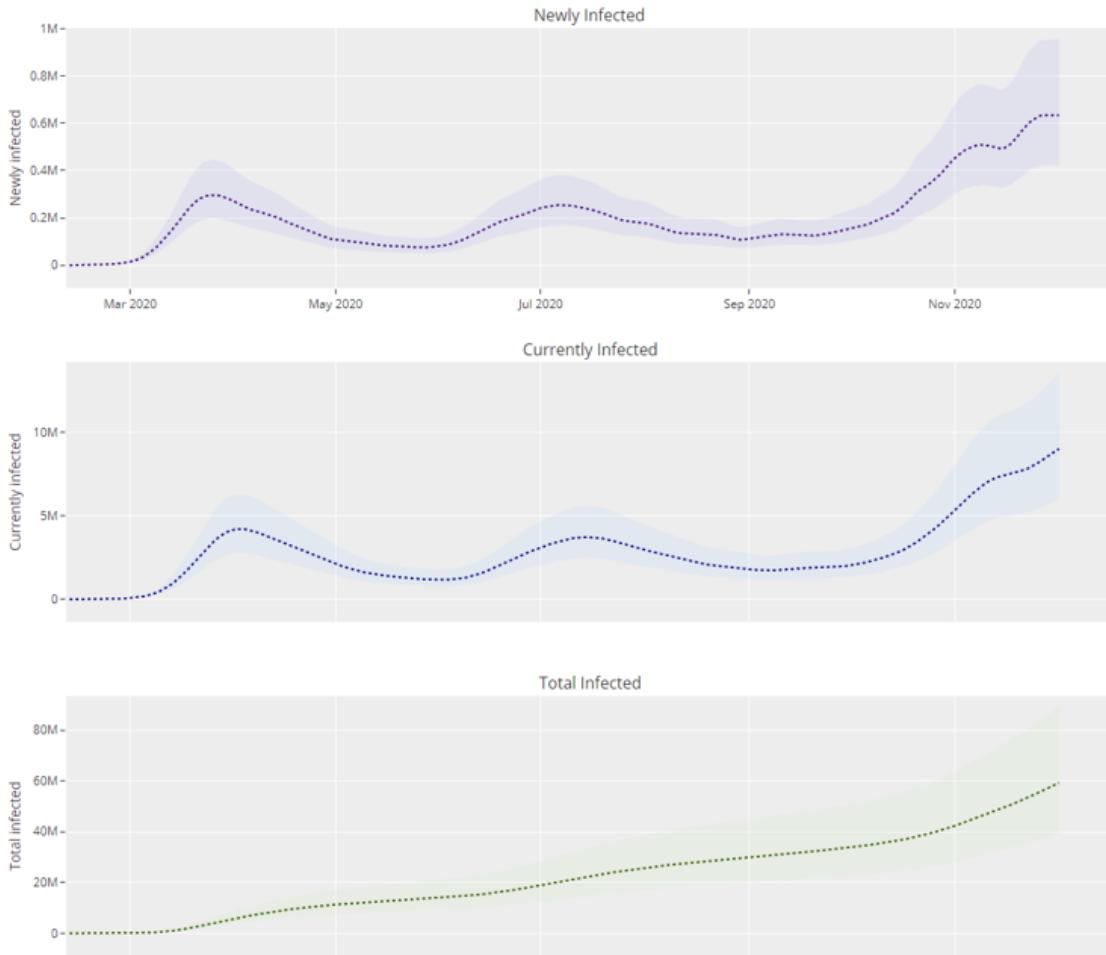
	Northeast	Midwest	South	West
10/15 to 10/22	2.95%	8.70%	7.85%	5.36%
10/22 to 10/28	3.68%	9.87%	8.58%	6.46%
10/29 to 11/4	4.28%	12.79%	8.86%	7.04%
11/5 to 11/11	5.56%	17.51%	9.89%	8.31%
11/12 to 11/18	6.99%	18.90%	11.64%	10.66%
11/19 to 11/25	7.00%	16.62%	10.41%	11.75%
11/26 to 12/2	8.38%	17.90%	12.45%	12.79%
12/3 to 12/9	10.47%	17.94%	13.70%	12.76%
12/10 to 12/16	10.15%	15.63%	15.91%	13.65%



I'm ready to believe the Midwest is headed in the right direction. I am not ready to believe the Northeast is following suit quite yet. That seems much more likely to be a blip and nothing more, although the snowstorm outside my window at the moment might slow things down a bit - being forced inside is not great but no one leaving the house at all is actively useful. Thus I think the baseline scenario is that the Midwest continues to drop. The South is the opposite case, with this number coming in a bit 'too high' so I expect it to level off a bit next week despite the true number of infections likely still rising a bit.

It is worth noting that California is now at 11.9% positive tests. The attempts to lock things down aren't working, presumably because they are not locking things down due to being sick of it all, and also because they're going after the wrong things too harshly with no end date in sight, and forcing people into all-or-nothing choices. That's also despite weather that shouldn't create as big a winter crisis as other areas. The counter-argument is that things might have been about to go the way of Arizona (32%), Colorado (28%), Idaho (47%) or Nevada (40%), but Oregon (5%) and Washington (8%) seem like better parallels and thus a strong counter-argument to that.

Covid Machine Learning Projections



Predicted total infected 17.9% on December 2, 634k new infections per day.

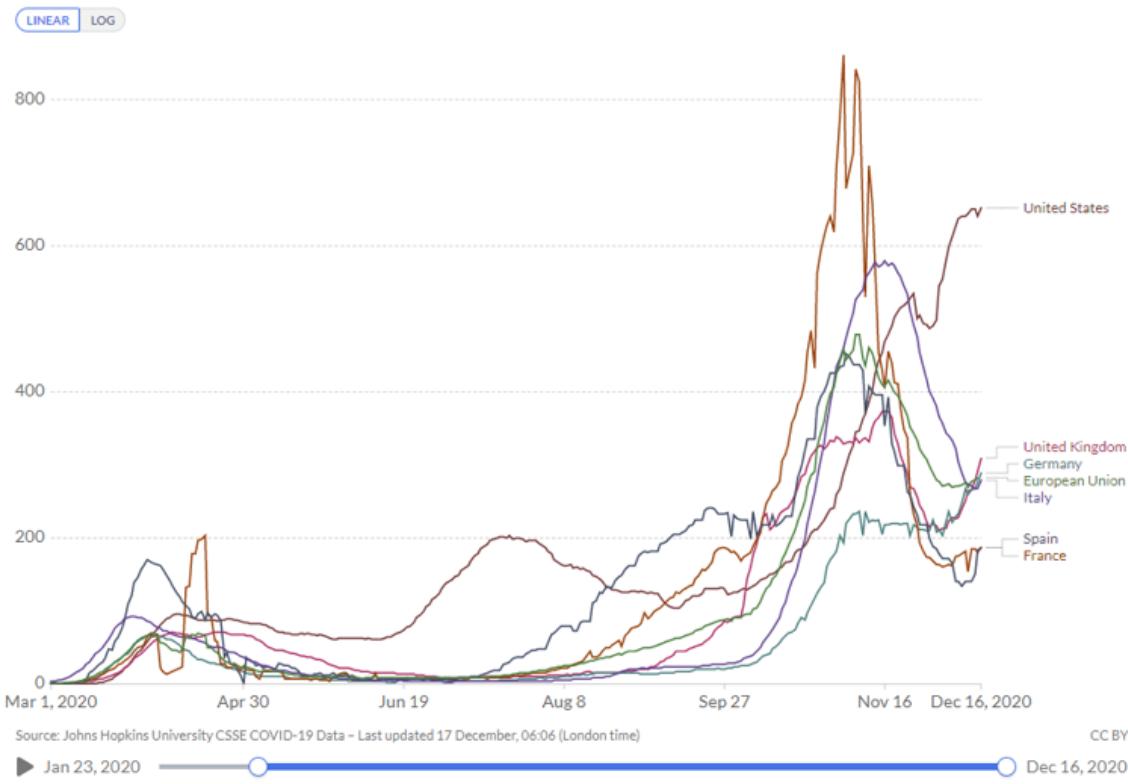
Remember, those numbers are several weeks behind, and in general I consider them to be soft lower bounds.

Europe

Daily new confirmed COVID-19 cases per million people

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

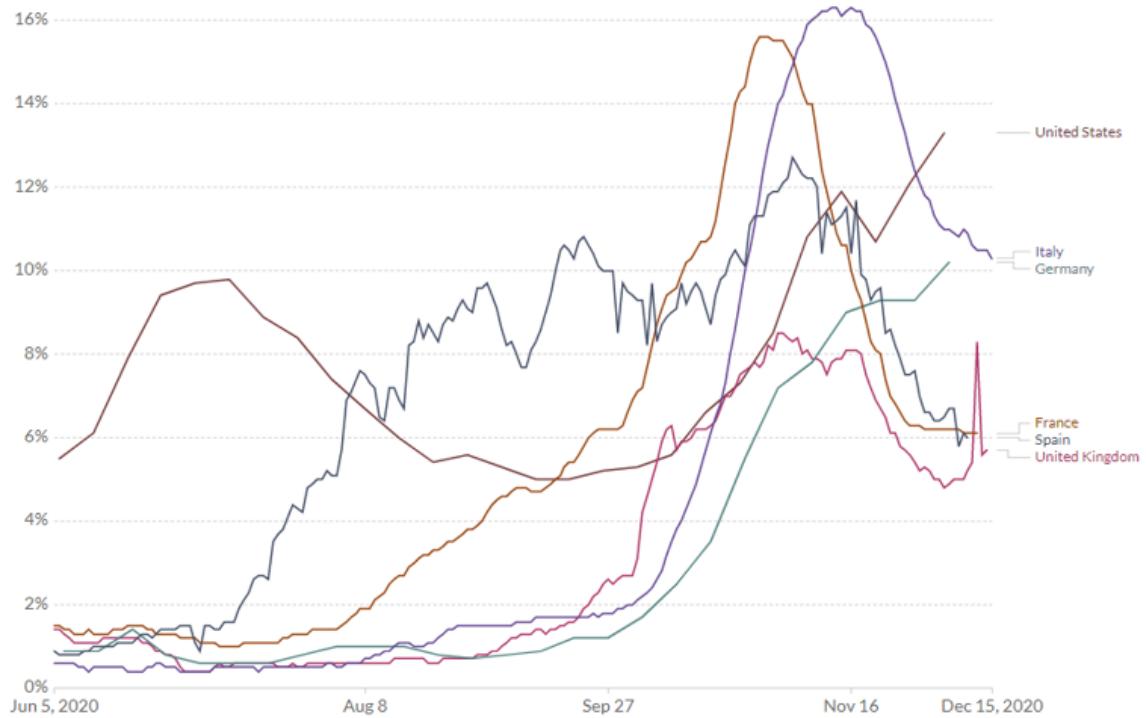
Our World
in Data



The share of daily COVID-19 tests that are positive

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

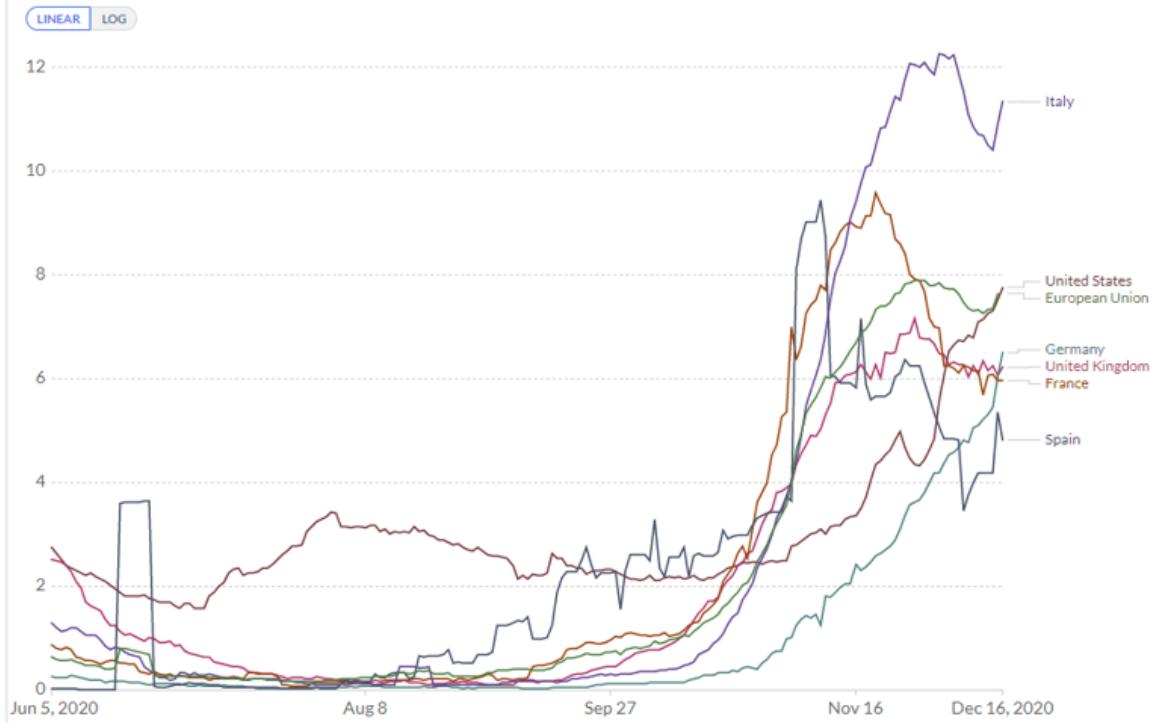
Our World
in Data



Daily new confirmed COVID-19 deaths per million people

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

Our World
in Data



America now has twice as many infections as any of the European nations tracked here. Germany and the United Kingdom once again seeing more infections. Italy rapidly improved on that front but is still seeing a lot of deaths due to lag. The European strategies seem to continue to oscillate between being harsh to get things under control, and at other times letting things get worse. This does not seem obviously worse or better than the de facto American strategy of doing nothing.

All I Want For Christmas is a Covid Vaccine

It seems all it took to get the FDA to approve the vaccine on Friday, only one day's worth of deaths [after their panel had taken its sweet time to meet and give its seal of approval](#), was for Mike Meadows to tell the head of the FDA to resign if they didn't give approval by the end of the day. It seems things were being held up because they had not prepared a proper "fact sheet" for the vaccine, which prompted Alex Tabarrok to calmly [suggest maybe getting that done in advance next time](#).

The FDA denies that the pressure impacted their timetable:

● SAT, DEC 12 2020 • 9:21 AM EST

FDA chief says Pfizer vaccine authorization not due to 'external pressure'

Dr. Stephen Hahn, commissioner of the U.S. Food and Drug Administration, emphasized during a press briefing that “science and data guided the FDA’s decision” to grant Pfizer’s vaccine emergency authorization and called reports that he would be fired if the drug wasn’t immediately authorized “inaccurate.”

“Science and data guided the FDA’s decision. We worked quickly based on the urgency of this pandemic, not because of any other external pressure,” Hahn said during a press briefing. “This decision was based on the strongest scientific integrity, and I’m so proud of the work that our career scientists have done.”

On Friday, [The Washington Post](#) first reported that White House chief of staff Mark Meadows told Hahn to [submit his resignation](#) if the agency didn’t clear the vaccine for emergency use by the end of the day. In a statement following the report, Hahn called the report “an untrue representation” of his conversation with Meadows.

“The representations in the press that I was threatened to be fired if we didn’t get it done by a certain date is inaccurate,” Hahn said at the briefing on Saturday.

— *Noah Higgins-Dunn*

He has to say that, because of the mindset ‘this was either due to science or due to pressure’ rather than the pressure being used to get around stupid red tape after the scientific job was already done (including lots of other stupid red tape), and also to save face. [I am very confident he is lying here.](#)

Calling for the head of the FDA if he doesn't find a way to assemble a fact sheet by end of business is exactly the type of thing one does when giving such questions the urgency they deserve. We still didn't vaccinate people until Monday, but things could have slipped even further. It's also worth noting that the vaccine didn't begin shipping until after every prior step was completed, which caused an additional unnecessary delay.

Early vaccinations seem to be running into some snags. [Here's a report from Florida](#). Hospitals are getting lots of doses, yet only using a few percent of them each day. That does not seem like the right sense of urgency. It also says that due to production concerns future shipments are on hold, which is a lot scarier. I haven't seen other indications of things being on hold, but that would be very bad news.

Other shipments seem to have issues with being... [too cold](#)? Not sure exactly what goes wrong when that happens.

Credit where credit is due: Looks like Moderna is being allowed to submit with 7-week instead of 8-week safety data, and without full verifications ([source](#)):

The initial EUA request was based on data from the pre-specified interim analysis (November 11, 2020 data cutoff) with a median follow-up duration of 7 weeks after dose 2; this interim analysis data is the primary basis of this EUA review and conclusions. Data and analyses from a November 25, 2020 data cut with a median duration of at least 2 months follow-up after completion of the 2-dose primary vaccination series was submitted as an amendment to the EUA request on December 7, 2020. The FDA has not independently verified the complete safety data from the primary analysis, aside from all new deaths (including those reported through December 3, 2020) and SAEs. No new safety concerns have been identified. The rates

Once again, it's all minor in the grand scheme, *provided this did not slow down production or vaccination.*

What matters is the long term path. [That seems to still mostly be on track for early spring](#). [We just bought another 100 million doses of the Moderna vaccine](#). By waiting, we gave ourselves the chance to save a tiny bit of money, and in exchange we slowed getting back to normal by a month. Not maxing out vaccine purchases was a supremely costly decision in terms of expected lives lost and lives ruined. I originally wrote it was a "mistake," but on reflection that implies things about decision makers that I don't endorse. Not a mistake.

Well, you should know this already, but FDA Delenda Est:

The FDA has reiterated its recommendation that patients receive two doses of the vaccine, warning against speculation that a single dose might provide sufficient protection.

Dr. Peter Marks, director of the Center for Biologics Evaluation and Research at the FDA, told reporters that those who participated in clinical trials received two doses, and the FDA's recommendations are based on the scientific data from studying the two-dose regimen.

"We spent so much time carefully reviewing the data and basing decisions on science that it seems pretty foolhardy to conjecture that one dose might be okay without knowing," Marks said.

"At least from an FDA perspective, we would be recommending that people complete the two-dose series so that we actually know they are truly protected at the rate of approximately 95% efficacy that was reported," Marks said.

● SAT, DEC 12 2020 • 9:46 AM EST

'Foolhardy' to assume one dose might be sufficient, FDA official says

[The full FDA report on the Moderna vaccine is here.](#)

Also, um, have you seen the one-dose efficacy data because, I mean, holy shit:

Table 15. Vaccine Efficacy^a of mRNA-1273 to Prevent COVID-19 From Dose 1 by Time Period in Participants Who Only Received One Dose, mITT Set

First COVID-19 Occurrence After Dose 1	Vaccine Group N=996	Placebo Group N=1079	VE (%) (95% CI)*
	Case n (%)	Case n (%)	
After dose 1	7/996 (87.5)	39/1079 (96.7)	80.2% (55.2%, 92.5%)
After dose 1 to 14 days after dose 1	5/996 (38.0)	11/1079 (41.1)	50.8% (-53.6%, 86.6%)
>14 days after dose 1**	2/983 (87.2)	28/1059 (96.2)	92.1% (68.8%, 99.1%)

*Surveillance time in person-years for given endpoint across all participants within each group at risk for the endpoint.

Guess who said [the Pfizer vaccine was safe and effective after one dose](#)? Yep. The FDA.

In what world is giving the second dose to the same person, raising them from 87% to 96% protected, a higher priority than vaccinating a second person?

WHAT THE HELL, HERO?

Guess who is the Only One Man brave enough to step up for this obviously correct strategy? Who is the most ridiculed of them all? It is, of course, [the quintessential Florida Man, Governor DeSantis](#). Remember when DeSantis was history's greatest villain for not closing beaches while Andrew Cuomo was a hero despite literally forcing nursing homes to take in people who were Covid-positive? Huh.

I do want to note the counterargument to this, which is that if an individual is 86% protected they have to decide whether to resume normal life while waiting for their second shot, whereas at 95% anyone not at high risk can safely take a lot more risk. At a societal level, of course, this if anything backfires further, but at the individual level there are definitely gains to concentrating your immunity.

You know what would be another way to get more people vaccinated? Use our entire supply of the vaccines! As in, the bottles provided are so overfilled that [pharmacists can extract one or sometimes even two extra doses out of each five-dose vial](#). That is quite a bit of extra vaccine going to waste by default! Which is fine for a vaccine or other medicine with abundant supply, where avoiding contamination and ensuring a buffer against spillage or other surprises is important, but is totally not fine here. Luckily for us, the FDA has decided to declare using these extra doses "acceptable," although as of the article's publication that guidance was not yet official, opening the door for the extra doses to be used.

There's still this, though:

Both Pfizer and FDA said that leftover vaccine from multiple vials should not be mixed, because of the contamination risk.

To which I would like to reply, *that's an interesting answer, can you please show your work?*

It seems to *imply strongly* that the risk of contamination for any given injection is sufficiently high that multiplying that risk by two or three makes it net negative to administer the vaccine. That would mean it is eating up a large fraction of the gains from vaccination, as opposed to us *essentially never hearing about contamination issues*.

If we presume that a vial right now has a random amount of leftover vaccine, then by default the average vial should have about half a dose remaining after administering all full doses. That means that if we average six doses otherwise, we can expand supply by another 8% or so by combining those doses, or 6% or so if we are willing to combine up to two vials but not three. We are throwing that away because it would increase "risk of contamination" and again I very much want to see their work.

I love this gem:

Only 2.2% of participants had evidence of prior infection at study enrollment, and there was only one COVID-19 case starting 14 days after dose 2 reported from this subgroup, which was in a participant in the placebo group. There is insufficient data to conclude on the efficacy of the vaccine in previously infected individuals.

I will gladly accept this particular demand for subgroup analysis if it keeps us from *utterly wasting* vaccinations on people who already have the immune response the vaccine is designed to create. Even if the vaccine was 100% effective in this group with zero side effects, that's obviously hugely wasteful. Yet it looks like hospitals are mandating that such people get vaccinated anyway.

How fast could we have gone? Remember how we learned last week how the Moderna vaccine was designed in two days? I was surprised by that because my father said he could have done it in one, [and it looks like Pfizer agrees with his timetable](#).

[COVID-19 Vaccine Allocation Dashboard by Benjamin Renton](#) tells us where America's vaccine doses are headed. The first shipment seems to be allocated by population, as expected, but later in December the numbers quickly diverge, and it's not obvious what is driving this. There isn't an obvious political economy story here, and one could plausibly credit a lot of the differences to shipments not having been catalogued yet. It seems like consistently 6-7% or so of the population is getting vaccinated in the places that are getting the most, and there are a lot of 'null' entries here, so the hope is that this represents roughly a 6.5% of population vaccination distributed smoothly by the end of December. That would match previously announced quantities. If we did get that, then combining it with growing immunity from other sources, the tide should turn rapidly in January, unless sufficient additional control systems set in to un-turn it.

In other vaccine news around the world, [The Oxford/AstraZeneca Vaccine Efficacy Data](#) is in. The whole process seems to have been quite the royal mess, making it hard to be at all confident in the efficacy numbers. The standard Very Serious Person response is to re-run the trials, including checking to see if the mistaken half-dose given in the UK was a random stroke of genius. There does seem to be a plausible theory of how that might be true. My guess is that efficacy is something like 75%, in between the two trial results, and the half-dose makes only a small difference either way, but given we have limited vaccine supply I'd definitely go with the initial half dose for now.

The big (alas mostly rhetorical) question remains, *why haven't we approved this vaccine?* The safety data seems very solid and I still can't believe people continue to fret about a single person out of tens of thousands getting sick with *something, anything* despite not even having a mechanistic story of how the vaccine could be related, let alone it happening often enough for us to want to care. Our confidence interval for effectiveness should presumably be something like at least 50% effective and at most 90% effective, and if it's 50% effective then *we should get on this yesterday*, obviously. Which we of course will not do.

Similarly, I was reminded this week that many arab countries have approved the Chinese vaccine, and was asked why this is so. I replied this was so *because it probably works*. There is no mystery to explain.

Perhaps we should consider trading vaccines with the Russians? [Their consumer watchdog is asking those who get vaccinated to refrain from alcohol for two months](#), as Russians, in Russia, during the winter. I totally get how this can help build up immunity but *come on*. This is what one calls 'not going to happen in a million years.' Bring them vodka or bring them death, and if it's both or neither for arbitrary values of vodka and relatively low levels of death, I'm pretty sure we already know their revealed preference.

My guess, of course, is that the watchdog is trying to use this as an excuse to get Russians to drink less, with the breakdown between ‘Russians drink way way too much and it’s worse for them than Covid-19 so getting them to cut down by any means is important’ and ‘Very Serious Person gets to take away only joy in people’s lives and is not about to miss that opportunity’ left as an exercise to the reader. Some actions are overdetermined.

I, on the other hand, have not had a drink of alcohol in years. May I suggest free trade?

They’re also experimenting with [combining the Sputnik vaccine with the one from AstraZeneca](#). Given the issues with the first dose building up immunity to the vector needed for the second dose, it makes sense that mixing and matching could be good. I’m glad they are trying this. However, this also points out the obvious, which is that combining vaccines is on priors *more effective* than only doing one, and if that’s true and the safety data is in, *why are we not taking all of the vaccines?*

[Sanofi vaccine given in insufficient doses in Phase I/II trials, and only generated what confidently looks like sufficient response in 18-49 year olds. Here’s a less detailed news report that also mentions delays in another vaccine.](#) So they’re going to redo the studies, and while doing so they will not be doing a Phase III on 18-49 year olds, even though there are plenty of other vaccines that can serve the older cohort while we give Sanofi’s vaccine to the young. On top of that, they tested a lower dose and it seemed to work in the young, so that lets us get more people vaccinated faster! Perhaps we should be going a step further, and considering that maybe younger people need smaller doses of *other* vaccines as well. That would be another way to get this done faster. Not that anyone will ever give any of that the slightest consideration.

I do genuinely feel for regulators sometimes. Even when they want to do the right thing, they face horrible incentives, blamed only for individual bad things that result from the things they do, without getting credit for anything good or being blamed in any meaningful way for holding things up. Consider [this throwing under the bus of India’s regulatory authority](#). It spends most of its time blaming them for approving things improperly, or letting trials happen that had individual adverse events. You see, these things ‘undermine confidence.’ It does not mention, *at all*, failure to run challenge trials or use other methods, including simply ‘run more and bigger trials sooner’ to get results faster, or its failure to actually do the job of ensuring the manufacturing of its full capacity of 3 billion doses per year of vaccine.

Everything Matters Versus Nothing Matters

A paper, [phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events](#), studies several such events. In particular, it seems that one conference held in Boston was responsible for a large percentage of all Covid-19 cases in the area for months afterwards, and subsequently for almost 2% of nationwide cases if you count all cases that derive from a virus variation that originated at the conference. That sounds like a really terrible conference.

The catch is that, like everything else in this pandemic, all those cases are intertwined with the control system. The counterfactual where the conference was cancelled does not contain 2% fewer Covid-19 infections or 2% fewer deaths. As people notice the extra infections, they adjust their behaviors, and governments adjust their rules. If you assume only fully naive SIR-model-style immunity effects from only detected cases, by now immunity has reduced new infections by 58%. If you started off 2% higher, those effects start off 2% stronger, and a majority of the increase cancels out. Then adjust that for the infections we miss, and the selection effects of who gets infected, and combine with the other aspects of the control system.

These are the two pandemic perspectives. On the one hand, every new infection *permanently increases the infection rate*, which then moves exponentially, so the conference

is responsible for 2% of infections. On the other hand, control systems, so it's not clear that the conference had a meaningful overall long term impact at all. In the extreme, if the control system was always destined to break down in November as the weather got colder, making things worse earlier can potentially build up immunity earlier, which prevents hospital overload and gives us a smaller peak number of infections, which minimizes overshooting after the corner gets turned and gets people overall better treatment. So being careless is either deeply irresponsible, actively saving lives, or something in between. Life is confusing like that.

Further Research Is Needed

Fast Grants and Marginal Revolution provide [suggestive early results that fluvoxamine, and perhaps SSRIs and sigma-1 receptors more broadly, could provide an effective treatment for Covid-19](#). Given robust safety records, these do look promising, with the downside that fluvoxamine is reported to be difficult to quit. This is the kind of thing we would have known in April if we had our act together. Instead, even if these treatments work, we won't be confident enough to use these treatments widely until most deaths have already happened. The sample and effect sizes here are quite small, so it could easily go either way.

You Should Know This Already (Inessential Reminders)

Yes, [Moderna's vaccine prevents transmission](#). One dose is good for reducing infection by 63%, two by over 90%.

CDC Delenda Est:

● SAT, DEC 12 2020 • 2:41 PM EST

People who contracted Covid should still get vaccinated, CDC says

Our government is still hard at work slowing down the vaccination process:

● SAT, DEC 12 2020 • 1:16 PM EST

U.S. to maintain reserve of Covid vaccine doses as safeguard against distribution problems

The U.S. will for now maintain a reserve of more than 50% of coronavirus vaccine doses manufactured in the event there is spoilage as Pfizer's vaccine begins being distributed nationwide this weekend, Army Gen. Gustave Perna said at a press conference.

As a reminder, if you only give someone one dose, *nothing bad happens*, and you can give out the second dose later with full benefits therefrom. This large of a reserve only has one purpose.

Reddit is asked: [why is there an explicit line between Phase 3 and roll out of a vaccine?](#) The responses point to bureaucratic, regulatory and "ethical" justifications aplenty, most of which readers will already be familiar with. Under normal circumstances they would even make sense.

[More confirmation of what is important to "bioethics."](#)

[Here's another confirmation](#) that what scares such people most is that things of value might be allocated to those who value them most, as measured by their willingness to pay, or who might go on to produce the most value. You see, it's a scandal and tragedy when we prioritize treating the President of the United States over more 'ethically deserving' others, and we have to watch out before we allocate resources where they might be useful. If we're not careful, next thing you know we'll have an entire economy full of producing useful things and allocating them where they are valued most and can produce the most value. That would be the worst.

Household gatherings are driving a lot of the spread, *but contact tracing works if and only if there are contacts one can easily trace, and assumes that the contact was the source.* It's still worth doing, and still provides useful information. But one must be careful with interpretation:

T



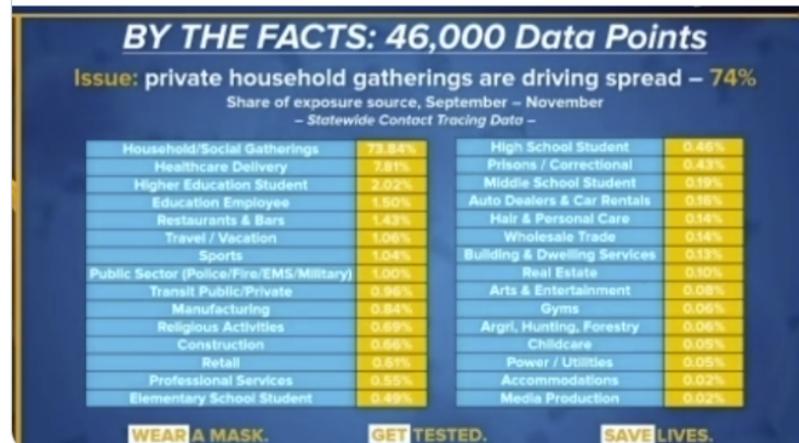
Nate Silver ✅ @NateSilver538 · Dec 11

I'm not sure I entirely buy this data. The issue is that if you caught COVID at home or from a friend, you'll know that contact also got COVID and so can identify the source of exposure. But if you caught it from another patron at, say, a restaurant, you likely wouldn't.

Jon Campbell ✅ @JonCampbellGAN · Dec 11

NEW: New York releases some statewide contact tracing data for the first time, breaking down the source of exposure.

Show this thread



92

129

1.1K

↑



Nate Silver ✅

@NateSilver538

...

Replying to @NateSilver538

In other words, household and social contacts might represent 74% of **known sources** of exposure, but that's potentially very different from their representing 74% of exposures.

12:09 PM · Dec 11, 2020 · Twitter Web App

58 Retweets 2 Quote Tweets 1K Likes

...

...

...

↑



Patrick Allison @dbarawn · Dec 11

Replying to @NateSilver538

Yup! Last time I saw contact tracing results, "community spread" = "no \$!\$!\$ing clue" was the most common result.

...

Thus, Nate and Patrick make strong points here, but it is even worse than the initial impression. Even if one has a contact that tests positive, and the timeline works out that they could have infected you, *that doesn't mean that was the source of your infection*. Thus,

it's not that 74% of known sources come from households and social gatherings, and then some cases have no known source. It's that of the cases where there is a known source, 74% of those involve households and social gatherings, but we don't know how often that is actually the source.

I do think it's right that social gatherings and exposure within households are the largest source of cases, and likely the majority of them, but the data we have is not sufficient to reach such conclusions.

Here is [an analysis of why the west failed so utterly at contact tracing](#) and other places succeeded. This all seems simple. Most nations did not attempt (the word 'try' is so overloaded I'm going to mostly taboo it) contact tracing. At maximum, they attempted to take symbolic action that they could cite as an attempt to do contact tracing. Between privacy concerns, distrust of government, and a complete unwillingness to ever compel anyone to go anywhere, do anything or reveal any information unless they've been arrested for a crime, *or to spend money on things like getting people places to safely and comfortably quarantine even when they wanted to do so*, and fearing (I think correctly) that if they did any of that people would not get tested and otherwise go crazy, they settled for symbolic action. There was not even an attempt to scale up to numbers of tracers that might plausibly have been adequate. Then case numbers quickly got out of hand, and one thing the places that succeeded via contact tracing had in common is they did so early, before things got out of hand.

So we did nothing, while Vietnam did this ([source](#) that includes more):



Eric Feigl-Ding
@DrEricDing

...

HOLY MOTHER OF CONQUERING A HOSPITAL
OUTBREAK—Vietnam quashed a big #COVID19 outbreak. Besides tracing 50,000+ people who were 1st, 2nd, 3rd degree contacts, they also quarantined & gave **FREE ROOM & BOARD to 27,893** F1 & F1 close contacts during quarantine! Wowzers!

There's also this [inessential lighter side twitter video link about the joys of contact tracing](#). I was amused.

Non-essentially, [more restating “ethical” and otherwise nonsensical reasons why we should not do the things we obviously should do](#), via MR. Not doing these things has caused and will cause many, many people to die, and many lives to be destroyed, and all of us to be dramatically worse off.

Should kids get snow days off during remote learning? New York City mayor DeBlasio, owner of a true [zero percent approval rating](#), has picked the side you would expect, saying 'snow days are over.' He's also saying [New York should prepare for a 'total shutdown' after Christmas.](#) Joy is the enemy and must everywhere be destroyed.



Jillian Jorgensen @Jill_Jorgensen 3h

One snag in the joy: The end of the snow day! With remote learning, even if school buildings close for snow, class will still be in session online (as it will be for grades already at home). But I suspect (and hope!) many kids will still find a way to enjoy it.

Jamie Stelter @JamieStelter

Totally agree with @bobhardt's sentiment just now on @NY1 that this snowstorm is gonna be more joyous than usual. You're stuck at home anyway! How do you feel about 8-14 inches?

Q 11 T 4 L 41 ...



Jason Haber

@jasonhaber

Replies to @Jill_Jorgensen

At my daughters school they announced snow days are a rite of childhood that shouldn't be taken away by covid. I thought that was awesome.

7:03am · 15 Dec 2020 · Twitter for iPhone

3 Replies 3 Retweets 37 Likes

Should those instructed at home get snow days? Depends on what snow days are, and what schools are. Rather than restate my views I'll let everyone ponder on their own.

Or, if you *don't* know the following particular thing already, [then we need to fix that, because the WHO says so](#):

[Thread](#)

 **Patrick Collison**  @patrickkc · 15h
Someone I know has had some quite useful COVID-related posts removed from Medium, LinkedIn, and Nextdoor—they've been deemed "COVID misinformation". (It's not what the WHO endorses!)

114 replies 427 retweets 2.5K likes

 **Patrick Collison**  @patrickkc · 15h
I think his posts overstate the case for the treatment he's arguing for. They aren't without justification, however. He has some considerable scientific evidence on his side (including a small RCT).

6 replies 4 retweets 271 likes

 **Patrick Collison**  @patrickkc
Replying to @patrickkc

This year, we've come to better appreciate the fallibility and shortcomings of numerous well-established institutions ("masks don't work")... while simultaneously entrenching more heavily mechanisms that assume their correctness ("removing COVID misinformation").

10:05 PM · Dec 14, 2020 · Twitter Web App

66 Retweets 9 Quote Tweets 793 Likes

 **Patrick Collison**  @patrickkc · 15h
Replies to @patrick
False claims about COVID (or any disease) are, of course, undesirable. But, leaving aside the merits of content removal as the response, have we really figured out a sensible applied epistemology for operationalizing such designations?

12 replies 14 retweets 457 likes

 **Patrick Collison**  @patrickkc · 15h
And, more broadly, aren't his efforts something close to a paradigmatic example of where society should benefit from the internet's broadly participatory nature? Is the equilibrium where that's stifled really the optimum?

7 replies 9 retweets 458 likes

 **Patrick Collison**  @patrickkc · 15h
Science is not a coherent, monolithic edifice. It changes daily and scientists don't all agree with each other. A lot of ex post correct views first show up as claims that look ridiculous by the standards of the time.
nintil.com/discoveries-ig...



Patrick Collison @patrickkc · 15h

...

These platforms have tough jobs, no doubt. But I'm worried that the embrace of "misinformation" as a newly illegitimate category may have costs that are considerably greater than what's apparent at the outset.

47

84

1.2K



Many platforms have a principle that if you disagree with the WHO or CDC (also known as 'the lying liars who said masks don't work and have similarly only admitted fact after fact months after they had become clear') that your posts and videos will be taken down. WordPress does not have such a principle, so I should be safe here, but if I did not have LessWrong as an automatic backup (and my drafts saved in Google as a second backup) I would be sure to stash additional copies. If I was posting this content to places like YouTube, Facebook or Medium, my expectation is that at least some posts would have been taken down.

Even the Internet Archive does not seem reliable, as they have announced plans to begin putting warnings on past content rather than offer up the web purely as it was. That pretty much signals that at some point, perhaps reasonably soon, they will start taking down records of things that powerful people sufficiently dislike.

Like most modern rules, this rule is not reliably and fairly enforced. If it was, it would be obvious that the rule was absurd and unacceptable. Instead, the rule is that when people are inclined to take something down, it is considered a valid excuse to cite disagreement with official sources, despite the known unreliability of those official sources.

If you tell people they aren't allowed to go to the office, they'll find offices where they may, such as [in the quiet cars of the Long Island Railroad](#).

When you don't let people get treatment until they prove they are sick enough for it, at which point the treatment is being given too late, you end up not using it:

3 hr 17 min ago

US has surplus of monoclonal antibody treatments for Covid-19, Azar says

From CNN's Shelby Lin Erdman

Not enough Covid-19 patients are asking for or receiving monoclonal antibody treatment, Health and Human Services Secretary Alex Azar said on Tuesday.

"We have a surplus of these monoclonal antibodies right now, the Regeneron and Lilly antibodies, and what's happening is people are waiting too long to seek out the treatments until they show up at the hospital and, by then, it may be too late in order to get the benefit of these antibody treatments that beat back the spread of the virus," Azar told CNBC in an interview.

In Other News

↻ Ted Knutson Retweeted



natasha lyonne ✅ @nlyonne

17h

Not cool.

sam greisman ✅ @SAMGREIS

All of us at 12:01 on New Years when
we realize we just woke up back at
January 1, 2020.



💬 290 ↗ 5k ❤️ 80k ⋮

[Russian Doll](#) is in fact excellent. Thought experiment: If this happened to you, other than making a killing in various markets, what would you do? Would you be able to stop the pandemic?

Non-Covid: FDA also approves [genetically modifying pigs](#). They don't fly.

FDA panel member [explains her vote against approval of Pfizer vaccine](#), potentially 'undermining confidence,' because the panel recommended it be given to 16-17 year olds in addition to those over 18, without having enough specific evidence for that, and the panel member fully supporting vaccinations in those over 18, and *the obvious fact that no one under 18 is going to get a vaccination for a long time either way*.

FAA generously decides to not attempt to murder its pilots at this time:

● SAT, DEC 12 2020 • 12:31 PM EST

FAA says pilots can take Pfizer's vaccine

In welcome news, Pelosi denounces all her policy positions:

● SAT, DEC 12 2020 • 9:37 AM EST

Pelosi calls for quick, fair and free distribution of Covid vaccine to as many Americans as possible

Carbone, which pre-pandemic was the only restaurant in New York that I both wanted to dine in and couldn't get a table at when I wanted one, [is taking outdoor dining to the next level](#). That next level appears to be 'indoor dining,' which they are claiming does not technically count so long as the ground it is built upon is legally a sidewalk.

39 min ago

Doctors in Northern Ireland treat patients from parked ambulances as hospitals pass full capacity

From CNN's Kara Fox

[Germany confirms they are engaging in triage](#) (in German).

1 hr 23 min ago

Singapore says almost half of all migrant workers living in dorms were infected with Covid-19

From CNN's Eric Cheung in Hong Kong

If you don't need to be quarantined, [you can have your own line at the airport](#).

Skin in the game:

1 hr 25 min ago

China will suspend inbound international flights if five people onboard test positive for coronavirus

From CNN's Eric Cheung in Hong Kong

China will suspend inbound international flight routes if five or more people test positive for Covid-19 when they land, the Civil Aviation Administration of China said in a statement on Wednesday.

Under the new rules, if five or more passengers test positive for Covid-19 after arriving in China on a flight, the airline company will be banned from running those flights for two weeks, it said.

A four-week suspension will be served if 10 or more passengers test positive for the virus, it said. Airlines will be allowed to resume flying one flight per week on the route once the suspension ends.

Previously, airlines had to serve a one-week suspension if five or more passengers on a flight tested positive for Covid-19. The tightened restrictions are implemented with immediate effect.

On Tuesday, the Civil Aviation Administration of China issued one-week suspension notices to Ethiopian Airlines, the Russian airline Pegas Fly, and Swiss International Air Lines after five or more passengers who traveled on their flights tested positive for Covid-19.

6 hr 5 min ago

South Korea warns of first potential lockdown as cases rise and ICU beds run out

From CNN's Jake Kwon, Gawon Bae and James Griffiths

San Mateo County in the San Francisco area declines to Sacrifice to the Gods and shut everything down in ways known to not be effective, [explains its reasoning in detail](#).

Not that anyone should expect people to obey the new California lockdowns, which is one reason why they definitely won't work ([source](#)):



Jonathan Blow
@Jonathan_Blow

o

I drove from San Francisco -> San Diego last Monday, and back from San Diego -> San Francisco yesterday.

As near as I can tell, nobody cares that the state's governor declared a new lockdown. It looks like people are just done with it.

12:45 PM · Dec 14, 2020 · Twitter Web App

In [a cool little piece rounding up views on techno-optimism](#), we get this perspective:

Matt Yglesias: "Some optimism about America's COVID response"

I'm pretty sure that it was the COVID vaccine that got us all thinking along techno-optimist lines. Other positive trends like solar and batteries had been going on for years, but the big push of the vaccine effort — the unprecedented speed and effectiveness, the public-private cooperation, the use of a novel technology (mRNA vaccines) to beat a novel pandemic. It was an inspiring moment.

And like many of us, Yglesias was inspired to think about than just vaccines. He writes:

I think it's clear that if we had to weather this pandemic with the tech of 20 years ago, we'd have had worse public health outcomes and worse economic outcomes. Two decades' worth of internet and mobile tech has proven its value, and more good things are to come if this period of experimentation ends up helping to accelerate the development of widely used remote work, telemedicine, and online education models...

and

[The vaccine is] a proof of concept for the kind of thing we could be doing in the clean energy space. Say an electric car that meets such-and-such specifications would get guaranteed orders to serve as government fleet vehicles. Or pre-commit to buying electric buses for schools and transit agencies. Nuclear micro-reactors for use on military bases or as backup systems for hospitals. The assurance that a market exists is a big stimulus to private investment, and when strong social consensus exists that innovation would be beneficial, we can get it done.

He's right, of course.

It is obviously correct that given the restrictions placed upon us, the disruptions we suffered in 2020 would be a drop in the bucket compared to attempting similar restrictions in the year 2000 for a similar length of time. Mobile and internet technology has proven its value many times over.

But that doesn't mean we would have had worse outcomes! It seems highly plausible we would have had *much better* outcomes on at least some fronts.

On the economic front, we would have had to choose either to *actually suppress the virus*, in which case we get much better outcomes all around, or to accept that the virus couldn't be stopped, *which also produces better economic outcomes*.

Our technological advancement gave us the choice to make massively larger sacrifices to the Gods rather than deal with the situation. And as we all know, [choices are bad](#). We also are, in my model, much more inclined to make such sacrifices now than we were in the past, even when the trade-offs are similar, which ties into my view that simulacra and maze levels

are higher, with a larger role played by fear of [motive ambiguity](#). We might have been willing to do challenge trials or other actual experiments, and have had a much better handle on things quicker on many levels.

Here's a quote from Tyler Cowen yesterday that illustrates the issue quite well, while also showing how much else we have lost:

Here is [my earlier Bloomberg column](#) rejecting the notion of forced quarantine of individuals for Covid-19, mostly on rights grounds, though I add some consequentialist arguments. I would not trade in the American performance for the Chinese anti-Covid performance if it meant we had to weld people inside their apartments without due process, for instance, as the Chinese (and Vietnamese and others) did regularly.

Seriously, what the hell, hero? You wouldn't make that trade, *in hindsight, knowing that their method would work and our method would be this awful?*

I think not making this trade, under these conditions of assured outcomes for both choices, and given only these choices, is *utterly completely insane*.

The civilization that chooses thus, does not survive long.

It is an open question whether being forced to give up on containment entirely, and accepting a lot of deaths, would have been a better outcome than over a year of Covid lockdowns of varying severity that we got instead. One could make a reasonable case for both sides even in hindsight.

But to outright say, no, we'd rather fail than use the means that work because we might do things without "due process"? Did you see anyone using "due process" this whole time for anything other than regulatory interference with those trying to solve the problem? I didn't. I saw a bunch of fiat restrictions of freedom with no legislative backing that were much bigger and didn't work, instead of restrictions that in total were much smaller and did work, together with lots of restrictions on doing anything useful.

One thing we definitely wouldn't have had in the year 2000 would be mRNA vaccines.

It is good to be reminded that the vaccine effort was extraordinarily fast, innovative, safe and effective. That doesn't cancel out the fact that it could have been much faster and much more effective if the people saying it couldn't be done had interfered less with the people doing it. And while that interference was a huge problem, the interference with other things like testing (or masks, or things as simple as ordinary life) was *far worse*.

Both are important. And there is a very optimistic take that combines both. If we did all of this *despite the best efforts of many of the most powerful people and organizations and dynamics*, then imagine what we could do if we took those gloves off!

I am not expecting this outcome, but we should keep shouting it from the rooftops, as soon as we are done shouting for one-dose vaccinations. If we can use the pandemic as an impetus to move from telling people it cannot be done and they are not allowed to do it, for most worthwhile new values of it, to providing subsidies or even giving people *the freedom to act at all*, across a variety of other domains, than perhaps the whole of 2020 will have been worthwhile after all.

Luna Lovegood and the Chamber of Secrets - Part 7

They were in the hospital wing.

"Who are you?" Harry Potter said to Luna.

"You-Know-Who. Hiss hiss hiss," Luna said. She held fingers down from her mouth and swayed like a snake.

"This is Loony—" Hermione said, "Luna Lovegood. She created the *BOY-WHO-LIVED GETS DRACO MALFOY PREGNANT* headline the summer before last. She believes there exists an unobservable Platform Nine and One-Half concealed by powerful magics. She claims you and she adventured together without me."

"Do you have any evidence I trusted you before I lost my memory?" Harry asked.

Luna glanced at Hermione.

"All my secrets are hers as well," Harry said.

"I know your real name, that you're a Parselmouth and that you've operated the Mirror of Atlantis," said Luna.

"My real name isn't something I'd tell you so you must have worked it out yourself and therefore does not constitute evidence of my trusting you in the past. There are at least three people who know I'm a Parselmouth so it's not surprising the information leaked out. These are the same people who know I've operated the Mirror so the Mirror adds no bits of information. An enemy would be at least as likely as an ally to possess this information," said Harry, "Did we discover an ancient artifact or something?"

"We found the Lost Diadem of Ravenclaw together," Luna showed it to him.

"This crown proves little since you could easily retcon a story around it. On the other hand, there is information I would tell someone I trusted if I wanted to guarantee recognition in case of obliviation," said Harry, "Do you have the recognition code?"

Luna recognized the logic of someone whose priors outweighed the available evidence. She fled the room.

The Marauder's Map showed Luna how to climb on top of the Great Hall. She hugged a stone gargoyle at the edge. It guided her tears down into the Hogwarts courtyard.

"Not now, Wanda," Luna said, "I need to feel this."

Luna detected quiet behind her. She continued staring at the students mingling below.

"I do not intend to jump," Luna said.

"I do not intend to stop you," Lady Yue said.

They listened to the wind. They smelled the gentle scent of trees.

"How does one make friends?" Luna asked.

"That is Professor Lockhart's field of expertise," Lady Yue said.

"Do you have a spell for loneliness?" Luna asked.

"今夜鄜州月，" Lady Yue sang from her teenage youth, "閨中只獨看。遙憐小兒女，未解憶長安。香霧雲鬟濕，清輝玉臂寒。何時倚虛幌，雙照淚痕乾？"

《月夜》 *Moon Night* by Du Fu 杜甫

今夜鄜州月，

Tonight Fuzhou moon,

閨中只獨看。

boudoir inside only alone see.

遙憐小兒女，

Distant tenderness small girl,

未解憶長安。

unsolved recall Chang'an.

香霧雲鬟濕，

Scent fog cloud hair wet,

清輝玉臂寒。

clear splendor jade shoulder cold.

何時倚虛幌，

When lean empty window curtain,

雙照淚痕乾。

both shine tear stains dry.

Luna Lovegood and the Chamber of Secrets - Part 10

Luna had been spending more and more time in the Forgotten Library. Half a year slipped by.

"Is there somewhere safer I can think?" Luna asked Lady Yue.

"Take us to Xi'an," said Lady Yue.

"Where is Xi'an?" asked Luna.

"34 degrees, 16 minutes and 37 seconds north," said Lady Yue, "108 degrees, 57 minutes and 42 seconds east."

Luna had three magical artifacts of incredible power. One was a map of Hogwarts. One put its user into a catatonic state. Luna withdrew her astrolabe.

An astrolabe displays the universe's location relative to itself. Luna set the latitude dials to $34^{\circ} 16' 37''$ N and the longitude dials to $108^{\circ} 57' 42''$ E.

Two witches stood on a sidewalk wider than a London street. Lady Yue guided Luna to the old city wall.



The original Xi'an was surrounded by a moat, giving it the appearance of resting on a motte. The walls were almost taller than the trees planted on the berm. The gatehouse used circular arches in contrast to Hogwarts' Gothic arches. The top was decorated with flags and lanterns the colors of Gryffindor. You could ride carriages four abreast along the top. Scattered Muggle tourists moseyed about.



Britain was a small, desolate rock as far from civilization as the Earth is from Sagittarius A*. Lady Yue gestured toward the real city.



Infantry in red scale armor patrolled the walls of Chang'an with repeating crossbows. Three wizards rappelled to perform maintenance on a giant retractable circular saw blade protruding horizontally two-thirds of the way up the wall. Witches in blue plate armed with spears and bungee cables dived off the machicolations to practice semi-aerial combat. An armored vehicle was positioned between every pair of guard towers,

which were armed with siege flamethrowers and rocket turrets. Chang'an had survived land wars in Asia.

"How long has it been Fidelius?" asked Luna.

"Since 1936," said Lady Yue.

"I guess you're not coming with me then," said Luna.

Lady Yue gazed longingly at old Xi'an.

"You don't look a century old," said Luna.

"I'm not," said Lady Yue.

Lady Yue handed Luna a slip of paper with English directions in black on the back, ancient runes in black on the front and a large Heirloom Seal of the Realm stamped in red on both sides. Two suits of armor with soldiers in them allowed Luna through the checkpoint.

Curious citizens swarmed Luna. They said words she couldn't understand and thrust strange foods into her arms. Luna showed her instructions to a especially helpful bicycle. It zipped Luna through the city.

Luna smelled oyster sauce and stinky tofu. She passed a self-woking food stall and old men wargaming on the sidewalk. Luna gripped the handlebars tighter as the bicycle showed off. The bicycle dinged happily. They screeched to a halt.

A gyrocopter was parked on the roof of the laboratory. A fission reactor stuck through the top of the garage. Six different styles of radio dish poked out of six open windows. Luna entered through the airlock. She bumped her head on a ceiling-mounted security camera.

Warning. You will be ejected in 5...4...3...—

"Stand down," said the halfling artificer to the security system, "It's nice to meet you. We don't often have visitors from outside. What is your name? My English name is Leet Haxor Zer0, or LZ (Lizzie) for short. The empress dowager says you are from Britain. I have heard all about Britain. Do all British Muggles travel through time and space in telephone booths or is public transit restricted to aristocrats? I have acquired three but none of them seem to work. I suspect magic in our air is interfering with the Muggle technology."

Luna didn't know what to say so she gave Lizzie a rice roll and a box of chicken feet.

"Thanks. I haven't eaten in three days. You know how it gets when you get deeply absorbed in a project," said Lizzie.

Lizzie guided Luna to the guest room, a planetarium not yet entirely filled with Muggle artifacts.

"You can store your clothes in the wardrobe. Leave space for the ghost inside that looks just like you. Don't let it freak you out. The ghost has never hurt anyone in the many times it has appeared over the decades," Lizzie said.

There was a siren.

"I have to adjust the reactor control rods," Lizzie said, "Holler if you need anything."

There were no windows in the planetarium. Light came from a giant projection of Saturn and the starfield behind it.

Luna exhaled. The Fidelius Charm would protect her from Rowena's basilisk. Luna hid in the wardrobe to bulwark causality.

"*Lumos*," said Luna, "I solemnly swear I am up to no good."

Luna found the location on the Marauder's map where the bones of Slytherin's basilisk lay. Luna monitored that place in space as she adjusted the time dial of the astrolabe backwards. The Ministry archaeologists came and went. Other wizards came and went earlier but did not pass the bones of the basilisk. The basilisk's killer appeared. Tom Riddle. Luna tuned her astrolabe back into causal time.

Tom Riddle had slain Slytherin's basilisk. Harry Potter was Tom Riddle. Harry Potter was a Parselmouth. Harry Potter possessed the secrets of Salazar Slytherin.

Luna spun the wheels of time on her astrolabe to between after Slytherin's basilisk was killed and before the Ministry archaeologists investigated it. Though the map's writing was tangled and confused, Luna could not mistake the name. Tom Riddle walked down the Chamber of Secrets passageway to the graveyard where David Monroe was reputed to have killed You-Know-Who. Luna read the date on the astrolabe. It was the date David Monroe was reputed to have killed You-Know-Who.

You-Know-Who's real name was Tom Riddle.

Announcement

There are three parts left to this story. Part 11 will be posted in the coming week. Parts 12 and 13 will be posted during the Less Wrong 2020/2021 New Year's Party.

Image Credits

The photos of Xi'an come from China Daily. [link](#)

The image of Chang'an comes from a 2016 historical documentary starring Matt Damon. [link](#)

Luna Lovegood and the Chamber of Secrets - Part 8

Luna had too many Calls to Adventure. If there was an author writing Luna's life, Luna would file a complaint. Luna had three magical artifacts of incredible power and at least two of them were maps. What kind of author gives her hero two magical maps? It's redundant.

"Are you a map of 10th century Wessex?" Luna asked the Diadem of Ravenclaw.

A princess's cute animal companion is supposed to be visible.

"Maybe you'll turn into a prince for me," she said to Wanda. Luna tried kissing Wanda but the Wrackspurt dodged and went for her ear instead, "Oh well."

Luna needed a mysterious old wizard to help guide her narrative.

Gilderoy Lockhart's office hours were always packed.

"How did you get past all the security to assassinate the Dark Lord of Berzerkistan?" a sixth-year girl asked.

"I can neither confirm nor deny whether I have ever set foot in Berzerkistan," Lockhart winked, "If, as rumored, I am indeed responsible for the Dark Lord of Berzerkistan's untimely demise then my methods must remain a state secret."

"Tell me how you got rid of the Bandon Banshee," a first-year girl said.

"It is all in my book," Lockhart said.

"How did you kill the Dementor sent to Hogwarts last year?" a third-year girl asked.

"I promised not to tell," Lockhart said.

Lockhart loved his little fans even though none of them cared how he had won *Witch Weekly's* Most Charming Smile Award. Lockhart patiently parried their questions until one last first-year remained. She had waited to ask him something in private.

"Umm. It's okay if you say no. But. Well. Can you teach me to be popular?" Luna asked.

Lockhart brandished his award-winning smile. Then his face fell.

"Such personal attention to a student might appear unseemly," Lockhart said genuinely.

"It's an interview for *The Quibbler*," Luna said.

Luna tried out for the Quidditch team. She crashed her broom into the stands. Luna tried out for the Gobstones team. She passed out from the fumes. Luna attempted to try out for Smite Club, a rumored underground continuation of Quirrel's battles. Luna

could not find where Smite Club held its meetings. Smite Club may have only ever existed in her imagination.

A month flew by. Luna was exhausted. She sleepwalked.

99% of other people never notice the things you fail at. 1% have forgotten by tomorrow morning.

—Lovegood, Luna. "Secrets of Gilderoy Lockhart." *The Quibbler*.

Luna started a Welters team. No other students showed up. Luna started a Wrackspurt training club. Wanda and Luna showed up, but nobody else. Luna started a Muggle Repelling Charms study group in the Forgotten Library.

Another month passed. Luna fell asleep in herbology. She wore shoes to bed.

People think about themselves 99% of the time. You are competing with all of wizardkind for the remaining 1%.

—Lovegood, Luna. "Deeper Secrets of Gilderoy Lockhart." *The Quibbler*.

Luna dreamed her classmates were casting a spell. Luna played along.

"*Somnium*."

Luna woke up. She had been sleepwalking.

"Congratulations Ms. Lovegood," said Gilderoy Lockhart, "You will be representing Ravenclaw at this year's dueling tournament."

You cannot predict how others will react when you do something out of the ordinary.

—Lovegood, Luna. "Who knew Gilderoy Lockhart possessed so many secrets?" *The Quibbler*.

My favorite essays of life advice

I start each of my [weekly reviews](#) by re-reading one of my favorite essays of life advice—a different one each week. It's useful for a few different reasons:

- It helps me get into the right reflective frame of mind.
- The best essays are dense enough with useful advice that I find new interesting bits every time I read them.
- Much good advice is [easy to understand, but hard to implement](#). So to get the most benefit from it, you should find whatever version of it most resonates you and then *re-read it frequently* to keep yourself on track.

I've collected my favorite essays for re-reading below. I'll keep this updated as I find more great essays, and I'd welcome other contributions—please suggest your own favorites in the comments!

There's a lot of essays here! If you'd like, I can email you one essay every weekend, so you can read it before your weekly review: ([sign up on site](#)).

Paul Graham, [Life is Short](#). Inspire yourself never to waste time on bullshit again:

Having kids showed me how to convert a continuous quantity, time, into discrete quantities. You only get 52 weekends with your 2 year old. If Christmas-as-magic lasts from say ages 3 to 10, you only get to watch your child experience it 8 times. And while it's impossible to say what is a lot or a little of a continuous quantity like time, 8 is not a lot of something. If you had a handful of 8 peanuts, or a shelf of 8 books to choose from, the quantity would definitely seem limited, no matter what your lifespan was.

Ok, so life actually is short. Does it make any difference to know that?

It has for me. It means arguments of the form “Life is too short for x” have great force. It’s not just a figure of speech to say that life is too short for something. It’s not just a synonym for annoying. If you find yourself thinking that life is too short for something, you should try to eliminate it if you can.

When I ask myself what I’ve found life is too short for, the word that pops into my head is “bullshit.” I realize that answer is somewhat tautological. It’s almost the definition of bullshit that it’s the stuff that life is too short for. And yet bullshit does have a distinctive character. There’s something fake about it. It’s the junk food of experience. [1]

If you ask yourself what you spend your time on that’s bullshit, you probably already know the answer. Unnecessary meetings, pointless disputes, bureaucracy, posturing, dealing with other people’s mistakes, traffic jams, addictive but unrewarding pastimes.

I’ve found that unless I’m vigilant, the amount of bullshit in my life only ever increases. Rereading *Life is Short* every so often gives me a kick in the pants to figure out what *really* matters and how to get the bullshit levels back down.

Derek Sivers, [There is no speed limit](#), in which he learns a semester's worth of music theory in an afternoon:

Within a minute, he started quizzing me. “If the 5-chord with the flat-7 has that tri-tone, then so does another flat-7 chord. Which one?”

“Uh... the flat-2 chord?”

“Right! So that’s a substitute chord. Any flat-7 chord can be substituted with the other flat-7 that shares the same tri-tone. So reharmonize all the chords you can in this chart. Go.”

The pace was intense, and I loved it. Finally, someone was challenging me — keeping me in over my head — encouraging and expecting me to pull myself up quickly. I was learning so fast, it felt like the adrenaline rush you get while playing a video game. He tossed every fact at me and made me prove that I got it.

In our three-hour lesson that morning, he taught me a full semester of Berklee’s harmony courses.

This was one of the major inspirations for [Be impatient](#). Every time I reread it, I think of at least one thing where I’m setting myself a speed limit for no reason!

Sam Altman, [How To Be Successful](#). Sam might have observed more successful people more closely than anyone else on the planet, and the advice is as good as you’d expect.

Focus is a force multiplier on work.

Almost everyone I’ve ever met would be well-served by spending more time thinking about what to focus on. It is much more important to work on the right thing than it is to work many hours. Most people waste most of their time on stuff that doesn’t matter.

Once you have figured out what to do, be unstoppable about getting your small handful of priorities accomplished quickly. I have yet to meet a slow-moving person who is very successful.

Almost always, the people who say “I am going to keep going until this works, and no matter what the challenges are I’m going to figure them out”, and mean it, go on to succeed. They are persistent long enough to give themselves a chance for luck to go their way.

... To be willful, you have to be optimistic—hopefully this is a personality trait that can be improved with practice. I have never met a very successful pessimistic person.

There are lots of different points here, so this one especially bears rereading!

R. W. Hamming, [You and your research](#). Hamming observed almost as many great scientists as Sam Altman did founders. He had some interesting conclusions:

At first I asked what were the important problems in chemistry, then what important problems they were working on, or problems that might lead to important results. One day I asked, "if what they were working on was not important, and was not likely to lead to important things, then why were they working on them?" After that I had to eat with the engineers!

About four months later, my friend stopped me in the hall and remarked that my question had bothered him. He had spent the summer thinking about the important problems in his area, and while he had not changed his research he thought it was well worth the effort. I thanked him and kept walking. A few weeks later I noticed that he was made head of the department. Many years later he became a member of the National Academy of Engineering. The one person who could hear the question went on to do important things and all the others—so far as I know—did not do anything worth public attention.

... Some people work with their doors open in clear view of those who pass by, while others carefully protect themselves from interruptions. Those with the door open get less work done each day, but those with their door closed tend not know what to work on, nor are they apt to hear the clues to the missing piece to one of their "list" problems. I cannot prove that the open door produces the open mind, or the other way around. I only can observe the correlation. I suspect that each reinforces the other, that an open door will more likely lead you to important problems than will a closed door.

There is another trait that took me many years to notice, and that is the ability to tolerate ambiguity. Most people want to believe what they learn is the truth: there are a few people who doubt everything. If you believe too much then you are not likely to find the essentially new view that transforms a field, and if you doubt too much you will not be able to do much at all. It is a fine balance between believing what you learn and at the same time doubting things. Great steps forward usually involve a change of viewpoint to outside the standard ones in the field.

While you are learning things you need to think about them and examine them from many sides. By connecting them in many ways with what you already know.... you can later retrieve them in unusual situations. It took me a long time to realize that each time I learned something I should put "hooks" on it. This is another face of the extra effort, the studying more deeply, the going the extra mile, that seems to be characteristic of great scientists.

Hamming is an unusual combination of (a) a great scientist himself, (b) curious and thoughtful about what makes others great, and (c) honest and open about his observations (it seems).

Anonymous, [Becoming a Magician](#)—on how to become a person that your current self would perceive as magical:

The description was about five or six handwritten pages long, and at the time, it was a manifestation of desperate longing to be somewhere other than where I

was, someone who felt free and cared for. At the time I saw that description as basically an impossibility; my life could never be so amazing in reality.

Fast forward about seven or ten years and I rediscovered the description when I was moving old notebooks and journals from one dusty storage spot to another. As I read through it, I discovered that 90% of the statements I had made in that description were true (or true in spirit). ... It was incredible to me, despite all the changes that had happened in my life since when I wrote the passage, that I had basically become the person whose life I had dreamed of living as a teenager.

That's pretty fucking cool.

And then came Sanatan Dinda. An Indian visual artist from Kolkata, he didn't even make the finals the first year he competed, and the next year he placed second with a style that broke half a dozen of the implicit rules of 'good artwork' at the competition. ... [T]he third year he came he won the entire competition by something like ten percent of the total awarded points over the next artist in second place.

... The thing that confused me though was this – I could not work out how he did it. Like, I had zero mental model of how he created that piece in the same timeframe we all had; how he came up with it, designed it, practiced it. Even though he placed first and I placed fifth and logically we both existed on a scale of 'competence at bodypainting' it seemed like the skills required were completely different.

The exercise they suggest is a really useful activity for weekly (or monthly or yearly) reviews. Highly recommended!

Dan Luu, [95th percentile isn't that good](#). Great for cultivating self-improvement mindset by reminding you how easy (in some sense) it is to make huge improvements at something:

Reaching 95%-ile isn't very impressive because it's not that hard to do. I think this is one of my most ridiculous ideas. It doesn't help that, when stated nakedly, that sounds elitist. But I think it's just the opposite: most people can become (relatively) good at most things.

Note that when I say 95%-ile, I mean 95%-ile among people who participate, not all people (for many activities, just doing it at all makes you 99%-ile or above across all people). I'm also not referring to 95%-ile among people who practice regularly. The "one weird trick" is that, for a lot of activities, being something like 10%-ile among people who practice can make you something like 90%-ile or 99%-ile among people who participate.

It's not weekly review material, but I also appreciate the bonus section on Dan's other most ridiculous ideas.

Suggest your own favorite life advice essays in the comments!

TAI Safety Bibliographic Database

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

Authors: Jess Riedel and Angelica Deibel

[Cross-posted to EA Forum](#)

In this post we present the first public version of our bibliographic database of research on the safety of transformative artificial intelligence (TAI). The primary motivations for assembling this database were to:

1. Aid potential donors in assessing organizations focusing on TAI safety by collecting and analyzing their research output.
2. Assemble a comprehensive bibliographic database that can be used as a base for future projects, such as a living review of the field.

The database contains research works motivated by, and substantively informing, the challenge of ensuring the safety of TAI, including both technical and meta topics. This initial version of the database has attempted comprehensive coverage only for traditionally formatted research produced in 2016-2020 by organizations with a significant safety focus (~360 items). The database also has significant but non-comprehensive coverage (~570 items) of earlier years, less traditional formats (e.g., blog posts), and non-safety-focused organizations. Usefully, we also have citation counts for essentially all the items for which that is applicable.

The core database takes the form of a [Zotero library](#). Snapshots are also available as [Google Sheet](#), [CSV](#), and [Zotero RDF](#). (Compact version for easier human reading: [Google Sheet](#), [CSV](#).)

The rest of this post describes the composition of the database in more detail and presents some high-level quantitative analysis of the contents. In particular, our analysis includes:

- Lists of the most cited TAI safety research for each of the past few years (Tables 2 and 3)
- A chart showing how written TAI safety research output has changed since 2016 (Figure 1).
- A visualization of the degree of collaboration on TAI safety between different research organizations (Table 4).
- A chart showing how the format of written research varied between organizations, e.g., manuscripts vs. journal articles vs. white papers (Figure 2).
- A comparison of the number of citations that different organizations have accumulated (Figure 4).

In 2020 we observe a year-over-year drop in technical safety research, but *not* meta safety research, which we do not understand (Figure 1). We suggest some possible causes, but without a convincing explanation we must caution against drawing strong conclusions from any of our data.

If you are interested in building on this work, we encourage you to contact us (or just grab the data from the links above). Please see the section “Feedback & Improvements”.

Composition

Inclusion & categorization

We use "paper" and "item" interchangeably to refer to any written piece of research, such as an article, book, blog post, or thesis. For this initial version of the database, we have divided all papers into two subject areas: technical safety (e.g. alignment, amplification, decision theoretic foundations) and meta safety (e.g., forecasting, governance, deployment strategy).

Our inclusion criteria do not represent an assessment of quality, but we do require that the intended audience is other researchers (as opposed to the general public). Our detailed criteria for including and categorizing papers can be found in the Appendix.

Safety organizations

Where appropriate, papers were associated with one or more of the following organizations that have an explicit focus, at least in part, on the safety of transformative artificial intelligence: AI Impacts, AI Safety Camp, BERI, CFI, CHAI, CLR, CSER, CSET, DeepMind, FHI, FLI, GCRI, GPI, Median Group, MIRI, Open AI, and Ought. We refer to all these as "safety organizations" hereafter.

Note that AI Impacts, AI Safety Camp, BERI, and MIRI use unconventional research/funding mechanisms and particular care must be taken when using our data to assess their impact. Further detail on this issue can be found in the section "Caveats" in the Appendix.

Coverage

The papers in our database can be partitioned by four binary properties:

- **Traditional / Web:** Traditionally formatted research intended for formal review (journal articles, books, white papers, manuscripts) vs. web content not intended for review (e.g., blog posts, forum posts, wikis).
- **Safety orgs / Other org:** Research associated with at least one of the safety organizations above vs. not.
- **Tech / Meta:** Technical safety vs. meta safety,
- **Recent / Older:** Dated 2016 and younger vs. 2015 and older.

For this initial version of our database we aimed for near comprehensive coverage of traditionally formatted research produced by safety organizations since 2016, i.e., "Traditional" AND "Safety org" AND "Recent" AND ("Tech" OR "Meta"). As discussed in the later subsection "Papers over time", we are probably nevertheless missing many papers from 2020 for various reasons, but we expect our coverage of 2020, and indeed of any given year, to improve over time.

	Safety orgs		Other orgs			
	Recent	Older	Recent	Older		
Traditional	Tech	238	47	Tech	94	35
	Meta	120	30	Meta	36	8
Web	Tech	124	8	Tech	34	6
	Meta	49	27	Meta	18	2

Table 1: We aimed for comprehensive coverage in the light blue region: “Traditional”, “Safety org”, “Tech” & “Meta”, “Recent”. Here, “Recent” means since 2016, inclusive.

Future versions of this database may aim for comprehensive coverage in other categories; see “Feedback & improvements” below.

Analysis

Our analysis in this section is largely restricted to the categories for which we are attempting comprehensive coverage, i.e., the ~360 items of TAI safety research produced since 2016 by safety organizations and intended for traditional review, indicated by the blue background in Table 1.

Warning: The drawbacks to quantifying research output by counting papers or citations are massive. At best, citations are a measure of the *amount of discussion* about a paper, and this is a poor measure of *importance* for many of the same reasons that the most discussed topics on Twitter are not the most important.

Top papers

To get a sense of the most read and discussed TAI safety research, we list in Tables 2 and 3 the most cited papers for each of the years 2016-2020.

The set of TAI safety papers at the top of the rankings is very dependent on one's criteria for considering whether a paper is *about* TAI safety. Some papers that are on the borderline of TAI safety (e.g., being more closely connected to near-term topics like autonomous vehicles and DL interpretability) have wider audiences and can dominate the citation counts.

Nevertheless, it seems useful to have an inclusive list of most cited papers to get a sense of what's out there. One can start at the top of the list and move down, ignoring those papers that are not sufficiently devoted to TAI safety.

Most Cited Technical Safety Items In 2016

Cites	Organization(s)	First Author	Title
724	Open-AI	Amodei	Concrete Problems in AI Safety
254	CHAI	Sadigh	Planning for Autonomous Cars that Leverage Effects on Human Actions
239	CHAI	Hadfield-Menell	Cooperative Inverse Reinforcement Learning
113	CHAI	Sadigh	Information gathering actions over human internal state
75	DeepMind; FHI	Orseau	Safely Interruptible Agents
65	FHI; Ought	Evans	Learning the preferences of ignorant, inconsistent agents
41	<Other org>	Greene	Embedding Ethical Principles in Collective Decision Support Systems
38	Open-AI	Clark	Faulty Reward Functions in the Wild
35	FHI	Bostrom	Fundamental issues of artificial intelligence
32	<Other org>	Babcock	The AGI Containment Problem

Most Cited Technical Safety Items In 2017

Cites	Organization(s)	First Author	Title
836	<Other org>	Doshi-Velez	Towards A Rigorous Science of Interpretable Machine Learning
822	DeepMind	Lakshminarayanan	Simple and Scalable Predictive Uncertainty Estimation using Deep Ense...
314	DeepMind; Open-AI	Christiano	Deep reinforcement learning from human preferences
310	Open-AI	Achiam	Constrained policy optimization
131	CHAI	Hadfield-Menell	Inverse Reward Design
130	DeepMind	Leike	AI Safety Gridworlds
101	<Other org>	Conitzer	Moral Decision Making Frameworks for Artificial Intelligence
62	CHAI	Basu	Do You Want Your Autonomous Car to Drive Like You?
52	CHAI	Hadfield-Menell	The Off-Switch Game
48	<Other org>	Eysenbach	Leave no Trace: Learning to Reset for Safe and Autonomous Reinforce...

Most Cited Technical Safety Items In 2018

Cites	Organization(s)	First Author	Title
455	<Other org>	Raghunathan	Certified Defenses against Adversarial Examples
180	DeepMind	Uesato	Adversarial Risk and the Dangers of Evaluating Against Weak Attacks
148	DeepMind	Rabinowitz	Machine Theory of Mind
145	DeepMind	Dvijotham	A Dual Approach to Scalable Verification of Deep Networks
125	<Other org>	Ruan	Reachability Analysis of Deep Neural Networks with Provable Guarantees
123	<Other org>	Koller	Learning-based Model Predictive Control for Safe Exploration
77	DeepMind	Farajtabar	More Robust Doubly Robust Off-policy Evaluation
73	<Other org>	Yu	Building Ethics into Artificial Intelligence
70	FHI; Ought	Saunders	Trial without Error: Towards Safe Reinforcement Learning via Human In...
64	DeepMind	Dvijotham	Training verified learners with learned verifiers

Most Cited Technical Safety Items In 2019

Cites	Organization(s)	First Author	Title
134	<Other org>	Lehman	The Surprising Creativity of Digital Evolution: A Collection of Anecd...
132	DeepMind	Nalisnick	Do Deep Generative Models Know What They Don't Know?
126	DeepMind	Ovadia	Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncerta...
94	DeepMind	Gowal	On the Effectiveness of Interval Bound Propagation for Training Verif...
55	CHAI	Li	Robust Multi-Agent Reinforcement Learning via Minimax Deep Determinis...
50	DeepMind	Ren	Likelihood Ratios for Out-of-Distribution Detection
39	DeepMind	Hendrycks	AugMix: A Simple Data Processing Method to Improve Robustness and Unc...
34	DeepMind	Qin	Adversarial Robustness through Local Linearization
33	DeepMind	Bahdanau	Learning to Understand Goal Specifications by Modelling Reward
32	<Other org>	Lütjens	Safe Reinforcement Learning with Model Uncertainty Estimates

Most Cited Technical Safety Items In 2020

Cites	Organization(s)	First Author	Title
58	MIRI	Taylor	Alignment for Advanced Machine Learning Systems
57	<Other org>	Adiwardana	Towards a Human-like Open-Domain Chatbot
34	GCRI	Baum	Social choice ethics in artificial intelligence
24	CHAI	Fisac	Pragmatic-Pedagogic Value Alignment
16	<Other org>	Marcus	The Next Decade in AI: Four Steps Towards Robust Artificial Intellige...
11	MIRI	Levinstein	Cheating Death in Damascus

Table 2. The top cited TAI technical safety research for years 2016-2020, including work produced by non-safety organizations. Sorry it's an image.

Most Cited Meta Safety Items In 2016

Cites	Organization(s)	First Author	Title
365	FHI	Müller	Future progress in artificial intelligence: A survey of expert opinion
120	<Other org>	Hanson	The Age of Em: Work, Love, and Life when Robots Rule the Earth
77	FHI	Armstrong	Racing to the precipice: a model of artificial intelligence development
64	<Other org>	Gurkaynak	Stifling artificial intelligence: Human perils
58	<Other org>	Pistono	Unethical Research: How to Create a Malevolent Artificial Intelligence
54	<Other org>	Yampolskiy	Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures
32	FLI	Asaro	The Liability Problem for Autonomous Artificial Agents
32	FHI	Yampolskiy	The Technological Singularity: Managing the Journey
28	FHI; GPI	Cotton-Barratt	Global Catastrophic Risks 2016
24	FHI	Bostrom	The Unilateralist's Curse and the Case for a Principle of Conformity

Most Cited Meta Safety Items In 2017

Cites	Organization(s)	First Author	Title
64	FHI	Bostrom	Strategic implications of openness in AI development
47	GCRI	Baum	On the promotion of safe and socially beneficial artificial intelligence
32	FHI	Callaghan	Technological Singularity
32	GCRI	Baum	A Survey of Artificial General Intelligence Projects for Ethics, Risk...
28	GCRI	Barrett	A model of pathways to artificial superintelligence catastrophe for risk...
21	FHI	Sandberg	That is not dead which can eternal lie: the aestivation hypothesis for...
20	<Other org>	Batin	Artificial Intelligence in Life Extension: from Deep Learning to Super...
20	CLTR	Sotala	Superintelligence As a Cause or Cure For Risks of Astronomical Suffering...
15	CLTR	Sotala	How feasible is the rapid development of artificial superintelligence?
11	FHI	Bostrom	Transhumanist FAQ 3.0

Most Cited Meta Safety Items In 2018

Cites	Organization(s)	First Author	Title
244	BERI; CFI; CSER; FHI; Open-AI	Brundage	The Malicious Use of Artificial Intelligence: Forecasting, Prevention...
117	<Other org>	Rahwan	Society-in-the-loop: programming the algorithmic social contract
59	FHI	Dafoe	AI governance: a research agenda
45	CFI; CSER	Cave	An AI Race for Strategic Advantage: Rhetoric and Risks
42	FHI	Ding	Deciphering China's AI dream
30	CSER	Avin	Classifying global catastrophic risks
29	<Other org>	Danzig	Managing Loss of Control as Many Militaries Pursue Technological Superiority
23	<Other org>	Hwang	Computational Power and the Social Impact of Artificial Intelligence
20	FHI	Bostrom	The vulnerable world hypothesis
18	FHI	Sandberg	Dissolving the Fermi Paradox

Most Cited Meta Safety Items In 2019

Cites	Organization(s)	First Author	Title
260	<Other org>	Cummings	On the referendum #31: Project Maven, procurement, lollapalooza results...
42	FHI	Zhang	Artificial Intelligence: American Attitudes and Trends
24	CFI	Whittlestone	The Role and Limits of Principles in AI Ethics: Towards a Focus on Technical...
22	CLTR; FHI; GCRI	Baum	Long-term trajectories of human civilization
16	CFI; CSER; FHI	Cave	Bridging near- and long-term concerns about AI
16	FHI	Garfinkel	How does the offense-defense balance scale?
13	BERI; FHI	Cihon	Standards for AI Governance: International Standards to Enable Global...

Most Cited Meta Safety Items In 2020

Cites	Organization(s)	First Author	Title
48	<Other org>	Turchin	Classification of global catastrophic risks connected with artificial...
43	<Other org>	Erdélyi	Regulating Artificial Intelligence: Proposal for a Global Solution
34	GCRI	Baum	Social choice ethics in artificial intelligence
22	CFI; CHAI; CSER; CSET; FHI; Open-AI	Brundage	Toward Trustworthy AI Development: Mechanisms for Supporting Verifiability
14	DeepMind	Mohamed	Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial...
14	FHI	Ord	The Precipice: Existential Risk and the Future of Humanity

Table 3. The top cited TAI meta safety research for years 2016-2020, including work produced by non-safety organizations. I really do feel bad about the fact

that it's an image.

If you want to go deeper into these lists, open up the compact version of the database and sort by year and number of cites: [Google Sheet](#), [CSV](#).

Papers over time

In this section we look at the number of TAI safety research papers produced each year in the span 2016-2020. As shown in Figure 1, we find a surprising drop in the number of technical safety works in 2020 relative to 2019. This potentially indicates an effect of the pandemic, a large number of missing papers from this year, a shift in research format, and/or something else; we're not sure. In any case, this lowered our confidence in the current state of this dataset and dissuaded us from looking at how the output of individual orgs changes over time until we can understand the drop better. The drop was seen in many organizations, including the largest contributors to TAI technical safety (CHAI, FHI, Deepmind, and Open AI), so it appears to be systemic.

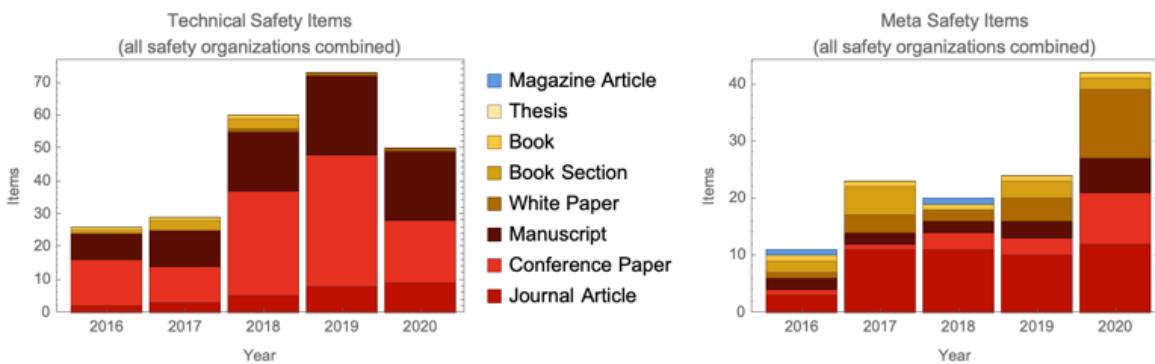


Figure 1. The number of traditionally formatted TAI safety research papers produced by safety organizations from 2016 to 2020.

Because we don't have a satisfactory explanation to this, we simply list some some complicating considerations:

- We only see this apparent drop in technical safety research, *not* in meta safety research.
- The magnitude of the drop from 2019 to 2020 diminishes substantially when we look at *all* technical safety papers rather than just those papers that were associated with safety organizations: From the safety organizations whose work is covered above, we have 73 traditionally formatted technical safety papers in 2019 and only 50 in 2020, but from all organizations we have 92 papers in 2019 and 82 in 2020.
- If we were to include web content (blog posts), the numbers would actually be higher for 2020 than 2019, as one might have originally expected: we have 125 such papers in 2019 and 161 in 2020, although this is hard to interpret since a typical blog post is a smaller unit of research than the typical paper, and our inclusion process for blog posts is very haphazard.
- Most of the 2019-to-2020 drop in technical-safety papers comes from diminished conference papers, but the number of meta-safety conference papers *increased*.
- As discussed in the subsection "Date" in the Appendix, our dates should be biased *forward* by the fact that the current year should include both new manuscripts (preprints) that were recently released as well as manuscripts from previous years that were published this year.
- As discussed in the subsection "Discoverability" in the Appendix, it may be substantially more difficult for us to find work in the past year because of delays in our discovery mechanisms. In particular, (1) we must wait for researchers to report new publications to their organization in order for the organization to give that information

to us, and (2) we can't find items in review articles that appeared after a review article was written. This effect may be large, but it does not account for why the 2019-to-2020 drop is seen in technical safety but not meta safety.

It's possible that there's some genuine trend here, with research moving away from the relatively narrow category of technical safety research published in non-blog-post form and supported by safety organizations, and instead towards meta-safety research, towards being published as blog posts, or towards being supported by other organizations or done without organizational support. It's also possible that the pandemic had an impact on the amount of research done (or published) in 2020 (the drop in technical safety conference papers from 2019 to 2020 makes this a tempting conclusion to jump to, but, again, meta safety conference papers actually increased). However, we don't think our data provides strong evidence on any of these points.

Separately, we didn't make a corresponding plot of citations in different years because these go *down* with time; younger papers haven't had time to accumulate as many cites. (Our automated solution for scraping Google Scholar for citation data means we don't have access to *when* citations are made.)

Collaboration between organizations

The following matrix illustrates the extent to which different AI-safety organizations collaborate on research.

	AI Safety Camp	BERI	CFI	CHAI	CLTR	CSER	CSET	Deep Mind	FHI	FLI	GCRI	GPI	Median Group	MIRI	Open AI	Ought
AI Safety Camp	1.0 1.0	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00
BERI	.00 .00	.57 .40	.00 .10	.29 .00	.00 .00	.00 .20	.00 .00	.00 .00	.00 .20	.00 .00	.00 .00	.00 .00	.14 .00	.00 .00	.00 .10	.00 .00
CFI	.00 .00	.00 .04	.67 .42	.00 .02	.00 .00	.33 .35	.00 .02	.00 .00	.00 .12	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .04	.00 .00
CHAI	.00 .00	.04 .00	.00 .11	.90 .44	.00 .00	.00 .11	.00 .11	.03 .00	.00 .11	.00 .00	.00 .00	.00 .00	.00 .00	.03 .00	.01 .11	.00 .00
CLTR	.00 .00	.00 .00	.00 .00	.00 .00	1.0 .82	.00 .00	.00 .00	.00 .00	.00 .09	.00 .00	.00 .09	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00
CSER	.00 .00	.00 .06	.43 .29	.00 .02	.00 .00	.57 .46	.00 .02	.00 .00	.00 .11	.00 .02	.00 .00	.00 .00	.00 .00	.00 .00	.00 .03	.00 .00
CSET	.00	.00	.11	.11	.00	.11	.44	.00	.11	.00	.00	.00	.00	.00	.11	.00
DeepMind	.00 .00	.00 .00	.00 .00	.04 .00	.00 .00	.00 .00	.00 .00	.88 .10	.06 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.01 .00	.00 .00
FHI	.00 .00	.00 .05	.00 .08	.00 .01	.00 .01	.00 .09	.00 .01	.09 .00	.77 .66	.00 .00	.00 .01	.00 .01	.00 .00	.05 .00	.00 .05	.09 .00
FLI	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .33	.00 .00	.00 .00	.00 .00	1.0 .67	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00
GCRI	.00 .00	.00 .00	.00 .00	.00 .00	.00 .05	.00 .00	.00 .00	.00 .00	.00 .05	.00 .00	1.0 .90	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00
GPI	.00	.00	.00	.00	.00	.00	.00	.00	.17	.00	.00	.83	.00	.00	.00	.00
Median group	.00	.50	.00	.00	.00	.00	.00	.00	.00	.00	.00	.50	.00	.00	.00	.00
MIRI	.00	.00	.00	.11	.00	.00	.00	.07	.00	.00	.00	.00	.82	.00	.00	.00
Open AI	.00 .00	.00 .10	.00 .10	.08 .05	.00 .00	.00 .10	.00 .05	.08 .00	.00 .20	.00 .00	.00 .00	.00 .00	.00 .00	.00 .00	.85 .40	.00 .00
Ought	.00	.00	.00	.00	.00	.00	.00	.00	.40	.00	.00	.00	.00	.00	.00	.60

Table 4. How the safety organizations collaborate on traditionally-formatted TAI safety research, 2016-2020. The left half of each off-diagonal box contains the fraction of all technical safety papers associated with the organization for that row that were also associated with the organization for that column. The right half is the same for meta safety. The intensity of the green color of each box is set by the fraction of all safety papers (both technical and meta combined).

Likewise, each on-diagonal box contains the fractions of papers where the organization did not collaborate with any of the other organizations. Box halves are left empty when the organization on that row had no papers of the corresponding type.

Output by organization

In Figure 2 we break down the research papers associated with each organization into different types: conference papers, books, manuscripts, etc. Recall that for this analysis we are ignoring web content not intended for review (e.g., blog posts).

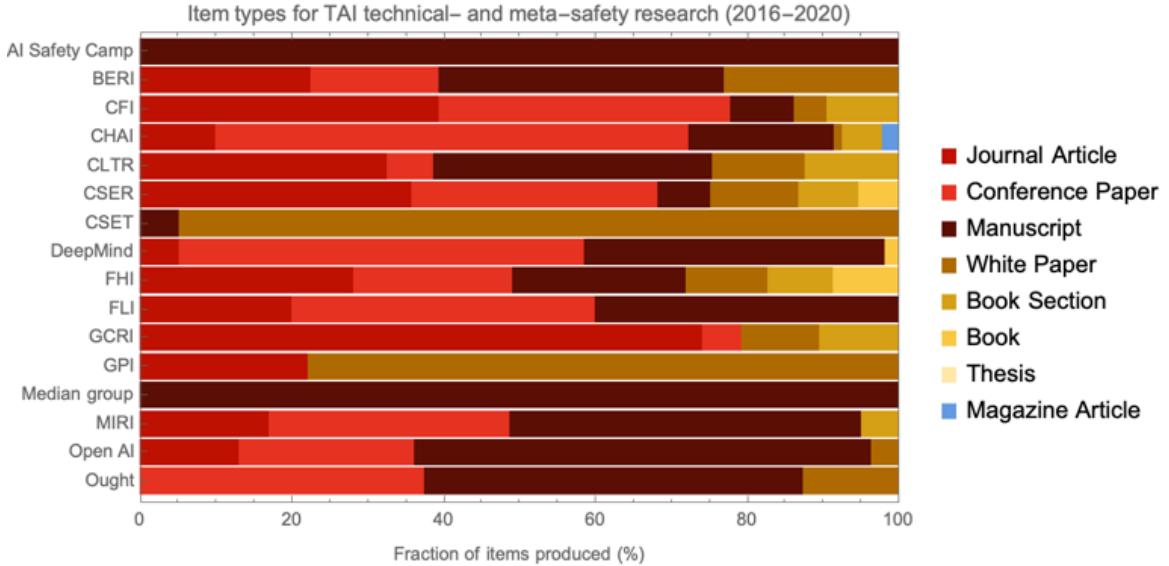


Figure 2. Type of item produced for traditionally-formatted TAI safety research by the safety organizations. Note that “manuscript” is the only type of item that has not (yet) undergone peer or editorial review.

In Figures 3 and 4 we display the total number of papers and citations associated with each organization. We call these “fractional papers” and “fractional cites” to emphasize that we have divided credit for each item and its citations equally to all the organizations that are associated with it. Needless to say, for any given item the support it received from different organizations is unlikely to be equal, but we didn’t have a way to account for that with a reasonable amount of work.

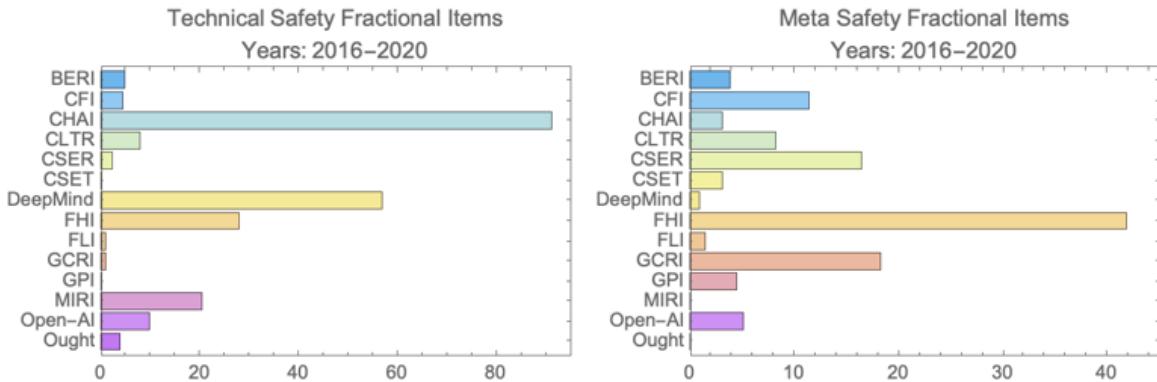


Figure 3. The total number of research papers produced by each safety organization between 2016 and 2020, with credit for multi-organization papers divided equally.

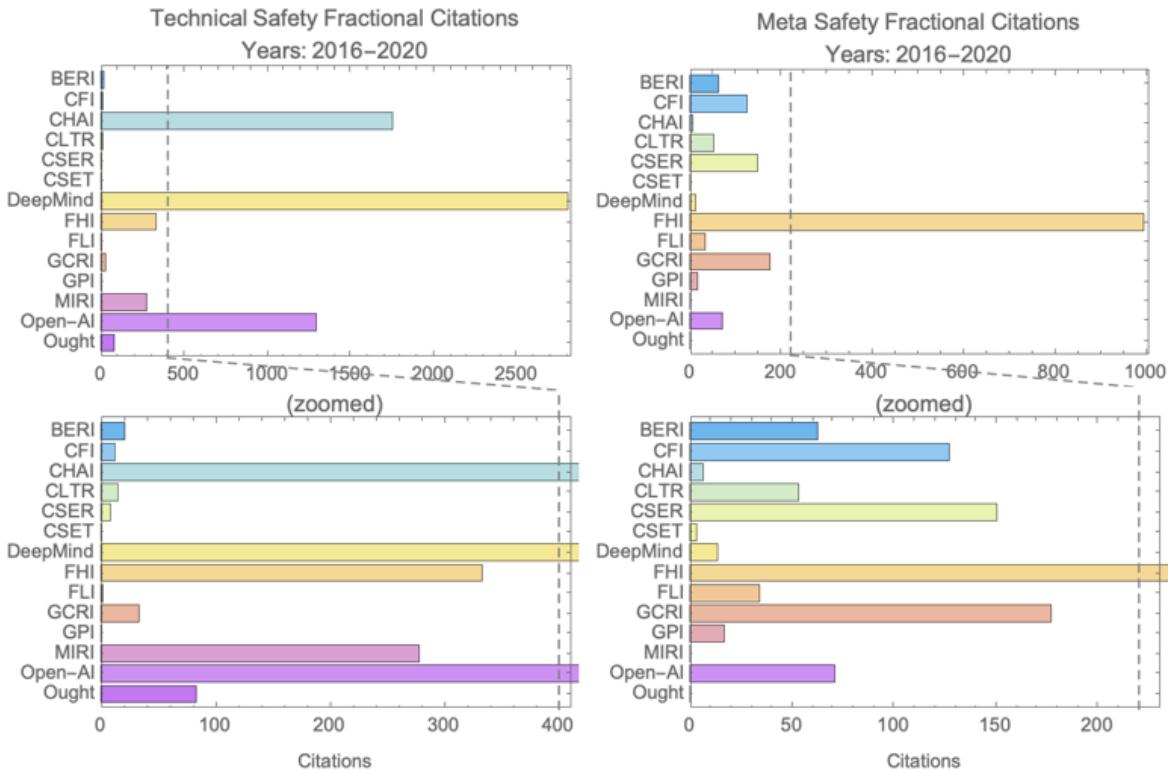


Figure 4. The total number of citations accumulated by each safety organization between 2016 and 2020, with credit for citations to multi-organization papers divided equally. Because of the large disparity in citation counts, the plots in the second row have been zoomed in to show detail.

This concludes our analysis.

Feedback & improvements

If there is sufficient interest, we would like to make some of the following improvements in the future:

- More granular categories (e.g., philosophy, governance, transparency, alignment, etc.).
- More comprehensive coverage of research that is not associated with any safety organization.
- More comprehensive coverage for 2015 and earlier.
- Commentary on individual papers and links to further discussion.

Want to help?

Accurately classifying 1,500+ papers is not an easy task, and there will be mistakes. Please bring them to our attention. You can [email](#) us or just [write down your suggestion here](#). If you are interested in looking at a batch of papers that we found ambiguous and recommending a classification, please contact us, especially if you are an AI safety researcher.

Please let us know if you find this database useful and how you would like it to be improved. We are certainly willing to adjust our inclusion criteria based on strong reasons, but it is important we do not let the database grow to an unmanageable size.

Simple comments like "this database informed my donation decision" or "I found an interesting AI Safety paper here that I didn't know about before" are very useful and help us decide whether to put more effort into this in the future.

Please also contact us if you are interested in helping to improve this database more generally, either one-off or in an on-going capacity. We think that maintaining and expanding the database for a year would be an excellent side-project for a grad student in AI safety to help them gain a broader perspective on the field and contribute to the community. In particular, there seems to be some interest for collecting brief summaries and assessments of the most important TAI safety papers in one place, and this database could be a foundation for that.

We also could use help modifying the [Zotero Scholar Citations plugin](#) to more intelligently avoid tripping Google Scholar's rate-limit ban.

Add your organization?

If you would like your organization included in future versions of the database, please send us a list of papers that ought to be associated with your organization, either as text file, Excel file, CSV file, or Google Sheet. It is not necessary to give us the complete bibliographic information, just enough for us to reliably uniquely identify the works. (For instance, a list of DOI and arXiv numbers is fine.) Optionally, you may also include suggested categorization and an explanation for why you believe a certain paper fits our criteria, which would be especially helpful for papers where that might be ambiguous.

Acknowledgements

We thank Andrew Critch, Dylan Hadfield-Menell, and Larks for feedback. We also thank the organizations who responded to our inquiries and helpfully provided additions and corrections to our database. We are, of course, responsible for the remaining mistakes.

Appendices

Inclusion & categorization

This section includes more details on what sort of research we include in our database, and how it is categorized.

These are our inclusion criteria:

- **The contents of the paper are directly motivated by, and substantively inform, the challenge of ensuring good outcomes for TAI.** The paper need not mention TAI explicitly, but it must be motivated by it, since there are far too many papers that are merely *relevant* to safety. Judging motivation is, unfortunately, inherently subjective, but this is necessary to avoid penalizing papers that do not explicitly mention TAI for appearance reasons, while also not including every paper on, e.g., adversarial examples (which are motivated by capabilities and near-term safety). If the paper would likely have been written even in the absence of TAI-safety concerns, it is excluded. Ultimately, we want to support researchers who are motivated by TAI safety and allow them to find each other's work.
- **There is substantive content on AI safety, not just AI capabilities.** That said, for more speculative papers it is harder to distinguish between safety vs. not safety, and between technical vs. meta, and we err on the side of inclusion. Articles on the

safety of autonomous vehicles are generally excluded, but articles on the foundations of decision theory for AGI are generally included.

- **The intended audience is the community of researchers.** Popular articles and books are excluded. Papers that are widely released but nevertheless have substantial research content (e.g., Bostrom's *Superintelligence*) are included, but papers that merely try to *recruit* researchers are excluded.
- **It meets a subjective threshold of seriousness/quality.** This is intended to be a very low threshold, and would, for instance, include anything that was accepted to be placed on the ArXiv. Web content not intended for review (e.g., blog posts) is only accepted if it has reached some (inevitably subjective) threshold of notability in the community. It is of course infeasible for us to document all blog posts that are about TAI safety, but we do not want to exclude some posts that have been influential but have never been published formally.
- **Peer review is not required.** White papers, preprints, and book chapters are all included.

Here are how research papers are classified as technical safety vs. meta safety:

- Technical safety concerns the design and understanding of TAI systems, e.g., alignment, agent foundations, corrigibility, amplification, robustness, verification, decision theoretic foundations, logical uncertainty.
- Meta safety concerns the higher-level details of ensuring that TAI is safe, e.g., planning, forecasting, governance, deployment strategy, and policy
- Academic review articles about technical safety research are classified as technical safety, not meta safety.
- Philosophy that speaks directly to the technical design of TAI (e.g., Newcomb's problem) is classified as technical safety. Philosophy that informs how TAI ought to be deployed is classified as meta safety. Philosophy narrowly addressing the moral worth of animals is excluded (even though it is relevant to safety because an AGI with bad goals regarding animals is a safety risk).

Note that near the popular level, or at the very speculative level, it's sometimes hard to distinguish meta safety and technical safety.

Here is how we define the different item types:

- **Journal article:** Published in a peer-reviewed academic journal.
- **Conference paper:** Published in the peer-reviewed proceedings of an academic conference.
- **Manuscript:** Intended for academic publishing or posted on a preprint server, but not peer-reviewed.
- **White paper:** Released with the endorsement of a particular academic or policy institution but without external peer-review.
- **Thesis:** Academic dissertation accepted by university graduate program.
- **Book:** Book published with editorial review from a publishing house.
- **Book section:** Part of a book.
- **Web content:** A catch-all term for all material released on the web that is not intended to be reviewed and not appearing on a preprint server. This includes wikis, blog posts, public forum posts, and press releases.
- **Magazine article:** Published in a magazine with editorial review.
- **Newspaper article:** Published in a newspaper with editorial review.

Criteria for our quantitative analysis

We err on the side of inclusion in our database, but when we analyze it quantitatively in this post we exclude some categories that are hard to cover comprehensively. This makes the conclusions we draw less dependent on the vagaries of what managed to make it into the database. In particular, the following items are excluded:

- **Web content not intended for review.** Currently, our coverage of blog posts, forum posts, and similar items is much too haphazard to be usefully analyzed in the aggregate. This may or may not change in the future.
- **Items dating from 2015 and before.** So far it has required too much work to get comprehensive coverage of older papers, especially since organizations often do not have complete records going arbitrarily far back.
- **Items not associated with any safety organization.** We can ask safety organizations for lists of the research they have produced, but TAI safety papers not associated with any organization can generally only be found by trawling through review papers and the rest of the literature. We believe this particularly noisy process would bias our quantitative analysis, so we ignore them for this purpose.

Dates

Our convention is to (1) date manuscripts (preprints) using the date on which they first appeared and (2) date published articles by the publication date. This has the unfortunate property that the date must be *updated* when a manuscript gets published, potentially biasing our data in weird ways. In particular, it makes judging whether a published article has a lot of citation given its age difficult, since a longer time between preprint and publication will yield more citations. (See for instance "[Cheating Death in Damascus](#)" by Levinstein & Soares, which was available since at least 2017 but published only this year.) It also means there will appear to be more papers produced in the last year relative to earlier years, even with constant research output in the field, since the number of papers attributed to a fixed year (say, 2017) can decrease over time as things are published and have their dates updated.

Unfortunately, this bias seems hard to avoid since the alternative is to track down the year that a preprint appeared for all published articles by hand, and using that year would be a non-standard convention in any case.

Discoverability

Our data might be less complete for 2020 compared to previous years because papers released recently by researchers at these organizations have not yet been reported to the organization's management; that would mean the papers are not listed on the organization website *and* the organizations don't know about them when we ask by email. (Almost all the organizations replied when we asked them for help filling in papers missing from our database.) A related factor is that many papers in our database were found in existing review articles and bibliographies (see "Sources" in the Appendix). This process necessarily only turns up papers released before those sources were written.

Unfortunately, some of this effect is probably unavoidable. It could potentially be reduced by finding (and confirming organization affiliation for) all papers by authors associated with each organization we cover, but this would be quite laborious, especially since there is a fair amount of author migration. Such effort would be mostly made useless the next year when authors had finished reporting their 2020 papers to the orgs.

Therefore, we have elected not to account for this effect. We encourage organizations to improve their collection of this data, and we caution the reader to put even less faith in the 2020 numbers than earlier years.

Organizations

Here are the organizations that we associated with papers in our database:

- [AI Impacts](#)
- [AI Safety Camp](#)

- [Berkeley Existential Risk Initiative](#) (BERI)
- [Leverhulme Centre for the Future of Intelligence](#) (CFI)
- [Center for Human-Compatible Artificial Intelligence](#) (CHAI)
- [Center on Long-Term Risk](#) (CLR), previously called the Foundational Research Institute
- [Centre for the Study of Existential Risk](#) (CSER)
- [Center for Security and Emerging Technologies](#) (CSET)
- [DeepMind](#)
- [Future of Humanity Institute](#) (FHI)
- [Future of Life Institute](#) (FLI)
- [Global Catastrophic Risk Institute](#) (GCRI)
- [Global Priorities Institute](#) (GPI)
- [Median Group](#)
- [Machine Intelligence Research Institute](#) (MIRI)
- [Open AI](#)
- [Ought](#)

In general, we associated a paper with an institution when, in the paper, the organization was listed as an author affiliation or was explicitly acknowledged as providing funding. In a few cases we associated a paper with an organization because the organization told us the author was highly likely to be supported by the organization during preparation of the paper, or removed such association because the organization told us the support provided was very minimal during preparation of the paper.

Many papers are associated with multiple organizations. When necessary to "count" the credit, we gave each organization equal weight. Of course we expect that sometimes one organization deserved much more credit, in some sense, but it wasn't feasible for us to determine this.

Caveats

Here are some comments on individual organizations that are important for interpreting our data, especially with respect to the organizations' overall impact:

- **AI Impacts:** Much of the research produced by this organization is divided between a [blog](#) and a large publicly readable, privately editable [wiki](#), neither of which are intended for formal publication. This is difficult to fit into our analysis, and we have elected to select only their [featured articles](#) for inclusion in our database. Please see their [2020 review](#) for more on the work they have done.
- **AI Safety Camp:** Some of the papers associated with this organization were produced by participants from the AI Safety Research Program 2019. That format was a spin-off from AI Safety Camp run by Czech effective altruists. AISRP served participants who were more advanced in their research career. On the other hand, AISC is aimed at early career aspirants who want to test their fit by trying to collaborate on a research priority. For reference the AISRP-related papers are tagged as such in our Zotero database, but our analysis includes all AI Safety Camp papers.
- **BERI:** This organization is neither a direct research organization nor a conventional grantmaker, and they do not require that the researchers acknowledge BERI support in their publications. Therefore, the papers that have been associated with BERI represent only one part of their impact.
- **MIRI:** As discussed a bit further in Lark's 2020 AI Alignment Literature Review and Charity Comparison, MIRI now by default does not publish their research. Other than some blog posts by their researchers, the only 2020 papers in our database from them were first released as pre-prints several years ago but were published this year.

Zotero library details

The information in this section may be useful if you are going through the details of our Zotero library.

Getting a copy

The [Zotero library is here](#). Unfortunately it is [not possible to export a copy of the library from the web interface](#). If you would like an up-to-date copy of the library, please contact us. If you are satisfied with a static snapshot from the day we released this post, you can download it as [Google Sheet](#), [CSV](#), and [Zotero RDF](#).

Tagging details

The actual Zotero library contains both papers that satisfy our inclusion criteria (tagged “TechSafety” and “MetaSafety”) and papers that we imported from other review articles and bibliographies but decided did not satisfy our criteria (“NotSafety”). We also marked papers that we decided on a first pass were borderline (tagged “AmbiguousSafety”) so that they can easily be re-considered in the future. Papers associated with a safety organization are tagged as such, while papers not associated with any safety organization are tagged “Other-org”. We also have some miscellaneous tags for a few blog posts that are unlikely to pass our criteria (tagged “non-notable”), AI Safety Camp papers produced during the AI Safety Research Program (tagged “AISRP2019”), and AI Impacts web pages that are not featured articles (tagged “AI-Impacts-NotFeatured”).

For technical reasons within Zotero, all web content is formally categorized as a “blog post”, and all white papers are formally categorized as a “report”.

Citation count scraping

To scrape citation count information from Google Scholar, we use the [Zotero Scholar Citation plugin](#) written by Anton Beloglazov and Max Kuehn. Citation information is recorded in the “extra” field and prefaced by “ZSCC:” (which stands for Zotero Scholar Citation Count). When this information is unavailable (“ZSCC: NoCitationData”) we take the data by hand and denote it with “ACC:” or “JCC:”. In a couple cases we noticed that the plugin was mistaking a paper from another published paper, so we used “JCCOverride:” to denote hand counts that are correct.

Copyright

Added May 29, 2021: We release the Zotero database under the [Creative Commons Attribution-ShareAlike 4.0 International License](#). In short, the means you are free to use, modify, and reproduce the database for anything so long as you cite us and release any derivative works under the same license.

Sources

The following sources are not traditional review articles, but are good sources of TAI safety research articles and summaries thereof (“maps”):

- [Larks, “2020 AI Alignment Literature Review and Charity Comparison”](#) (2020). (See also previous years going back to 2016.)
- Jérémie Perret, [“Resources for AI Alignment Cartography”](#) (2020).
- Rohin Shah, [Alignment Newsletter](#) (2020). (See especially the [spreadsheet of summaries](#).)
- CHAI, [“Annotated Bibliography of Recommended Materials”](#) (2016).
- Victoria Krakovna, [“AI safety resources”](#) (2020).

Here are some AI Safety review articles:

- Critch & Krueger, “[AI Research Considerations for Human Existential Safety \(ARCHEs\)](#)” (2020).
- Richard Ngo, “[AGI safety from first principles](#)” (2020).
- Everitt, Lea, & Hutter, “[AGI Safety Literature Review](#)” (2018).
- Soares & Fallenstein, “[Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda](#)” (2017).
- Dewey, Russell, Tegmark, “[A survey of research questions for robust and beneficial AI](#)” (2016).
- Russell, Dewey, & Tegmark, “[Research Priorities for Robust and Beneficial Artificial Intelligence](#)” (2015).
- Steinhardt, “[Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems](#)” (2015).

Here are the online lists of sponsored publications or researchers maintained by the safety organizations, a few of which are filtered by AI safety.

- [AI Impacts](#)
- [AI Safety Camp](#)
- [CFI](#)
- [CHAI](#)
- [CLR](#)
- [CSET](#)
- [Deepmind](#)
- [FHI](#)
- [FLI](#)
- [GCRI](#)
- [GPI](#)
- [Median Group](#)
- [MIRI](#)
- [Open AI](#)
- [Ought](#)

Unfortunately, they are only sporadically updated and difficult to consume using automated tools. We encourage organizations to start releasing machine-readable bibliographies to make our lives easier.

The Best Visualizations on Every Subject

Edit: the list is now [a public GitHub repository](#), with all that implies. Added sections by media type. Last update: 6 Jan 2021

Motivation

This is [The Best Textbooks on Every Subject](#), but for visualizations. I greatly adore good visualizations, chiefly because there are so many visualizations that are so terrible. I have seen many such tools mentioned here, but always in passing.

The actual motivator is re-reading the posts [Exercises in Comprehensive Information Gathering](#) and [Fact Posts: How and Why](#). While there is no substitute for the wrench-time they recommend, I think these kinds of tools make the process more efficient and lend themselves to insights which are difficult to acquire through reading alone; in my experience scale and distance are both easier to grasp in a visual medium, for example.

Also there is a non-trivial sense in which they are beautiful in their own right. If we are able to compare many examples, people in the community might even be able to help advance the art.

Submission Rules

One nomination per comment; please include an explanation of why you nominated it. Contra the best textbooks list we won't require comparison with other visualizations because there are so few authoritative ones.

Current List

WEB:

History

- ORBIS: The Geospatial Network Model of the Roman World:
<https://orbis.stanford.edu/>
- Data Visualization and the Modern Imagination:
<https://exhibits.stanford.edu/dataviz>
- A Simulated Dendrochronology of Immigration 1790-2016:
<https://web.northeastern.edu/naturalizing-immigration-dataviz/>

Math

- Byrne's Euclid: The First Six Books of the Elements of Euclid With Coloured Diagrams and Symbols: <https://www.c82.net/euclid/>
- The Empirical MetaMathematics of Euclid and Beyond:
<https://writings.stephenwolfram.com/2020/09/the-empirical-metamathematics-of-euclid-and-beyond/>
- An Interactive Introduction to Fourier Transforms:
<http://www.jezzamon.com/fourier/>
- Better Explained: <https://betterexplained.com/>

- Jason Davies: Math and Cartography: <https://www.jasondavies.com/>
- Bit-player: "An amateur's outlook on computation and mathematics": <http://bit-player.org/>

Economics

- The Observatory of Economic Complexity: <https://oec.world/>
- Our World in Data: <https://ourworldindata.org/>

Machine Learning

- Distill: <https://distill.pub/>

Miscellaneous

- Explorable Explanations: <https://explorabl.es/>
- Complexity Explorables: <https://complexity-explorables.org/>
- The Pudding Visual Essays: <https://pudding.cool/>

VIDEO:

BOOKS:

- An Illustrated Theory of Numbers: <http://illustratedtheoryofnumbers.com/>
- Visual Group Theory: <https://bookstore.ams.org/clm-32/>
- Dataclysm: <https://www.amazon.com/Dataclysm-Identity-What-Online-Offline-Selves/dp/0385347391>

JOURNALISM:

- Europe's HIV Divide: <https://www.politico.eu/article/aids-european-state-of-play/>
- Explore 175 Years of Word Usage in Scientific American:
<https://www.scientificamerican.com/article/explore-175-years-of-words-in-scientific-american/>

Why Neural Networks Generalise, and Why They Are (Kind of) Bayesian

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

Currently, we do not have a good theoretical understanding of how or why neural networks actually work. For example, we know that large neural networks are sufficiently expressive to compute almost any kind of function. Moreover, most functions that fit a given set of training data will not generalise well to new data. And yet, if we train a neural network we will usually obtain a function that gives good generalisation. What is the mechanism behind this phenomenon?

There has been some recent research which (I believe) sheds some light on this issue. I would like to call attention to this blog post:

[Neural Networks Are Fundamentally Bayesian](#)

This post provides a summary of the research in these three papers, which provide a candidate for a theory of generalisation:

<https://arxiv.org/abs/2006.15191>
<https://arxiv.org/abs/1909.11522>
<https://arxiv.org/abs/1805.08522>

(You may notice that I had some involvement with this research, but the main credit should go to Chris Mingard and Guillermo Valle-Perez!)

I believe that research of this type is very relevant for AI alignment. It seems quite plausible that neural networks, or something similar to them, will be used as a component of AGI. If that is the case, then we want to be able to reliably predict and reason about how neural networks behave in new situations, and how they interact with other systems, and it is hard to imagine how that would be possible without a deep understanding of the dynamics at play when neural networks learn from data. Understanding their inductive bias seems particularly important, since this is the key to understanding everything from [why they work in the first place](#), to phenomena such as [adversarial examples](#), to the risk of [mesa-optimisation](#). I hence believe that it makes sense for alignment researchers to keep an eye on what is happening in this space.

If you want some more stuff to read in this genre, I can also recommend these two posts:

[Recent Progress in the Theory of Neural Networks](#)
[Understanding "Deep Double Descent"](#)

EDIT: Here is a second post, which talks more about the "prior" of neural networks:

[Deep Neural Networks are biased, at initialisation, towards simple functions](#)